RATIONAL APPROACHES IN LANGUAGE SCIENCE

EDITED BY: Matthew W. Crocker, Gerhard Jäger, Gina Kuperberg,
Hannah Rohde, Elke Teich and Rory Turnbull

PUBLISHED IN: Frontiers in Psychology and Frontiers in Communication

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-765-8
DOI 10.3389/978-2-88974-765-8

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

RATIONAL APPROACHES IN LANGUAGE SCIENCE

Topic Editors:
Matthew W. Crocker, Saarland University, Germany
Gerhard Jäger, University of Tübingen, Germany
Gina Kuperberg, Tufts University, United States
Hannah Rohde, University of Edinburgh, United Kingdom
Elke Teich, Saarland University, Germany
Rory Turnbull, Newcastle University, United Kingdom

Citation: Crocker, M. W., Jäger, G., Kuperberg, G., Rohde, H., Teich, E., Turnbull, R., eds. (2022). Rational Approaches in Language Science. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-765-8

Table of Contents

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

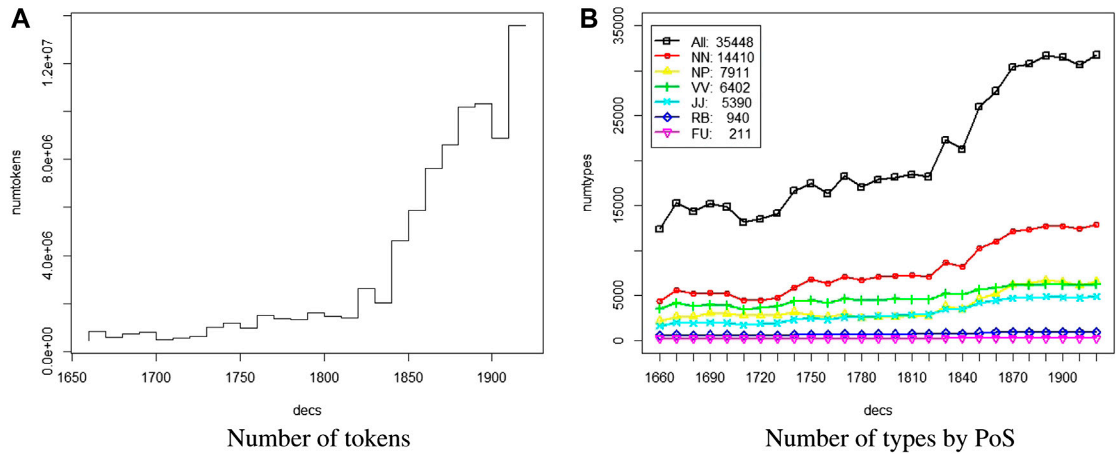


FIGURE 1 | (A) Number of tokens and **(B)** number of types by PoS per decade.

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:

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

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

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)

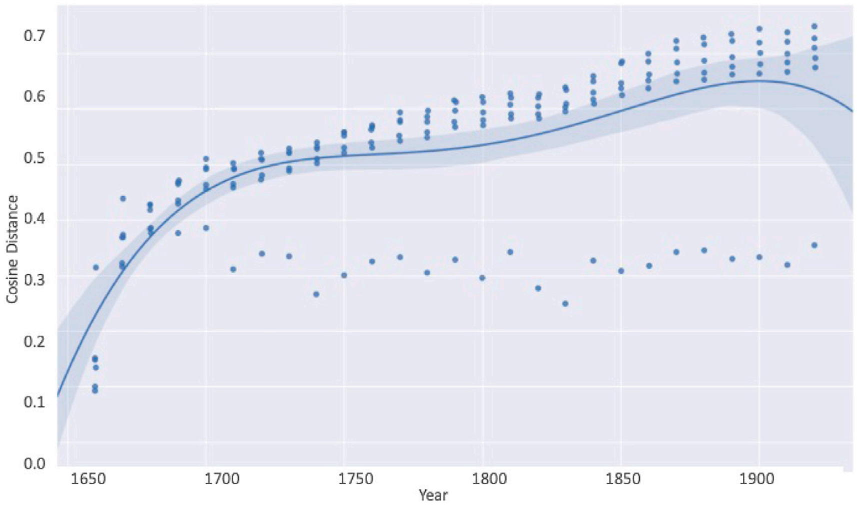
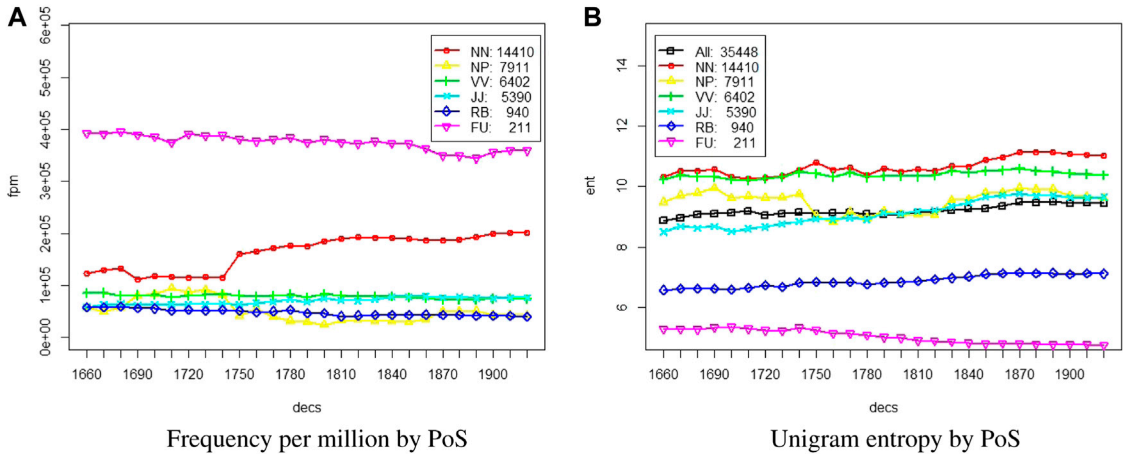


FIGURE 2 | Average cosine distances of words between randomly chosen groups of words for lexical words **(upper points)** and function words **(lower points)**.

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.

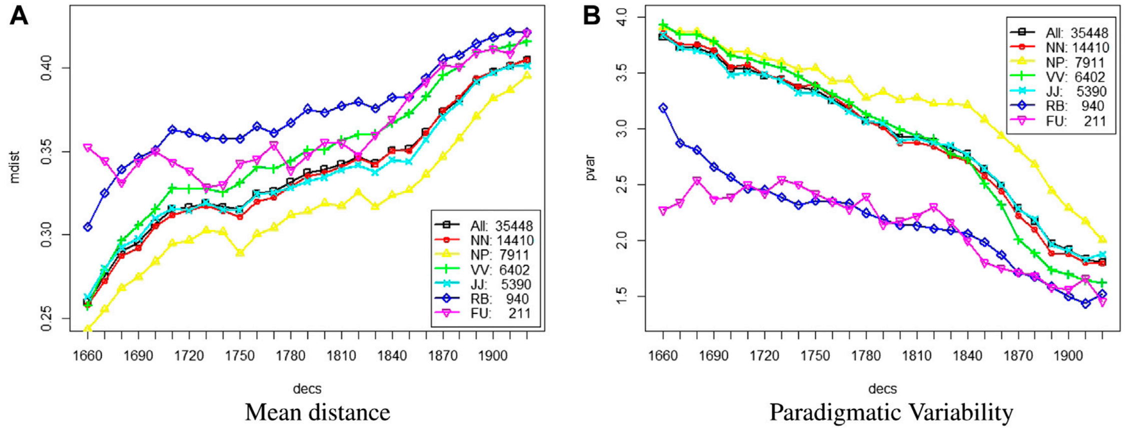
Correlating with overall increasing distance, paradigmatic variability decreases over time as a general trend. **Figure 4** shows mean distance and paradigmatic variability by major parts of speech. Function words (FU) and adverbs (RB) are



Frequency per million by PoS

Unigram entropy by PoS

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



Mean distance

Paradigmatic Variability

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.

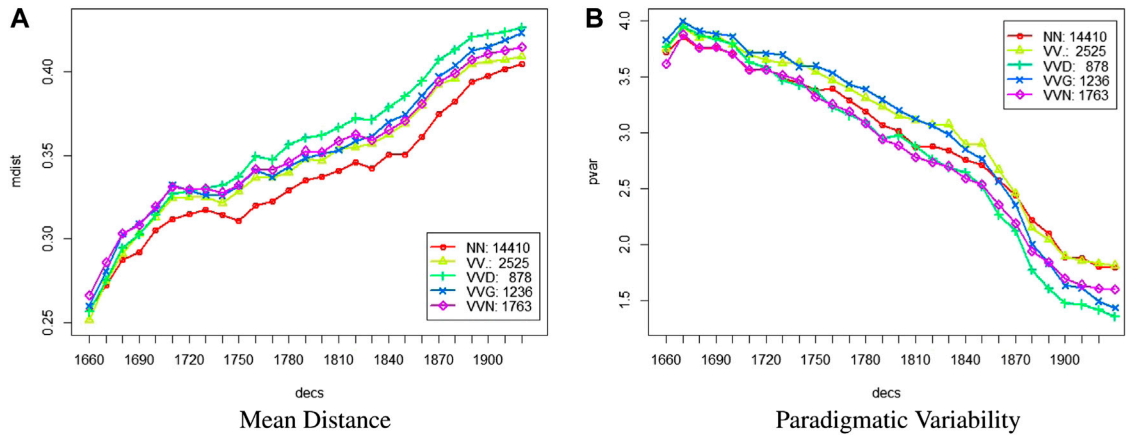


FIGURE 5 | Diachronic trends of nouns (NN) vs. verbal parts of speech (VVG, VVD, VVN) in terms of **(A)** Mean Distance (mdist) and **(B)** Paradigmatic Variability (pvar) by decade.

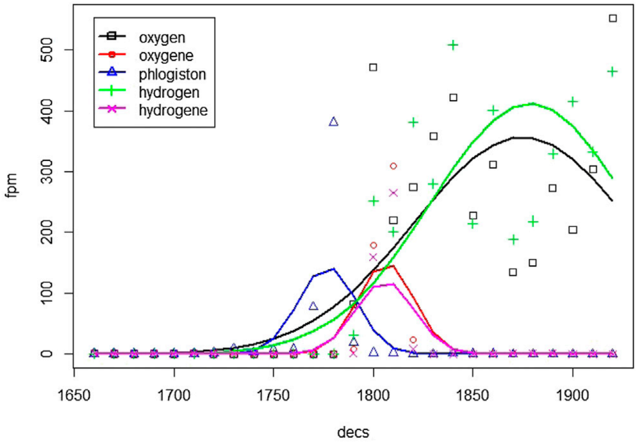


FIGURE 6 | Conventionalization by substitution for *oxygene*

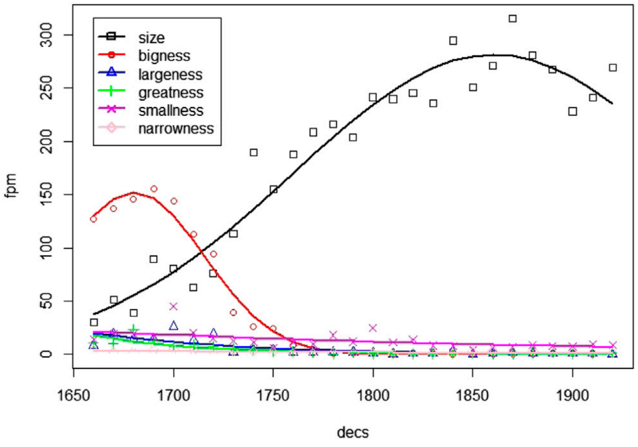


FIGURE 7 | Conventionalization by convergence on *size*.

4.2 Microanalysis: Linguistic Patterns of Paradigmatic Reduction

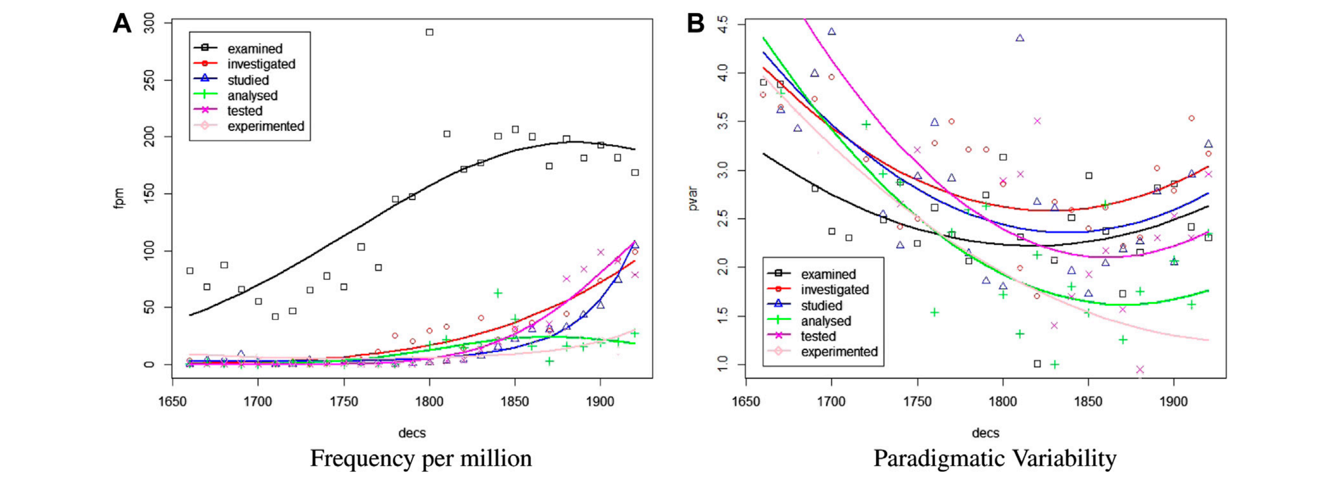
We observe two main (non-exclusive) mechanisms for limiting paradigmatic variability over time: CONVENTIONALIZATION—a word becoming the dominant choice within its neighbourhood by frequency (convergence), possibly replacing other, alternative words (substitution)—and DIVERSIFICATION, i.e., words within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

[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.



Frequency per million

Paradigmatic Variability

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.

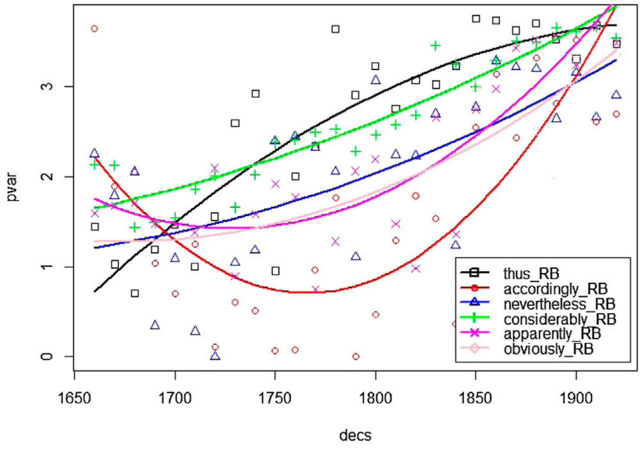


FIGURE 9 | Adverbs with increasing Paradigmatic Variability (pvar+).

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 28 January 2021
doi: 10.3389/fpsyg.2021.589930

Rational Adaptation in Using Conceptual Versus Lexical Information in Adults With Aphasia

Haley C. Dresang[1,2,3], Tessa Warren[4,5], William D. Hula[1,3] and Michael Walsh Dickey[1,2,3]*

[1] Department of Communication Science and Disorders, University of Pittsburgh, Pittsburgh, PA, United States, [2] Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA, United States, [3] VA Pittsburgh Healthcare System, Pittsburgh, PA, United States, [4] Department of Psychology, University of Pittsburgh, Pittsburgh, PA, United States, [5] Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA, United States

The information theoretic principle of rational adaptation predicts that individuals with aphasia adapt to their language impairments by relying more heavily on comparatively unimpaired non-linguistic knowledge to communicate. This prediction was examined by assessing the extent to which adults with chronic aphasia due to left-hemisphere stroke rely more on conceptual rather than lexical information during verb retrieval, as compared to age-matched neurotypical controls. A primed verb naming task examined the degree of facilitation each participant group received from either conceptual event-related or lexical collocate cues, compared to unrelated baseline cues. The results provide evidence that adults with aphasia received amplified facilitation from conceptual cues compared to controls, whereas healthy controls received greater facilitation from lexical cues. This indicates that adaptation to alternative and relatively unimpaired information may facilitate successful word retrieval in aphasia. Implications for models of rational adaptation and clinical neurorehabilitation are discussed.

Keywords: aphasia, rational adaptation, adaptation, verb naming, priming, event knowledge, co-occurrence statistics

OPEN ACCESS

Edited by:
Gina Kuperberg,
Tufts University, United States

Reviewed by:
Dave Kush,
Norwegian University of Science and Technology, Norway
Jennifer E. Mack,
University of Massachusetts Amherst, United States

****Correspondence:***
Haley C. Dresang
DresangH@Einstein.edu

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

***Received:** 31 July 2020*
***Accepted:** 05 January 2021*
***Published:** 28 January 2021*

Citation:
Dresang HC, Warren T, Hula WD and Dickey MW (2021) Rational Adaptation in Using Conceptual Versus Lexical Information in Adults With Aphasia. Front. Psychol. 12:589930. doi: 10.3389/fpsyg.2021.589930

INTRODUCTION

The language-processing system has often been viewed as relatively static and context-invariant, particularly by sentence comprehension models (e.g., Frazier, 1987; Bornkessel and Schlesewsky, 2006). However, recent evidence indicates that successful language processing, including sentence comprehension, is accomplished by an adaptive system (Ellis and Larsen-Freeman, 2009 for review; Gibson et al., 2013). There is growing evidence that the language system flexibly takes advantage of a wide array of sources of information to guide performance. These may include linguistic representations (grammatical categories, thematic roles, and lexical co-occurrence probabilities), contextual constraints, and knowledge of the relationships between words and real-world events (e.g., Caramazza and Zurif, 1976; McRae and Matsuki, 2009; Gibson et al., 2013; Kuperberg and Jaeger, 2016; Dresang et al., 2018). According to information theory, reliance on these information sources is governed by *the principle of rational adaptation* (Anderson, 1991; Howes et al., 2009), which states that a system can modify the degree to which it relies on different information sources in order to optimize behavior under different experimental conditions (e.g., Gibson et al., 2013) or disease states (Caramazza and Zurif, 1976; Gibson et al., 2015; Warren et al., 2017).

Language performance in individuals with aphasia provides a unique way to evaluate hypotheses regarding the adaptive use of information sources during language processing. People with aphasia have impairments in accessing and using linguistic information, but their stored conceptual-semantic knowledge is usually less impaired. The assumption that people with aphasia therefore rely more heavily on conceptual-semantic information undergirds both classic accounts of aphasic sentence processing (Caramazza and Zurif, 1976; Goodglass, 1976) and efficacious speech-language treatments (e.g., Boyle, 2010; Wambaugh et al., 2014; Edmonds, 2016). However, it remains unclear whether individuals with aphasia show evidence of rational adaptation during production tasks. The current study looks for evidence of rational adaptation during verb retrieval by people with aphasia. In doing so, it is one of few to investigate aphasic rational adaptation in reliance on stored representations of linguistic versus conceptual knowledge (see also Caramazza and Zurif, 1976), rather than in reliance on bottom-up linguistic input (e.g., Gibson et al., 2015; Warren et al., 2017). Verb-retrieval deficits are important to study because they are frequently observed in 70 percent of individuals with aphasia, across severity levels and syndrome classification types (Mätzig et al., 2009).

The rational adaptation principle is key to the noisy channel, or rational inference, account of sentence comprehension. According to this account, comprehenders perceive a sentence and immediately compute the probabilities associated with its possible intended messages. Their estimations of these probabilities adapt quickly to changes in the amount of noise or the reliability of cues in the context (Gibson et al., 2013). Gibson et al. (2013) demonstrated that increasing the rate of typos in an experiment led participants to rely less on linguistic form during sentence interpretation. Similarly, increasing the proportion of implausible sentences in the experiment led participants to rely less on meaning to guide sentence interpretation. Gibson et al. (2015) extended this work to adults with aphasia. They tested the hypothesis that during language comprehension, people with aphasia should rely more heavily on conceptual knowledge than healthy adults, because their linguistic impairments are more likely to introduce noise into their representations of the bottom-up linguistic input. In this study, like the 2013 one, sentence plausibility was crossed with sentence structure in such a way as to create implausible sentences that differed from plausible sentences by a small edit, and vice versa. For example, the implausible sentence *The mother gave the candle the daughter* is a single dropped *to* from the plausible sentence *The mother gave the candle to the daughter.* Greater reliance on conceptual knowledge would be shown by a stronger tendency to interpret implausible sentences like *The mother gave the candle the daughter* as if they were plausible near neighbors like *The mother gave the candle to the daughter*. This is because plausibility is conceptually driven. Gibson and colleagues showed that, like controls, people with aphasia were sensitive to the likelihood that a particular sentence structure would be distorted into its near neighbor (for example, they were more likely to stick with the literal interpretation of sentences with structures that were higher frequency or required an insertion rather than a deletion to become a plausible near neighbor). But across multiple types of sentences, people with aphasia were more likely than controls to interpret implausible sentences as their plausible near neighbors. That is, participants with aphasia showed a stronger influence of plausibility on their sentence interpretations than control participants did. This suggests they had rationally adapted to rely more heavily on conceptual knowledge, e.g., plausibility, than control participants. Warren et al. (2017) extended and replicated these findings using a different paradigm and a larger sample of people with aphasia.

These findings from experiments testing noisy channel processing in aphasia point to a flexible language processing system that is sensitive to aphasia-related changes in the reliability of cues to interpretation, including the likelihood of input distortion. But these studies have been relatively narrowly focused, in that the only language-related cue that has been tested is the form of the input, and the only outcome measure has been the ultimate interpretation of the sentence. A study by Hayes et al. (2016) tested a different kind of language-related cue, namely verb-argument requirements, during incremental comprehension. They pitted verb-argument information against plausibility in a visual-world study testing the anticipatory processing of event locations (e.g., "The child put/rode the bicycle in the park/pool."). They found that both the argument structure requirements of verbs and the plausibility of the event location guided the anticipatory processing of neurotypical adults across the lifespan, but only plausibility influenced anticipatory processing in adults with aphasia. This is consistent with aphasia increasing reliance on conceptual plausibility knowledge. However, the small size of their sample of participants with aphasia raises concerns about power, and this evidence (like that of Gibson et al., 2015 and Warren et al., 2017) speaks only to whether rational adaptation characterizes comprehension performance in aphasia.

The current study builds on a series of studies reported in Willits et al. (2015) that investigated unimpaired language users' reliance on language knowledge versus event knowledge across multiple tasks. The form of language knowledge they focused on is word co-occurrence frequency (Hale, 2001; Levy, 2008). We know that healthy language users utilize their stored knowledge of word co-occurrence in both comprehension and production (e.g., Wasow, 1997; Reali and Christiansen, 2007). There is also evidence that people with aphasia make use of lexical frequency and word co-occurrence information. In Gahl (2002), participants with fluent and anomic aphasia types showed sensitivity to lexical verb biases in a sentence plausibility judgment task. In a subsequent set of experiments, Dede (2013a,b) observed that the effects of lexical verb bias were greater in adults with aphasia than controls in an on-line self-paced reading task. These results suggest that word co-occurrence can influence sentence comprehension in aphasia. However, it remains unknown whether individuals with aphasia make use of word co-occurrence to facilitate naming.

Willis and colleagues (Willits et al., 2015) also tested the influence of event knowledge on language performance. In healthy adults, priming experiments have demonstrated that memory is structured such that multiple types of single-word cues allow immediate access to event knowledge (Ferretti et al., 2001;

McRae et al., 2001, 2005; Hare et al., 2009). In particular, verbs prime nouns that commonly fill their event-related thematic roles (agents, patients, instruments; Ferretti et al., 2001) and vice versa (McRae et al., 2001, 2005). In addition, Hare et al. (2009) found that nouns that denote common events (e.g., trip, accident) primed objects and agents typically involved in that event (trip–luggage; accident–policeman), and that location and instrument nouns primed event-related object and agent targets. Taken together, this evidence indicates that isolated verbs, event nouns, and thematic role/participant nouns activate conceptual event knowledge, resulting in facilitated naming of related concepts. This kind of direct event-related priming has not previously been tested in people with aphasia, but Dresang et al. (2019) found an indirect relation between event knowledge and verb naming. They found that conceptual knowledge of events positively predicted performance on verb naming and argument structure production tests in a sample of people with aphasia.

These two types of knowledge, word co-occurrence and event knowledge, are not always independent given that language is used to communicate information about events in the real world. But they can be dissociated. Willits et al. (2015) conducted two corpus analyses and found that past progressive verbs co-occur more frequently with locations than do past perfect verbs. However, this varied across individual verbs. Willits et al. (2015) capitalized on this variability to create verb-location stimuli with three levels: event related pairs with high co-occurrence probability, event related pairs with low co-occurrence probability, and unrelated pairs with low co-occurrence probability. These stimuli were tested in four behavioral tasks, to investigate whether young neurotypical adults lean more heavily on different sources of information under different task conditions. In two semantic tasks, plausibility judgment ("Rate how likely it is that the event or action described typically takes place in this location.") and semantic judgment ("Is this a location?"), results were driven by event knowledge. But in two language-production-focused tasks, primed verb naming ("Say the target word aloud.") and sentence completion ("Mary was visiting. . ."), effects were driven by word co-occurrence patterns. These findings support the notion that healthy adults prioritize conceptual event versus word co-occurrence information to different degrees depending on the task demands.

The current study extends this work with the goal of investigating rational adaptation in aphasia by testing the hypothesis that: because language impairment reduces the reliability of linguistic information for people with aphasia, they will rely more heavily on event knowledge and less heavily on linguistic knowledge as compared to unimpaired adults. Given that Willits et al. (2015) found that young neurotypical participants relied heavily on word co-occurrence information in a naming task, the current study used a naming task in people with aphasia. We expected to replicate Willits and colleagues' finding that healthy control participants exhibit stronger effects of word co-occurrence than event-relatedness on naming. But we further predicted that people with aphasia would show the opposite pattern and exhibit a larger facilitative effect of event relatedness than word co-occurrence on naming. The current study breaks new ground because evidence for rational adaptation in aphasia to date is limited to auditory sentence comprehension (Caramazza and Zurif, 1976; Schwartz et al., 1980, 1987; Gibson et al., 2015; Hayes et al., 2016; Warren et al., 2017). This study also has practical import because rational adaptation could be a mechanism behind the apparent efficacy of speech-language therapies that treat verb-retrieval deficits in people with aphasia by strengthening conceptual-semantic networks around verbs (e.g., Verb Network Strengthening Treatment [VNeST]; see Edmonds, 2016, for review). Demonstrating rational adaptation in verb naming would be a first step in showing that it may underlie these efficacious speech-language treatments and might be leveraged to develop more targeted neurorehabilitation methods, by determining what information cue types and experimental (learning) conditions facilitate verb retrieval. Finally, it contributes to studying a common, but relatively understudied, aspect of aphasia. 70 percent of individuals with aphasia experience chronic verb-retrieval deficits (Mätzig et al., 2009).

MATERIALS AND METHODS

Participants

Participants were 17 individuals with chronic aphasia due to unilateral left hemisphere stroke and 15 age-matched neurotypical controls. All participants were (1) native English speakers, (2) able to provide informed consent, (3) 25–85 years old, (4) (premorbidly) right-handed, (5) had no significant hearing loss or vision impairment that prevented them from completing the experimental tasks, (6) had no pre-existing or subsequent brain injury/stroke (e.g., to right-hemisphere regions for individuals with aphasia), and (7) had no history of progressive neurological or psychiatric disease, drug, or alcohol dependence, or significant mood or behavioral disorder.

In addition, all neurotypical participants passed a line-bisection visual screening, a binaural pure-tone hearing screening (0.5, 1, 2, and 4 KHz at 40 dB), a Mini-Mental State Examination cognitive screen (required 27/30; Folstein et al., 1975), and Raven's Colored Progressive Matrices non-linguistic cognitive screen (required 30/36; Raven, 1965). All individuals with aphasia were more than 6 months post-onset (range: 19–265 months; M = 95.8, SD = 62 months), had a Comprehensive Aphasia Test (CAT; Swinburn et al., 2004). Naming Modality T-score $\geq$ 40, and an overall mean T-score $<$ 70. Cognitive screening and general language assessment measures, including the CAT, were already available for the participants with aphasia, who all participated in Hula et al. (2020). Participants were not recruited if their T scores were less than 30 for the CAT Cognitive Screening semantic memory or recognition memory subtests. T scores under 30 would be indicative of frank auditory, visual, motor speech, or general cognitive deficits. Demographic participant characteristics are reported in **Table 1** for participants with aphasia and **Table 2** for age-matched controls.

Institutional Review Board approval was obtained, and all participants provided informed written consent and were compensated for their time.

TABLE 1 | Demographic characteristics of participants with aphasia.

Participant ID	Age	Sex	Education level	Years of education	Months post-onset	Years post-onset
7201	59	F	Graduate degree	20	132	11
7202	63	M	Bachelor's degree	14	265	22.08
7203	61	F	Master's degree	17	60	5
7204	55	M	High school	12	53	4.42
7205	52	M	High school	12	136	11.33
7206	78	F	Some graduate	13	114	9.5
7207	70	F	Some college	14	45	3.75
7208	76	M	Some college	14	138	11.5
7209	77	M	Law degree	19	53	4.42
7210	54	M	Bachelor's degree	16	83	6.92
7211	71	M	Some college	14	26	2.17
7212	55	M	Bachelor's degree	16	19	1.58
7213	68	M	High school	12	184	15.33
7214	53	F	Bachelor's degree	17	81	6.75
7215	71	M	Bachelor's degree	16	87	7.25
7216	72	M	Some college	14	60	5
7217	72	M	Some college	15	93	7.75
Summary	M = 65.12	5 F; 12 M		M = 15	M = 95.82	M = 7.99
	SD = 9.11			SD = 2.35	SD = 62	SD = 5.17

Materials

Experimental stimuli were adapted from existing normed stimuli for agent-, patient-, instrument-, and location-verb pairs (McRae et al., 2005). We developed items that paired 48 target verbs from McRae et al. (2005) with each of three kinds of noun primes. In the event-related condition, the primes were nouns that were strongly associated with the target verb's event but rarely appeared within four words of the verb in COCA's Wikipedia corpus (*pencil–WRITE*). Event-related primes were drawn from McRae et al. (2005) or from the USF Free Association Norms (Nelson et al., 1998) and consisted of agents, patients, instruments, or locations strongly associated with the target verb's event. Only seven event-related primes were among the top 100 noun collocates for their target verb (Maximum = 50th, M = 65th). In the lexical co-occurrence condition, the primes were nouns that co-occurred frequently with the target verbs but were not strongly associated with the target verb's event (*name–WRITE*). Lexical co-occurrence primes were selected from the nouns that most frequently appear within four words of the target verb in COCA's Wikipedia corpus (Davies, 2008). We chose the highest-ranked (M = 7–8th, range: 1st–25th) collocate that: (1) was not a paradigmatic participant in the verb's event (i.e., did not appear in McRae et al., 2005 or Nelson et al., 1998 norms), (2) did not form a compound with the verb (e.g., *board-WALK; school-WORK*), and (3) was not a high collocate of many verbs. Two of the authors confirmed these via independent judgments. In the baseline control condition, the primes were nouns that were neither associated with the verb's event nor often appeared near the verb (*water–WRITE*). They were generated by reassigning event-related primes to targets such that semantic relationships were minimized. Semantic distance between cue and target words was calculated using snaut semantic distance measure (Landauer and Dumais, 1997; Mandera et al., 2017) to confirm that lexical co-occurrence and baseline conditions were matched for lexical-semantic relatedness between cue and target words (t-statistic = −0.41; p-value = 0.68). Prime noun word length was balanced across conditions (all p's > 0.26). Following a Latin square design, conditions were counterbalanced and pseudorandomized across three presentation lists. See **Supplementary Materials** for a complete stimulus list that includes individual item properties.

TABLE 2 | Demographic characteristics of age-matched control participants.

Participant ID	Age	Sex	Education level	Years of education
7001	42	M	Tech college	14.5
7002	59	M	High school	12
7003	74	M	Bachelor's degree	16
7004	52	M	Bachelor's degree	16
7005	54	M	Bachelor's degree	16
7006	57	M	High school	12
7007	72	F	Master's degree	18
7008	64	F	Master's degree	18
7010	74	M	Master's degree	20
7011	68	F	Master's degree	22
7012	72	M	Bachelor's degree	16
7013	65	M	Law degree	19
7014	71	F	Master's degree	17
7015	52	M	Master's degree	22
7016	69	F	Master's degree	18
Summary	M = 63	5 F; 10 M		M = 17.1
	SD = 9.82			SD = 3

Testing Procedures

Each participant completed all three presentation lists, interleaved with other behavioral experiments with different tasks. Every presentation list began with six practice trials, followed by 48 experimental trials. Each trial began with a central fixation cross displayed for 25 milliseconds, followed by a noun prime (in lower-case blue letters) for 450 milliseconds, followed by a central mask (&&&&&&&) for 50 milliseconds, and then the verb naming target (in upper-case black letters) remained on the screen until the participant provided a response or indicated inability to do so. An audio click was presented simultaneously with the target verb for the purpose of manual measurement of naming latencies. Because naming is challenging for people with aphasia and they do not always process incoming linguistic information efficiently (Goodglass and Wingfield, 1997; Faroqi-Shah et al., 2010; Silkes et al., 2020), we used a relatively long prime duration (longer than the standard 200 milliseconds for lexical decision tasks). In addition, within each presentation list, we blocked items according to whether the primes most naturally preceded the verb (i.e., event prime agents and instruments; preceding collocates; e.g., *actor–PERFORM, ax–CHOP*) or followed it (i.e., event prime patients and locations; following collocates; e.g., *customer–SERVE, gym–EXERCISE*). Following McRae et al. (2005), trials were separated by a 1,500-millisecond blank screen. Participants were instructed to name the target verb aloud as quickly and accurately as possible. An external microphone recorded naming responses in Audacity® , and accuracy and latency measurements were coded by hand.

Accuracy and response time were the dependent variables. Trained raters followed procedures outlined by the Philadelphia Naming Test (Roach et al., 1996) in order to determine the first complete attempt, which was then scored for both accuracy and latency. Accuracy was coded as correct or incorrect. Participants with aphasia who had concomitant motor speech impairments (e.g., dysarthria, speech apraxia) were allowed one sound omission, addition, or substitution per response when considering correctness (Roach et al., 1996). Response time (latency) was measured in milliseconds from the time in which the target word was displayed (with audio click) until the participant began to produce their first complete response. These scoring procedures followed the conventional procedures used for the Philadelphia Naming Test (Roach et al., 1996). Two raters measured the critical time points and calculated the naming latency for each trial. They had 93.77 percent agreement on a randomly selected sample of 10 percent of the items (ratings within 50 milliseconds of each other constituted agreement). The raters discussed these discrepancies and reached 100 percent agreement. The degree of priming was measured by comparing the latency of event and lexically related word pairs to baseline, unrelated trials.

Analyses

Data were analyzed using Bayesian mixed effects regression models, which were created in the Stan computational framework (Carpenter et al., 2017; http://mc-stan.org/) accessed with the brms package (Bürkner, 2017). Trial-level naming accuracy served as the outcome variable for two logit-link bernoulli family models, and trial-level naming response time served as the outcome variable for two ex-gaussian family models. Model 1 examined naming accuracy between participant groups; Model 2 examined naming response time between groups; Model 3 examined accuracy in participants with aphasia; and Model 4 examined response time in participants with aphasia. Estimates of facilitation under each prime condition (baseline, event-related, and lexical co-occurrence) were assessed in terms of the assumptions of normality, homoscedasticity, linearity, and the presence of outliers. To address outliers and to achieve model convergence, latency observations above the 95th percentile for each group were trimmed. From 3,200 trials, 89 trials were trimmed (2.8% of the original data), resulting in a total of 3,111 observations across both groups. Finally, only accurate trials were examined in Model 2 and Model 4, for which response time was the dependent variable (Forster, 1976).

The model structures are discussed below. Each parameter was given dispersed starting values and a vague prior, thus allowing the Bayesian estimation process to explore the full parameter space and provide conservative estimates of posterior distributions (McElreath, 2020). For each model, four Hamiltonian Markov chain Monte Carlo (MCMC) chains were run for 20,000 samples, with half of the iterations discarded as warm-up and 10,000 iterations monitored for convergence and parameter estimation. There was no thinning and no divergent transitions for any of the models. For each model, MCMC convergence was assessed graphically by inspection of the autocorrelation and trace plots, as well as statistically using the Gelman-Rubin potential scale reduction statistic (R̂) and the number of effective samples. The R̂ statistic is a ratio of the variance within each chain to the variance pooled across chains. $\hat{R}$ values close to 1 indicate satisfactory convergence of the chains to a stable distribution (Gelman et al., 2013). ESS factors out the autocorrelation in the observed MCMC chains and estimates the number of independent samples that would achieve the same degree of precision for the parameter estimates (Carpenter et al., 2017). Large ESS values indicate satisfactory convergence. The posterior distributions are summarized by the estimated parameters and 95% highest density credible intervals (HDI). The HDI is comparable to the frequentist confidence interval and is determined as the narrowest interval containing the assigned proportion of the posterior distribution's probability mass within which all values have a higher probability density than any values outside the interval (see Fergadiotis et al., 2019 for further explanation).

Post hoc pairwise comparisons were conducted using the emmeans package in R (Lenth, 2020) for Models 1–2 in order to evaluate the reliability of every potential condition-specific priming effect for both groups of participants.

First, naming accuracy was compared between participant groups with and without aphasia. The outcome variable was trial-level verb naming accuracy. Fixed effects were group assignment (participants with aphasia coded as 0 versus neurotypical controls coded as 1) and prime condition (event-relatedness versus baseline; and lexical co-occurrence versus baseline), and two interaction effects (group *x* event condition; group *x* lexical condition). The effect of prime condition was dummy coded with the baseline condition as the reference level. Specifically, the prime condition fixed effect was coded with two contrasts across the three levels of the variable, such that each condition of interest was compared to the baseline prime condition. Random intercepts were included for subjects and items.

Random slopes were included for condition within subjects and group within items. More complex random effects structures failed to converge.

Second, naming response time was compared between participant groups with and without aphasia. Fixed effects and random effects structures were the same as for Model 1, but the outcome variable was response time (latency) from word presentation to participant response, in milliseconds.

Third, naming accuracy was examined in greater detail for participants with aphasia. The outcome variable was trial-level verb naming accuracy. Each prime condition was a fixed effect. Prime conditions were coded the same way as for Models 1 and 2. Random intercepts were included for subjects and items. Random slopes were included for condition within subjects and aphasia severity within items.

Fourth, naming response time was examined in greater detail for participants with aphasia. Fixed effects and random effects structures were the same as for Model 3, but the outcome variable was response time (latency) from word presentation to participant response, in milliseconds.

RESULTS

Descriptive statistics of group level accuracy and response time across each prime condition are reported in **Table 3**. The trace plots for all parameters demonstrated rapid convergence and were stationary relative to the parameter means. The autocorrelation plots corroborated this assessment and showed minimal autocorrelation for all four models. These plots and all posterior predictive checks are provided in the **Supplementary Material**. The $\hat{R}$ statistic and number of effective samples for each parameter indicated satisfactory convergence and MCMC mixing. These statistics are reported in **Tables 4–7**. **Tables 4–7** also provide the point estimates and 95% credible intervals for each parameter. The posterior predictive checks and histograms of the posterior distributions for the estimates of interest are provided below. Only differences where less than 20% of the posterior probability distributions did not overlap zero are interpreted below (Hair et al., 2009; Hazelrigg, 2009).

Model 1: Primed Naming Accuracy Between Participant Groups

Group (aphasia versus control) reliably predicted trial-level primed naming accuracy (β = 5.41, EE = 1.22, and 95% HDI = [3.06, 7.83]), with participants with aphasia (M = 0.790, SD = 0.407) performing less well than controls (M = 0.995, SD = 0.067). **Figure 1** shows the posterior probability distribution for the group effect. Furthermore, group interacted with prime condition in predicting naming accuracy, such that aphasia amplified the facilitation of event-related cues (β = −1.32, EE = 0.88, and 95% HDI = [−3.13, 0.37]) but lack of aphasia (i.e., the control group) amplified the effect of lexical co-occurrence cues (β = 1.35, EE = 1.51, and 95% HDI = [−1.32, 4.46]). Although both of these credible intervals overlap with zero, there is a 94.57 percent chance that the interaction between group and event facilitation is less than zero (**Figure 2**), and an 82.62 percent chance that the group and lexical co-occurrence interaction is greater than zero (**Figure 3**). This suggests that the observed interaction between group and event facilitation was robust, but the interaction between group

TABLE 3 | Descriptive statistics of group level accuracy (percent correct) and response time (seconds) across prime conditions.

Prime Condition		Participants with aphasia		Control participants		Grand total	
		M	SD	M	SD	M	SD
Baseline	Accuracy	0.774	0.419	0.996	0.064	0.885	0.242
	Latency	0.779	0.271	0.598	0.139	0.689	0.205
Event	Accuracy	0.805	0.397	0.990	0.098	0.898	0.248
	Latency	0.822	0.271	0.596	0.159	0.709	0.215
Lexical	Accuracy	0.792	0.406	0.999	0.037	0.896	0.222
	Latency	0.822	0.280	0.589	0.127	0.706	0.204
Grand total	Accuracy	0.790	0.407	0.995	0.067	0.893	0.237
	Latency	0.814	0.269	0.593	0.132	0.701	0.208

These values reflect the descriptive statistics after excluding outliers.

TABLE 4 | Model 1 primed naming accuracy population-level effects for participants with aphasia and age-matched control participants.

	Estimate	Est. error	Lower 95% HDI	Upper 95% HDI	$\hat{R}$	Bulk ESS	Tail ESS
(Intercept)	1.84	0.54	0.79	2.94	1	1836	3811
Group	5.41	1.22	3.06	7.83	1	3055	4291
Event-related prime	0.3	0.21	−0.14	0.69	1	8274	6879
Lexical co-occurrence prime	0.26	0.26	−0.22	0.80	1	6277	5352
Group: Event prime	−1.32	0.88	−3.13	0.37	1	6724	5599
Group: Lexical prime	1.35	1.51	−1.32	4.46	1	8408	5913

HDI, Highest density credible interval. $\hat{R}$ = The potential scale reduction factor on split chains (at convergence, $\hat{R}$ = 1). ESS, Effective sample size.

TABLE 5 | Model 1 naming accuracy pairwise comparisons.

Contrast	Estimate	Lower 95% HDI	Upper 95% HDI
Aphasia baseline – control baseline	−5.337	−7.831	−3.060
Aphasia baseline – aphasia event	−0.293	−0.686	0.144
Aphasia baseline – control event	−4.327	−6.557	−2.349
Aphasia baseline – aphasia lexical	−0.251	−0.799	0.217
Aphasia baseline – control lexical	−6.851	−10.345	−3.896
Control baseline – aphasia event	5.040	2.844	7.576
Control baseline – control event	0.979	−0.623	2.794
Control baseline – aphasia lexical	5.074	2.791	7.621
Control baseline – control lexical	−1.485	−4.736	1.068
Aphasia event – control event	−4.029	−6.3	−2.140
Aphasia event – aphasia lexical	*0.041*	*−0.519*	*0.549*
Aphasia event – control lexical	−6.545	−10.089	−3.635
Control event – aphasia lexical	4.075	2.046	6.284
Control event – control lexical	*−2.453*	*−5.5*	*0.040*
Aphasia lexical – control lexical	−6.598	−10.092	−3.616

HDI = Highest density credible interval. Median point estimates displayed. Results are given on the log odds ratio scale.

TABLE 7 | Model 2 naming response time pairwise comparisons.

Contrast	Estimate	Lower 95% HDI	Upper 95% HDI
Aphasia baseline – control baseline	0.0484	0.0237	0.0746
Aphasia baseline – aphasia event	−0.0026	−0.0073	0.0019
Aphasia baseline – control event	0.0488	0.0238	0.0754
Aphasia baseline – aphasia lexical	−0.0008	−0.0054	0.0038
Aphasia baseline – control lexical	0.0501	0.0244	0.0757
Control baseline – aphasia event	−0.0510	−0.0762	−0.0250
Control baseline – control event	0.0002	−0.0046	0.0049
Control baseline – aphasia lexical	−0.0492	−0.0758	−0.0248
Control baseline – control lexical	0.0015	−0.0035	0.0062
Aphasia event – control event	0.0513	0.0243	0.0753
Aphasia event – aphasia lexical	*0.0017*	*−0.0029*	*0.0065*
Aphasia event – control lexical	0.0525	0.0264	0.0778
Control event – aphasia lexical	−0.0495	−0.0759	−0.0244
Control event – control lexical	*0.0013*	*−0.0038*	*0.0061*
Aphasia lexical – control lexical	0.0508	0.0247	0.0758

HDI, Highest density credible interval. Median point estimates displayed.

and lexical co-occurrence facilitation was relatively unreliable. Based on *post hoc* pairwise comparisons, neurotypical controls received greater priming following lexical co-occurrence cues than event-related cues (β = −2.45, 95% HDI = [−5.5, 0.04]). This comparison did not show robust differences in participants with aphasia. The full set of results is reported in **Table 4**. The full set of pairwise comparisons is reported in **Table 5**.

Model 2: Primed Naming Response Time Between Participant Groups

Group (aphasia versus control) reliably predicted trial-level primed naming response time (β = −0.274, EE = 0.072, and 95% HDI = [−0.415, −0.133]; **Figure 4**), with participants with aphasia (M = 0.814 s, SD = 0.269) performing slower than controls (M = 0.593 s, SD = 0.132). The main effects of the prime conditions and their interactions with group were small and not credibly different from zero. Based on *post hoc* pairwise comparisons, neither neurotypical controls nor participants with aphasia showed robust differences in response time following lexical co-occurrence versus event-related cues. The full set of results is reported in **Table 6**. The full set of pairwise comparisons is reported in **Table 7**.

Model 3: Primed Naming Accuracy in Participants With Aphasia

Both prime conditions predicted naming accuracy in participants with aphasia, with individuals producing more correct responses after both event-related (M = 0.805, SD = 0.397) and lexical co-occurrence primes (M = 0.792, SD = 0.406), as compared to unrelated baseline (M = 0.774, SD = 0.419). Although the 95% credible intervals for both of these effects overlap with zero, 94.69 percent of the posterior probability distribution for event primes (β = 0.36, EE = 0.23, and 95% HDI = [−0.10, 0.78], **Figure 5**) and 95.02 percent of the posterior distribution for lexical primes (β = 0.41, EE = 0.27, and 95% HDI = [−0.11, 0.94], **Figure 6**) exceed zero. The full set of results is reported in **Table 8**.

Model 4: Primed Naming Response Time in Participants With Aphasia

No reliable priming in response time was observed in participants with aphasia for either event-related (β = 0.017, EE = 0.020, and

TABLE 6 | Model 2 primed naming response time population-level effects for participants with aphasia and age-matched control participants.

	Estimate	Est. error	Lower 95% HDI	Upper 95% HDI	R̂	Bulk ESS	Tail ESS
(Intercept)	−0.272	0.052	−0.045	−0.009	1	1928	3531
Group	−0.274	0.072	−0.415	−0.133	1	1795	3438
Event-related prime	0.008	0.018	−0.027	0.0431	1	4452	6071
Lexical co-occurrence prime	0.000	0.018	−0.034	0.036	1	4360	6273
Group : Event prime	−0.011	0.015	−0.041	0.016	1	12723	8405
Group : Lexical prime	−0.009	0.014	−0.036	0.019	1	12647	7708

HDI, Highest density credible interval. R̂ = The potential scale reduction factor on split chains (at convergence, R̂ = 1). ESS, Effective sample size.

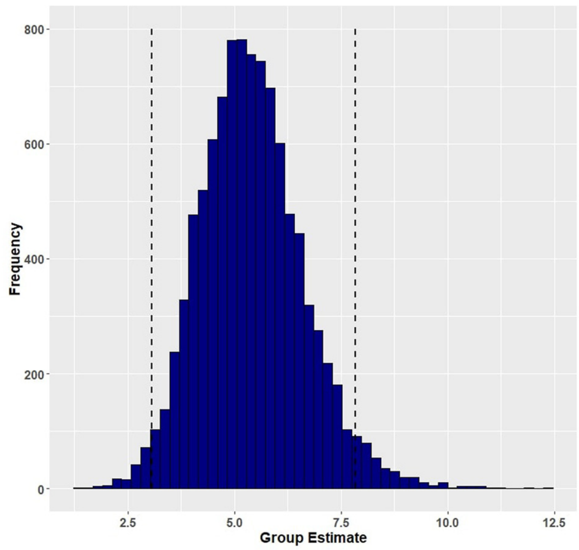


FIGURE 1 | Posterior distribution and 95% highest density intervals (HDIs) of the fixed effect of group from Model 1 (primed accuracy for participants with aphasia and healthy controls). Dashed lines mark the 95% highest density intervals (HDIs) for the posterior distribution.

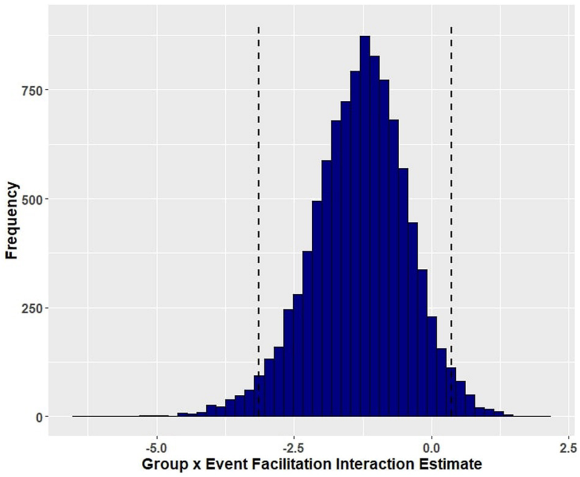


FIGURE 2 | Posterior distribution and 95% highest density intervals of the interaction effect of group and event-related facilitation from Model 1 (primed accuracy for participants with aphasia and healthy controls). Dashed lines mark the 95% highest density intervals (HDIs) for the posterior distribution.

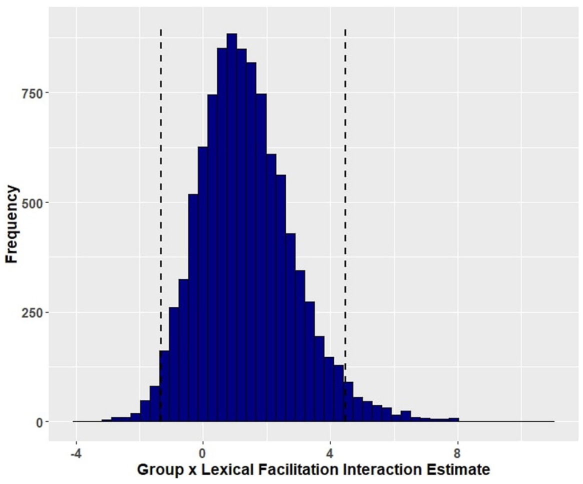


FIGURE 3 | Posterior distribution and 95% highest density intervals of the interaction effect of group and lexical co-occurrence facilitation from Model 1 (accuracy for participants with aphasia and healthy controls). Dashed lines mark the 95% highest density intervals (HDIs) for the posterior distribution.

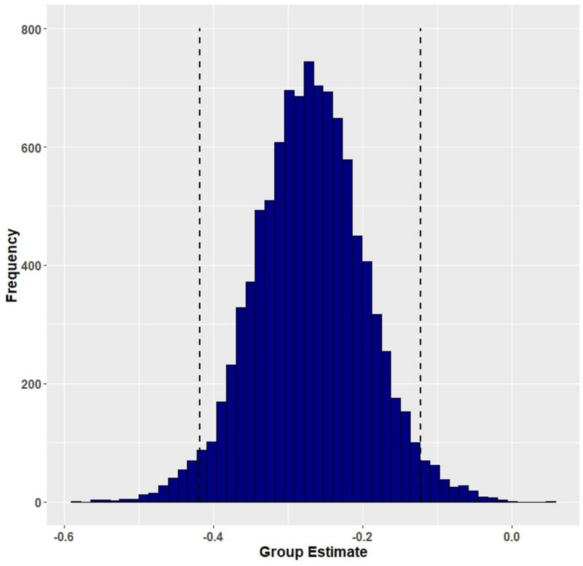


FIGURE 4 | Posterior distribution and 95% highest density intervals (HDIs) of the fixed effect of group from Model 2 (primed response time for participants with aphasia and healthy controls). Dashed lines mark the 95% highest density intervals (HDIs) for the posterior distribution.

95% HDI = [−0.024, 0.056]) or lexical co-occurrence conditions (β = 0.010, EE = 0.017, and 95% HDI = [−0.024, 0.044]). The full set of results is reported in **Table 9**.

DISCUSSION

The purpose of this study was threefold. First, it aimed to replicate and extend findings from Willits et al. (2015) that indicate naming is a language-focused task in which healthy language users prioritize knowledge of word co-occurrence over conceptual event relatedness. Second, it examined the hypothesis, grounded in rational adaptation, that during verb naming adults with aphasia would rely more heavily on conceptual event-related cues and less heavily on lexical co-occurrence cues, compared to neurotypical controls. Third, aphasic behavior was examined more closely to assess

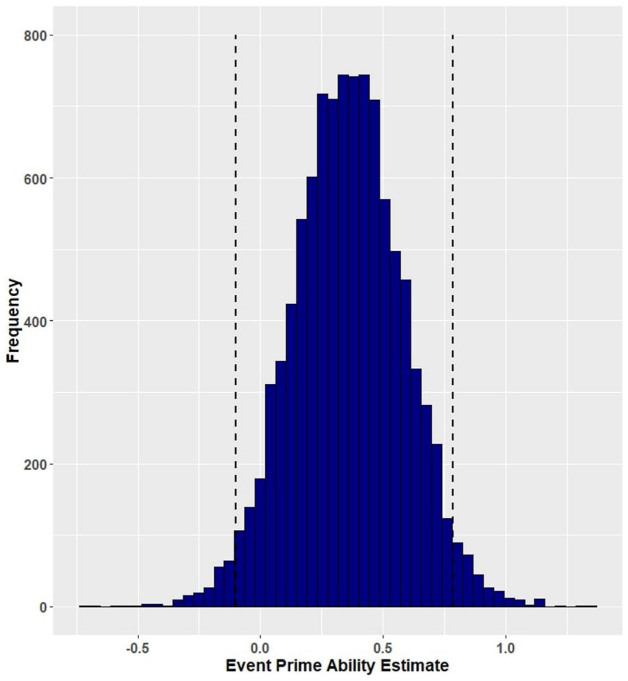


FIGURE 5 | Posterior distribution and 95% highest density intervals (HDIs) of the fixed effect of event-related facilitation from Model 3 (primed accuracy for participants with aphasia). Dashed lines mark the 95% highest density intervals (HDIs) for the posterior distribution.

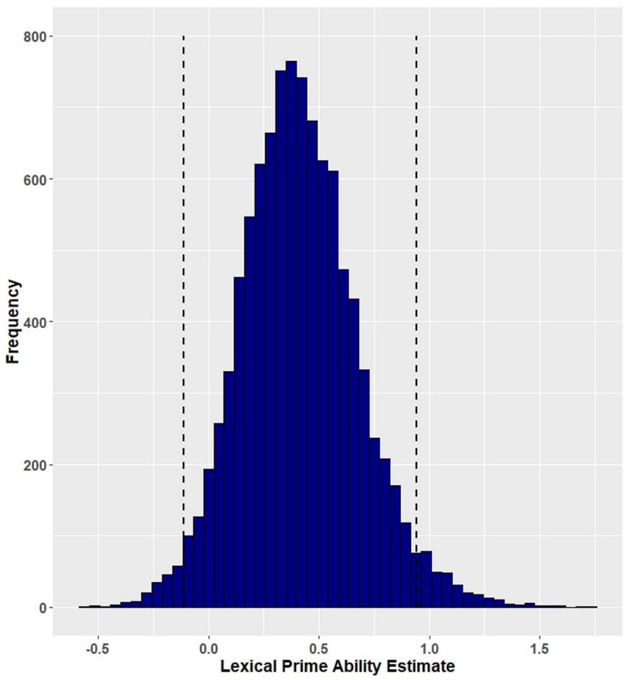


FIGURE 6 | Posterior distribution and 95% highest density intervals (HDIs) of the fixed effect of lexical co-occurrence facilitation from Model 3 (primed accuracy for participants with aphasia). Dashed lines mark the 95% highest density intervals (HDIs) for the posterior distribution.

differences in conceptual versus lexical facilitation within the sample of individuals with aphasia. The findings are summarized below, and their implications are discussed in relation to rational adaptation hypotheses and potential clinical directions moving forward.

First, our results from neurotypical controls were broadly consistent with findings from Willits et al. (2015), who observed that participants showed robust facilitation from frequently co-occurring words in naming tasks. The current sample of older neurotypical adults showed similar patterns to Willits and colleagues' college-aged participants, with greater facilitation of naming in lexical-prime conditions compared to event-related conditions. This is confirmed by the pairwise comparison results. However, in the current study, these patterns appeared in accuracy rather than latency measures. Our speculation is that this might be driven by a speed-accuracy trade off, given both the high variability in latency in the current sample and previous evidence that older adults are likely to prioritize accuracy over speed (Ratcliff et al., 2004; Starns and Ratcliff, 2010). These findings suggest that unimpaired language users prioritize linguistic information (specifically, word co-occurrence frequency information) more than conceptual cues when performing naming tasks. This is consistent with findings from language production studies showing that wordform retrieval is especially sensitive to lexical frequency effects (e.g., Jescheniak and Levelt, 1994) and that high-frequency word collocations speed processing (McDonald and Shillcock, 2003; Arnon and Snider, 2010; Smith and Levy, 2013). Our results are also consistent with evidence supporting task-based rational adaptation, which contends that language users rely on the most informative source of knowledge to optimize their behavior on the task at hand (Anderson, 1991; Howes et al., 2009).

Next, we examined the effect of aphasia on primed verb naming. As expected, adults with aphasia consistently named verbs more slowly and less accurately than controls for all prime conditions. This is consistent with a large body of literature that demonstrates verb-retrieval deficits in individuals with aphasia (e.g., Berndt et al., 1997; Jonkers and Bastiaanse, 2007; Rofes et al., 2015). Response latencies showed no other effects, but verb retrieval accuracy did. Importantly, presence of aphasia interacted with prime condition in predicting verb retrieval accuracy. Participants with aphasia received an amplified facilitation effect, or greater priming, from conceptual event-related cues compared to the control group. This group by conceptual priming interaction effect was strongly reliable, with approximately 95% of the posterior probability distribution >0. There was a weaker effect in the opposite direction for lexical co-occurrence (83% of the posterior probability distribution >0): the control group received somewhat greater priming from lexical co-occurrence cues compared to participants with aphasia. However, models that examined performance only in participants with aphasia found robust facilitation effects of both conceptual event and lexical co-occurrence cues. These accuracy results extend evidence from healthy adults to individuals with aphasia: nouns prime verbs that denote events in which the nouns are commonly involved (e.g., McRae et al., 2005, 2001). This extension is critical because it highlights the importance of conceptual event knowledge in disordered language processing, which is consistent with the hypothesized mechanisms underlying efficacious speech-language treatments targeting verbs (e.g., VNeST: Edmonds, 2016; see further discussion below). Of note, the relatively unreliable interaction suggesting that lexical co-occurrence priming might be stronger in the control group than in participants with aphasia is not consistent with previous evidence suggesting that aphasia may magnify the effects of lexical frequency on language performance (Gahl, 2002; Dede, 2013a,b).

Taken together, the findings of this experiment are consistent with previous evidence of rational adaptation in aphasia and suggest that the evidence base may extend beyond sentence comprehension to verb naming. In contrast to previous investigations of rational adaptation in aphasia, this study examined stored knowledge of linguistic representations – specifically, stored knowledge of word co-occurrences – rather than bottom-up linguistic input, such as the literal sentence form (Gibson et al., 2015; Warren et al., 2017). Another critical contribution of this study is that it separately examines automatic facilitatory effects of linguistic and conceptual information types, which are independent of one another in this study design. Much of the previous evidence that is consistent with rational adaption in aphasia could be explained by the fact that people with aphasia show less reliance on linguistic knowledge than neurotypicals (e.g., Hayes et al., 2016; Warren et al., 2017). This prediction is not unique to rational adaptation, nor is it surprising given that aphasia, by definition, impairs language. For example, although linguistic and conceptual knowledge were also independent in the study by Hayes et al. (2016), they only found evidence that people with aphasia relied less on linguistic knowledge than neurotypical controls did. The current study goes beyond this in showing an increase in the use of conceptual knowledge for people with aphasia. Although overall naming performance was poorer in people with aphasia, they showed greater priming from conceptually related words than neurotypical controls did. To be clear, this finding does not necessitate rational adaptation; it could be the case that impairing one type of knowledge could change the relative utility of other types of knowledge for a structural reason, for example because one source of knowledge had been inhibiting another. Still, rational adaptation provides a straightforward and elegant account of these data.

If rational adaptation is driving these effects, assessing the mechanisms that underlie it and the potential tradeoffs between conceptual and lexical information will be informative as to what cognitive processes or routes rational adaptation might be operating over. For example, it could be reweighting different routes to lexical access, or alternatively, successive stages of lexical access. If it is reweighting lexical-access routes, the current findings may be evidence that the conceptual system – which some grounded-cognition-inspired models of meaning (Kelter and Kaup, 2012) and highly interactive/interconnected connectionist models of lexical representation (Plaut et al., 1996) have argued provides an indirect, alternate, and typically less efficient route to access lexical wordform information – is a relatively more efficient route to wordform access for people with aphasia. If rational adaptation is re-weighting inputs to successive stages of lexical access, then the nature of a lexical-access deficit may affect how successful rational adaptation is. Individuals with aphasia can experience deficits to different stages of lexical access, affecting either conceptual-to-lexical or lexical-to-phonological mapping, or both (Foygel and Dell, 2000). Individuals with more impaired conceptual-to-lexical mapping (s-weight) might receive less priming from conceptual event-related cues than individuals with relatively spared lexical-semantic processing. Of note, the degree of lexical-semantic or lexical-phonological impairment is associated with neurological variability such as lesion site and white-matter connectivity (Dell et al., 2013; Hula et al., 2020); this neurological variability may underlie person-level variation in degree of conceptual priming. Further research is needed to assess potential mechanisms that underlie the role of conceptual information in aphasic language processing.

In addition, rational adaptation predicts that increased damage to the language system would result in increased

TABLE 8 | Model 3 primed naming accuracy population-level effects for participants with aphasia.

	Estimate	Est. error	Lower 95% HDI	Upper 95% HDI	$\hat{R}$	Bulk ESS	Tail ESS
(Intercept)	1.66	0.60	0.47	2.87	1	1347	2349
Event-related prime	0.36	0.23	−0.10	0.78	1	8469	6814
Lexical co-occurrence prime	0.41	0.27	−0.11	0.94	1	5683	5312

HDI, Highest density credible interval. $\hat{R}$ = The potential scale reduction factor on split chains (at convergence, $\hat{R}$ = 1). ESS, Effective sample size.

TABLE 9 | Model 4 primed naming response time population-level effects for participants with aphasia.

	Estimate	Est. error	Lower 95% HDI	Upper 95% HDI	$\hat{R}$	Bulk ESS	Tail ESS
(Intercept)	−0.155	0.083	−0.048	0.022	1	1613	3030
Event-related prime	0.017	0.020	−0.024	0.056	1	9456	6504
Lexical co-occurrence prime	0.010	0.017	−0.024	0.044	1	10178	7234

HDI, Highest density credible interval. $\hat{R}$ = The potential scale reduction factor on split chains (at convergence, $\hat{R}$ = 1). ESS, Effective sample size.

adaptive reliance on conceptual information types in people with aphasia. Applying this prediction to the current study, we would expect that aphasia severity would interact with information cue type, such that greater severity would amplify the facilitation effects of conceptual event-related cues but reduce the effects of lexical cue facilitation on verb naming. In the current investigation, overall aphasia severity was not included as a covariate predictor due to its multicollinearity with fixed effects of greater theoretical interest, such as the degree of facilitation from different information cue types. Because including aphasia severity in our models attenuated the magnitude of facilitation effects, our analyses were unable to test this prediction in the current (limited) sample. This potential limitation and the relatively small magnitude effects highlight the need for larger samples of participants with aphasia in future studies.

It is also the case, as suggested in Silkes et al. (2020), that the level of linguistic task complexity could also contribute to whether and to what degree an individual with aphasia might rely on conceptual information. Silkes et al. (2020) hypothesized that more complex tasks may be associated with decreased efficiency in engaging linguistic representations, prompting greater recruitment of more broadly distributed representations such as conceptual ones. Future work might therefore examine linguistic tasks that vary in complexity, for example comparing potential adaptation during (speeded) primed verb naming to untimed sentence completion tasks (Willits et al., 2015). The mechanisms underlying rational adaptation may be informed by a more thorough characterization of the locus and severity of behavioral and neurological impairments in individuals who receive facilitation from conceptual information during lexical access. In addition, future research might examine whether adults with aphasia show evidence of rational adaptation during language production with higher ecological validity, such as connected discourse.

Finally, the current findings may provide new evidence for mechanisms involved in efficacious aphasia interventions. A key finding from this study is that participants with aphasia exhibited a greater degree of naming facilitation from conceptual cues than neurotypical controls did. This result has critical implications for aphasia rehabilitation, because it aligns with the hypothesized mechanism of action for speech-language treatments like Semantic Feature Analysis (SFA; Boyle, 2010), SFA for Actions (Wambaugh et al., 2014), and Verb Network Strengthening Treatment (VNeST; Edmonds, 2016). Specifically, these treatments systematically activate information conceptually related to target words, based on evidence for bidirectional facilitation effects between event-related verbs and thematic roles (Ferretti et al., 2001; McRae et al., 2005). These interventions promote improved lexical retrieval ability for treated nouns (SFA) and verbs (SFA for Actions, VNeST), and there is evidence that improvements can generalize beyond trained items to the lexical retrieval of untreated words, sentences, and performance in connected discourse (e.g., Rider et al., 2008; Edmonds, 2016; Quique et al., 2019). Our rational adaption findings thus demonstrate the likely mechanism driving conceptual/semantic-based aphasia rehabilitation: If people with aphasia already exhibit reliance on conceptual information to retrieve words, then treatment can take advantage of this established mechanism by strengthening conceptually driven activation/retrieval processes. Future efforts to characterize the specific psycholinguistic and neurocognitive systems involved in this adaptation and to identify the types of patients who are most likely to engage adaptive strategies to rely more on conceptual knowledge will advance both our theoretical and clinical approaches to aphasia rehabilitation.

CONCLUSION

This study found evidence suggesting that individuals with aphasia may rationally adapt to their language impairments by relying on conceptual cues to a greater extent than healthy controls do. Participants with aphasia received an amplified facilitation effect from conceptual event-related cues compared to the control group, whereas naming in the control group showed a tendency to be more facilitated by lexical co-occurrence information, consistent with previous findings regarding neurotypical reliance on lexical information in verb naming (e.g., Willits et al., 2015). These findings suggest that adaptation to alternative and relatively unimpaired information types may facilitate successful word retrieval in adults with aphasia. Further work should continue to assess potential mechanisms that might underlie rational adaptation in aphasic language, as well as the specific psycholinguistic mechanisms by which conceptual information sources may facilitate verb retrieval. This line of research will ultimately help advance neurorehabilitation and speech-language interventions.

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 University of Pittsburgh Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

HD, TW, and MD conceived of the presented idea. HD and TW developed the experimental materials. HD carried out the experiment, supervised by MD and HD performed the computations and WH verified the analytical methods. HD wrote the manuscript with support from TW and MD.

All authors discussed the results and contributed to the final manuscript.

FUNDING

This research was supported through funding received from the National Institutes of Health (NIH) National Institute on Deafness and Other Communication Disorders (NIDCD) F31DC017896; the William Orr Dingwall Foundation's Dissertation Fellowship in the Cognitive, Clinical, and Neural Foundations of Language; the Council of Academic Programs in Communication Sciences and Disorders (CAPCSD) Ph.D. Scholarship; the Audrey Holland Endowed Student Resource Fund; and the University of Pittsburgh's School of Health and Rehabilitation Sciences (SHRS) Ph.D. Student Award for dissertation research.

ACKNOWLEDGMENTS

The authors thank the Language and Brain Lab for experimental support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.589930/full#supplementary-material

REFERENCES

Anderson, J. (1991). Is human cognition adaptive? *Behav. Brain Sci.* 14, 471–517. doi: 10.1017/s0140525x00070801

Arnon, I., and Snider, N. (2010). More than words: frequency effects for multi-word phrases. *J. Mem. Lang.* 62, 67–82. doi: 10.1016/j.jml.2009.09.005

Berndt, R. S., Mitchum, C. C., and Wayland, S. (1997). Patterns of sentence comprehension in aphasia: a consideration of three hypotheses. *Brain Lang.* 60, 197–221. doi: 10.1006/brln.1997.1799

Bornkessel, I., and Schlesewsky, M. (2006). The extended argument dependency model: a neurocognitive approach to sentence comprehension across languages. *Psychol. Rev.* 113:787. doi: 10.1037/0033-295x.113.4.787

Boyle, M. (2010). Semantic feature analysis treatment for aphasic word retrieval impairments: what's in a name? *Topics Stroke Rehabil.* 17, 411–422. doi: 10.1310/tsr1706-411

Bürkner, P. (2017). brms: an R Package for bayesian multilevel models using stan. *J. Stat. Softw.* 80, 1–28. doi: 10.18637/jss.v080.i01

Caramazza, A., and Zurif, E. B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: evidence from aphasia. *Brain Lang.* 3, 572–582. doi: 10.1016/0093-934X(76)90048-1

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

Davies, M. (2008). *The Corpus of Contemporary American English (COCA): 560 Million Words, 1990-Present*. Available online at: https://corpus.byu.edu/coca/. doi: 10.1017/9781108770750.002 (accessed May 3, 2018).

Dede, G. (2013a). Verb transitivity bias affects on-line sentence reading in people with aphasia. *Aphasiology* 27, 326–343. doi: 10.1080/02687038.2012.725243

Dede, G. (2013b). Effects of verb bias and syntactic ambiguity on reading in people with aphasia. *Aphasiology* 27, 1408–1425. doi: 10.1080/02687038.2013.843151

Dell, G. S., Schwartz, M. F., Nozari, N., Faseyitan, O., and Coslett, H. B. (2013). Voxel-based lesion-parameter mapping: identifying the neural correlates of a computational model of word production. *Cognition* 128, 380–396. doi: 10.1016/j.cognition.2013.05.007

Dresang, H. C., Dickey, M. W., and Warren, T. (2018). "Event-referent activation in the visual world: persistent activation is guided by both lexical and event representations," in *Proceedings of the 31st Annual CUNY Sentence Processing Conference*, (Davis, CA).

Dresang, H. C., Dickey, M. W., and Warren, T. C. (2019). Semantic memory for objects, actions, and events: a novel test of event-related conceptual semantic knowledge. *Cogn. Neuropsychol.* 36, 313–335. doi: 10.1080/02643294.2019.1656604

Edmonds, L. A. (2016). A review of verb network strengthening treatment: theory, methods, results, and clinical implications. *Topics Lang. Disord.* 36, 123–135. doi: 10.1097/tld.0000000000000088

Ellis, N. C., and Larsen-Freeman, D. (2009). *Language as a Complex Adaptive System*. Hoboken, NJ: John Wiley & Sons.

Faroqi-Shah, Y., Wood, E., and Gassert, J. (2010). Verb impairment in aphasia: a priming study of body-part overlap. *Aphasiology* 24, 1377–1388. doi: 10.1080/02687030903515362

Fergadiotis, G., Hula, W. D., Swiderski, A. M., Lei, C.-M., and Kellough, S. (2019). Enhancing the efficiency of confrontation naming assessment for aphasia using computer adaptive testing. *J. Speech Lang. Hear. Res. JSLHR* 62, 1724–1738. doi: 10.1044/2018_JSLHR-L-18-0344

Ferretti, T. R., McRae, K., and Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *J. Mem. Lang.* 44, 516–547. doi: 10.1006/jmla.2000.2728

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198.

Forster, K. I. (1976). "Accessing the mental lexicon," in *New Approaches to Language Mechanisms*, eds R. J. Walker, and F. Wales, (Amsterdam: North-Holland).

Foygel, D., and Dell, G. S. (2000). Models of impaired lexical access in speech production. *J. Mem. Lang.* 43, 182–216. doi: 10.1006/jmla.2000.2716

Frazier, L. (1987). "Sentence processing: a tutorial review," in *Attention and Performance 12: The Psychology of Reading*, ed. M. Coltheart, (New Jersey: Lawrence Erlbaum Associates, Inc), 559–586.

Gahl, S. (2002). Lexical biases in aphasic sentence comprehension: an experimental and corpus linguistic study. *Aphasiology* 16, 1173–1198. doi: 10.1080/02687030244000428

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis*, 3rd Edn. Boca Raton, FL: CRC Press.

Gibson, E., Bergen, L., and Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proc. Natl. Acad. Sci. U.S.A.* 110, 8051–8056. doi: 10.1073/pnas.1216438110

Gibson, E., Sandberg, C., Fedorenko, E., Bergen, L., and Kiran, S. (2015). A rational inference approach to aphasic language comprehension. *Aphasiology* 30, 1341–1360. doi: 10.1080/02687038.2015.1111994

Goodglass, H. (1976). Agrammatism. *Stud. Neurolinguistics* 1, 237–260.

Goodglass, H., and Wingfield, A. (1997). *Anomia: Neuroanatomical and Cognitive Correlates*. Amsterdam: Elsevier, doi: 10.1016/B978-0-12-289685-9.X5000-5

Hair, J. F. Jr., Black, W. C., Babin, B. J., and Anderson, R. E. (2009). *Multivariate Data Analysis*, 7th Edn. Upper Saddle River, NJ: Prentice-Hall.

Hale, J. (2001). "A probabilistic earley parser as a psycholinguistic model," in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*. NAACL 2001, (New York, NY: ACM).

Hare, M., Jones, M., Thomson, C., Kelly, S., and McRae, K. (2009). Activating event knowledge. *Cognition* 111, 151–167. doi: 10.1016/j.cognition.2009.01.009

Hayes, R. A., Dickey, M. W., and Warren, T. (2016). Looking for a location: dissociated effects of event-related plausibility and verb-argument information on predictive processing in aphasia. *Am. J. Speech Lang. Pathol.* 25, S758–S775.

Hazelrigg, L. (2009). "Inference," in *The Handbook of Data Analysis*, eds M. Hardy, and A. Bryman, (Thousand Oaks, CA: Sage).

Howes, A., Lewis, R. L., and Vera, A. (2009). Rational adaptation under task and processing constraints: implications for testing theories of cognition and action. *Psychol. Rev.* 116, 717–751. doi: 10.1037/a0017187

Hula, W. D., Panesar, S., Gravier, M., Yeh, F.-C., Dresang, H. C., Dickey, M. W., et al. (2020). Structural white matter connectometry of word production in aphasia. *Brain* 143, 2532–2544. doi: 10.1093/brain/awaa193

Jescheniak, J. D., and Levelt, W. J. M. (1994). Word frequency effects in speech production: retrieval of syntactic information and of phonological form. *J. Exp. Psychol. Learn. Mem. Cogn.* 20, 824–843. doi: 10.1037/0278-7393.20.4.824

Jonkers, R., and Bastiaanse, R. (2007). Action naming in anomic aphasic speakers: effects of instrumentality and name relation. *Brain Lang.* 102, 262–272. doi: 10.1016/j.bandl.2007.01.002

Kelter, S., and Kaup, B. (2012). "Conceptual knowledge, categorization, and meaning," in *Semantics: An International Handbook of Natural Language Meaning*, eds C. Maienborn, K. von Heusinger, and P. Portner, (Berlin: Walter de Gruyter), 2775–2804.

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

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. doi: 10.1037/0033-295X.104.2.211

Lenth, R. (2020). *emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.5.3.* Boston, MA: RStudio.

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

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.

Mätzig, S., Druks, J., Masterson, J., and Vigliocco, G. (2009). Noun and verb differences in picture naming: past studies and new evidence. *Cortex* 45, 738–758. doi: 10.1016/j.cortex.2008.10.003

McDonald, S. A., and Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychol. Sci.* 14, 648–652. doi: 10.1046/j.0956-7976.2003.psci_1480.x

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton, FL: Chapman and Hall/CRC.

McRae, K., Hare, M., Elman, J. L., and Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Mem. Cogn.* 33, 1174–1184. doi: 10.3758/bf03193221

McRae, K., Hare, M., Ferretti, T., and Elman, J. L. (2001). "Activating verbs from typical agents, patients, instruments, and locations via event schemas," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, (Austin, TX: Cognitive Science Society), 617–622.

McRae, K., and Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Lang. Linguist. Compass* 3, 1417–1429. doi: 10.1111/j.1749-818X.2009.00174.x

Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. *Behav. Res. Methods Instrum. Comput.* 36, 402–407. doi: 10.3758/bf03195588

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychol. Rev.* 103, 56–115. doi: 10.1037/0033-295X.103.1.56

Quique, Y. M., Evans, W. S., and Dickey, M. W. (2019). Acquisition and generalization responses in aphasia naming treatment: a meta-analysis of semantic feature analysis outcomes. *Am. J. Speech Lang. Pathol.* 28, 230–246. doi: 10.1044/2018_AJSLP-17-0155

Ratcliff, R., Thapar, A., Gomez, P., and McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychol. Aging* 19, 278–289. doi: 10.1037/0882-7974.19.2.278

Raven, J. C. (1965). *Guide to using the Coloured Progressive Matrices: Sets A, Ab, B.* Livingston: William Grieve & Sons.

Reali, F., and Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *J. Mem. Lang.* 57, 1–23. doi: 10.1016/j.jml.2006.08.014

Rider, J. D., Wright, H. H., Marshall, R. C., and Page, J. L. (2008). Using semantic feature analysis to improve contextual discourse in adults with aphasia. *Am. J. Speech Lang. Pathol.* 17:161. doi: 10.1044/1058-0360(2008/016)

Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., and Brecher, A. (1996). The Philadelphia naming test: scoring and rationale. *Clin. Aphasiol.* 24, 121–134.

Rofes, A., Capasso, R., and Miceli, G. (2015). Verb production tasks in the measurement of communicative abilities in aphasia. *J. Clin. Exp. Neuropsychol.* 37, 483–502. doi: 10.1080/13803395.2015.1025709

Schwartz, M. F., Linebarger, M. C., Saffran, E. M., and Pate, D. (1987). Syntactic transparency and sentence interpretation in aphasia. *Lang. Cogn. Process.* 2, 85–113. doi: 10.1080/01690968708406352

Schwartz, M. F., Saffran, E. M., and Marin, O. S. (1980). The word order problem in agrammatism: I. comprehension. *Brain Lang.* 10, 249–262. doi: 10.1016/0093-934x(80)90055-3

Silkes, J. P., Baker, C., and Love, T. (2020). The time course of priming in aphasia. *Topics Lang. Disord.* 40, 54–80. doi: 10.1097/TLD.0000000000000205

Smith, N. J., and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition* 128, 302–319. doi: 10.1016/j.cognition.2013.02.013

Starns, J. J., and Ratcliff, R. (2010). The effects of aging on the speed-accuracy compromise: boundary optimality in the diffusion model. *Psychol. Aging* 25, 377–390. doi: 10.1037/a0018022

Swinburn, K., Porter, G., and Howard, D. (2004). *CAT: Comprehensive Aphasia Test*. London: Psychology Press.

Wambaugh, J. L., Mauszycki, S., and Wright, S. (2014). Semantic feature analysis: application to confrontation naming of actions in aphasia. *Aphasiology* 28, 1–24. doi: 10.1080/02687038.2013.845739

Warren, T., Dickey, M. W., and Liburd, T. L. (2017). A rational inference approach to group and individual-level sentence comprehension performance in aphasia. *Cortex* 92, 19–31. doi: 10.1016/j.cortex.2017.02.015

Wasow, T. (1997). End-Weight from the speaker's perspective. *J. Psycholinguist. Res.* 26, 347–361. doi: 10.1023/A:1025080709112

Willits, J. A., Amato, M. S., and MacDonald, M. C. (2015). Language knowledge and event knowledge in language use. *Cogn. Psychol.* 78, 1–27. doi: 10.1016/j.cogpsych.2015.02.002

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 Psychology

ORIGINAL RESEARCH
published: 11 February 2021
doi: 10.3389/fpsyg.2021.615538

Neurobehavioral Correlates of Surprisal in Language Comprehension: A Neurocomputational Model

Harm Brouwer, Francesca Delogu, Noortje J. Venhuizen and Matthew W. Crocker*

Department of Language Science and Technology, Saarland University, Saarbrücken, Germany

OPEN ACCESS

Edited by:
Gabriella Vigliocco,
University College London,
United Kingdom

Reviewed by:
Peter beim Graben,
Humboldt University of Berlin,
Germany
Olaf Hauk,
University of Cambridge,
United Kingdom
Stefan Frank,
Radboud University Nijmegen,
Netherlands

****Correspondence:***
Harm Brouwer
brouwer@coli.uni-saarland.de

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

Received: *09 October 2020*
Accepted: *11 January 2021*
Published: *11 February 2021*

Citation:
Brouwer H, Delogu F, Venhuizen NJ and Crocker MW (2021) Neurobehavioral Correlates of Surprisal in Language Comprehension: A Neurocomputational Model. Front. Psychol. 12:615538. doi: 10.3389/fpsyg.2021.615538

Expectation-based theories of language comprehension, in particular Surprisal Theory, go a long way in accounting for the behavioral correlates of word-by-word processing difficulty, such as reading times. An open question, however, is in which component(s) of the Event-Related brain Potential (ERP) signal Surprisal is reflected, and how these electrophysiological correlates relate to behavioral processing indices. Here, we address this question by instantiating an explicit neurocomputational model of incremental, word-by-word language comprehension that produces estimates of the N400 and the P600—the two most salient ERP components for language processing—as well as estimates of "comprehension-centric" Surprisal for each word in a sentence. We derive model predictions for a recent experimental design that directly investigates "world-knowledge"-induced Surprisal. By relating these predictions to both empirical electrophysiological and behavioral results, we establish a close link between Surprisal, as indexed by reading times, and the P600 component of the ERP signal. The resultant model thus offers an integrated neurobehavioral account of processing difficulty in language comprehension.

Keywords: event-related potentials (ERPs), N400, P600, language comprehension, surprisal theory

1. INTRODUCTION

In language comprehension, an interpretation is incrementally constructed on a more or less word-by-word basis, where some words incur more processing difficulty than others. Expectation-based theories of comprehension, in particular Surprisal Theory (Hale, 2001, 2003; Levy, 2008), have become influential in explaining word-by-word processing difficulty. Surprisal Theory asserts that the effort incurred by a word is proportional to its expectancy in context: difficulty$(w_t) \approx -\log P(w_t|w_1 \ldots w_{t-1}, \text{CONTEXT})$, where CONTEXT denotes the extra-sentential context. Indeed, Surprisal estimates derived from language models go a long way in accounting for behavioral correlates of processing difficulty, in particular reading times (e.g., Boston et al., 2008; Demberg and Keller, 2008; Smith and Levy, 2008, 2013; Frank, 2009; Roark et al., 2009; Brouwer et al., 2010). As such, a natural, yet thus far unanswered question is: What are the electrophysiological indices of Surprisal? More specifically, what component(s) of the Event-Related brain Potential (ERP) signal index(es) Surprisal, and what is their relationship to behavioral indices of processing difficulty?

While previous work has sought to answer this question by correlating Surprisal estimates derived from language models with the amplitude of relevant ERP components on a word-by-word basis (Frank et al., 2015), we here take a different approach. Specifically, we build upon two recent computational models of incremental, word-by-word language comprehension. The first is the model of "comprehension-centric" Surprisal by Venhuizen et al. (2019a) that goes beyond typical language models in that Surprisal is derived directly from the interpretations that are constructed during comprehension—rich, probabilistic representations instantiating situation models—thereby rendering it sensitive both to linguistic experience (like language models), but crucially, also to knowledge about the world, which enables the model to also account for "world knowledge"-driven effects on processing (e.g., Albrecht and O'Brien, 1993; Morris, 1994; Myers and O'Brien, 1998; Cook and Myers, 2004; Knoeferle et al., 2005; van Berkum et al., 2005, among others). We here employ these meaning representations in a neurocomputational model by Brouwer et al. (2017) that instantiates the Retrieval-Integration account of the electrophysiology of language comprehension (Brouwer et al., 2012; Brouwer and Hoeks, 2013), thereby offering a mechanistic account of the modulation pattern of the N400 and the P600—the two most salient ERP components for language processing—that explains key data on semantic processing (as reviewed in Kuperberg, 2007; Bornkessel-Schlesewsky and Schlesewsky, 2008; Brouwer et al., 2012).

The resultant model produces, on a word-by-word basis, estimates of the N400, reflecting the contextualized retrieval of word meaning, estimates of the P600, reflecting the integration of retrieved word meaning into the unfolding utterance interpretation, as well as estimates of "comprehension-centric" Surprisal, reflecting the likelihood of a change in interpretation. Critically, while both retrieval and integration are predicted to be sensitive to a notion of expectation, retrieval processes are modulated by the expectancy of *word meaning*, while integration processes are modulated by the expectancy of *utterance meaning*. In order to identify how "comprehension-centric" Surprisal, taken to be indexed by reading times, relates to electrophysiological indices, we require empirical evidence that bears upon these different types of expectancy.

A recent study by Delogu et al. (2019), henceforth DBC, employs a context manipulation design in which they manipulated word meaning expectancy (retrieval/N400) through semantic association (henceforth *association*), and utterance meaning expectancy (integration/P600) through *plausibility*. More specifically, they manipulated the association and plausibility of a target word in German mini-discourses, across three conditions:

Baseline [+plausible, +associated]
Johann betrat das Restaurant. Wenig später öffnete er die Speisekarte und [...]
"*John entered the restaurant. Before long, he opened the menu and [...]*"

Event-related [−plausible, +associated]
Johann verließ das Restaurant. Wenig später öffnete er die Speisekarte und [...]
"*John left the restaurant. Before long, he opened the menu and [...]*"

Event-unrelated [−plausible, −associated]
Johann betrat die Wohnung. Wenig später öffnete er die Speisekarte und [...]
"*John entered the apartment. Before long, he opened the menu and [...]*"

Figure 1 shows the plausibility judgments (left) and association ratings (middle) found by DBC. In both the event-related and the event-unrelated condition, the target word (e.g., "Speisekarte"/"menu") rendered the entire mini-discourse implausible relative to baseline. In addition, there was also a difference in plausibility between the event-related and event-unrelated condition. Further, the event-related and the event-unrelated conditions differed in the degree of association between the target word and its prior context; that is, in the event-unrelated condition the target word is unassociated with the context, while in the event-related (and baseline) condition it is associated with the context. **Figure 1** (right) shows the Cloze probabilities of the target words in all three conditions, as determined based on completions of two-sentence discourses up to and including the determiner preceding the target word. Crucially, the Cloze probabilities—which quantify the expectancy of the critical words in context, and the negative logarithm of which determines their Surprisal—show a qualitatively similar pattern to the plausibility ratings with all conditions differing from each other.

In what follows, we will first derive an explicit neurocomputational model of comprehension that produces explicit N400, P600, and Surprisal estimates for these conditions. Subsequently, we will outline the predictions of the model, the ERP results obtained by DBC, as well as the reading time results from replication of this study using a self-paced reading (SPR) paradigm. Our results suggest a strong qualitative link between "comprehension-centric" Surprisal, as indexed by reading times, and the integration processes underlying the P600 component of the ERP signal. While this conclusion differs from previous findings linking Surprisal to the N400 component, we discuss how these results can be reconciled within the Retrieval-Integration framework, thereby offering a more integrated neurobehavioral account of processing difficulty in language comprehension.

2. A NEUROCOMPUTATIONAL MODEL

To model both estimates of ERP components (N400 and P600), as well as estimates of Surprisal (reading times), we start from the neurocomputational model of the N400 and P600 by Brouwer et al. (2017), and augment it with the rich, probabilistic situation model representations used by Venhuizen et al. (2019a). Critically, by replacing the thematic role assignment representations used in Brouwer et al. (2017) with these richer meaning representations—which naturally capture probabilistic knowledge about the world—the resultant model produces N400, P600, and "comprehension-centric" Surprisal estimates on a word-by-word basis.

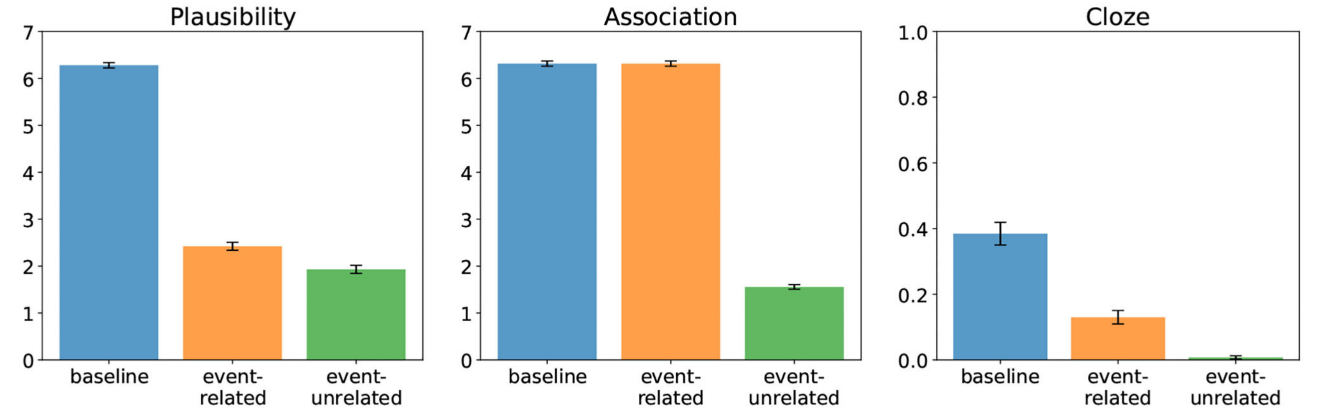


FIGURE 1 | Offline ratings from Delogu et al. (2019) for plausibility **(left)** and association **(middle)**, and estimated Cloze probability of the target **(right)** in all three conditions.

2.1. Architecture

The neurocomputational model of language electrophysiology by Brouwer et al. (2017) instantiates the Retrieval-Integration (RI) account of the N400 and the P600 (Brouwer et al., 2012; Brouwer and Hoeks, 2013; Delogu et al., 2019). The RI account postulates that incremental, word-by-word comprehension proceeds in cycles consisting of the Retrieval of word meaning, the ease of which is reflected in N400 amplitude (retrieval of word meaning is facilitated if it is expected given the preceding context), and the subsequent Integration of this word meaning into the unfolding utterance representation, the effort incurred by which is indexed by P600 amplitude (integration difficulty increases as a function of the degree to which integrating retrieved word meaning renders the interpretation unexpected, unclear, or implausible).

Mechanistically, the processing of a word can be conceptualized as a function *process*, which maps an acoustically or orthographically perceived word w_t (word form), and the context as established after processing words $w_1 \dots w_{t-1}$ (utterance context), onto an utterance interpretation spanning words $w_1 \dots w_t$ (utterance representation):

process: (word form, utterance context) → utterance representation

This mapping is, however, indirect in that the *process* function is itself composed of a *retrieve* and an *integrate* function, which are hypothesized to underlie the N400 and the P600 components, respectively. The *retrieve* function maps the incoming word form w_t onto a representation of its meaning (word meaning), while taking into account the context in which it occurs (utterance context):

retrieve: (word form, utterance context) → word meaning [∼N400]

The result of this *retrieve* function (word meaning) serves as input for the *integrate* function, which maps the meaning of w_t (word meaning) and its prior context (utterance context) onto an updated utterance interpretation (utterance representation):

integrate: (word meaning, utterance context) → utterance representation [∼P600]

The resultant, updated interpretation determines the context for the retrieval and integration of a next word.

Formally, the neurocomputational model is a recurrent, artificial neural network model that instantiates the *process* function, broken down into its *retrieve* and *integrate* subprocesses. **Figure 2** provides a schematic overview of the model architecture. The model consists of five layers of artificial neurons, implementing the input to the model (**input**), a Retrieval module (**retrieval** and **retrieval_output**), and an Integration module (**integration** and **integration_output**). As artificial neurons, we used leaky rectified linear units, the activation function of which is defined as follows (for the leak parameter we used $\alpha = 0.3$):

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha x & \text{otherwise} \end{cases} \tag{1}$$

Units in the **retrieval_output** and **integration_output** are capped at 1.0—i.e., $f'(x) = min(f(x), 1.0)$—as the representations that the model is trained to recover at these layers are binary representations (see below). To facilitate learning, however, units are not capped at zero, allowing a small positive gradient for inactive units.

Time in the model is discrete, and at each processing timestep t, activation flows from the **input** layer, through the **retrieval** layer to the **retrieval_output** layer, and from **retrieval_output** layer through the **integration** layer to the **integration_output** layer. To allow for context-sensitive retrieval and integration, the **retrieval** and the **integration** layer both also receive input from the activation pattern in the **integration** layer as established at the previous timestep $t - 1$, effectuated through an additional context layer (**integration_context**; see Elman, 1990). Prior to feed-forward propagation of activation from the **input** to the **integration_output** layer, this **integration_context** layer receives a copy of the **integration** layer (at timestep $t = 0$, the activation value of each unit in the **integration_context** layer is set to 0.5). Finally, all layers except the **input** and **integration_context** layer also receive input from a bias unit, the activation value of which is always 1.

As will be detailed below, the model is trained to incrementally, on a word-by-word basis, map sequences of (orthographic or acoustic) word forms, presented at the **input** layer, onto an utterance meaning representation at the **integration_output** layer, thus instantiating the *process* function at each time tick. Crucially, the mapping from word forms onto an utterance representation is not direct; it is broken down into the *retrieve* and *integrate* sub-processes. Provided a localist representation of an incoming word w_t (**input**), encoding its perceived orthographic/acoustic form, and the unfolding context (**integration_context**), the **retrieval** layer serves to activate a word meaning representation of w_t in the **retrieval_output** layer. Hence, the function of the **retrieval** layer is to retrieve word meaning representations, which take the form of distributed, binary semantic feature vectors (derived from the training sentences using the Correlated Occurrence Analogue to Lexical Semantics, COALS, Rohde et al., 2005; see below). The effort involved in retrieval is taken to be reflected in the N400 amplitude, which is estimated as the degree to which the activation pattern of the **retrieval** layer changes as a result of processing the incoming word:

$$\mathrm{N400}(w_t) = \mathrm{dist}(\mathbf{retrieval}_t, \mathbf{retrieval}_{t-1}) \tag{2}$$

where $\mathrm{dist}(x, y) = 1.0 - \cos(x, y)$[1]. The logic is that if the model finds itself in a state in which the meaning of an incoming word is expected, there will be little change in **retrieval** layer from $t-1$ to t, and the estimated N400 amplitude will be small. If, on the other hand, the meaning of an incoming word is unexpected, this will induce a larger change, and a larger estimated N400 amplitude.

The **integration** layer, in turn, combines the retrieved word meaning representation (**retrieval_output**) with the unfolding utterance context (**integration_context**), into an updated utterance representation (**integration_output**). The **integration** layer thus serves to integrate word meaning into the unfolding interpretation. The effort involved in updating the interpretation with the meaning contributed by the incoming word is taken to be reflected in the P600 amplitude, which is estimated as the degree to which the activation pattern of the **integration** layer changes from $t-1$ to t:

$$\mathrm{P600}(w_t) = \mathrm{dist}(\mathbf{integration}_t, \mathbf{integration}_{t-1}) \tag{3}$$

where again $\mathrm{dist}(x, y) = 1.0 - \cos(x, y)$. If the interpretation is expected, given the linguistic experience of the model and/or its knowledge about the world, integration of the meaning contributed by the incoming word should be relatively effortless, and hence induce a relatively small change in the **integration** layer, thus producing a small estimated P600 amplitude. Conversely, if the interpretation is unexpected, the change in the **integration** layer will be larger, and so will the estimated P600 amplitude.

The utterance meaning representations that the model produces—at its **integration_output** layer—are rich "situation model"-like meaning representations that encode meaning as points in a Distributed Situation-state Space (DSS; Frank et al., 2003, 2009; for a recent reconceptualization of these representations grounded in formal semantics, see Venhuizen et al., 2019c). DSS offers distributed representations that allow for encoding world knowledge, and that are both compositional and probabilistic (see section 2.2.3 below for more detail). Crucially, the probabilistic nature of the DSS representations allows for deriving Surprisal estimates directly from the meaning vectors (Frank and Vigliocco, 2011). In particular, Venhuizen et al. (2019a) define an online, comprehension-centric notion of Surprisal that is sensitive to both linguistic experience and world knowledge, and that derives directly from a change in interpretation from time-step $t-1$ to t:

$$\begin{aligned}&\mathrm{Surprisal}(w_t) = \\ &-\log \mathrm{P}(\mathbf{integration\_output}_t | \mathbf{integration\_output}_{t-1})\end{aligned} \tag{4}$$

That is, the more likely the interpretation at t given the interpretation at $t-1$, the lower the Surprisal induced by word w_t (see Venhuizen et al., 2019b, for a similar DSS-derived conceptualization of Entropy).

To summarize, the model processes utterances on an incremental word-by-word basis, and produces N400, P600, and Surprisal estimates for every word. More specifically, for a given incoming word form (**input**), and a given context (**integration_context**), the **retrieval** layer retrieves a word meaning representation (**retrieval_output**). Ease of retrieval is reflected in the estimated N400 amplitude. Subsequently, the **integration** layer serves to integrate this retrieved word meaning representation into the unfolding utterance meaning representation (**integration_context**), to produce an updated utterance interpretation (**integration_output**). Ease of integration is reflected in the estimated P600 amplitude, and Surprisal estimates reflect the likelihood of the updated interpretation given the previous interpretation. The model thus predicts a strong correlation between the P600 and Surprisal.

2.2. Representations

2.2.1. Word Form Representations

The acoustic/orthographic word form for each of the unique words in the training set is represented as a 16-dimensional localist representation, such that each unit uniquely identifies a single word.

2.2.2. Word Meaning Representations

In line with influential theories of word meaning (see McRae et al., 2005, for a review), our model employs feature-based semantic representations as word meaning representations, in which related concepts may share semantic features. Specifically, like in the Brouwer et al. (2017) model, the semantics associated with individual words are distributed, binary feature-vectors

[1] Linking hypotheses such as these, between model behavior and the electrophysiological signal, are also known as "synthetic ERPs" (Barrès et al., 2013, see also beim Graben et al., 2008; Crocker et al., 2010; Rabovsky et al., 2018; Fitz and Chang, 2019, among others).

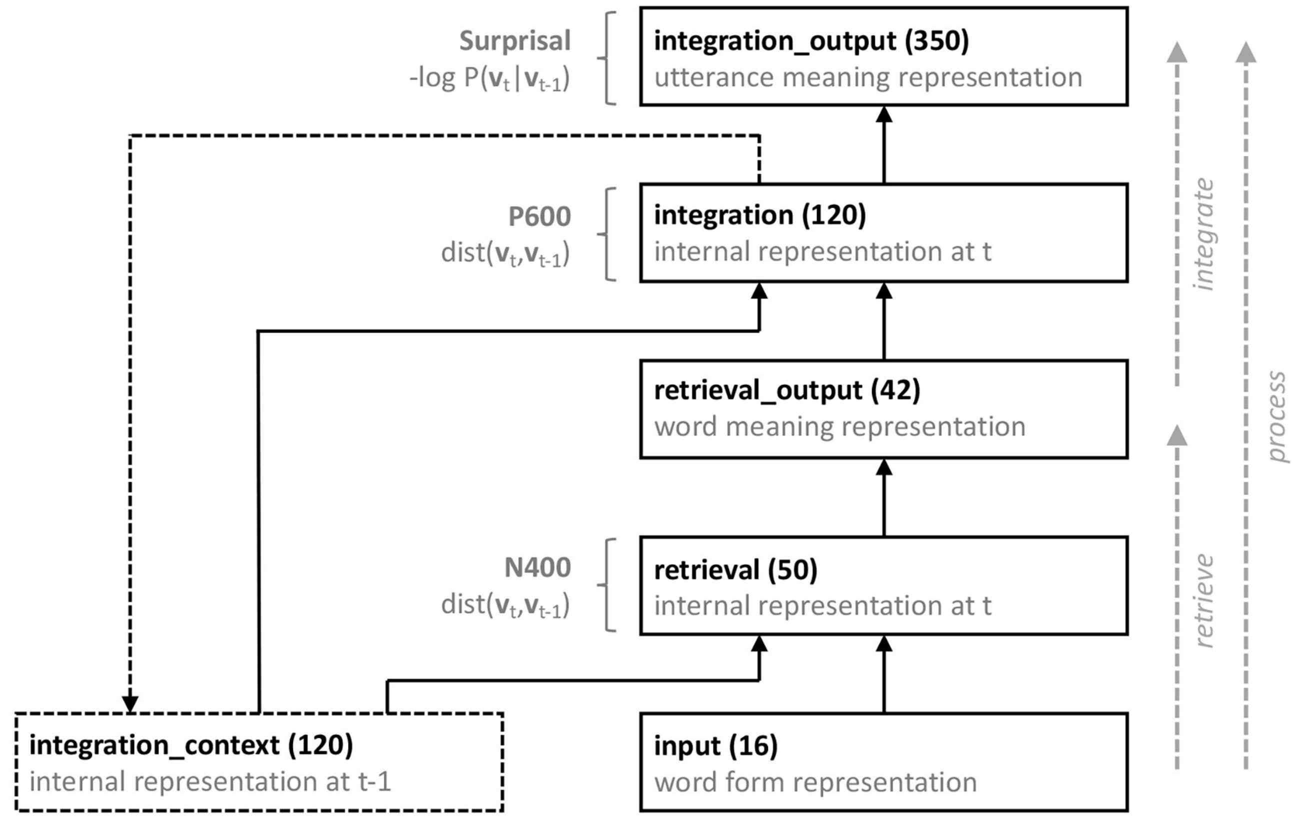

Surprisal
-log P($\mathbf{v}_t$|$\mathbf{v}_{t-1}$)
integration_output (350)
utterance meaning representation
P600
dist($\mathbf{v}_t$,$\mathbf{v}_{t-1}$)
integration (120)
internal representation at t
retrieval_output (42)
word meaning representation
N400
dist($\mathbf{v}_t$,$\mathbf{v}_{t-1}$)
retrieval (50)
internal representation at t
integration_context (120)
internal representation at t-1
input (16)
word form representation
integrate
retrieve
process

FIGURE 2 | Schematic illustration of the neurocomputational model. Each rectangle represents a layer of artificial (leaky rectified linear) neurons, and each solid arrow represents full connectivity between each neuron in a projecting layer and each neuron in a receiving layer. The dashed rectangle is a context layer, and the dashed arrow represents a copy projection, such that prior to feed-forward propagation the **integration_output** layer receives a copy of the **integration** layer. All groups except the **input** and **integration_context** layer also receive input from a bias unit (not shown). See text for details.

derived using the Correlated Occurrence Analogue to Lexical Semantics (COALS; Rohde et al., 2005). While Brouwer et al. (2017) derived COALS representations from a large corpus of newspaper text, we here derive them directly from the training data in order to exert more control over the resulting vectors. That is, our objective here is to arrive at distributed, partially overlapping semantic feature vectors, and not necessarily at feature vectors that reflect human similarity judgments (see Brouwer et al., 2017, for discussion). While these vectors could in principle be constructed by hand, the COALS method allows us to automatically derive them from our training sentences. Critically, an artifact of applying the COALS method to a data set of such small size, is that one may obtain identical vectors for two or more words. We mitigate this by concatenating the resulting COALS vectors with an identifier that assures that each word meaning vector is unique.

First, we computed a co-occurrence matrix using a 1-word window. We then converted the co-occurrence frequencies into pairwise correlations. Following the COALS procedure, we then discard negative correlations by setting them to zero, and we reduce the distance between weak and strong positive correlations by replacing them with their square root. Finally, as the training set contains 16 lexical items, we derived 16-dimensional binary word meaning vectors by replacing non-zero values with 1. To assure unique vectors for all words, the 16-dimensional vectors were concatenated with a 26-unit identifier containing two hot bits, resulting in 42-dimensional unique word meaning representations.

2.2.3. Utterance Meaning Representations

Following Venhuizen et al. (2019a), the semantics associated with the training sentences presented to the model are derived from the Distributed Situation-state Space model (DSS, Frank et al., 2003, 2009; see also the formalization in terms of Distributional Formal Semantics described in Venhuizen et al., 2019c). In DSS, utterance meaning vectors are derived from a meaning space that defines co-occurrences of individual propositional meanings across a set of observations (formalized as formal semantic models in Venhuizen et al., 2019c). For the current meaning space, a set of propositions was generated using the predicates *enter(p,l)*, *leave(p,l)*, and *go_to(p,g)*, in combination with arguments that identify a person (*p*), location (*l*), and goal (*g*) (see **Table 1**). In addition, the meaning space contains the unary predicates *entity* and *event* that assert the existence of

TABLE 1 | Propositions described in the current meaning space and their arguments.

Type	Variable	Instantiation
proposition	–	enter(*p*,*l*), leave(*p*,*l*), go_to(*p*,*g*), entity(*p*), entity(*l*), entity(*g*), entity(*r*), event(enter), event(leave), event(go_to)
person	*p*	kevin
location	*l*	church, cinema, farm, school
goal (*church*)	*g*	bible
goal (*cinema*)	*g*	cash_register
goal (*farm*)	*g*	cows
goal (*school*)	*g*	classroom
goal	*g*	bus_stop, parking, toilet, tram
referent (*church*)	*r*	candle, hymn_book
referent (*cinema*)	*r*	popcorn_machine, seat
referent (*farm*)	*r*	farmer, pitchfork
referent (*school*)	*r*	teacher, rector

In the first column, location names in brackets indicate that certain goals and referents are associated with particular locations, triggering presupposed entities (see text for details).

referential entities and events, respectively, in the observations that constitute the meaning space: predicate names (*enter*, *leave*, and *go_to*) instantiate arguments for *event* propositions, and persons, locations and goals, together with a set of location-specific referents (r) instantiate arguments for *entity* propositions. This resulted in a total of 40 atomic propositions.

Based on this set of propositions $\mathcal{P}$, a meaning space is constructed using an incremental, inference-driven probabilistic sampling algorithm (see Venhuizen et al., 2019c). The sampling algorithm uses a set of hard and probabilistic constraints to derive a set of models $\mathcal{M}$ that describe states-of-affairs in terms of combinations of propositions in $\mathcal{P}$. Together, these models (i.e., observations) define a meaning space. The hard constraints used to derive the current meaning space restrict observations to describe a single *enter* or *leave* event, and at most one *go_to* event. In addition, predicates always co-occur with explicit referential introductions of each of their arguments and the denoted event [e.g., *enter(kevin,cinema)* always co-occurs with *entity(kevin)*, *entity(cinema)*, and *event(enter)*]. Moreover, in order for the comprehension model to learn to associate locations to particular entities, certain propositions are constrained to always co-occur with certain presuppositions: locations always co-occur with their location-specific referents (selected based on the Cloze ratings from the DBC study), and each goal necessarily co-occurs with its associated location (as well as the associated presupposed referents). Probabilistically, the meaning space is constructed in such a way that goals occur more often with their related location than with any other location (following the plausibility ratings from the DBC study; see below).

Based on these constraints, we constructed a meaning space consisting of $3,000$ observations, which was reduced to 350 dimensions using the dimension selection algorithm described in Venhuizen et al. (2019a). The resulting meaning space defines meaning vectors for each of the propositions in $\mathcal{P}$; the meaning of proposition $p \in \mathcal{P}$ is defined as the vector $\vec{v}(p)$, such that $\vec{v}_i(p) = 1$ if p is true in model $M_i \in \mathcal{M}$, and $\vec{v}_i(p) = 0$ otherwise. These vectors can be compositionally combined in order to derive meaning vectors for logically complex expressions. In particular, the meaning of the conjunction between propositions p and q is defined as the point-wise multiplication of the meaning vectors $\vec{v}(p)$ and $\vec{v}(q)$: $\vec{v}(p \wedge q) = \vec{v}(p)\vec{v}(q)$ (Frank et al., 2003; Venhuizen et al., 2019a). The meaning vectors that are derived from the meaning space are also inherently probabilistic, as they define the fraction of models in which a proposition (or combination thereof) is true. More generally, given a meaning space of n observations, we can describe the probability of any point a in the meaning space (which may describe a proposition, a logical combination thereof, or any point in meaning space that cannot be directly expressed in terms of a logical combination of propositions) as follows (Frank et al., 2003; Venhuizen et al., 2019a):

$$P(a) = \frac{1}{n}\sum_i a_i \qquad (5)$$

Given the compositional nature of meaning vectors defined above, we can directly derive the conditional probability of any point in meaning space a given another point b in meaning space, that is, $P(a|b) = P(a \wedge b)/P(b)$, which in turn can be used to derive the comprehension-centric notion of Surprisal (see Equation 4).

2.3. Training

2.3.1. Training Sentences

To obtain model predictions for the conditions from the DBC study, we trained the model on a set of sentence-semantics pairs that were constructed based on a subset of the stimuli used for the DBC study (in German, but for clarity we here report the English equivalents). All sentences presented to the model are of the form "Kevin entered/left [LOC] went_to [REL-TGT/UNREL-TGT]," which are associated with the semantics $enter(kevin, \text{LOC}) \wedge go\_to(kevin, \text{REL-TGT}/\text{UNREL-TGT})$ and $leave(kevin, \text{LOC}) \wedge go\_to(kevin, \text{REL-TGT}/\text{UNREL-TGT})$, respectively. **Table 2** shows the combinations of location (LOC) and target (REL-TGT/UNREL-TGT) that constitute sentences from the baseline/event-related condition ("Kevin entered/left [LOC] went_to [REL-TGT]") and the event-unrelated condition ("Kevin entered [LOC] went_to [UNREL-TGT]"). In addition, to balance plausibility across the *enter*/*leave* sentences, we also created a set of counterbalance sentences with plausible completions for the *leave* event, based on the Cloze completions from the DBC study ("Kevin left [LOC] went_to [REL-TGT]").

The model is taught that any combination of verb–location–target is in principle possible (following Brouwer et al., 2017), but that sentences from the baseline condition are more frequent than other *enter*–location–target combinations ($13:1$), and that counterbalance sentences are more frequent ($4:1$) than other *leave*–location–target combinations. This results in a total of 160 training sentences, with 64 unique semantics, half of which constitute *enter* sentences and the other half *leave* sentences. All locations occur equally often across the entire training set

TABLE 2 | Verb-Location-Target pairs used for constructing the training data.

VERB	LOC	REL-TGT	UNREL-TGT
enter	cinema	cash_register	bible
enter	farm	cows	classroom
enter	school	classroom	cash_register
enter	church	bible	cows
leave	[LOC]	bus_stop/parking/tram/toilet	–

Related targets (REL-TGT) are used for constructing the baseline and event-related (and counterbalance) sentences, and Unrelated targets (UNREL-TGT) are used for constructing the event-unrelated sentences (see text for details).

(40×), as well as all targets (20×). In terms of the probabilistic structure of the DSS meaning vectors derived for these sentences, the conjunctive semantics associated with the sentences from the baseline condition have a higher probability ($M = 0.04$, $N = 4$) than the semantics of both the event-related ($M = 0.009$, $N = 4$) and the event-unrelated ($M = 0.005$, $N = 4$) conditions.

2.3.2. Training Procedure

We used bounded gradient descent (Rohde, 2002), a modification of the standard backpropagation algorithm (Rumelhart et al., 1986), to train the model. Moreover, following Brouwer et al. (2017), we trained the model in two stages. In the first stage, we trained the integration module only; that is, the entire model modulo the **input** and **retrieval** layers. The integration module is trained to map sequences of word meaning representations onto utterance meaning representations. The model was trained for 2,000 epochs, using a momentum coefficient of 0.9 and a learning rate of 0.1, which was scaled down by 10% after every 500 epochs. In the second stage, the weights of the integration module are frozen, and the **input** and **retrieval** layer are added back into the model. The entire model is then trained to map sequences of word form representations onto utterance meaning representations. In this second stage, the model was again trained for 2,000 epochs, with a momentum coefficient of 0.5 and a learning rate of 0.025 (which was again scaled down by 10% after every 500 epochs). To assure generalizability of our results, we trained 10 instances of the model, each with different initial weight matrices. After training, we evaluated the models in terms of mean squared error, output-target similarity, and overall comprehension performance. Overall, performance of the models was very good (mean squared error: $M = 0.11$; $SD = 0.03$, output-target similarity: $M = 0.96$; $SD = 0.01$; Recall@1 = 100%, comprehension score: $M = 0.65$; $SD = 0.03$).

3. NEUROBEHAVIORAL CORRELATES OF SURPRISAL

3.1. Modeling Predictions

To obtain model predictions, we computed N400, P600, and Surprisal estimates for the three conditions of the DBC experiment. **Figure 3** shows the estimated N400 and P600 effects for the event-related relative to baseline contrast, and the event-unrelated relative to baseline contrast. While increased N400 and P600 estimates are positive distances in the **retrieval** and **integration** layers of the model, respectively, we plot the estimated N400-effects downward to signify the negative direction of the corresponding effects in the ERP signal. Note that the inputs and outputs of the retrieval and integration processes differ fundamentally and as consequence, the internal representations that the model develops at the **retrieval** and **integration** layers will also differ. Therefore, the absolute magnitudes of the N400 and P600 estimates should not be directly compared, and also do not directly map onto scalp-recorded voltages; that is, only the relative distances between the conditions in the **retrieval** and **integration** layers are of interest.

The predicted N400 estimates (**Figure 3**, left) show that while the model predicts a larger N400 amplitude for the event-unrelated condition relative to baseline, it predicts little to no difference between baseline and the event-related condition. Indeed, the N400 estimates pattern with the association manipulation, showing that a higher degree of association of a target word to its context leads to more facilitated retrieval of its meaning. The P600 estimates (**Figure 3**, right), in turn, reveal that relative to baseline, both the event-related and the event-unrelated condition produce larger estimated P600 amplitudes in the model. Here, the results pattern with the plausibility ratings and the Cloze probabilities. That is, the more implausible a target word is in a given context, and the lower its Cloze probability, the higher the P600 estimate it induces, reflecting increased effort in integrating its meaning into the unfolding utterance interpretation.

The Surprisal estimates (**Figure 4**) also follow the plausibility ratings and Cloze probabilities: the more implausible a word is in context, and the lower its Cloze probability, the higher its Surprisal according to the model. This means that integrating an implausible, unexpected word yields an interpretation—a point in situation-state space—that is improbable given the interpretation constructed prior to encountering it. Crucially, the Surprisal estimates clearly align with the P600 estimates, and not with the N400 estimates, suggesting a link between Surprisal and the P600. Indeed, while P600 amplitude in the model reflects the effort involved in updating the unfolding interpretation with the meaning contributed by the incoming word, that is, the work involved in actually traversing from one point to the next in situation-state space, Surprisal estimates reflect the likelihood of this traversal.

In sum, relative to baseline, the model predicts an N400-effect for the event-unrelated, but not for the event-related condition. The N400 estimates thus pattern with the association ratings. As for the P600 and Surprisal estimates, the model predicts an effect for both the event-related and the event-unrelated condition relative to baseline. Both the P600 and Surprisal estimates thus follow the plausibility ratings and Cloze probabilities.

3.2. Electrophysiological Results

DBC report on the electrophysiological responses associated with the event-related and event-unrelated conditions. **Figure 5** shows the ERP results in the N400 (300–500 ms, left column) and P600 (600–1,000 ms, right column) time windows, for the

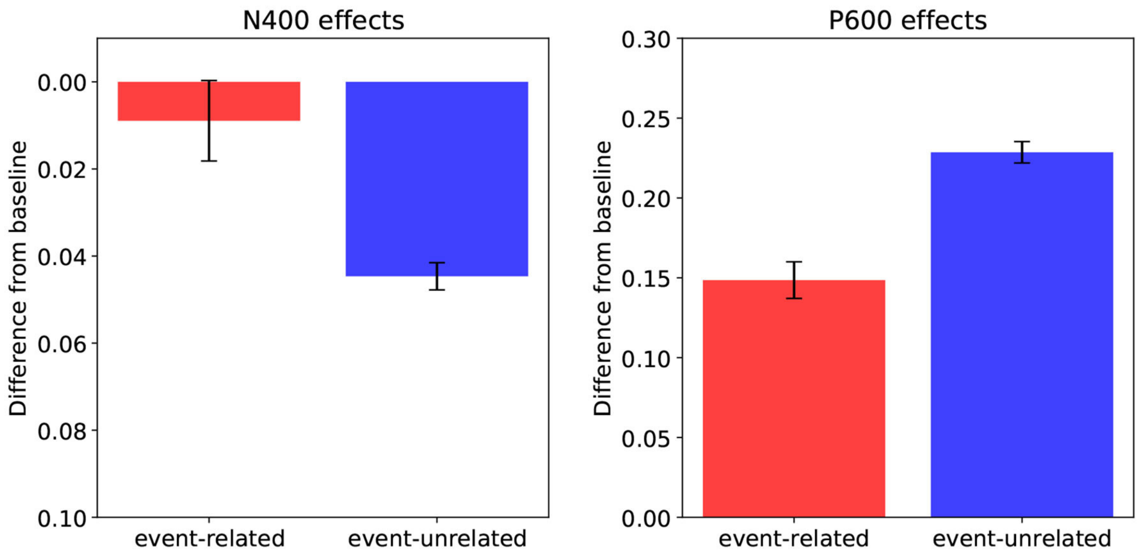


FIGURE 3 | Model predictions: N400-effects (**left**, plotted downwards; see text) and P600-effects **(right)**, for the event-related condition relative to baseline, and for the event-related condition relative to baseline. Error bars show standard errors.

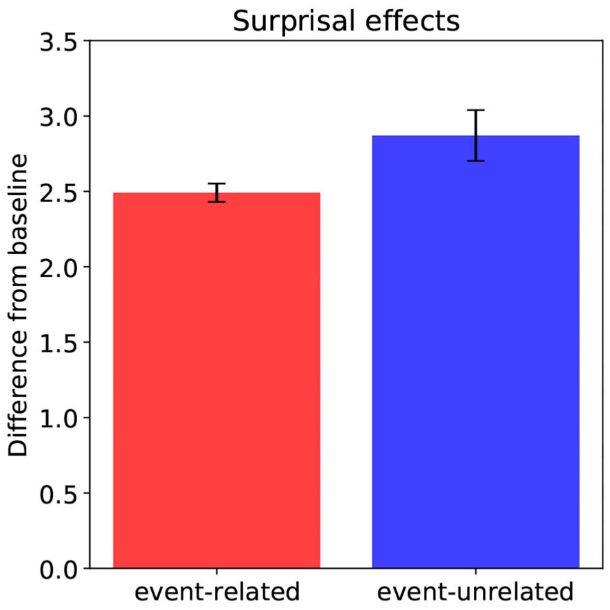


FIGURE 4 | Model predictions: Surprisal effects for the event-related condition relative to baseline, and for the event-unrelated condition relative to baseline. Error bars show standard errors.

event-related and event-unrelated conditions relative to baseline. The event-related condition, which only differs from baseline in plausibility, produced no difference in the N400 time window (top left), but a clear positive effect in the P600 time window (top right). The event-unrelated condition, in turn, which differs from baseline in both association and plausibility, produced a clear negative effect in the N400 time-window (bottom left), which sustained into P600 time window, albeit more frontally pronounced (bottom right). Indeed, while the overall pattern of results in the N400 time window support the view that association is manifest in N400 amplitude, which is in line with the predictions from the model, the results in the P600 time window are less clear. That is, while the results for the event-related condition support the view that plausibility is reflected in the P600, consistent with the model, the results for the event-unrelated condition seem to go against this.

Crucially, DBC argue that the P600 results may be reconciled if one factors in spatiotemporal overlap between the N400 and the P600; that is, they argue that P600 amplitude for the event-unrelated condition in the P600 time window is attenuated by spatiotemporal overlap with the N400. DBC substantiate this explanation by pointing out that—as would be predicted when spatiotemporal component overlap is at play—the broad negativity observed in the N400 time window becomes more frontally pronounced in the P600 time window, where a significant positivity arises at the occipital electrodes. This issue of spatiotemporal component overlap in interpreting ERP data is generally acknowledged (see Hagoort, 2003; Brouwer and Crocker, 2017, for discussions specific to language comprehenion), but as it affects the signal prior to recording, it presents a problem that is notoriously hard to mitigate; that is, given that the N400 and the P600 sum into a single scalp-recorded voltage, isolating their contribution requires a technique that allows for decomposing this voltage into its relevant constituent, latent voltages.

Brouwer et al. (2020) have recently shown that regression-based ERP (rERP) waveform estimation, as proposed by Smith and Kutas (2015a,b), allows for such a decomposition of scalp-recorded voltages. In an rERP analysis, linear regression models are fitted for each subject, time point, and electrode separately, using predictors that instantiate stimulus properties for each trial.

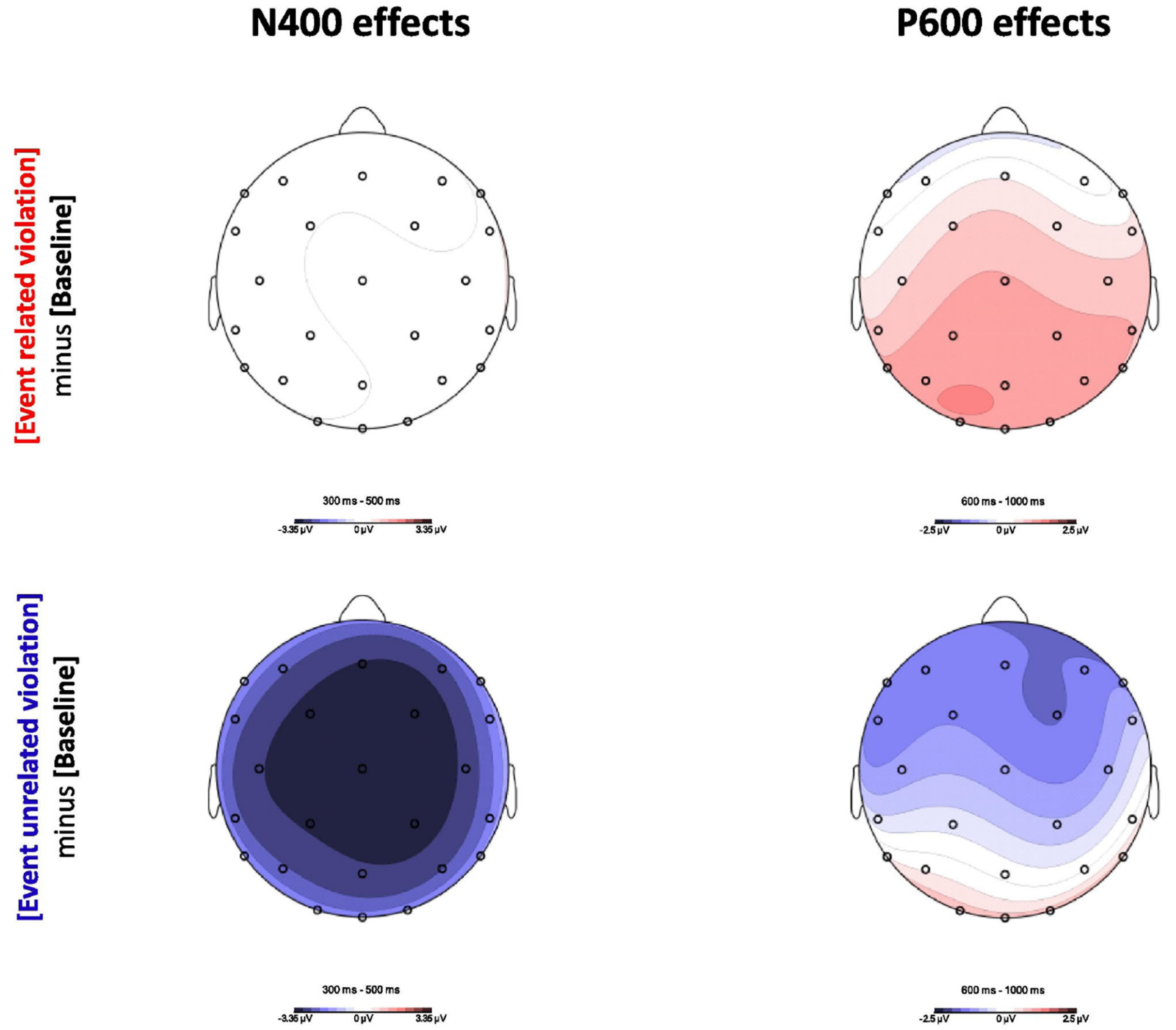


FIGURE 5 | Topographic maps of the ERP effects in the N400 time window (300–500 ms, left column) and the P600 time window (600–1, 000 ms, right column). The upper panel shows the difference between the event-related condition and the baseline. The lower panel shows the difference between the event-unrelated condition and the baseline. Reproduced with permission (CC BY-NC-ND 4.0) from Delogu et al. (2019).

Brouwer et al. (2020) derive an rERP analysis of the DBC data using *plausibility* and *association* as predictors. That is, for each subject, time point, and electrode, they fit the following linear regression model to the data:

$$y_i = \beta_0 + \beta_1 \text{plausibility} + \beta_2 \text{association} + \epsilon_i \qquad (6)$$

where β_0 is an intercept, β_1 the slope for *plausibility* predictor, and β_2 the slope for *association* predictor. For a given trial *i*, the predicted value y_i is the estimated voltage, the residual ϵ_i is the difference between the observed voltage and this estimate, and the predictors *plausibility* and *association* are set to their relevant values for the stimulus presented at this trial. Given a set of trials $y_1 \ldots y_n$, the β coefficients are then fitted by minimizing total squared residuals ($\sum_i^n \epsilon_i^2$) across trials.

Using these fitted models, an rERP data set can be computed in which each observed voltage is replaced by an estimated voltage. Brouwer et al. (2020) show that the resultant rERP data set adequately mimics the observed ERP data, both in terms of residuals (by examining grand-average residuals for each electrode and time point) and in terms of variance (by subjecting the rERP data to the same statistical analysis as the ERP data; that is, by effectively treating it as a replication study). Crucially, as each estimated voltage is now a linear combination of *plausibility* and *association*, the individual contribution of one predictor can be isolated by neutralizing the other (e.g., by setting it to its mean

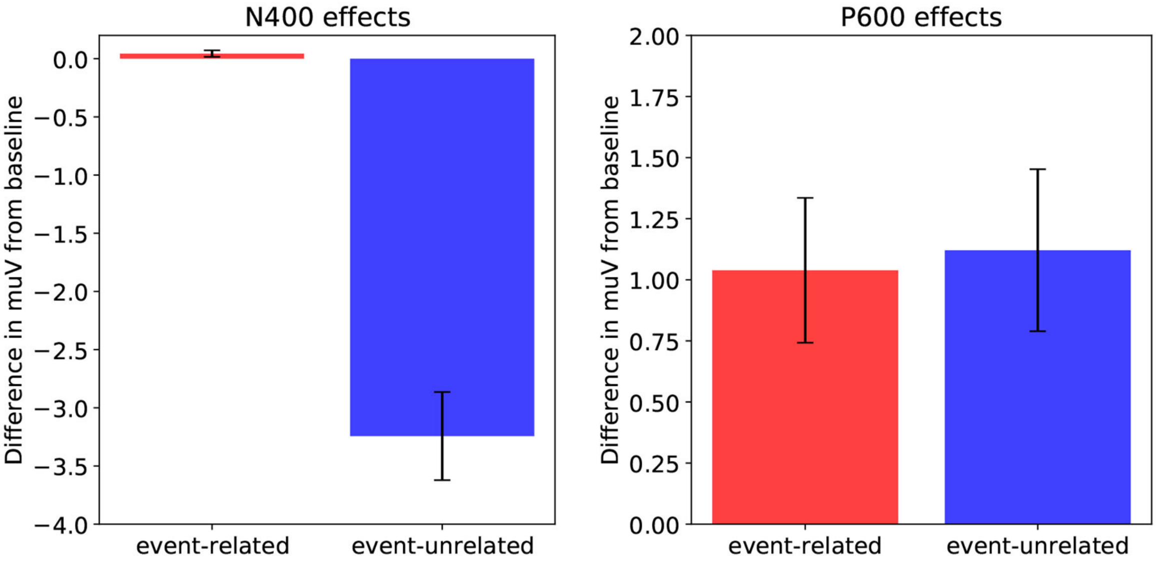


FIGURE 6 | Effects as estimated using regression-based ERP (rERP) estimation: the isolated effects of association in the N400 time-window (300–500 ms, left), and the isolated effects of plausibility in the P600 time-window (600–1, 000 ms, right) for the event-related condition relative to baseline, and for the event-unrelated condition relative to baseline. Error bars show standard errors.

value across trials). This allows us to obtain an clear view on what is going on in the N400 and P600 time-windows.

Starting with the N400 time window, we observe that the results align with the association manipulation. That is, we observe a difference between event-unrelated and baseline, which differ in association, and not between event-related and baseline, which do not differ in association. Moreover, as both the event-related and event-unrelated condition are more implausible than baseline, there is no possible constellation in which plausibility drives the N400, but gets attenuated in the event-related condition through association (as their is no difference in association between event-related and baseline). Finally, given that we do not observe a difference between event-related and baseline, plausibility seems to have little to no effect on the N400 results. **Figure 6** (left) shows the N400-effects in the rERP data when the influence of *association* is isolated (by neutralizing *plausibility*). As in the ERPs, there is no difference between the event-related condition and baseline, while there is a large N400-effect for the event-unrelated condition relative baseline. Indeed, neutralizing the effect of *plausibility* has little effect on the results in the N400 time-window, confirming that the N400 results are driven by association.

As for the P600 time window, it is clear that the P600-effect for the event-related condition relative to baseline must be driven by plausibility, as these conditions do not differ in association. The question here, however, is how association and plausibility combine to explain the results for the event-unrelated condition relative to baseline. **Figure 6** (right) show the P600-effects in the rERP data when the influence of *plausibility* is isolated (by neutralizing *association*). This shows the expected P600-effect for event-related relative to baseline, but critically, also a P600-effect for event-unrelated relative to baseline. Indeed, this suggests that the negativity that was observed for event-unrelated relative to baseline in the ERP data, can be explained by association and plausibility pulling in opposite directions, and association being the stronger force. Crucially, as association seems to drive the N400, and plausibility the P600, this thus suggests that the increase in P600 amplitude for the event-unrelated condition—which we revealed by isolating the effect of plausibility—is attenuated by spatiotemporal overlap with a sustained N400 driven by association.

In sum, when spatiotemporal component overlap between the N400 and the P600 is taken into account, the electrophysiological results of DBC align closely with the predictions of the model (compare **Figures 3**, **6**): an N400-effect for event-unrelated relative to baseline, and a P600-effect for both the event-related and the event-unrelated conditions relative to baseline.

3.3. Behavioral Study

Surprisal has been typically linked to reading times (Levy, 2008). To investigate the behavioral cost associated with the implausible (and therefore higher in Surprisal) conditions from the study reported in DBC, and how this cost relates to the observed ERP responses, we have replicated the DBC study as a self-paced reading (SPR) experiment. Previous work investigating the effects of both plausibility and lexical association on reading times in sentence or discourse contexts has shown robust effects of plausibility, while the effects of lexical association are weaker and appear to be modulated by the global context (see Ledoux et al., 2006). For example, using eye-tracking, Camblin et al. (2007) found effects of discourse congruence on both the target and spillover regions of their stimuli, while effects of association were only observed in the target region for incongruent words. Moreover, Frank (2017) has argued that any effect of semantic relatedness on reading times may be due to a confound with word predictability. Based on these findings, we expect reading times to be mainly affected by plausibility on both the target and spillover regions. In particular, we expect longer reading times for critical words that are lexically associated with the preceding context but implausible, compared to associated and plausible targets.

3.3.1. Method

3.3.1.1. Participants

Thirty-one participants from Saarland University took part in the experiment. All had normal or corrected-to-normal vision, and none had participated in the DBC study. All were native German speakers, gave a written informed consent and were paid to take part in the experiment.

3.3.1.2. Materials

The materials were the same as those used in the DBC study. There were 90 two-sentence discourses in German in three conditions (baseline, event-related implausible, event-unrelated implausible) intermixed with 90 filler passages. Experimental items and fillers were arranged in three counterbalanced lists (see, for details Delogu et al., 2019, p. 3–4).

3.3.1.3. Procedure

The procedure was maintained as close as possible to the procedure in the ERP study by DBC. The context sentence in each pair was presented as a whole. Then a fixation cross appeared in the center of the screen. Participants had to press the space bar on the keyboard to proceed. Next the target sentence appeared word-by-word in the center of the screen. Participants controlled the rate of presentation of each word by pressing the space bar. At the end of each trial participants were asked to judge the plausibility of the mini-discourse by pressing one of two keys on the keyboard. The position of the plausible and implausible keys was counterbalanced across participants.

3.3.1.4. Analysis

Statistical analyses were performed on two critical regions, the target word (*menu*) and a spillover region corresponding to the function word following the target (*und*)[2]. We present the results for the two regions separately and also for the two regions combined into a single one, in order to decrease noise. Prior to statistical analysis, reading times (RTs) shorter than 80 ms and longer than $2,500$ ms were discarded for each region (for the combined region, we discarded RTs shorter than 160 ms and longer than $5,000$ ms)[3]. Linear mixed-effects regression models (LMMs) were fitted to log-transformed RTs, with condition (three levels: baseline, event-related implausible, event-unrelated implausible), as the fixed effects, and participants and items as random effects. The condition variable was effect-coded. Contrasts were used to compare the two implausible conditions with the baseline (effect of plausibility) and the event-related with the event-unrelated conditions (effect of association in the implausible conditions). In evaluating the models, we started with the maximal structure of random effects, which included random intercepts and slopes for both subjects and items. The random structures were then simplified by progressively excluding the effects explaining the least amount of variability in the model (following Bates et al., 2015). For each statistical model, we report effect coefficients (β), standard errors (SEs), and t-values (t). If the absolute value of t exceeded 2.5, the coefficient was judged to be significant.

3.3.2. Results

3.3.2.1. Plausibility Judgements

Participants judged the baseline condition to be more plausible than the event-related and the event-unrelated conditions (baseline: 91%; event-related: 24%; event-unrelated: 8%). These results closely mirror the offline plausibility ratings and online judgments reported in the DBC study.

3.3.2.2. Reading Times

Figure 7 shows the results[4]. At the target word, participants were slower to read both in the event-related (M $=$ 434.8 ms, SD $=$ 182.9) and the event-unrelated (M $=$ 450.6 ms, SD $=$ 221.8) conditions compared to the baseline (M $=$ 416.8 ms, SD $=$ 175.3). The results of the LMM analysis revealed a significant effect of plausibility ($\beta = 0.035$, SE $= 0.013$, t $= 2.64$) and no difference between the two implausible conditions ($\beta = 0.018$, SE $= 0.019$, t $= 0.985$).

The same reading time pattern emerged at the spillover word. Participants were slower to read both in the event-related (M $=$ 377.9 ms, SD $=$ 89.0) and the event-unrelated (M $=$ 389.7 ms, SD $=$ 95.5) conditions compared to the baseline (M $=$ 359.5 ms, SD $=$ 84.5). While the effect of plausibility was significant ($\beta = 0.05$, SE $= 0.013$, $t = 3.960$), the difference between the event-related and the event-unrelated conditions was not ($\beta = 0.022$, SE $= 0.014$, $t = 1.61$).

LMMs on the region including both the target and the spillover word showed an effect of plausibility ($\beta = 0.051$, SE $= 0.012$, $t = 4.299$) and a marginal difference between the two implausible conditions ($\beta = 0.028$, SE $= 0.014$, $t = 2.004$).

To summarize, in the analysis of the target and spillover regions, both the event-related and the event-unrelated conditions took longer to read than the baseline, suggesting that reading times were sensitive to plausibility rather than association. However, the event-unrelated condition was numerically slower than the event-related condition, possibly suggesting an additive effect of association and plausibility. To further investigate the relative contribution of these factors in predicting reading times, we fitted LMMs to log-transformed RTs in the merged target and spillover region, with plausibility and association ratings (and their interaction) as continuous predictors, and participants and items as random factors. Both plausibility and association were inverted and z-transformed prior to analysis (see Brouwer et al., 2020). Model selection procedure was the same as in the previous analysis. There was no effect of association ($\beta = 0.005$, SE $= 0.010$, $t = 0.49$), and no interaction of association and plausibility ($\beta = 0.006$,

[2] The precritical region did not show any significant difference between conditions.

[3] The reading time data is available at: https://github.com/hbrouwer/dbc2019rerps.

[4] We did not exclude trials from the analyses on the basis of the results from the plausibility judgments.

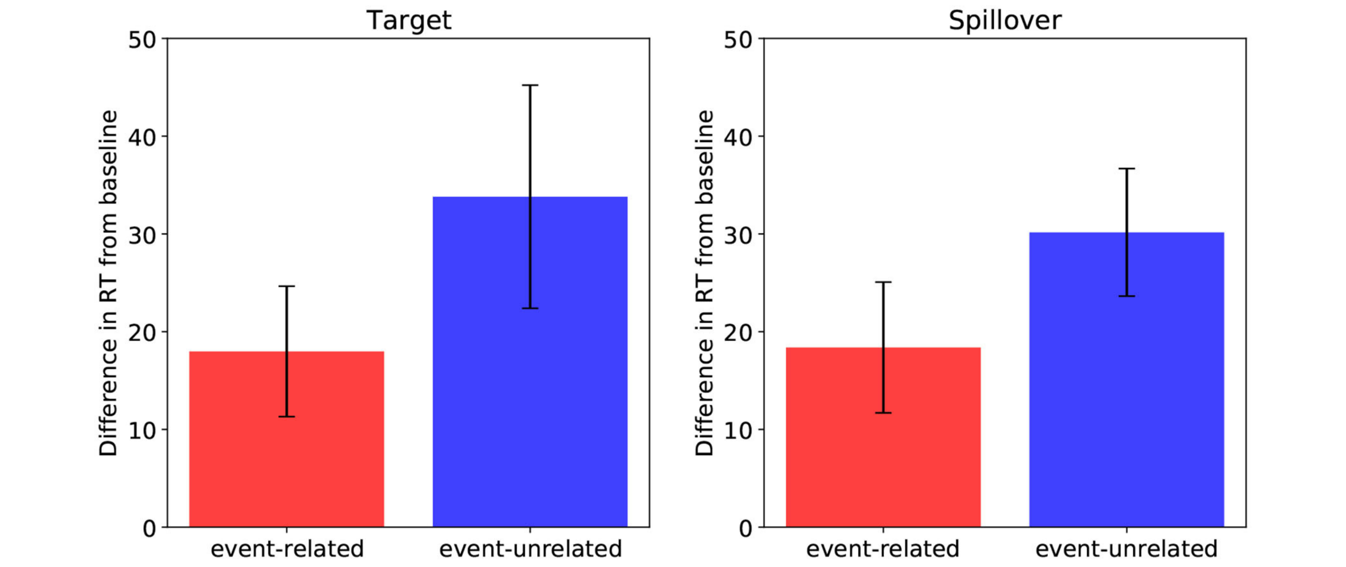


FIGURE 7 | Self-paced reading times (RTs) effects in the target region **(left)** and the spillover region **(right)**, for the event-related condition relative to baseline, and for the event-unrelated condition relative to baseline. Error bars show standard errors.

SE = 0.011, t = 0.53). Plausibility, however, significantly predicted reading times in this region ($\beta = 0.025$, SE = 0.009, t = 2.717). Thus, plausibility appears to be a more robust predictor of reading times than association in the target and spillover region.

In sum, the behavioral results show increased reading times for both the event-related and event-unrelated condition relative to baseline, and no effect of association, consistent with previous findings showing a reading time cost for implausible targets (e.g., Ledoux et al., 2006). These results pattern with the P600 results from DBC (compare **Figure 7** to **Figure 6**), as well as with the P600 and Surprisal estimates from the model (compare **Figure 7** to **Figures 3**, **4**).

4. DISCUSSION

We have presented a neurocomputational model of incremental, word-by-word language comprehension that produces N400, P600, and Surprisal estimates for each word. In this model, which integrates the neurocomputational model of the Retrieval-Integration account (Brouwer et al., 2017) with a "comprehension-centric" model of Surprisal (Venhuizen et al., 2019a), N400 amplitude is hypothesized to reflect the effort involved in the context-dependent retrieval of word meaning, P600 amplitude is hypothesized to index the work required to integrate this retrieved word meaning into the unfolding utterance interpretation, and Surprisal is taken to reflect the likelihood of the resultant interpretation, given the interpretation prior to integrating the meaning contributed by the incoming word. We set out to test a key prediction of the model: The P600, and not the N400, indexes "comprehension-centric" Surprisal. To investigate this link, we obtained model predictions for a recent study by Delogu et al. (2019, DBC), which directly investigated the electrophysiological correlates of plausibility-induced Surprisal. We found that—when spatiotemporal overlap between the empirically observed N400 and P600 is taken into account—the predictions of the model closely align with the empirical ERP data, showing that while the N400 is driven by association between a target word and its context, plausibility drives the P600. Further, to assess the alignment of the Surprisal estimates of the model with behavioral indices of processing difficulty, we presented the results from a self-paced reading replication of the DBC study. These empirical results again align closely with the model predictions, showing increases in reading times that are predominantly driven by plausibility. Taken together, our results thus support the conclusion that the P600 is an index of "comprehension-centric" Surprisal.

While we have focused on plausibility-induced semantic Surprisal, this conclusion is consistent with the proposal that the P600 is an overarching index of compositional semantic processes (Brouwer et al., 2012), which is sensitive to syntax (e.g., Osterhout and Holcomb, 1992; Hagoort et al., 1993; Gouvea et al., 2010), semantics (e.g., Kutas and Hillyard, 1980; Kolk et al., 2003; Hoeks et al., 2004), and pragmatics (e.g., Burkhardt, 2006; van Berkum et al., 2007; Dimitrova et al., 2012). Moreover, by establishing a link between the P600 and expectancy, as quantified through Surprisal, an interesting question arises, namely if the P600 is indeed an instance of the P300, and in particular of the late P3b subcomponent that has been shown to be sensitive to the detection of salient "oddball" stimuli (for recent discussion, see Sassenhagen and Fiebach, 2019; Leckey and Federmeier, 2020). On the one hand, the proposed link between the P600 and expectancy may be tentatively be taken to suggest that the integrative processes underlying this component are similar to the hypothesized context-updating mechanisms underlying the P300 (Donchin and Coles, 1988). On the other hand, the P300 is strongly dependent on task-demands, and while the P600 is sensitive to the task at hand, the presence of an explicit task it not a prerequisite for its elicitation (Kolk et al., 2003). Hence, while the "P600-as-P3 hypothesis" (Sassenhagen et al., 2014) poses interesting question, our results do not further elucidate this relationship.

Importantly, the conclusion that the P600 indexes comprehension-centric Surprisal is fully consistent with results showing a reliable correlation between Surprisal and the N400 (e.g., Frank et al., 2015, who employ word Surprisal estimates derived from a language model). In fact, it follows from the architecture of the model that the unfolding interpretation should influence the retrieval of word meaning—which modulates the N400 estimates—through lexical and contextual priming (see Kutas and Federmeier, 2000; Lau et al., 2008; van Berkum, 2009; Brouwer et al., 2012; Brouwer and Hoeks, 2013, for detailed discussions on how these factors may influence retrieval). Indeed, the N400 is effectively a function of the degree to which the memory system anticipates the conceptual knowledge associated with an incoming word, and in general, anticipation in the memory system tends to correlate with the expectancy of a word, as quantified through its Cloze probability (Kutas et al., 1984; Thornhill and Van Petten, 2012). In these cases, N400 amplitude patterns with interpretation-level Surprisal, but is not a direct reflection of it. Crucially, studies such as those by DBC underline this indirectness, as they show that the semantic association of a target word to its context can overrule its unexpectedness, thereby producing no difference between expected and unexpected targets in the N400; also see the literature on Semantic Illusions (e.g., Kuperberg, 2007; Bornkessel-Schlesewsky and Schlesewsky, 2008; Brouwer et al., 2012, for reviews). It should be noted, however, that unlike in many of the Semantic Illusion studies, the DBC study rules out an explanation in which the absence of an N400-effect for unexpected, but associated targets is due to "shallow" integrative processing—as assumed in models in which the N400 is itself a direct index of integrative semantic processing (e.g., Rabovsky et al., 2018)—because the robust P600-effect for this condition, as well as high accuracy in behavioral implausibility judgments, show that comprehenders are explicitly aware of the unexpectedness of the target (see also Sanford et al., 2011). Further, given that the target sentences of the DBC stimuli were globally and locally unambiguous, this observed P600-effect cannot be explained by models that attribute the increase in P600 amplitude to index syntactic repair or reanalysis (e.g., Fitz and Chang, 2019).

We have qualitatively established the P600 as a direct index of "comprehension-centric" Surprisal by showing that its estimated amplitude increases in response to surprising, implausible target words, relative to unsurprising, plausible ones. An open question remains if the P600 is also a quantitative index of Surprisal; that is, if its amplitude is sensitive to expectancy in a graded manner. The experiment by DBC was not designed to address this question. We do observe, however, in both the electrophysiological and the behavioral results that the event-related condition at least numerically incurs less processing difficulty than the event-unrelated condition. Indeed, this is in line with the offline plausibility ratings and Cloze ratings, in which the event-related condition is rated as more plausible and expected than the event-unrelated condition, respectively. While this may suggest a graded difference in Surprisal between these conditions, we believe these ratings to be confounded by association; that is, in the event-related condition, the strong semantic association of a target word to its context, leads people to judge them as slightly more plausible, than the unassociated, implausible target words in the event-unrelated condition.

Interestingly, however, the model predicts the same graded pattern, both in its P600 estimates and in its Surprisal estimates, as observed in the empirical data. Crucially, in constructing the meaning space—from which the utterance meaning representations that the model recovers in processing are derived—we did not explicitly induce any probabilistic difference between the semantics associated with the two implausible conditions. Yet, we do observe a difference in that the semantics associated with the event-related sentences are slightly more probable than the semantics associated with the event-unrelated sentences. This difference can be explained by the structure of the meaning space, which is defined in terms of probabilistic co-occurrences. Indeed, given that the baseline and event-related condition share many of the same presuppositions, as instantiated by *entity* predicate (see above), their semantics occupy parts of the same region of the overall meaning space. The event-unrelated semantics, by contrast, trigger a different set of presuppositions, thereby constituting a different part of the meaning space. As during processing the model navigates the meaning space on a word-by-word basis, this spatial organization directly affects its behavior, as reflected in its P600 and Surprisal estimates; that is, the target word in event-unrelated sentences triggers a larger transition in meaning space than the target word in event-related sentences, thereby explaining the difference in P600 and Surprisal estimates. Hence, it is the presence of referential presuppositions, which serve to associate specific targets with specific contexts, that explains the graded pattern in the model. On a speculative note, the model thus effectively predicts plausibility to be confounded with association, which numerically aligns with the offline ratings and empirical results.

In sum, while our results support a qualitative link between Surprisal and the P600, it remains an open question if this extends to a quantitative one, in that, like reading times, the P600 is sensitive to expectancy in a graded manner. Given the issue of spatiotemporal component overlap, however, addressing this question may be challenging, as manipulating expectancy in a graded manner may also yield graded N400 results, thereby rendering it non-transparent what is going on in the P600 (e.g., see Thornhill and Van Petten, 2012). In future work, this can be addressed by using rERP analyses, which allow for disentangling the N400 and the P600 in space and time, on results from co-registered reading time and ERP studies.

5. CONCLUSION

We have presented a neurocomputational model of incremental, word-by-word language comprehension that produces N400, P600, and "comprehension-centric" Surprisal estimates at each word in a sentence. In the model, estimated N400 amplitude reflects the effort involved in the contextualized retrieval of the meaning of an incoming word, while estimated P600 amplitude indexes the effort involved in integrating this retrieved word meaning into the unfolding utterance interpretation. Surprisal estimates, in turn, reflect the likelihood of an updated interpretation, given the interpretation prior to updating it. By testing it on an experimental design that directly tests "world-knowledge"-induced Surprisal, we have shown that the predictions of the model align with empirical electrophysiological results—when spatiotemporal component overlap between the N400 and P600 is taken into account—as well as with behavioral reading times. We find a close relationship between Surprisal, which we take to be reflected by reading times, and P600 amplitude, thereby supporting the interpretation of the P600 as the ERP component that indexes "comprehension-centric" Surprisal. Future work must determine if this link is only qualitative, or if it also holds quantitatively, in that the P600, like reading times, is sensitive to graded manipulations of expectancy. Overall, we believe that this theory-driven linkage of electrophysiological and behavioral correlates of processing difficulty, through explicit neurocomputational modeling, provides an important step toward an integrated neurobehavioral theory of language comprehension.

DATA AVAILABILITY STATEMENT

The ERP and reading time data is available at: https://github.com/hbrouwer/dbc2019rerps.

ETHICS STATEMENT

The studies involving human participants were approved by Deutsche Gesellschaft für Sprache (DGfS). The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

HB and NV conducted the computational modeling. FD conducted data analysis. All authors contributed equally to the writing of the manuscript.

FUNDING

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 232722074—SFB 1102.

ACKNOWLEDGMENTS

The authors would like to thank Ludmilla Borisenkov for her help in collecting the self-paced reading times data. We also thank the three reviewers for constructive comments on an earlier version of this manuscript.

REFERENCES

Albrecht, J. E., and O'Brien, E. J. (1993). Updating a mental model: maintaining both local and global coherence. *J. Exp. Psychol.* 19, 1061–1070. doi: 10.1037/0278-7393.19.5.1061

Barrés, V., Simons, A. III, and Arbib, M. (2013). Synthetic event-related potentials: a computational bridge between neurolinguistic models and experiments. *Neural Netw*. 37, 66–92. doi: 10.1016/j.neunet.2012.09.021

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*. Available online at: https://arxiv.org/abs/1506.04967

Beim Graben, P., Gerth, S., and Vasishth, S. (2008). Towards dynamical system models of language-related brain potentials. *Cogn. Neurodyn*. 2, 229–255. doi: 10.1007/s11571-008-9041-5

Bornkessel-Schlesewsky, I., and Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Res. Rev*. 59, 55–73. doi: 10.1016/j.brainresrev.2008.05.003

Boston, M. F., Hale, J. T., Kliegl, R., Patil, U., and Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: an evaluation using the Potsdam Sentence Corpus. *J. Eye Mov. Res*. 2, 1–12. doi: 10.16910/jemr.2.1.1

Brouwer, H., and Crocker, M. W. (2017). On the proper treatment of the N400 and P600 in language comprehension. *Front. Psychol.* 8:1327. doi: 10.3389/fpsyg.2017.01327

Brouwer, H., Crocker, M. W., Venhuizen, N. J., and Hoeks, J. C. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cogn. Sci*. 41, 1318–1352. doi: 10.1111/cogs.12461

Brouwer, H., Delogu, F., and Crocker, M. W. (2020). Splitting event-related potentials: modeling latent component using regression-based waveform estimation. *Eur. J. Neurosci*. doi: 10.1111/ejn.14961. [Epub ahead of print].

Brouwer, H., Fitz, H., and Hoeks, J. (2010). "Modeling the noun phrase versus sentence coordination ambiguity in Dutch: evidence from surprisal theory," in *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics* (Uppsala: Association for Computational Linguistics), 72–80.

Brouwer, H., Fitz, H., and Hoeks, J. (2012). Getting real about semantic illusions: rethinking the functional role of the P600 in language comprehension. *Brain Res*. 1446, 127–143. doi: 10.1016/j.brainres.2012.01.055

Brouwer, H., and Hoeks, J. C. (2013). A time and place for language comprehension: mapping the N400 and the P600 to a minimal cortical network. *Front. Hum. Neurosci*. 7:758. doi: 10.3389/fnhum.2013.00758

Burkhardt, P. (2006). Inferential bridging relations reveal distinct neural mechanisms: evidence from event-related brain potentials. *Brain Lang*. 98, 159–168. doi: 10.1016/j.bandl.2006.04.005

Camblin, C. C., Gordon, P. C., and Swaab, T. Y. (2007). The interplay of discourse congruence and lexical association during sentence processing: evidence from ERPs and eye tracking. *J. Mem. Lang*. 56, 103–128. doi: 10.1016/j.jml.2006.07.005

Cook, A. E., and Myers, J. L. (2004). Processing discourse roles in scripted narratives: the influences of context and world knowledge. *J. Mem. Lang*. 50, 268–288. doi: 10.1016/j.jml.2003.11.003

Crocker, M. W., Knoeferle, P., and Mayberry, M. R. (2010). Situated sentence processing: the coordinated interplay account and a neurobehavioral model. *Brain Lang*. 112, 189–201. doi: 10.1016/j.bandl.2009.03.004

Delogu, F., Brouwer, H., and Crocker, M. W. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain Cogn*. 135:103569. doi: 10.1016/j.bandc.2019.05.007

Demberg, V., and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109, 193–210. doi: 10.1016/j.cognition.2008.07.008

Dimitrova, D. V., Stowe, L. A., Redeker, G., and Hoeks, J. C. (2012). Less is not more: neural responses to missing and superfluous accents in context. *J. Cogn. Neurosci*. 24, 2400–2418. doi: 10.1162/jocn_a_00302

Donchin, E., and Coles, M. G. (1988). Is the P300 component a manifestation of context updating? *Behav. Brain Sci*. 11, 357–374. doi: 10.1017/S0140525X00058027

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci*. 14, 179–211. doi: 10.1207/s15516709cog1402_1

Fitz, H., and Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cogn. Psychol*. 111, 15–52. doi: 10.1016/j.cogpsych.2019.03.002

Frank, S. L. (2009). "Surprisal-based comparison between a symbolic and a connectionist model of sentence processing," in *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (Austin, TX: Cognitive Science Society), 1139–1144.

Frank, S. L. (2017). "Word embedding distance does not predict word reading time," in *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (Austin, TX: Cognitive Science Society), 385–390.

Frank, S. L., Haselager, W. F., and van Rooij, I. (2009). Connectionist semantic systematicity. *Cognition* 110, 358–379. doi: 10.1016/j.cognition.2008.11.013

Frank, S. L., Koppen, M., Noordman, L. G., and Vonk, W. (2003). Modeling knowledge-based inferences in story comprehension. *Cogn. Sci*. 27, 875–910. doi: 10.1207/s15516709cog2706_3

Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain Lang*. 140, 1–11. doi: 10.1016/j.bandl.2014.10.006

Frank, S. L., and Vigliocco, G. (2011). Sentence comprehension as mental simulation: an information-theoretic perspective. *Information* 2, 672–696. doi: 10.3390/info2040672

Gouvea, A. C., Phillips, C., Kazanina, N., and Poeppel, D. (2010). The linguistic processes underlying the P600. *Lang. Cogn. Process*. 25, 149–188. doi: 10.1080/01690960902965951

Hagoort, P. (2003). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *J. Cogn. Neurosci*. 15, 883–899. doi: 10.1162/089892903322370807

Hagoort, P., Brown, C., and Groothusen, J. (1993). The Syntactic Positive Shift (SPS) as an ERP measure of syntactic processing. *Lang. Cogn. Process*. 8, 439–483. doi: 10.1080/01690969308407585

Hale, J. (2001). "A probabilistic Earley parser as a psycholinguistic model," in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (Stroudsburg, PA: Association for Computational Linguistics), 159–166. doi: 10.3115/1073336.1073357

Hale, J. T. (2003). The information conveyed by words in sentences. *J. Psycholinguist. Res*. 32, 101–123. doi: 10.1023/A:1022492123056

Hoeks, J. C. J., Stowe, L. A., and Doedens, G. (2004). Seeing words in context: the interaction of lexical and sentence level information during reading. *Cogn. Brain Res*. 19, 59–73. doi: 10.1016/j.cogbrainres.2003.10.022

Knoeferle, P., Crocker, M. W., Scheepers, C., and Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. *Cognition* 95, 95–127. doi: 10.1016/j.cognition.2004.03.002

Kolk, H. H. J., Chwilla, D. J., van Herten, M., and Oor, P. J. W. (2003). Structure and limited capacity in verbal working memory: a study with event-related potentials. *Brain Lang*. 85, 1–36. doi: 10.1016/S0093-934X(02)00548-5

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: challenges to syntax. *Brain Res*. 1146, 23–49. doi: 10.1016/j.brainres.2006.12.063

Kutas, M., and Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends Cogn. Sci*. 4, 463–470. doi: 10.1016/S1364-6613(00)01560-6

Kutas, M., and Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* 207, 203–205. doi: 10.1126/science.7350657

Kutas, M., Lindamood, T. E., and Hillyard, S. A. (1984). "Word expectancy and event-related brain potentials during sentence processing," in *Preparatory States & Processes: Proceedings of the Franco-American Conference* (Ann Arbor, MI: Lawrence Erlbaum), 217. doi: 10.4324/9781315792385-11

Lau, E. F., Phillips, C., and Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nat. Rev. Neurosci*. 9, 920–933. doi: 10.1038/nrn2532

Leckey, M., and Federmeier, K. D. (2020). The P3b and P600(s): positive contributions to language comprehension. *Psychophysiology* 57:e13351. doi: 10.1111/psyp.13351

Ledoux, K., Camblin, C. C., Swaab, T. Y., and Gordon, P. C. (2006). Reading words in discourse: the modulation of lexical priming effects by message-level context. *Behav. Cogn. Neurosci. Rev*. 5, 107–127. doi: 10.1177/1534582306289573

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

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behav. Res. Methods* 37, 547–559. doi: 10.3758/BF03192726

Morris, R. K. (1994). Lexical and message-level sentence context effects on fixation times in reading. *J. Exp. Psychol*. 20, 92–102. doi: 10.1037/0278-7393.20.1.92

Myers, J. L., and O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discour. Process*. 26, 131–157. doi: 10.1080/01638539809545042

Osterhout, L., and Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *J. Mem. Lang*. 31, 785–806. doi: 10.1016/0749-596X(92)90039-Z

Rabovsky, M., Hansen, S. S., and McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nat. Hum. Behav*. 2, 693–705. doi: 10.1038/s41562-018-0406-4

Roark, B., Bachrach, A., Cardenas, C., and Pallier, C. (2009). "Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Vol. 1* (Stroudsburg, PA: Association for Computational Linguistics), 324–333. doi: 10.3115/1699510.1699553

Rohde, D. L. (2002). *A connectionist model of sentence comprehension and production* (Ph.D. thesis). School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, United States.

Rohde, D. L., Gonnerman, L. M., and Plaut, D. C. (2005). An improved model of semantic similarity based on lexical co-occurrence. *Commun. ACM* 8:116. Available online at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.9401&rep=rep1&type=pdf

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Sanford, A. J., Leuthold, H., Bohan, J., and Sanford, A. J. S. (2011). Anomalies at the borderline of awareness: an ERP study. *J. Cogn. Neurosci*. 23, 514–523. doi: 10.1162/jocn.2009.21370

Sassenhagen, J., and Fiebach, C. J. (2019). Finding the P3 in the P600: decoding shared neural mechanisms of responses to syntactic violations and oddball targets. *NeuroImage* 200, 425–436. doi: 10.1016/j.neuroimage.2019.06.048

Sassenhagen, J., Schlesewsky, M., and Bornkessel-Schlesewsky, I. (2014). The P600-as-P3 hypothesis revisited: single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. *Brain Lang*. 137, 29–39. doi: 10.1016/j.bandl.2014.07.010

Smith, N. J., and Kutas, M. (2015a). Regression-based estimation of ERP waveforms: I. the rERP framework. *Psychophysiology* 52, 157–168. doi: 10.1111/psyp.12317

Smith, N. J. and Kutas, M. (2015b). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology* 52, 169–181. doi: 10.1111/psyp.12320

Smith, N. J., and Levy, R. (2008). "Optimal processing times in reading: a formal model and empirical investigation," in *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (Austin, TX: Cognitive Science Society), 595–600.

Smith, N. J., and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition* 128, 302–319. doi: 10.1016/j.cognition.2013.02.013

Thornhill, D. E., and Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: frontal positivity and N400 ERP components. *Int. J. Psychophysiol*. 83, 382–392. doi: 10.1016/j.ijpsycho.2011.12.007

van Berkum, J. J. A. (2009). "The 'neuropragmatics' of simple utterance comprehension: an ERP review," in *Semantics and Pragmatics: From Experiment to Theory* (Basingstoke: Palgrave Macmillan), 276–316.

van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *J. Exp. Psychol*. 31, 443–467. doi: 10.1037/0278-7393.31.3.443

van Berkum, J. J. A., Koornneef, A. W., Otten, M., and Nieuwland, M. S. (2007). Establishing reference in language comprehension: an electrophysiological perspective. *Brain Res*. 1146, 158–171. doi: 10.1016/j.brainres.2006.06.091

Venhuizen, N. J., Crocker, M. W., and Brouwer, H. (2019a). Expectation-based comprehension: modeling the interaction of world knowledge and linguistic experience. *Discour. Process*. 56, 229–255. doi: 10.1080/0163853X.2018.1448677

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

Venhuizen, N. J., Hendriks, P., Crocker, M. W., and Brouwer, H. (2019c). "A framework for distributional formal semantics," in *International Workshop on Logic, Language, Information, and Computation* (Berlin; Heidelberg: Springer), 633–646. doi: 10.1007/978-3-662-59533-6_39

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.

PERSPECTIVE
published: 25 March 2021
doi: 10.3389/fpsyg.2021.647076

Rational and Flexible Adaptation of Sentence Production to Ongoing Language Experience

*Malathi Thothathiri**

Department of Speech, Language and Hearing Sciences, The George Washington University, Washington, DC, United States

Whether sentences are formulated primarily using lexically based or non-lexically based information has been much debated. In this perspective article, I review evidence for rational flexibility in the sentence production architecture. Sentences can be constructed flexibly via lexically dependent or independent routes, and rationally depending on the statistical properties of the input and the validity of lexical vs. abstract cues for predicting sentence structure. Different neural pathways appear to be recruited for individuals with different executive function abilities and for verbs with different statistical properties, suggesting that alternative routes are available for producing the same structure. Together, extant evidence indicates that the human brain adapts to ongoing language experience during adulthood, and that the nature of the adjustment may depend rationally on the statistical contingencies of the current context.

Keywords: statistical learning, cue validity, executive function, dorsal stream, ventral stream, individual differences, verb bias

OPEN ACCESS

Edited by:
Elke Teich,
Saarland University, Germany

Reviewed by:
Francesca Delogu,
Saarland University, Germany
Anouschka Foltz,
University of Graz, Austria

****Correspondence:***
Malathi Thothathiri
malathi@gwu.edu

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

Received: *29 December 2020*
Accepted: *01 March 2021*
Published: *25 March 2021*

Citation:
Thothathiri M (2021) Rational and Flexible Adaptation of Sentence Production to Ongoing Language Experience.
Front. Psychol. 12:647076.
doi: 10.3389/fpsyg.2021.647076

INTRODUCTION

Sentence production involves converting thoughts into structured sequences of words. The representations and processes used to formulate these structured sequences are subject to theoretical debate (see e.g., Konopka and Bock, 2009; Lane and Ferreira, 2010). Consider, for example, a situation where a speaker would like to describe to a listener the information that Amelia had given a bag to John. The speaker's brain could accomplish this communicative act by choosing an abstract structural frame associated with transfer events [e.g., <Agent> <transfer verb> <theme> to <recipient> or Noun-Phrase (NP) Verb (V) NP Preposition NP] and subsequently filling in the specific verb and the other words (e.g., *give, bag*). Alternatively, the structured sequence could be formulated by first choosing the core verb (e.g., *give*) and then accessing the structural information associated with that verb (e.g., where the different arguments of *give* can be placed in a sentence). This debate is often posed as a dichotomy but it is possible that both routes to sentence production are available and can be chosen under different circumstances. The perspective put forth in this paper is that the path to sentence formulation can be rational and flexible i.e., depending on the statistical properties of ongoing language experience, the brain can come to rely on either verb-specific or verb-general representations for sentence production in a given context. This process is rational because the choice is tuned to the statistical contingencies of the current context. It is flexible because the architecture adapts to changing statistical contingencies throughout the lifespan.

Under a rationalist view, learning to understand and produce sentences involves learning which sentence structures are the most likely to be used in the future based on past experience. The human brain can encode and use past experience at different granularities, including all prior input, the most recent input, and input tied to specific cues and contexts (Ellis, 2006). Probability-based tuning is rational because past experience is a good predictor of future occurrence. Additionally, how language is used differs across speakers, dialects, and modalities. Therefore, continual tuning *post*-acquisition allows the language user to adapt appropriately to the current context (Fine et al., 2013). But does sentence formulation adjust rationally and flexibly to ongoing input in this way? Below, I first describe independent evidence for the verb-general and verb-specific routes to sentence production before turning to how the choice between the two adapts to current statistical properties.

Structural priming studies are a predominant source of evidence for the debate between frame-based or abstract syntactic accounts and lexicalist accounts of sentence production. Comprehending or producing a syntactic structure (e.g., a prepositional-object dative like *The wealthy widow gave her Mercedes to the church*) increases the likelihood of speakers using the same structure again with unrelated verbs and nouns (e.g., *The grandfather is reading a story to his grandson*). Such priming, independent of lexical overlap, suggests a role for abstract sentential frames that are not tied to specific lexical items (Bock and Griffin, 2000; Konopka and Bock, 2009; inter alia). Even idiomatic phrases, which are widely assumed to be lexicalized, show abstract priming (Konopka and Bock, 2009). Other non-priming evidence from stem-exchange errors (e.g., "hates the record" becomes "records the hate") suggests that the production of syntactic-category-consistent stress (e.g., REcord vs. reCORD) is influenced by abstract syntax rather than by lexical selection, consistent with frame-based theories (Lane and Ferreira, 2010).

However, lexical influences on sentence production have also been noted. Structural priming shows a "lexical boost" when the verb repeats between prime and target sentences (Pickering and Branigan, 1998; Hartsuiker et al., 2008). This suggests that structural information tied to specific lexical items can be primed. In naturalistic speech, some verbs (e.g., *give*) can appear in two alternative structures while others are grammatical in only one of the two options [e.g., *donate* is acceptable in prepositional-object (PO) datives like *Laila donated money to the church* but not in double-object (DO) datives like **Laila donated the church money*]. Thus, sentence production can be sensitive to the usage pattern of a specific verb (hereafter referred to as "verb bias").

Earlier evidence had led some researchers to suggest a difference between sentence comprehension and production such that the former is guided more strongly by the lexicon and the latter by abstract syntax (e.g., Arai et al., 2007). However, a recent study compared the two modalities directly and found similar effects, leading the authors to conclude in favor of shared mechanisms for understanding and formulating sentences (Tooley and Bock, 2014). In particular, both abstract structural priming and a lexical boost were detected, indicating that the brain uses structural information stored at lexically independent as well as lexically dependent levels.

If both routes to sentence production are available, how does the brain choose which one to use when? Artificial languages are a useful way to control the language input of participants whose real-life language experiences may be variable. Though these paradigms tap learning a new language, the findings are relevant for natural language use (Wonnacott et al., 2008; Romberg and Saffran, 2010). Further, in the present perspective, language *use* is intricately tied to *learning* the context-appropriate properties of the input. Therefore, I begin by reviewing evidence from artificial language studies before describing the findings for natural language. To preview, this emerging evidence supports the idea of flexibility by showing that:

(1) speakers learn and use new verb biases from short lab-based input sessions not only in an artificial language but also in their native language (Wonnacott et al., 2008; Thothathiri and Rattinger, 2016; Thothathiri et al., 2017. See also Ryskin et al., 2017).
(2) the brain differentially uses alternative processing streams for producing the same structural output for verbs with different statistical properties (Thothathiri and Rattinger, 2015).
(3) frontal executive function regions are recruited differentially in different individuals and for different verb biases (Thothathiri, 2018).

The adaptation appears to be rational, as evidenced by:

(1) sensitivity to verb-specific or verb-general cues depending on the predictive validity of those cues (Thothathiri and Rattinger, 2016; Thothathiri and Braiuca, 2020. See also Perek and Goldberg, 2017).
(2) a division of labor between neural pathways such that effortful semantic processing is engaged only when simpler contingencies are unavailable (Thothathiri and Rattinger, 2015).

Rational and Flexible Adaptation of Sentence Production in an Artificial Language

In a seminal artificial language study, Wonnacott et al. (2008) showed that adult learners tracked both verb-specific and verb-general statistics and used these sources of information in a rational manner that was dependent on the distribution of verbs and verb types in the input language. Specifically, sentence production after language exposure showed a more lexically specific pattern for high frequency verbs and/or if most verbs in the language were biased toward one or another structure and did not appear in both structures (making individual verbs useful predictors for how they should be used). Conversely, verbs were more likely to be generalized to a structure that they had not appeared in if they were low frequency (providing insufficient verb-specific information) or if the language predominantly contained alternating verbs that appeared in both structures (biasing toward verb-general patterns). The authors concluded that the findings were consistent with a rational Bayesian approach to learning (see also Perfors et al., 2010).

Thothathiri and Rattinger (2016) extended these findings to different types of cues, namely verb-specific syntactic distribution and verb-general semantics-to-structure mappings. They demonstrated that adults could learn which cue was a better predictor of structures heard in the input and prioritize the cue with higher validity for guiding subsequent language use. In Experiment 1, participants were exposed to an artificial language where two alternative structures (Agent-Patient vs. Patient-Agent order) were used equally often to describe transitive actions, making the event semantics non-predictive. Ten out of 12 verbs were biased to appear in one of the two structures, making the verb cue highly predictive of the structure heard during input. Under these conditions, participants' free-choice sentence production in a subsequent test showed a verb-specific pattern, with higher Patient-Agent order produced for verbs that were heard in that order than for verbs that were not. Experiments 2 and 3 (with new participants) made the verb-general semantic cue more predictive than the verb cue by associating two different word orders with two different kinds of events (an event involving an instrument vs. a modifier). Notably, 10 out of 12 verbs were still biased to appear in one of the two structures. Thus, the verb was still highly (but not 100%) predictive. However, the competing semantic cue—namely, whether the observed event involved an instrument or a modifier—was even more (100%) predictive. Under these conditions, speakers overrode verb-specific statistics and used the structure that was appropriate for the event semantics. The authors concluded that sentence production need not be exclusively lexically conservative or generalized. Instead, it can be guided flexibly and rationally by different representations depending on the predictive validities of different cues (Bates and MacWhinney, 1987, 1989; Goldberg et al., 2005; MacWhinney, 2013).

Rational and Flexible Adaptation of Sentence Production in the Speakers' Native Language

Subsequent studies using a similar methodology in the speakers' native language (English) showed that language users maintain some flexibility in adulthood (Thothathiri et al., 2017; Thothathiri, 2018; Thothathiri and Braiuca, 2020). English speakers learned to use new biases for known dative verbs and a new semantic cue for known dative structures in a manner consistent with cue validity. This is remarkable given the extent of prior English exposure for a speaker who is 18 years or older. The results highlight the fact that language continues to adapt past the childhood stage of acquisition (see also Kamide, 2012; Kroczek and Gunter, 2017; Ryskin et al., 2017) and that the brain rationally learns to use cues that are highly predictive in the current context.

In Thothathiri and colleagues' natural language experiments, participants were provided with lab-based English input containing dative sentences (Thothathiri et al., 2017; Thothathiri, 2018; Thothathiri and Braiuca, 2020). As before, different verbs were biased to appear in different structures, with some appearing exclusively in DO, others exclusively in PO, and yet others equally in both. The assignment of different dative verbs to different bias conditions was counterbalanced across lists. Would native English speakers adapt flexibly to these new biases for known verbs? Thothathiri et al. (2017) found that they did. Across this and other studies below, DO datives were uniformly less common than PO, suggesting that it was the harder structure (note: these DO datives contained full-noun-phrase objects, which occur less commonly in a DO structure than pronouns). Within this overarching tendency, there was differentiation between bias conditions: speakers were most likely to produce DO with verbs that had been heard only in that structure during lab-based exposure and least likely to do so with verbs that had been heard only in the competing PO (with Equi or equal-DO-PO verbs in between), resulting in a significant linear pattern (DO-only > Equi > PO-only).

In a subsequent study, Thothathiri and Braiuca (2020) investigated whether adaptation to new input depends rationally on the relative validity of verb-specific vs. general semantic cues. As before, participants were exposed to lab-based dative input with different verbs assigned to different bias conditions. However, the new experiments included a 100% predictive semantic cue—complete transfer actions where the theme successfully reached the recipient were always described using DO while incomplete transfers were always described using PO. Will event semantics override verb-specific statistics because it has higher cue validity (as in the artificial language experiments in Thothathiri and Rattinger, 2016)? The results presented a nuanced picture. Sentence structure choice and utterance characteristics showed an influence of event semantics when the semantic cue was much more predictive than individual verbs (100 vs. 60 or 70%) but not when the two cues were closer in their validities (100 vs. 90%). In fact, there was a reliable effect of the verb and not the semantic cue in the latter case despite the fact that the verb cue had lower validity. These patterns led the authors to conclude that prior knowledge about the relevance of the verb cue for English datives could mean that it continues to influence native language sentence production under new input conditions. Although the human brain can track and use statistical associations rationally, it is subject to selection biases because some cues might be attended to more selectively and weighted more heavily than is warranted by their predictive validity (see Ellis, 2006 for discussion of similar issues within second language acquisition).

Neural Mechanisms

Functional magnetic resonance imaging (fMRI) studies provide complementary evidence for rational and flexible adaptation at the level of neural mechanisms. Prior research has suggested that the brain rationally employs "division of labor" between semantic and non-semantic processes for language processing (Plaut et al., 1996; Ueno et al., 2014). In the context of sentence production, the brain flexibly weights the ventral (semantic) and dorsal (non-semantic) streams differently for producing the same dative structure for verbs with different statistical properties. The weightings appear to be rational, favoring effortful semantic processing only when necessary i.e., when there are no easier contingencies present in the input for a given verb.

Thothathiri and Rattinger (2015) first demonstrated flexibility and rational division of labor in an artificial language paradigm. After exposure to the language (as described above), participants' brains were scanned during sentence production in a separate session. The analyses focused on whether producing the harder word order (Patient-Agent) compared to the common one (Agent-Patient) recruited different regions for verbs with different biases (Agent-Patient only, Patient-Agent only, or Equi). The results showed greater bilateral temporal lobe activation and greater functional connectivity between speech motor areas and the right temporal lobe for Equi verbs than for verbs that had appeared in a single order during the input phase. Thus, there was increased involvement of the ventral stream for Equi verbs, which could have resulted from competition between multiple structures for the same verb and deeper semantic processing for identifying meaning-to-order mappings. By contrast, verbs encountered in a single consistent mapping may have been directly associated with their corresponding structures without extensive semantic processing[1]. More broadly, the results showed that the brain can accomplish the same structural output using different alternative pathways.

The brain might also rationally adapt by using different resources in individuals with different cognitive profiles. The relevant studies have focused on frontal-cortex-supported executive function because of its documented role in adaptive, context-appropriate behavior (Koechlin, 2016). Thothathiri and Rattinger (2015) found that better executive function as measured by the Stroop task correlated with a higher proportion of the harder Patient-Agent order for Equi verbs but not for verbs that appeared in a single order. Thus, input statistical properties (verb bias condition) interacted with learner characteristics (Stroop performance) in predicting sentence production choices. This finding was later corroborated by Thothathiri et al. (2017), who examined native language production using English dative structures and found a correlation between individuals' Stroop performance and their production of the harder DO dative for Equi but not for other verbs. A subset of the participants in the latter study took part in a subsequent fMRI session where their brains were scanned during free-choice dative sentence production (Thothathiri, 2018). When producing the harder DO dative after the easier PO dative, participants with better Stroop performance activated the anterior cingulate cortex (ACC) more than those with poorer performance. Furthermore, there was an interaction between learner characteristics and input statistical properties such that individual differences in ACC activation were maximal for PO-only verbs produced in the opposite DO, smallest for DO-only verbs produced in DO, and in between the two for Equi verbs. Functionally, ACC activation was correlated with increased DO production over time for Equi and decreased DO production for PO-only verbs (there was no correlation for DO-only verbs). This suggests that the ACC influences language production in different ways for different verbs in a manner that is consistent with recent experience. It can help boost the production of a difficult sentence structure that is in competition with an easier structure if that structure is sanctioned by recent statistical experience (as in the case of Equi verbs)[2]. Conversely, it can help suppress the production of that same structure if recent experience suggests that the structure is not sanctioned (as for PO-only verbs). Together, these findings raise the intriguing possibility that ACC (and other frontal regions) might be involved in rational and flexible adaptation of language based on speaker, input and context characteristics.

TABLE 1 | Open questions.

Open questions for future research
Input statistical factors
(1) What is the effect of prior knowledge about the validities of different cues? Under what conditions, if any, do speakers override prior knowledge?
(2) What are the relevant grains of prior knowledge? Does the brain track predictive validities separately for different structural alternations within a language?
(3) Are there conditions (e.g., discourse contexts) under which speakers ignore predictive validities entirely? What features might such conditions share?
Brain regions and mechanisms
(1) What are the relevant individual differences in cognitive abilities for sentence production? Are these differences and their effects stable over time?
(2) What is the division of labor between ventral and dorsal streams for different structures and input conditions?
(3) Is executive function necessary or merely facilitative for flexibly choosing between alternative routes to sentence production?
(4) What mechanisms are used to consolidate prior and ongoing language experiences?

DISCUSSION

The proposed perspective is consistent with longstanding ideas in the study of language, including cue validity (Bates and MacWhinney, 1987), constraint-based sentence processing (MacDonald et al., 1994; Trueswell and Tanenhaus, 1994), division of labor (Plaut et al., 1996; Ueno et al., 2014), and Bayesian learning (Perfors et al., 2010). The available evidence is intriguing but many open questions remain, which are summarized in **Table 1**.

For example, Thothathiri and Braiuca (2020) suggested that prior knowledge about the relevance of verb bias for English datives could have continued to affect speakers' sentence production in the new context. The nature of the relevant prior knowledge as well as the mechanisms used to consolidate prior and ongoing language experiences remain to be fleshed out (but see Chang et al., 2006; Fine et al., 2013). Multiple studies suggest flexibility in the cues and pathways used for sentence production (Thothathiri and Rattinger, 2015, 2016; Thothathiri, 2018) but additional work is needed to build a comprehensive theoretical framework that explains (a) how predictive validity might rationally change the weighting of different brain regions, and (b) how executive function may be used to select sentence structures

[1] This is analogous to reading aloud regular words, whose letters can be translated directly to the corresponding sounds, without lexical semantic processing.

[2] DO-biased verbs appeared repeatedly and only in the DO structure. This statistical association facilitates DO production for these verbs without much competition from the alternative PO structure.

under different conditions and for different individuals. Going beyond these questions that are closely related to the perspective described here, it is also important to investigate how context-specific the effects of exposure are and how long they last (Wells et al., 2009; Kamide, 2012).

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.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board at GWU. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MT supervised all the research reported here and wrote all versions of the manuscript.

REFERENCES

Arai, M., Van Gompel, R. P., and Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cogn. Psychol.* 54, 218–250. doi: 10.1016/j.cogpsych.2006.07.001

Bates, E., and MacWhinney, B. (1987). "Competition, variation, and language learning," in *Mechanisms of Language Acquisition*, ed B. MacWhinney (Hillsdale, NJ: Erlbaum), 157–193.

Bates, E., and MacWhinney, B. (eds.). (1989). "Functionalism and the competition model," in *The Cross-Linguistic Study of Sentence Processing* (Cambridge: Cambridge University Press), 10–75.

Bock, K., and Griffin, Z. M. (2000). The persistence of structural priming: transient activation or implicit learning? *J. Exp. Psychol. General* 129, 177–192. doi: 10.1037/0096-3445.129.2.177

Chang, F., Dell, G. S., and Bock, K. (2006). Becoming syntactic. *Psychol. Rev.* 113, 234–272. doi: 10.1037/0033-295X.113.2.234

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Appl. Linguist.* 27, 1–24. doi: 10.1093/applin/ami038

Fine, A. B., Jaeger, T. F., Farmer, T. A., and Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE.* 8:e77661. doi: 10.1371/journal.pone.0077661

Goldberg, A. E., Casenhiser, D. M., and Sethuraman, N. (2005). The role of prediction in construction-learning. *J. Child Lang.* 32, 407–426. doi: 10.1017/S0305000904006798

Hartsuiker, R. J., Bernolet, S., Schoonbaert, S., Speybroeck, S., and Vanderelst, D. (2008). Syntactic priming persists while the lexical boost decays: evidence from written and spoken dialogue. *J. Memory Lang.* 58, 214–238. doi: 10.1016/j.jml.2007.07.003

Kamide, Y. (2012). Learning individual talkers' structural preferences. *Cognition* 124, 66–71. doi: 10.1016/j.cognition.2012.03.001

Koechlin, E. (2016). Prefrontal executive function and adaptive behavior in complex environments. *Curr. Opin. Neurobiol.* 37, 1–6. doi: 10.1016/j.conb.2015.11.004

Konopka, A. E., and Bock, K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cogn. Psychol.* 58, 68–101. doi: 10.1016/j.cogpsych.2008.05.002

Kroczek, L. O., and Gunter, T. C. (2017). Communicative predictions can overrule linguistic priors. *Sci. Rep.* 7, 1–9. doi: 10.1038/s41598-017-17907-9

Lane, L. W., and Ferreira, V. S. (2010). Abstract syntax in sentence production: Evidence from stem-exchange errors. *J. Memory Lang.* 62, 151–165. doi: 10.1016/j.jml.2009.11.005

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychol. Rev.* 101, 676–703. doi: 10.1037/0033-295X.101.4.676

MacWhinney, B. (2013). "The logic of the unified model," in *Handbook of Second Language Acquisition*, eds S. Gass and A. Mackey (New York, NY: Routledge), 211–227.

Perek, F., and Goldberg, A. E. (2017). Linguistic generalization on the basis of function and constraints on the basis of statistical preemption. *Cognition* 168, 276–293. doi: 10.1016/j.cognition.2017.06.019

Perfors, A., Tenenbaum, J. B., and Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *J. Child Lang.* 37, 607–642. doi: 10.1017/S0305000910000012

Pickering, M. J., and Branigan, H. P. (1998). The representation of verbs: evidence from syntactic priming in language production. *J. Memory Lang.* 39, 633–651. doi: 10.1006/jmla.1998.2592

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychol. Rev.* 103, 56–115. doi: 10.1037/0033-295X.103.1.56

Romberg, A. R., and Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdiscipl. Rev. Cogn. Sci.* 1, 906–914. doi: 10.1002/wcs.78

Ryskin, R. A., Qi, Z., Duff, M. C., and Brown-Schmidt, S. (2017). Verb biases are shaped through lifelong learning. *J. Exp. Psychol. Learn. Memory Cogn.* 43, 781–794. doi: 10.1037/xlm0000341

Thothathiri, M. (2018). Statistical experience and individual cognitive differences modulate neural activity during sentence production. *Brain Lang.* 183, 47–53. doi: 10.1016/j.bandl.2018.06.005

Thothathiri, M., and Braiuca, M. C. (2020). Distributional learning in English: the effect of verb-specific biases and verb-general semantic mappings on sentence production. *J. Exp. Psychol. Learn. Memory Cogn.* 47, 113–128. doi: 10.1037/xlm0000814

Thothathiri, M., Evans, D. G., and Poudel, S. (2017). Verb bias and verb-specific competition effects on sentence production. *PLoS ONE.* 12:e0180580. doi: 10.1371/journal.pone.0180580

Thothathiri, M., and Rattinger, M. (2015). Ventral and dorsal streams for choosing word order during sentence production. *Proc. Natl. Acad. Sci. U.S.A.* 112, 15456–15461. doi: 10.1073/pnas.1514711112

Thothathiri, M., and Rattinger, M. G. (2016). Acquiring and producing sentences: whether learners use verb-specific or verb-general information depends on cue validity. *Front. Psychol.* 7:404. doi: 10.3389/fpsyg.2016.00404

Tooley, K. M., and Bock, K. (2014). On the parity of structural persistence in language production and comprehension. *Cognition* 132, 101–136. doi: 10.1016/j.cognition.2014.04.002

Trueswell, J., and Tanenhaus, M. (1994). "Toward a lexical framework of constraint-based syntactic ambiguity resolution," in *Perspectives on Sentence Processing*, eds C. J. Clifton, L. Frazier, and K. Rayner (Hillsdale, NJ: Erlbaum), 155–179.

Ueno, T., Saito, S., Saito, A., Tanida, Y., Patterson, K., and Lambon Ralph, M. A. (2014). Not lost in translation: generalization of the primary systems hypothesis to Japanese-specific language processes. *J. Cogn. Neurosci.* 26, 433–446. doi: 10.1162/jocn_a_00467

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., and MacDonald, M. C. (2009). Experience and sentence processing: statistical learning and relative clause comprehension. *Cogn. Psychol.* 58, 250–271. doi: 10.1016/j.cogpsych.2008.08.002

Wonnacott, E., Newport, E. L., and Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: distributional learning in a miniature language. *Cogn. Psychol.* 56, 165–209. doi: 10.1016/j.cogpsych.2007.04.002

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 Psychology

ORIGINAL RESEARCH
published: 14 April 2021
doi: 10.3389/fpsyg.2021.622873

Rational Adaptation in Lexical Prediction: The Influence of Prediction Strength

Tal Ness[1]* *and Aya Meltzer-Asscher*[1,2]

[1] Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel, [2] Linguistics Department, Tel Aviv University, Tel Aviv, Israel

Recent studies indicate that the processing of an unexpected word is costly when the initial, disconfirmed prediction was strong. This penalty was suggested to stem from commitment to the strongly predicted word, requiring its inhibition when disconfirmed. Additional studies show that comprehenders rationally adapt their predictions in different situations. In the current study, we hypothesized that since the disconfirmation of strong predictions incurs costs, it would also trigger adaptation mechanisms influencing the processing of subsequent (potentially) strong predictions. In two experiments (in Hebrew and English), participants made speeded congruency judgments on two-word phrases in which the first word was either highly constraining (e.g., "climate," which strongly predicts "change") or not (e.g., "vegetable," which does not have any highly probable completion). We manipulated the proportion of disconfirmed predictions in highly constraining contexts between participants. The results provide additional evidence of the costs associated with the disconfirmation of strong predictions. Moreover, they show a reduction in these costs when participants experience a high proportion of disconfirmed strong predictions throughout the experiment, indicating that participants adjust the strength of their predictions when strong prediction is discouraged. We formulate a Bayesian adaptation model whereby prediction failure cost is weighted by the participant's belief (updated on each trial) about the likelihood of encountering the expected word, and show that it accounts for the trial-by-trial data.

Edited by:
Gina Kuperberg,
Tufts University, United States

Reviewed by:
Anouschka Foltz,
University of Graz, Austria
Jean Baratgin,
Université Paris 8, France
Emily Morgan,
UC Davis, United States

****Correspondence:***
Tal Ness
talness@mail.tau.ac.il

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

Received: *29 October 2020*
Accepted: *16 March 2021*
Published: *14 April 2021*

Citation:
Ness T and Meltzer-Asscher A (2021)
Rational Adaptation in Lexical
Prediction: The Influence of
Prediction Strength.
Front. Psychol. 12:622873.
doi: 10.3389/fpsyg.2021.622873

Keywords: prediction, adaptation, language processing, bayesian adaptation, prediction error

INTRODUCTION

Despite the seemingly inexhaustible capabilities of the human brain, cognitive research has shown time and again that in some respects, our processing resources are limited. For example, although our brain can store over 10^9 bits of information over our lifetime (Von Neumann, 1958), the processing of visual objects or linguistic input is limited to no more than a few items at once (e.g., "the magical number seven" suggested by Miller, 1956, "the magic number four," Cowan, 2010; Green, 2017, or even fewer items, as suggested by McElree, 2001). It is therefore often assumed (explicitly or implicitly) that successful language processing requires efficient resource allocation.

One core aspect of language processing, which may seem somewhat contradictory to this assumption, is prediction. Over the past decades, accumulating evidence provided strong support for the idea that during language processing, we engage in actively anticipating upcoming input, rather than passively waiting for the input in order to process it as it unfolds. This anticipatory processing is evidenced in reduced processing difficulty for predictable relative to unpredictable words, manifested in reduced reading times or reaction times (RT; Ehrlich and Rayner, 1981; Schwanenflugel and Shoben, 1985; Traxler and Foss, 2000) and reduced amplitudes of the N400 event-related potentials (ERP) component (e.g., Kutas and Hillyard, 1984; DeLong et al., 2005; Wlotko and Federmeier, 2012). Notably, evidence suggests that this anticipation of upcoming input is, at least under certain circumstances, as specific as predicting the exact word that is expected to appear, including its phonological form, grammatical gender, etc. (e.g., Wicha et al., 2004; DeLong et al., 2005; van Berkum et al., 2005; Martin et al., 2013; Nieuwland et al., 2018; Nicenboim et al., 2019; Szewczyk and Wodniecka, 2020). For example, Wicha et al. (2004) examined ERPs elicited when Spanish native speakers read a determiner (el/la, *un/una*, and *las/los*), which appears prior to the noun and has to agree with the noun's grammatical gender. Their results show that in sentences that lead to a highly probable noun, determiners with a gender feature that does not match the predictable noun elicit enhanced positivity. These results indicate that the predictions generated were beyond the conceptual level, such that the specific noun was predicted, including its grammatical features.

Allocating resources to generate predictions, especially such specific predictions, intuitively seems to be a very wasteful processing strategy. We use language to communicate information, and in order for an utterance to be informative it has to be unpredictable to some extent (i.e., no new information would be gained by the listener, if they had, in advance, all the information needed in order to predict the utterance with 100% certainty prior to perceiving it). Why, then, generate predictions that will inevitably have some likelihood of being incorrect, when we can instead merely process the input as it is perceived? This question becomes even more puzzling when taking into account evidence of prediction failure costs. Predictability is often measured using the cloze task, in which participants are given the beginning of a sentence or a phrase, and are asked to provide the first completion that comes to mind. From this task, the predictability of a word is reflected in the word's cloze probability, defined as the proportion of participants who provided this word as a completion. Additionally, the constraint of a context is also calculated, defined as the cloze probability of it most common completion. It is considered to reflect the extent to which the context can lead to a strong prediction. Recent studies indicate that the processing of an unexpected (low cloze probability) word entails additional costs when it is presented in a high constraint sentence, i.e., when the initial prediction was strong. These costs are not incurred when processing a similarly unexpected word if no strong prediction was formed in the first place. This increased difficulty is mostly evidenced in the frontal post-N400 positivity (f-PNP), an ERP component that is elicited by unexpected words, only when a highly probable prediction was initially available (e.g., Federmeier et al., 2007). Since these costs are not incurred by unexpected words in low constraint contexts, they cannot be attributed to the processing of an unexpected word in and of itself. They are therefore attributed to the need to handle the incorrect prediction. This prediction failure cost was suggested to stem from a commitment made to the initial (strong) prediction, requiring its inhibition or suppression in order to integrate the actual input (e.g., Kutas, 1993; Ness and Meltzer-Asscher, 2018a,b; Kuperberg et al., 2020). We note that such inhibition can be needed at different levels of representation, namely, prediction failure costs can be incurred by a need to inhibit a low-level representation of the word in the lexicon, a higher-level representation of the sentence or the message, or both (see further discussion of this distinction in the General discussion). However, regardless of the specific nature of prediction failure costs, engagement in prediction is "wasteful" in processing resources not only due to the resources needed for the generation of predictions, but also due to the resources needed to handle the disconfirmation of strong predictions.

Several reasons have been suggested for the use of prediction as a language processing strategy despite its "wastefulness," explaining why engaging in prediction constitutes a sensible use of resources after all. For example, prediction may be helpful in reducing the ambiguity that exists in most linguistic input, either due to semantically/grammatically ambiguous utterances or due to perceptual ambiguity (e.g., arising from noisy input and production variation), by constraining the interpretation of the input to more probable meanings/representations. Additionally, prediction has been suggested to provide an effective learning mechanism based on prediction error signals. It has also been argued to enable coordinated "turn taking" during dialog (for discussion of motivations for prediction see Huettig, 2015). Prediction thus serves important functions, meaning that allocating resources for prediction is not inefficient. Notably, however, even though in general it is presumably useful to engage in prediction, the mere fact that prediction bears costs means that situations can differ in how beneficial prediction is. For example, if prediction is indeed helpful in disambiguating perceptually ambiguous input, then it may be more effective to allocate resources to generate strong predictions in a noisy environment than in a quiet one. If prediction is needed to coordinate "turn taking," we may engage more in prediction during a conversation than during passive listening (e.g., listening to a lecture or watching a movie). Moreover, regardless of the specific reason(s) that make prediction a useful processing strategy, the costs of prediction failure may outweigh the benefits derived from successful predictions, in a situation where unexpected input is often encountered. Hence, while it is reasonable to allocate resources for prediction, it is inefficient to always do so to the same extent, regardless of the situation.

Indeed, several previous studies have shown that prediction can be adapted to different situations (e.g., Neely, 1977; Schwanenflugel and Shoben, 1985; Hutchison, 2007; Lau et al., 2013; Brothers et al., 2017, 2019). Most commonly this is demonstrated as a relatedness proportion effect, i.e., the facilitation due to relatedness between a prime and a target in a prime-target lexical/semantic decision task increases when the proportion of related prime-target pairs increases (e.g., Lau et al., 2013). This indicates that when the prime and the target are often related, participants increase their reliance on semantic relatedness to the prime word in anticipating the target word (this maximization of the use of contextual information can be explained by several frameworks, in particular relevance theory, see Sperber and Wilson, 1996). Recently, this adaptation was shown to fit a Bayesian model, in which participants repeatedly update their belief about the likelihood of a related prime-target pair, and this belief is used in order to weigh the relative influence of relatedness, relative to general word frequency (Delaney-Busch et al., 2019). Namely, when the likelihood of a related prime-target pair is low, participants do not adopt a prediction strategy, and reaction times are mostly influenced by word frequency; then, as participants accumulate evidence of a high likelihood of relatedness between primes and targets, they adopt a prediction strategy that relies more on semantic relatedness, and these predictions become stronger the greater the participant's belief that related prime-target pairs are likely to appear.

Thus, prediction requires processing resources, different situations differ in how beneficial prediction is and what the optimal prediction strategy is, and evidence suggests that we have means to adapt our prediction mechanisms accordingly. This state of affairs poses two questions:

- When do comprehenders alter their prediction strategies (i.e., can other factors, besides proportion of related prime-target pairs, trigger changes to prediction strategies)? Specifically, the current study aims to test whether comprehenders alter their prediction strategies when they experience failure of strong predictions.
- How do comprehenders optimize their prediction strategies (i.e., which processes or mechanisms are susceptible to transient changes, and what are these changes)? Specifically, the current study aims to test whether comprehenders can alter their tendency to commit to strong predictions, in order to achieve an optimal balance between the benefits of successful prediction and the costs of prediction failure.

Thus, the current study focuses on the role of prediction strength and prediction failure in adaptation of prediction. As discussed above, the disconfirmation of strong predictions incurs prediction failure costs associated with a need to inhibit the falsely predicted word, due to some form of commitment made to the strong prediction. Thus, in the current study, we hypothesized that the disconfirmation of strong predictions serves as a trigger for adaptation, and that this adaptation influences subsequent predictions by decreasing the tendency to commit to strong predictions, in order to avoid prediction failure costs.

A previous study provides indication that prediction failure costs can be affected by adaptation. Schwanenflugel and Shoben (1985) have conducted a series of experiments, which showed that prediction failure costs were increased when a participant encountered a large proportion of high constraint sentences in which the most predictable word appeared, but not when they encountered a large proportion of low constraint sentences in which the most predictable word appeared. This indicates that repeated confirmation of predictions leads to increased costs when a prediction is disconfirmed. Notably, in this study, the manipulation was conducted by the addition of fillers, which were high\low constraint trials in which the most predictable word is presented (keeping constant the number of trials in which an unexpected word appeared instead of the predicted word). Namely, in this experiment, successful predictions served as the trigger for adaptation. Thus, this study indicates that prediction failure costs are influenced by adaptation, but not that prediction failure can serve as a trigger for adaptation.

Additionally, this design does not allow to isolating the contribution of prediction strength to this adaptation. Trials in which the most predictable word is presented in a high vs. low constraint inevitably differ not only in the constraint of the context, but also in the cloze probability of the presented word, since the most predictable word in low constraint contexts is not as predictable as the most predictable word in high constraint contexts. As inherent to the definition of cloze probability, a word with 80% cloze probability was provided as the first completion that came to mind by 80% of the participants in the cloze task, reflecting that it would likely be the strongest prediction for ~80% of the population or ~80% of the time for a given individual. Likewise, a word with 30% cloze probability would likely be the strongest prediction for ~30% of the population or ~30% of the time for a given individual. This means that the "most predictable word" would indeed be the participant's current prediction (in that trial) in a larger proportion of the high constraint trials compared to the low constraint ones. Thus, a participant who encounters a large proportion of trials in which the most predictable word is presented in high constraint contexts will experience confirmation of their prediction more often than a participant who encounters a large proportion of trials in which the most predictable word is presented in low constraint contexts. It is therefore not possible to determine whether adaptation was triggered by the mere repeated confirmation of a participant's prediction, or whether the strength of the confirmed prediction also played a role in the adaptation mechanism.

The current study thus aims to test whether adaptation is influenced by prediction strength. In order to do so, we focus on adaptation due to prediction failure (discouraging further prediction), rather than due to successful prediction (encouraging further prediction). This allows us to manipulate prediction strength independently of the predictability of the presented word, i.e., by presenting low cloze words in high vs. low constraint, we manipulate the strength of the initial prediction (strong or weak, respectively), while keeping the presented word equally unpredictable in both cases. In this way, we test whether adaptation is specifically triggered by unexpected words that appear in a context where an initially strong prediction could be generated (i.e., high constraint), relative to similarly unexpected words that appear in a context where no strong prediction was available (i.e., low constraint). Two experiments were conducted, in which the proportion of disconfirmed strong predictions was manipulated between participants, and we tested the influence of this proportion on prediction failure costs throughout the experiment. As stated above, our hypothesis was that disconfirmation of strong predictions serves as a trigger for adaptation, decreasing the tendency to commit to strong predictions in order to avoid prediction failure costs. If our hypothesis is correct, prediction failure costs should decrease as the experiment progresses, as the participants experience disconfirmation of strong predictions. Crucially, the greater the proportion of disconfirmed strong predictions a participant encounters, the more their prediction failure costs should be reduced, which should result in smaller prediction failure costs overall, as well as a greater rate of decrease in these costs throughout the experiment. In addition, we formulate a Bayesian adaptation model and show that it accounts for the trial-by-trial adaptation of prediction.

EXPERIMENT 1

Methods

The design and analyses for this study were pre-registered on the open science framework (OSF). The pre-registration report for Experiment 1 can be found at: https://osf.io/hwdq4/?view_only=516dcfb53b814d7483bdff03e61c271e. Data and analysis code can be found at: https://osf.io/d9s8g/?view_only=3123cc4830db42bc80ed31a5c5ed029f.

Participants

Participants were 120 Tel-Aviv University students (42 males), all native Hebrew speakers, with an average age of 24.33 (range: 18–36). Participants were given course credit or were paid 15 NIS (~4.5$) for their participation. The experiment was approved by the Ethics Committee at Tel Aviv University. Ten additional participants completed the experiment but were excluded from the analysis due to low accuracy in the task (the pre-registered exclusion criterion was below chance performance in either the congruent or the anomalous trials).

Materials

The materials were in Hebrew. They consisted of two-word phrases in which the first word was either highly constraining (i.e., had a highly probable completion) or not (i.e., did not have any highly probable completion), based on a cloze questionnaire (described below). The second word was always unexpected (i.e., a low cloze probability word), as determined by the cloze questionnaire results. This created two trial types: high constraint context – low cloze probability completion (High-Low, HL), and low constraint context – low cloze probability completion (Low-Low, LL). See **Table 1** for example materials.

Twelve critical trials from each condition were presented to all participants. Filler trials were used in order to manipulate the proportion of HL and LL trials between participants: half of the participants encountered 72 additional HL trials, and half encountered 72 additional LL trials (see **Table 2**). The trials from each type (including the fillers) were distributed throughout the experiment in a pseudo-randomized order (different for each participant). Twenty-four anomalous filler items (e.g., "socks cake") were also included, in order to enable the task (anomaly detection, see Procedure).

The LL and HL items were matched for length and frequency of the second word, overall (Length: HL mean = 4.66, LL mean = 4.89, p = 0.493, length was measured in number of letters; frequency: HL mean = 51.02, LL mean = 37.52, p = 0.519, frequency was taken from the corpus of Linzen, 2009), and for the 12 critical trials (Length: HL mean = 4, LL mean = 4.65, p = 0.191; frequency: HL mean = 30.67, LL mean = 17.83, p = 0.202). The critical trials were also matched for basic RTs for the second word, i.e., RTs in a lexical decision task for the second word in each item (without the presentation of the first word in the phrase) were similar in both conditions (HL mean = 578.84, LL mean = 579.07, p = 0.860). The basic RTs were collected from 20 participants, different from those in the main experiment.

Cloze probability questionnaires were conducted in order to assess constraint and cloze probability for each item. Participants (different from those in the main experiment) were presented with the first word of an item, and were instructed to provide the first completion that comes to mind. Each item was presented to 30–35 participants. Presentation order was randomized for each participant. High constraint items had a constraint of 65% or higher, low constraint items had constraint of 35% or lower. The average constraint was 83.03% in the high constraint items (87.03% in the 12 critical HL trials), and 24.51% in the low constraint items (19.82% in the 12 critical LL trials). HL and LL items were matched for cloze probability of the second word, with average cloze

TABLE 1 | Example materials for Experiment 1.

Trial type	First word	Second word	Second word with highest cloze probability (not presented in the experiment)
High constraint, Low cloze probability (HL)	*bu'ot* bubbles	*avir* air Cloze probability: 3.2% Translation of the phrase: "Air bubbles"	*sabon* soap Cloze probability: 93.5% Translation of the phrase: "Soap bubbles"
Low constraint, Low cloze probability (LL)	*šulxan*	*kafe* coffee Cloze probability: 3.0% Translation of the phrase: "Coffee table"	*oxel* food Cloze probability: 30.3% Translation of the phrase: "Dining table"

TABLE 2 | Trial composition in each list in Experiments 1 and 2.

Experiment 1 (Hebrew)		Experiment 2 (English)		
Low-low list	**High-low list**	**Low-low list**	**Mixed list**	**High-low list**
		15 HH trials	15 HH trials	15 HH trials
		3 Anomalies	3 Anomalies	3 Anomalies
12 HL critical trials	12 HL critical trials	12 HL critical trials	12 HL critical trials	12 HL critical trials
12 LL critical trials	12 LL critical trials	12 LL critical trials	12 LL critical trials	12 LL critical trials
72 LL filler trials	**72 HL filler trials**	12 HH critical trials	12 HH critical trials	12 HH critical trials
24 Anomalies	24 Anomalies	**60 LL filler trials**	**30 HL filler trials**	**60 HL filler trials**
		24 Anomalies	**30 LL filler trials**	24 Anomalies
			24 Anomalies	

Presentation order of the trials listed in each cell of the table was pseudo-randomized for each participant (keeping each trial type evenly distributed). The trials that differ between lists (in each experiment) are marked in bold.

probability of 4.40% in the HL trials, and 4.46% in the LL trials, overall ($p = 0.964$), and in the 12 critical trials: 1.97 and 2.06% in the HL and LL trials, respectively ($p = 0.865$).

Procedure

Stimuli were presented using the E-prime 2.0 software (Psychology Software Tools, Pittsburgh, PA). Each trial was preceded by a 200 ms fixation cross. The two-word phrases were presented word-by-word in the middle of the screen. The first word was presented for 750 ms, with a 350 ms ISI. The second word was presented until the participant made a response, or up to 4 s (i.e., if the participant did not make a response within 4 s, the trial was terminated). Participants were instructed to press a green or a red button to indicate whether or not the phrase was congruent (respectively), as quickly as possible once the second word appears. Reaction times were recorded. After each trial, a string of hash keys (####) appeared on the screen and the participants pressed a button when they were ready to start the next trial. Prior to the experiment, participants completed a practice block of six trials.

Data Analysis

Reaction times were analyzed with linear mixed-effects models. Analyses were conducted using the lmerTest package (Kuznetsova et al., 2014) in the R software environment. Only the data from the critical trials was included in the initial analysis (data from all non-anomalous trials was included in the Bayesian adaptation model, see below). Trials with errors (i.e., trials in which the participant pressed the red button, indicating that the phrase is incongruent) were excluded. Outliers were trimmed by replacing data points exceeding 2.5 SDs from each participant's mean with the value of 2.5 SDs from that participant's mean (affecting 2.9% of the data). RTs were logarithmically transformed before being entered into the model. The model included the factors List (HL list and LL list, with LL list as the reference level), Trial type (HL and LL, with LL as the reference level), and Trial number (the position of the trial throughout the experiment). The binary factors (List and Trial type) were coded for simple contrasts (one level of the factor coded as 0.5, and the other as −0.5). All models initially included random intercepts for participants and items and were fully crossed (including all factors and their interaction as random slopes for items, and Trial type, Trial number, and their interactions as random slopes for participants; List was not included as random slope for participants since each participant belongs to only one level of this factor). However, all random slopes had to be removed in order to achieve convergence (this was done by iteratively removing the random slope associated with the smallest variance, Barr et al., 2013).

Results

Accuracy

As mentioned above, the performance of all participants included in the analysis was above chance in both the congruent and the anomalous trials (separately). The average accuracy in the critical trials was 95.1% (SD = 4.30%), with high performance across conditions (LL list: LL trials – 99.2%, HL trials – 89.9%; HL list: LL trials – 98.6%, HL trials – 92.5%). Accuracy was analyzed using a logistic mixed-effects model, with the factors List (HL and LL, with LL list as the reference level) and Trial type (HL and LL, with LL as the reference level). There was an effect of Trial type such that accuracy was higher in the LL trials than in the HL trials (Estimate = −1.81, SE = 0.40, $z = -4.54$, $p < 0.001$). There was no significant effect of list (Estimate = 0.12, SE = 0.26, $z = 0.47$, $p = 0.637$), nor an interaction between Trial type and List (Estimate = 0.75, SE = 0.52, $z = 1.45$, $p = 0.146$).

Linear Regression Analysis: Pre-registered Analysis

The full results of the analyses are reported in **Table 3**. Reaction times are displayed in **Figure 1**. The results (Model 1) showed an effect of Trial type such that RTs (for the critical trials) where longer for HL trials than for LL trials ($p < 0.001$), reflecting prediction failure costs. There was also an effect of List such that RTs were shorter in the HL list relative to the LL list ($p = 0.002$). These two effects were qualified by a significant interaction between List and Trial type, such that the difference between HL and LL trials was reduced in the HL list relative to the LL list ($p = 0.048$), indicating that frequent disconfirmation of strong predictions led to reduced prediction failure costs. There was also an effect of Trial number, such that RTs decreased as the experiment progressed ($p < 0.001$). Notably, we expected a three-way interaction between Trial type, List, and Trial number, indicating that throughout the experiment, the rate at which reaction times for HL trials decreased was greater for participants in the HL list than in the LL list. However, no interaction involving Trial number reached significance (see Discussion for a possible reason). We therefore formulated an adaptation model in order to capture the trial-by-trial dynamics.

Bayesian Adaptation Model: Exploratory Analysis

In order to account for the trial-by-trial data, we formulated a Bayesian adaptation model whereby inhibition cost at each trial was modeled as μ*PE, such that:

1. μ is a point estimate for the participant's belief about the likelihood of encountering the expected word (i.e., their current estimation of predictive validity). This value is defined as the mean of a beta distribution, updated on each trial, with an initial prior of beta(1, 1). Updating occurs whenever the participant encounters an HL trial: beta(1, 1 + number of HL trials encountered). This has the effect of lowering the estimated predictive validity with more encountered instances of failed prediction.
2. PE is the prediction error, defined as the difference between the constraint of the item and the cloze probability of the second word.

The inhibition index (μ*PE), reflecting inhibition costs for a trial, therefore is large: (i) when μ is large, i.e., the participant believes they will encounter the expected word (since they have not experienced many prediction failures); and/or (ii) when PE is large – the first word is highly constraining, and the second word is highly unpredictable.

The inhibition index was calculated for each trial, experimental and filler.[1] As can be seen in **Figure 2**, the calculated inhibition index was higher for HL trials than for LL trials, since the prediction error is smaller in the LL trials. In addition, the calculated inhibition index decreases as the experiment progresses, as μ becomes smaller with the accumulation of more HL trials, and more so for the HL trials. Importantly, this decrease is greater and faster in the HL list, as in this list, which includes more HL trials, μ becomes smaller at a faster rate.

TABLE 3 | Mixed-effects regression models coefficients for Experiment 1.

	Estimate	SE	df	*t*-value	*p*-value
Model 1					
List	−0.0493	0.0160	219.6	−3.081	0.002*
Trial type	0.0440	0.0094	2,612	4.656	<0.001*
Trial number	−0.0006	0.0001	2,612	−9.069	<0.001*
List × Trial type	−0.0372	0.0189	2,611	−1.970	0.048*
List × Trial number	0.0002	0.0002	2,612	1.460	0.144
Trial type × Trial number	0.0001	0.0002	2,612	−0.980	0.327
List × Trial type × Trial number	0.0003	0.0003	2,612	1.057	0.290
Model 2					
Cloze probability	−0.0033	0.0006	166.4	−5.654	<0.001*
Inhibition index	0.0040	0.0002	10,610	18.738	<0.001*
Model 3					
List	−0.0179	0.0151	216.8	−1.190	0.235
Trial type	−0.0007	0.0107	907.1	−0.066	0.947
Trial number	−0.0004	0.00006	10,780	−5.797	<0.001*
Inhibition index	0.0027	0.0003	10,860	8.558	<0.001*
List × Trial type	0.0165	0.0147	10,880	1.110	0.267
List × Trial number	−0.00003	0.0001	10,740	−0.309	0.758
Trial type × Trial number	0.0003	0.0001	10,760	2.345	0.019*
List × Trial type × Trial number	−0.0002	0.0002	10,730	−0.989	0.323

*$p < 0.05$.

In order to test whether this inhibition index is a significant predictor of the data, we entered it into a linear mixed-effect regression. Note that the inhibition index only reflects the expected costs of prediction failure, but does not account for facilitatory effects of correct predictions. Namely, for a given HL or LL item, the majority of participants would not have predicted the low cloze word that was presented, and the costs associated with this scenario are modeled in the inhibition index. However, a portion of the participants (which correlates to the word's cloze probability) would have predicted the presented word and would have therefore experienced facilitation, which is not accounted for by the inhibition index. To account for these facilitatory effects, we included the cloze probability of the presented second word as a predictor in the model, in addition to inhibition index (Model 2, see **Table 3**). The results showed that the inhibition index was a significant predictor of reaction times ($p < 0.001$).

Furthermore, the inhibition index was entered as an additional predictor in the initial model (Model 1 above) in order to test whether it is a significant predictor of reaction times above and beyond List, Trial type, and Trial number (Model 3; **Table 3**). The inhibition index remained a significant predictor of reaction times in this model ($p < 0.001$), indicating that it explains variance in reaction times beyond the original factors. The performance of the Bayesian adaptation model (Model 2) was also compared to alternative models, which include similar (or the same) information to the information that went into the calculation of the inhibition index, but as separate factors (i.e., without the assumptions of the adaptation model): (1) a model which included PE (the difference between constraint and cloze probability), Trial number, and the interaction between these factors. (2) A model which included PE, the number of HL trials encountered, and the interaction between these factors. The Bayesian adaptation model outperformed the alternatives (Bayesian model: AIC = −16,990, BIC = −16,903, Log likelihood = 8507.1; Alt1: AIC = −16,712, BIC = −16,712, Log likelihood = 8388.6; Alt2: AIC = −16,926, BIC = −16,874, Log likelihood = 8388.6; $p < 0.001$). These results indicate that the assumptions of the Bayesian adaptation model

[1] We note that the inhibition index reflects the expected cost of inhibition, which we expect to only take place in (and be affected by) high constraint trials. However, for practical reasons, we had to decide how to handle LL trials in the analyses, which include the inhibition index. Treating the value of the inhibition index for all LL trials as missing value was not possible, since this is not a situation of "missing at random" (i.e., there would be a systematic difference between trials with a "missing" inhibition index value and trials with actual values), which would distort the regression results. For consistency, we therefore chose to have a uniform formula for the calculation of all trials, with the assumption that the inhibition index for LL trials would not contribute much to the explanatory power of the model in any case, as it is low and relatively invariable (due to small prediction error). The alternative would be to set the value of the inhibition index to zero in all LL trials, representing the lack of prediction failure and no inhibition. In order to ensure that our results and conclusions do not hinge on the decision to compute an inhibition index for LL trials rather than set it to zero, we ran the analyses for both experiments again, but with the inhibition index set to zero in all LL trials. This modification had very little effect on the results. Crucially, none of the significant results in the original analyses became non-significant or vice versa.

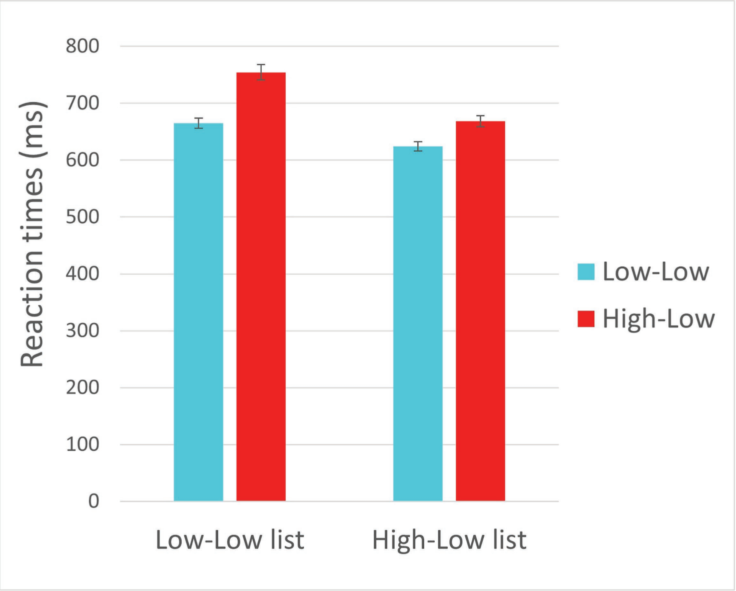


FIGURE 1 | Reaction times in the critical trials in Experiment 1.

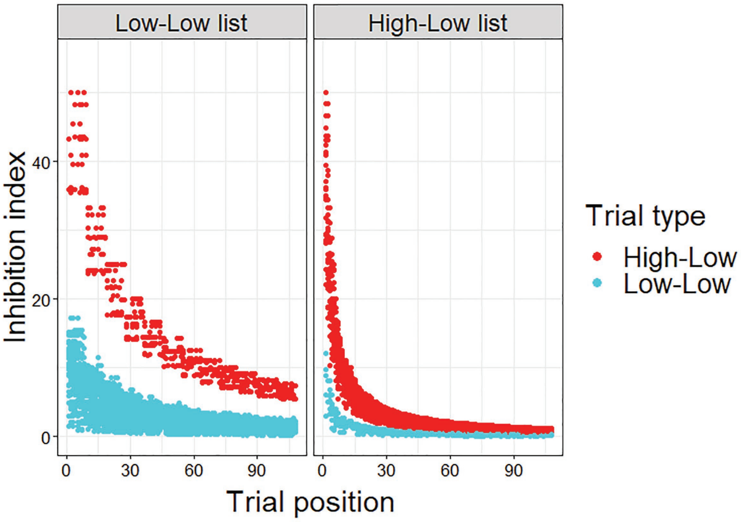


FIGURE 2 | Calculated inhibition index (μ*PE) in Experiment 1.

indeed increase its explanatory power, relative to models including the basic information entered into its calculations, but without its further assumptions. Namely, the calculation of the inhibition index increases the variance explained by the model, relative to models that include the same data but without this calculation.

Discussion

The current experiment manipulated the proportion of disconfirmed strong predictions (HL trials) throughout the experiment, and tested the influence of this proportion on prediction failure costs. First, the results showed increased reaction times in the HL trials relative to LL trials. Since these conditions did not differ in the predictability of the second word in the phrase (i.e., cloze probability did not differ between these conditions), this result provides additional evidence for the incurrence of prediction failure costs (see General Discussion). Moreover, the results showed that this increase in reaction times in the HL relative to LL trials was smaller in the HL list than in the LL list, indicating that participants who experienced disconfirmation of strong predictions more often adapted to the experimental context by reducing their engagement in strong prediction. Since the filler items that differed between lists did not contrast in how predictable the presented words were (i.e., cloze probability), but only in the strength of the initially available prediction (i.e., constraint), this result supports our main hypothesis that the disconfirmation of strong predictions, rather than simply the occurrence of unpredictable words, triggers adaptation.

We additionally expected a three-way interaction between Trial type, List, and Trial number, reflecting that throughout the experiment the rate at which reaction times for HL trials decreased was greater for participants in the HL list than in the LL list. However, we did not find this interaction. We believe, based on examination of the data that adaptation in the HL list occurred too quickly to be detectable in our experiment. The proportion of HL trials in the HL list was very high – seven HL trials for every LL and anomaly trial. In addition, the experiment did not include high constraint trials in which the predicted word appeared. Given this, adaptation, namely learning that strong predictions are extremely likely to be disconfirmed in the experiment, may have taken place prior to any critical trials, or after very few of them.

In the absence of the predicted three-way interaction, in order to better account for the trial-by-trial dynamics, we formulated a Bayesian adaptation model. We showed that this model, which takes into consideration the ongoing updating of the participant's belief about the likelihood of encountering a predictable word (i.e., their estimate of predictive validity), can capture the data.

EXPERIMENT 2

In Experiment 1, the Bayesian model and the related analyses were conceived after data collection, and were thus exploratory. Since the addition of unplanned analyses greatly increases the likelihood of false positives, we then followed up with a replication experiment (Experiment 2), for which the Bayesian model and related analyses were pre-registered. In this experiment, we also included high constraint – high cloze probability (HH) trials, in an attempt to slow down adaptation. The Bayesian adaptation model was therefore extended to include such trials (see below). Additionally, in this experiment we included three lists (instead of two), in order to manipulate the proportion of HL trials more gradually.

In addition, while Experiment 1 was a lab-based experiment with Hebrew speakers, Experiment 2 was in English, conducted online with native English speakers. This was done due to considerations of participant recruitment, and was not predicted to affect the results. However, the use of new materials in a different language, and a different participant population, does contribute to the generalizability of our findings.

Methods

The design and analyses for this study were pre-registered on the OSF. The pre-registration report for Experiment 2 can be found at: https://osf.io/3k6am/?view_only=2bd9dc5c43c2459385bead7cf03978f6. Data and analysis code can be found at: https://osf.io/5h9tv/?view_only=c2f47d6d3adf405297b1c863b88b3818.

Participants

Participants were 150 (69 males) native English speakers, born and living in the United States, with an average age of 31.11 (range: 20–45). The participants were recruited *via* Prolific and were paid 1.5 GBP (~2$) for their participation. The experiment was approved by the Ethics Committee in Tel Aviv University. Fourteen additional participants completed the experiment but were excluded from the analysis: 12 due to low accuracy in the task, and two due to mean RTs that exceeded 2.5 SD from the group's mean RT (based on the pre-registered exclusion criteria).

Materials

As in Experiment 1, the materials included 12 HL and 12 LL critical trials that were presented to all participants. Additionally, 12 high constraint, high cloze probability (HH) critical trials were included. Constraint and cloze probability were determined based on a cloze questionnaire, as described below. See **Table 4** for example materials. The HH items were introduced in the current experiment in order to slow down adaptation, by indicating to the participant that predictions can be confirmed in the experimental context. Filler trials were manipulated between participants, such that one third of the participants encountered 60 additional HL trials, one third encountered 60 additional LL trials, and one third encountered 30 additional HL trials and 30 additional LL trials. The different trial types were distributed throughout the experiment in a pseudorandomized order. However, 15 additional HH trials were presented to all participants at the beginning of the experiment, in order to make sure all participants could initially assume that forming predictions is beneficial in the experimental context. Twenty-four anomalous filler items (e.g., "socks cake") were also included, in order to enable the task (anomaly detection, see Procedure). The trial composition in each list is summarized in **Table 2**.

TABLE 4 | Example materials for Experiment 2.

Trial type	First word	Second word	Second word with highest cloze probability (not presented in the experiment in HL and LL trials)
High constraint, Low cloze probability (HL)	Rearview	camera Cloze probability: 6.7%	mirror Cloze probability: 93%
Low constraint, Low cloze probability (LL)	Desert	storm Cloze probability: 6.8%	island Cloze probability: 14%
High constraint, High cloze probability (HH)	Peanut	butter Cloze probability: 83%	

The LL and HL items were matched for length and frequency of the second word, overall (Length: HL mean = 6.05, LL mean = 6.23, p = 0.591, length was measured in number of letters; frequency: HL mean = 78.03, LL mean = 92.70, p = 0.470, frequency was taken from the Corpus of Contemporary American English, COCA, Davies, 2009). The LL, HL, and HH items were matched for length and frequency of the 12 critical trials (Length: HL mean = 5.59, LL mean = 6.58, HH mean = 5.75, HL vs. LL: p = 0.312, HH vs. LL: p = 0.791, HH vs. HL: p = 0.842; frequency: HL mean = 100.76, LL mean = 86.80, HH mean = 113.7, HL vs. LL: p = 0.450, HH vs. LL: p = 0.780, HH vs. HL: p = 0.789).

Cloze probability questionnaires were conducted in order to assess constraint and cloze probability of each item. Each item was presented to 30 participants (different from those in the main experiment). Presentation order was randomized for each participant. High constraint items had a constraint of 50% or higher and low constraint items had a constraint of 25% or lower. The average constraint was 73.13% in the high constraint items (76.94% in the 12 critical HL trials, 72.48% in the 12 critical HH trials), and 14.64% in the low constraint items (14.44% in the 12 critical LL trials). HH and HL items were matched for constraint (p = 0.321). HL and LL items were matched for cloze probability (p = 0.450 overall, p = 0.316 for the critical items), with average cloze probability of 3.28% in the HL trials, and 2.73% in the LL trials (in the 12 critical trials: 6.94 and 4.72% in the HL and LL trials, respectively).

Procedure and Data Analysis

The procedure was as detailed for Experiment 1, except that the experiment was built in PsychoPy 2 (Peirce et al., 2019) and was run online on the Pavlovia platform.[2] Data analysis was identical to Experiment 1, except that the factor Trial type included HH trials (i.e., HH, HL, and LL, coded for simple contrasts, with LL as the baseline level), and the factor List included three levels rather than two (this factor was treated as ordinal/continuous, since the three levels of this factor are ordered on a scale of the proportion of HL trials; thus, the three levels were included as one numerical variable: LL list = 1, mixed list = 2, and HL list = 3).

Results

Accuracy

As mentioned above, the performance of all participants included in the analysis was above chance in both the congruent and the anomalous trials (separately). The average accuracy in the critical trials was 96.7% (SD = 2.72%), with performance high across conditions (LL list: HH trials – 99.7%, LL trials – 98.2%, HL trials – 90.0%; Mixed list: HH trials – 99.3%, LL trials – 99.2%, HL trials – 93.2%; HL list: HH trials – 99.2%, LL trials – 98.8%, HL trials – 93.0%). Accuracy was analyzed using a logistic mixed-effects model, with the factor Trial type (HH, LL and HL, with LL as the reference level) and List (HL, Mixed, LL, as an ordinal variable). There were effects of Trial type such that accuracy was higher in the HH trials than in the LL trials (Estimate = 2.87, SE = 1.08, z = 2.66, p = 0.008), and higher in the LL trials than in the HL trials (Estimate = −1.16, SE = 0.44, z = −2.63, p = 0.009). Additionally, there was an interaction between List and Trial type at the levels of HH vs. LL, such that the difference in accuracy between the HH and LL trials was smaller the higher the proportion of HL trials was (Estimate = −0.97, SE = 0.43, z = 2.27, p = 0.023). There was no significant effect of List (Estimate = 0.15, SE = 0.15, z = 0.97, p = 0.332), and the difference in accuracy between HL and LL trials did not differ significantly between lists, (Estimate = −0.24, SE = 0.21, z = −1.14, p = 0.255).

Linear Regression Analysis (Pre-registered)

The full results of the analyses are reported in **Table 5**. Reaction times are displayed in **Figure 3**. The results (Model 1) showed effects of Trial type such that RTs (for the critical trials) were shorter for HH trials than for LL trials (p < 0.001), reflecting facilitation due to higher predictability in the HH trials; and longer for HL trials than for the LL trials (p < 0.001), and reflecting prediction failure costs. Additionally, there was a significant interaction between List and Trial type at the levels of HL vs. LL, such that the difference between HL and LL trials decreased the more HL trials the list included (p = 0.012). There was also an effect of Trial number, such that RTs decreased as the experiment progressed (p < 0.001), as well as an interaction between Trial number and Trial type at the levels of HL vs. LL such that the decrease in RTs as the experiment progressed was greater for HL trials than for LL trials (p = 0.011). Again, the three-way interaction between Trial type (HL vs. LL), List and Trial number did not reach significance.

Bayesian Adaptation Model (Pre-registered)

The Bayesian adaptation model was similar to that of Experiment 1, modified for the inclusion of HH trials. Thus, in the current model, updating of the participant's belief about predictive validity occurred whenever the participant encountered a high constraint trial, such that a HL trial lowered the estimated predictive validity (as in Experiment 1), and a HH trail raised the estimated predictive validity: beta(1 + number of HH trials encountered, 1 + number of HL trials encountered). The inhibition index (μ*PE) was calculated for each trial (**Figure 4**), and entered into a linear mixed-effect regression with cloze probability as an additional predictor (Model 2). The results showed that the inhibition index was a significant predictor of reaction times (p < 0.001).

The inhibition index was then entered as an additional predictor in the initial model (Model 1 above) in order to test whether it is a significant predictor of reaction times above and beyond List, Trial type, and Trial number (Model 3). The inhibition index remains a significant predictor of reaction times in this model (p < 0.001), indicating that it explains variance in reaction times beyond the original factors. Again, the performance of the Bayesian adaptation model (Model 2) was also compared to alternative models, which include similar (or the same) information to the information that went into the calculation of the inhibition index, but as separate factors

[2] pavlovia.org

TABLE 5 | Mixed-effects regression models coefficients for Experiment 2.

	Estimate	SE	df	*t*-value	*p*-value
Model 1					
List	0.0017	0.0074	204.1	0.230	0.818
Trial type (HH vs. LL)	−0.1228	0.0195	1,075	−6.308	<0.001*
Trial type (HL vs. LL)	0.0635	0.0158	4,193	4.035	<0.001*
Trial number	−0.0005	0.0001	13,520	−4.928	<0.001*
List × Trial type (HH vs. LL)	−0.0004	0.0079	13,590	−0.054	0.957
List × Trial type (HL vs. LL)	−0.0169	0.0067	13,740	−2.516	0.012*
List × Trial number	0.0001	0.00004	13,520	1.568	0.117
Trial type (HH vs. LL) × Trial number	0.00003	0.0002	13,520	0.112	0.911
Trial type (HL vs. LL) × Trial number	0.0005	0.0002	13,520	−2.538	0.011
List × Trial type (HH vs. LL) × Trial number	0.0001	0.0001	13,520	1.073	0.283
List × Trial type (HL vs. LL) × Trial number	0.0002	0.0001	13,520	1.740	0.082
Model 2					
Cloze probability	−0.1349	0.0149	137.9	−9.034	<0.001*
Inhibition index	0.1303	0.0105	1,471	12.446	<0.001*
Model 3					
List	0.0061	0.0075	208.5	0.824	0.411
Trial type (HH vs. LL)	−0. 1,058	0.0183	873.8	−5.796	<0.001*
Trial type (HL vs. LL)	−0.0286	0.0218	1,564	−1.307	0.191
Trial number	−0.0004	0.0001	12,780	−4.134	<0.001*
Inhibition index	0.1573	0.0265	255.2	5.931	<0.001*
List × Trial type (HH vs. LL)	−0.0014	0.0079	13,610	−0.177	0.859
List × Trial type (HL vs. LL)	−0.0058	0.0071	2,578	−0.827	0.408
List × Trial number	0.0001	0.00004	12,930	2.342	0.019*
Trial type (HH vs. LL) × Trial number	0.00001	0.0002	13,490	0.039	0.969
Trial type (HL vs. LL) × Trial number	−0.0003	0.0002	10,330	−1.643	0.101
List × Trial type (HH vs. LL) × Trial number	0.0001	0.0001	13,520	0.883	0.377
List × Trial type (HL vs. LL) × Trial number	0.0002	0.0001	11,150	2.444	0.015*

*$p < 0.05$.

(i.e., without the assumptions of the adaptation model): (1) A model which included PE (the difference between constraint and cloze probability), Trial number, and the interaction between these factors. (2) A model which included PE, the number of HL trials encountered, the number of HH trials encountered, and the interaction between these factors. The Bayesian adaptation model outperformed the alternatives (Bayesian model: AIC = −20,685, BIC = −20,549, Log likelihood = 10,360; Alt1: AIC = −20,623, BIC = −20,540, Log likelihood = 10,323; Alt2: AIC = −20,589, BIC = −20,537, Log likelihood = 10,302; $p < 0.001$), indicating that the assumptions of the Bayesian adaptation model indeed increase its explanatory power, relative to other models including the basic information entered into its calculations, but without its additional assumptions.

Discussion

Experiment 2 replicated and extended the results of Experiment 1. First, the results showed increased reaction times in the HL trials relative to LL trials, providing additional evidence for the incurrence of prediction failure costs. In addition, the results showed that this increase in reaction times in the HL relative to LL trials was smaller the more HL trials the participant encountered, providing additional evidence that participants who encounter the disconfirmation of strong predictions more often adapt by reducing their engagement in strong prediction. This result thus provides additional support for our main hypothesis that the disconfirmation of strong predictions, rather than simply the occurrence of unpredictable words, triggers adaptation.

The Bayesian adaptation model was again shown to capture the trial-by-trial data, corroborating the results of the exploratory analysis in Experiment 1. Importantly, in Experiment 2 this model and the related analyses were pre-registered, alleviating the increased risk of false positives in an exploratory analysis.

GENERAL DISCUSSION

In the current study, we hypothesized that the disconfirmation of strong predictions serves as a trigger for adaptation, influencing subsequent processing by decreasing the participant's tendency to commit to strong predictions, in order to avoid prediction failure costs. This hypothesis was tested in two experiments by manipulating the proportion of disconfirmed strong predictions encountered during the experiment and measuring the influence of this proportion on prediction failure costs.

First, the results of both experiments showed increased reaction times in trials consisting of a highly constraining word followed by an unpredictable word (HL trials), relative to trials where an unpredictable word appeared after a word which was not constraining (LL trials). Since these conditions did not differ in the predictability of the second word in the phrase (i.e., cloze probability did not differ between these

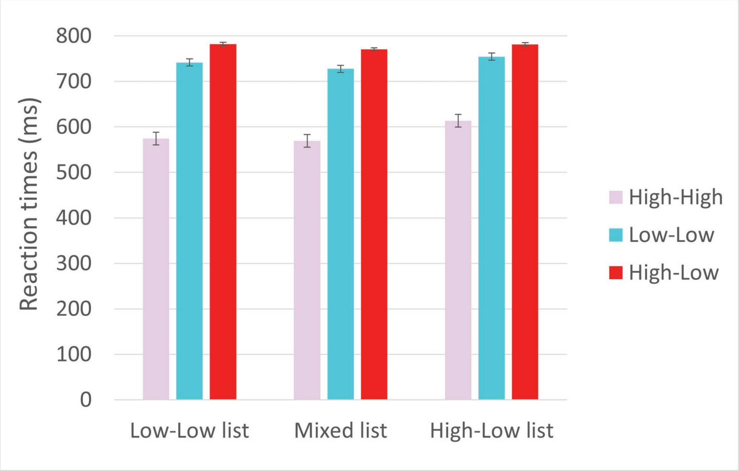


FIGURE 3 | Reaction times in the critical trials in Experiment 2.

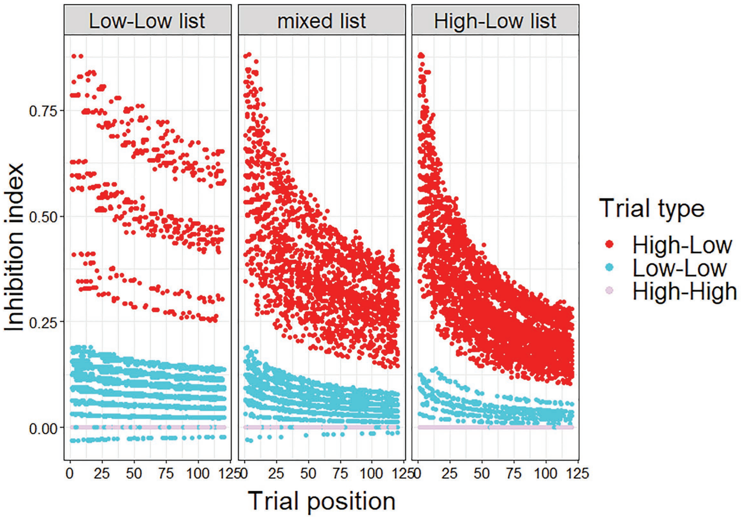


FIGURE 4 | Calculated inhibition index (μ*PE) in Experiment 2.

conditions), this result provides evidence for prediction failure costs, i.e., costs that are incurred due to the initially formed prediction rather than due to the processing of an unpredictable word, in and of itself. This result is particularly interesting in light of recent evidence regarding the f-PNP ERP component. As discussed in the Introduction, prediction failure costs were often demonstrated in ERP studies showing a late frontal positivity (f-PNP) elicited by unexpected words only in high constraint contexts (e.g., Federmeier et al., 2007). However, in a recent study, Brothers et al. (2020) have tested the effect of context length on the f-PNP component. Their results showed a significant f-PNP effect elicited by unpredictable words in high constraint contexts, only when the context was rich and globally constraining, but not when the strong lexical prediction could only be generated based on a single word immediately preceding the target word. For example, the f-PNP was not observed in a sentence such as "(…) James unlocked the… door/laptop," when constraint was purely reliant on a single word ("unlocked"). Similarly, a f-PNP was not observed by Lau et al. (2016), with materials consisting of a one-word context (a prenominal adjective). These results may thus suggest that impoverished contexts do not give rise to prediction failure costs, which is seemingly inconsistent with our current results, demonstrating prediction failure costs in two-word phrases (i.e., single word contexts). Crucially, however, there are several factors in the current materials and design, which may reconcile the current results with the results of Brothers et al. (2020). First, in the current study, we used a relatively slow presentation rate (the SOA was 1,000 ms, while the SOA in the experiments of Brothers et al., 2020, was 550 ms). The long SOA provided participants with more time to form strong and specific predictions (see e.g., Ito et al., 2016), which may have contributed to the incurrence of prediction failure costs. Additionally, the task in the current study was a speeded anomaly judgment task, while participants in the Brothers et al. (2020) study discussed above read for comprehension and then gave a non-speeded judgment, and participants in the Lau et al. (2016) study were not required to provide any response during the trials (a memory recognition test was administered after each block). Thus, the task in the current study may have provided further encouragement to generate predictions, in order to respond as quickly as possible once the second word appeared. Indeed, the f-PNP component was shown to be greater when prediction is encouraged by task demands (Brothers et al., 2017). Moreover, the average constraint in the current study was relatively high (87% in Exp. 1 and 77% in Exp.2, compared to 63% in the minimal context materials of Brothers et al., 2020), which could have significant influence on prediction failure costs, considering that the f-PNP component is only elicited in high constraint contexts. Thus, while the use of two-word phrases in the current study perhaps had some diminishing influence on prediction failure costs, the other factors discussed above may have outweighed this influence, allowing the manifestation of prediction failure costs nonetheless.

Importantly, the results also showed that this increase in reaction times in the HL relative to LL trials was smaller the higher the proportion of HL trials was in the experiment, indicating that participants who experienced disconfirmation of strong predictions more often adapted by reducing their engagement in strong prediction. Since the lists did not differ in how predictable the presented words were (i.e., cloze probability), but only in the strength of the initially available predictions (i.e., constraint), this result supports our main hypothesis that the disconfirmation of strong predictions, rather than simply the occurrence of unpredictable words, triggers adaptation.

We formulated a Bayesian adaptation model in order to account for the trial-by-trial adaptation dynamics. In this model, the comprehender iteratively updates their belief about predictive validity in the current situation. The comprehender's estimate of predictive validity decreases when an unexpected word appears in a high constraint context (i.e., a HL trial), and increases when the predictable word appears in a high constraint context (i.e., a HH trial). This estimate of predictive validity is then used to weigh the strength of the subsequent prediction, thus alleviating prediction failure costs when the comprehender believes predictive validity is low and it is not beneficial to engage in strong prediction. This model was shown to be a significant predictor of reaction times in both experiments, first in an exploratory analysis in Experiment 1, and then in a pre-registered analysis in Experiment 2.

As discussed in the Introduction, processing resources are known to be limited and prediction can be considered a "wasteful" processing strategy, requiring the generation of predictions and the handling of disconfirmed predictions. The current study provides support for the notion that processing resources are nonetheless allocated efficiently, in that prediction is not always employed to the same extent. Instead, when situations differ in how beneficial prediction is, comprehenders rationally adapt their processing strategies, to increase or decrease the reliance on strong predictions.

Prior Beliefs About Predictive Validity

In the current study, the main aim of the Bayesian model was to account for adaptation by modeling the change in participants' beliefs about predictive validity throughout the experiment, and its influence on processing prediction failure. Although our focus was on changes in the estimated predictive validity, the model had to include an initial prior, representing the participant's expected predictive validity when they arrive at the experiment, prior to any trials. The prior that we chose, beta(1,1), implies that the participant begins the experiment with a belief that the predictive validity is 50%, i.e., when encountering a predictive first word (a high constraint item) there is a 50% chance that the predicted word will be presented. This is not necessarily an accurate assumption. However, we chose to use this standard prior since determining a more accurate prior requires non-trivial decisions on parameters that we cannot assess. Essentially, the participants' estimate of predictive validity at the beginning of the experiment should reflect the predictive validity in their accumulated linguistic experience, i.e., the likelihood of encountering the predicted word following a high constraint context. Namely, the prior should match the mean constraint of "high constraint" contexts in the language. However, we do not know the distribution of constraint in the language. Moreover, we do not know what constitutes a "high constraint" context. That is, while, we do believe that there is a qualitative difference in the processing of high and low constraint contexts (see section "The role of prediction failure in adaptation" below), we do not know where the threshold between the two lies. Thus, we cannot achieve a better estimation for the participants' belief about predictive validity at the beginning of the experiment.

Additionally, we chose a weak prior (reflected in the sum of the two parameters to the beta distribution), since we assume that when participants approach an experimental task, they are relatively "prone to adaptation." When engaging in conversation in everyday life it is reasonable for a comprehender to be relatively confident that they can rely on their previous experience, and they are therefore likely to give more weight to previous experience and need more evidence in order to adapt. In contrast, an experimental setting is either a new situation for the participant (for inexperienced participants) or a situation which the participant knows varies significantly between occurrences (i.e., upcoming input in a new experiment is not expected to resemble previous, unrelated, experiments that the participant may have participated in). Therefore, participants are likely not to put a lot of weight on their prior belief (i.e., have a weak initial prior).

It may be interesting to consider the influence that alternative priors would have on the output of the model. A prior which represents a higher initial estimate of predictive validity would result in a greater decrease in the estimated predictive validity with every HL trial encountered early in the experiment, leading to even faster adaptation than the current model predicts. Of course, lower initial estimates of predictive validity would have the opposite effect (i.e., slower adaptation). Additionally, the higher the weight of the initial prior, the slower the adaptation would be, since more evidence would be needed in order to outweigh previous experience. Although it is possible to try and determine the initial prior that would provide the best fit for the current data, we did not explore this issue further in this study, as this prior would mostly indicate how participants approach the experimental situation, and is not necessarily generalizable to real-life situations. Importantly, these considerations about the initial prior are orthogonal to our main aim and conclusions in the current paper, since we manipulated the proportion of disconfirmed strong predictions between lists, and participants were randomly assigned a list, i.e., there is no ground to assume a systematic difference between lists in the initial prior participants arrive with.

The Role of Prediction Failure in Adaptation

The current results provide evidence for the importance of prediction failure as a trigger for adaptation of prediction. Namely, the manipulation in the current study was achieved by presenting either HL fillers, or LL fillers (or both), which differ in constraint but not in cloze probability. Thus, the adaptation, we observed is driven by prediction failure, i.e., by the disconfirmation of highly probable predictions. This conclusion accords with the prevalent notion that prediction errors are crucial for implicit learning, as they signal the need to update future predictions (e.g., Shanks, 1995; Schultz et al., 1997; Schultz and Dickinson, 2000). A basic principle in numerous learning/adaptation models, inherent to prominent frameworks such as reinforcement learning and Bayesian adaptation, is that the extent of learning/adaptation exerted by a given input depends on the prediction error experienced. For example, Jaeger and Snider (2013) have shown that syntactic alignment increases as a function of the prediction error experienced, while processing the prime structure, i.e., the same syntactic structure can exert stronger or weaker syntactic priming depending on how surprising it was when it appeared as a prime. Notably, their results show that the extent of adaptation depends on both prior and recent experience. Specifically, they show that syntactic alignment is stronger when the prime's structure is unexpected given the verb's bias (i.e., when prediction error is large based on prior experience), but also when the prime's structure was infrequent in previous trials in the experiment (i.e., when prediction error is large based on recent experience). The influence of both prior and recent experience on the extent of adaptation is also evidenced in the current study, and implemented in our adaptation model. First, HL trials, in which the participant can experience a significant prediction error, induce adaptation, while LL trials do not. This is an influence of prior experience, i.e., a low cloze word in a high constraint context incurs larger prediction error than in a low constraint context, based on the participant's accumulated knowledge regarding the cloze probability distribution (or some representation of it). Additionally, in a Bayesian adaptation model, the more improbable an input is given the prior, the greater the update it causes. This is implemented in the calculation of the participant's estimated predictive validity (μ) in our model: a HL trial encountered early in the experiment, when the estimated predictive validity is higher, induces a greater change to the participant's belief about predictive validity (and thus a greater change to the behavior in subsequent trials) than a HL trial encountered later in the experiment, when the estimated predictive validity is lower (and vice versa for a HH trial). This is an influence of recent experience, i.e., despite the participant's prior knowledge regarding the cloze probability distributions, the prediction error experienced when a low cloze word appears in a high constraint context has less of an effect as the participant learns not to expect the high cloze word.

We note that although in the current study, we take the approach of formulating a Bayesian adaptation model, and the results show that this model accounts for reaction times in our experiments, the same data can potentially be compatible with models based on other frameworks (e.g., reinforcement learning). However, the choice to model Bayesian adaptation is motivated by the vast literature employing such models to account for a myriad of phenomena in different domains, such as formal semantics (e.g., Lassiter and Goodman, 2015), reasoning (e.g., Heit, 1998), Bayesian pragmatics (e.g., Werning et al., 2019), and, most relevantly, priming effects in language processing (e.g., Myslín and Levy, 2016; Delaney-Busch et al., 2019). Importantly, the performance of the Bayesian adaptation model in the current study indicates that any model that would account for the data should implement the basic notions that adaptation is initiated by the incompatibility of the input with the participant's predictions (i.e., prediction error) and that the extent of adaptation at each trial is dependent on how incompatible the trial is with the predictions generated, which leads to the non-linear adaptation throughout the experiment (i.e., greater adaptation in earlier trials).

Pre-updating, Commitment, and Inhibition

The current results provide additional evidence indicating that prediction failure costs can be influenced by adaptation (as also demonstrated by Schwanenflugel and Shoben, 1985, see Introduction). This raises the question of how prediction failure costs are reduced, i.e., which process (or processes) is made easier, or is even eliminated, when adaptation occurs.

As discussed in the Introduction, prediction failure costs were suggested to stem from a need to inhibit the falsely predicted word due to commitment made to the strong prediction (e.g., Ness and Meltzer-Asscher, 2018a,b). This commitment was recently suggested to be the result of a prediction mechanism termed "pre-updating" (Lau et al., 2013; Kuperberg and Jaeger, 2016; Ness and Meltzer-Asscher, 2018b), which involves not only the activation of the predicted content, but its actual integration into the sentence's representation being built in working memory. Since a pre-updated prediction is integrated into the sentence representation, if it is then disconfirmed, inhibition is required in order to "override" the integrated representation and allow integration of the actual input instead. Interestingly, overriding an integrated representation may require inhibition or suppression at different levels of representation (Kuperberg et al., 2020). Ultimately, the high-level representation of the sentence or the event being conveyed by the sentence (and preceding context) needs to be corrected to no longer include the wrong prediction. This correction of the high-level representation entails suppression of the incorrectly predicted event, and may or may not require inhibition of the lower-level representation of the predicted word or its semantic features. Indeed, recent experiments employing the cross-modal lexical priming (CMLP) paradigm provided indication that inhibition of the wrongly predicted word can be observed when a (congruent) unexpected word is presented in a highly constraining sentence, and that this inhibition may be correlated with the f-PNP component (Ness and Meltzer-Asscher, 2018a). Thus, prediction failure costs (and the f-PNP component) may encompass processes at multiple levels of representation.

Due to these costly processes that are needed when a pre-updated prediction is disconfirmed, pre-updating constitutes a strong form of prediction, which can occur only when a highly probable (highly pre-activated) prediction is available. Pre-updating was recently suggested to be initiated by an activation threshold, i.e., when the activation level of a predicted word passes a threshold, this word will be pre-updated (Ness and Meltzer-Asscher, 2018b, 2021). Thus, we propose that the underlying mechanism by which prediction failure costs are modulated is the adjustment of the threshold for pre-updating. When the estimated predictive validity is decreased, the threshold for pre-updating is raised, leading to a lower tendency to pre-update. In such a situation, when pre-updating is avoided, the disconfirmation of a high cloze prediction would not require inhibition, alleviating prediction failure costs. In the opposite situation, when the estimated predictive validity is increased, the threshold is lowered, leading to a higher tendency to pre-update. In such a situation, if a strong prediction is then disconfirmed, prediction failure costs will be increased, since the disconfirmed prediction is more likely to have been pre-updated, requiring inhibition when revealed not to be correct.

CONCLUSION

As discussed in the introduction, the current study aimed at addressing two questions regarding the adaptation of prediction. First, what triggers it; and second, which aspects of prediction are adaptable. The current study addressed these questions with regard to prediction failure, providing evidence that prediction failure can serve as a trigger for adaptation, and that prediction failure costs are adaptable (i.e., can be influenced by adaptation). We show that a Bayesian adaptation model can account for the trial-by-trial dynamics, and propose that the adaptation of prediction failure costs is achieved *via* a thresholding mechanism adjusting the tendency to commit to strong predictions.

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: OSF: https://osf.io/d9s8g/?view_only=3123cc4830db42bc80ed31a5c5ed029f and https://osf.io/5h9tv/?view_only=c2f47d6d3adf405297b1c863b88b3818.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee at Tel Aviv University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TN and AM-A contributed to the conceptualization and design of the study, and the writing of the manuscript. TN conducted the experiments and the analyses. AM-A supervised and provided funding and resources. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the Israel Science Foundation grant #216/18 (P.I.: AM-A) and by the German-Israeli Foundation grant I-2496-104.4/2018 (P.I.: AM-A).

ACKNOWLEDGMENTS

We thank the reviewers and audience at the conference on Rational Approaches in Language Science (RAILS 2019) for their helpful questions and comments. We also thank Dr. Tamara Swaab for valuable input and discussions throughout the development of this work.

REFERENCES

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, 255–278. doi: 10.1016/j.jml.2012.11.001

Brothers, T., Dave, S., Hoversten, L. J., Traxler, M. J., and Swaab, T. Y. (2019). Flexible predictions during listening comprehension: speaker reliability affects anticipatory processes. *Neuropsychologia* 135:107225. doi: 10.1016/j.neuropsychologia.2019.107225

Brothers, T., Swaab, T. Y., and Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *J. Mem. Lang.* 93, 203–216. doi: 10.1016/j.jml.2016.10.002

Brothers, T., Wlotko, E. W., Warnke, L., and Kuperberg, G. R. (2020). Going the extra mile: effects of discourse context on two late positivities during language comprehension. *Neurobiol. Lang.* 1, 135–160. doi: 10.1162/nol_a_00006

Cowan, N. (2010). The magical mystery four: how is working memory capacity limited, and why? *Curr. Dir. Psychol. Sci.* 19, 51–57. doi: 10.1177/0963721409359277

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): design, architecture, and linguistic insights. *Int. J. Corpus Linguist.* 14, 159–190. doi: 10.1075/ijcl.14.2.02dav

Delaney-Busch, N., Morgan, E., Lau, E., and Kuperberg, G. R. (2019). Neural evidence for bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition* 187, 10–20. doi: 10.1016/j.cognition.2019.01.001

DeLong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* 8, 1117–1121. doi: 10.1038/nn1504

Ehrlich, S. F., and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *J. Verbal Learn. Verbal Behav.* 20, 641–655. doi: 10.1016/S0022-5371(81)90220-6

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., and Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Res.* 1146, 75–84. doi: 10.1016/j.brainres.2006.06.101

Green, C. (2017). Usage-based linguistics and the magic number four. *Cogn. Linguis.* 28, 209–237. doi: 10.1515/cog-2015-0112

Heit, E. (1998). "A bayesian analysis of some forms of inductive reasoning" in *Rational models of cognition*. eds. M. Oaksford and N. Chater (Oxford, United Kingdom: Oxford University Press), 248–274.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Res.* 1626, 118–135. doi: 10.1016/j.brainres.2015.02.014

Hutchison, K. A. (2007). Attentional control and the relatedness proportion effect in semantic priming. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 645–652. doi: 10.1037/0278-7393.33.4.645

Ito, A., Corley, M., Pickering, M. J., Martin, A. E., and Nieuwland, M. S. (2016). Predicting form and meaning: evidence from brain potentials. *J. Mem. Lang.* 86, 157–171. doi: 10.1016/j.jml.2015.10.007

Jaeger, T. F., and Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition* 127, 57–83. doi: 10.1016/j.cognition.2012.10.013

Kuperberg, G. R., Brothers, T., and Wlotko, E. W. (2020). A tale of two Positivities and the N400: distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *J. Cogn. Neurosci.* 32, 12–35. doi: 10.1162/jocn_a_01465

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

Kutas, M. (1993). In the company of other words: electrophysiological evidence for single-word and sentence context effects. *Lang. Cogn. Process.* 8, 533–572. doi: 10.1080/01690969308407587

Kutas, M., and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307, 161–163. doi: 10.1038/307161a0

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2014). lmerTest: tests for random and fixed effects for linear mixed effect models. R package version 2.0–11. Available at: https://cran.r-project.org/package=lmerTest

Lassiter, D., and Goodman, N. D. (2015). How many kinds of reasoning? Inference, probability, and natural language semantics. *Cognition* 136, 123–134. doi: 10.1016/j.cognition.2014.10.016

Lau, E. F., Holcomb, P. J., and Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *J. Cogn. Neurosci.* 25, 484–502. doi: 10.1162/jocn_a_00328

Lau, E., Namyst, A., Fogel, A., and Delgado, T. (2016). A direct comparison of N400 effects of predictability and incongruity in adjective-noun combination. *Collabra* 2, 1–19. doi: 10.1525/collabra.40

Linzen, T. (2009). *Corpus of blog postings collected from the Israblog website*. Tel Aviv: Tel Aviv University.

Martin, C. D., Thierry, G., Kuipers, J. R., Boutonnet, B., Foucart, A., and Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *J. Mem. Lang.* 69, 574–588. doi: 10.1016/j.jml.2013.08.001

McElree, B. (2001). Working memory and focal attention. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 817–835. doi: 10.1037/0278-7393.27.3.817

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63, 81–97. doi: 10.1037/h0043158

Myslín, M., and Levy, R. (2016). Comprehension priming as rational expectation for repetition: evidence from syntactic processing. *Cognition* 147, 29–56. doi: 10.1016/j.cognition.2015.10.021

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: roles of inhibitionless spreading activation and limited-capacity attention. *J. Exp. Psychol. Gen.* 106, 226–254. doi: 10.1037/0096-3445.106.3.226

Ness, T., and Meltzer-Asscher, A. (2018a). Lexical inhibition due to failed prediction: behavioral evidence and ERP correlates. *J. Exp. Psychol. Learn. Mem. Cogn.* 44, 1269–1285. doi: 10.1037/xlm0000525

Ness, T., and Meltzer-Asscher, A. (2018b). Predictive pre-updating and working memory capacity: evidence from event-related potentials. *J. Cogn. Neurosci.* 30, 1916–1938. doi: 10.1162/jocn_a_01322

Ness, T., and Meltzer-Asscher, A. (2021). From pre-activation to pre-updating: a threshold mechanism for commitment to strong predictions. *Psychophysiology* 8:e13797. doi: 10.1111/psyp.13797

Nicenboim, B., Vasishth, S., and Rösler, F. (2019). Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia* [Preprint] doi: 10.31234/osf.io/2atrh

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., et al. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *Elife* 7:e33468. doi: 10.7554/eLife.33468

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., et al. (2019). PsychoPy2: experiments in behavior made easy. *Behav. Res. Methods* 51, 195–203. doi: 10.3758/s13428-018-01193-y

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599. doi: 10.1126/science.275.5306.1593

Schultz, W., and Dickinson, A. (2000). Neuronal coding of prediction errors. *Annu. Rev. Neurosci.* 23, 473–500. doi: 10.1146/annurev.neuro.23.1.473

Schwanenflugel, P. J., and Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. *J. Mem. Lang.* 24, 232–252. doi: 10.1016/0749-596X(85)90026-9

Shanks, D. R. (1995). *The psychology of associative learning. Vol. 13*. Cambridge, UK: Cambridge University Press.

Sperber, D., and Wilson, K. (1996). Fodor's frame problem and relevance theory-response. *Behav. Brain Sci.* 19, 530–532. doi: 10.1017/S0140525X00082030

Szewczyk, J. M., and Wodniecka, Z. (2020). The mechanisms of prediction updating that impact the processing of upcoming word: an event-related potential study on sentence comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* 46, 1714–1734. doi: 10.1037/xlm0000835

Traxler, M. J., and Foss, D. J. (2000). Effects of sentence constraint on priming in natural language comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 1266–1282. doi: 10.1037//0278-7393.26.5.1266

van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 443–467. doi: 10.1037/0278-7393.31.3.443

Von Neumann, J. (1958). *The computer and the brain*. New Haven, CT: Yale University Press.

Werning, M., Unterhuber, M., and Wiedemann, G. (2019). "Bayesian pragmatics provides the best quantitative model of context effects on word meaning in EEG and cloze data" in *Proceedings of the 41st annual conference of the cognitive science society*. eds. A. Goel, C. Seifert and C. Freska. July 24–27, 2019 (Austin, TX: Cognitive Science Society).

Wicha, N. Y., Moreno, E. M., and Kutas, M. (2004). Anticipating words and their gender: an event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *J. Cogn. Neurosci.* 16, 1272–1288. doi: 10.1162/0898929041920487

Wlotko, E. W., and Federmeier, K. D. (2012). So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *Neuroimage* 62, 356–366. doi: 10.1016/j.neuroimage.2012.04.054

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 Psychology

ORIGINAL RESEARCH
published: 30 April 2021
doi: 10.3389/fpsyg.2021.626023

Default Inheritance in Modified Statements: Bias or Inference?

*Corina Strößner**

Department of Philosophy II, Ruhr-University of Bochum, Bochum, Germany

It is a fact that human subjects rate sentences about typical properties such as "Ravens are black" as very likely to be true. In comparison, modified sentences such as "Feathered ravens are black" receive lower ratings, especially if the modifier is atypical for the noun, as in "Jungle ravens are black". This is called the *modifier effect*. However, the likelihood of the unmodified statement influences the perceived likelihood of the modified statement: the higher the rated likelihood of the unmodified sentence, the higher the rated likelihood of the modified one. That means the modifier effect does not fully block *default inheritance* of typical properties from nouns to modified nouns. This paper discusses this inheritance effect. In particular, I ask whether it is the direct result of composing concepts from nouns, that is, a bias toward "black" when processing "raven". I report a series of experiments in which I find no evidence for a direct inheritance from composition. This supports the view that default inheritance is rather an inference than a bias.

Keywords: modifier effect, default inheritance, prototype theory, compositionality, rational reasoning

OPEN ACCESS

Edited by:
Matthew W. Crocker,
Saarland University, Germany

Reviewed by:
Lena Nadarevic,
Universität Mannheim, Germany
Thomas Spalding,
University of Alberta, Canada

****Correspondence:***
Corina Strößner
corina.stroessner@rub.de

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

Received: *04 November 2020*
Accepted: *10 March 2021*
Published: *30 April 2021*

Citation:
Strößner C (2021) Default Inheritance in Modified Statements: Bias or Inference? Front. Psychol. 12:626023.
doi: 10.3389/fpsyg.2021.626023

1. INTRODUCTION

One of the central questions of cognitive science concerns the status of prototypes. During the twentieth century, it became clear that most of our concepts are not definable in terms of necessary and jointly sufficient features. The late Wittgenstein's discussion of "game" is a well-known example (c.f. Wittgenstein, 1953).[1] Prototype theory (c.f. Rosch and Mervis, 1975; Rosch, 1978) offered the alternative idea that concepts are determined by prototypes. These are highly central exemplars or summary representations of typical properties associated with the concept. While prototype theory has been a very successful research paradigm within psychology, there remain doubts whether concepts should really be understood in terms of prototypes. One prominent voice against the prototype view was Jerry Fodor. In several philosophical works (e.g., Fodor and Lepore, 1996; Fodor, 1998), he argued that concepts should not be identified with prototypes. He accepts that concepts are *associated to* prototypes but denies that they are part of what the concept essentially is. His main critique is that prototypes lack compositionality. The meaning of composed concepts such as "pet fish" is not a straightforward composition of "pet" and "fish". Different versions of the compositionality criterium have been developed that are more compatible with prototype concepts (Hampton, 1987; Smith et al., 1988; Hampton and Jönsson, 2012; Strößner, 2020). While these versions depart from a very strict reading of compositionality, they still hold that the typical features of a concept such as "raven" influence complex concepts such as "jungle raven" or "feathered raven". *Prima facie* the typical properties of the concept (e.g., blackness) are inherited by the complex concepts, unless the modifier speaks against inheritance of the typical property (e.g., "albino raven").

[1] Wittgenstein's original argumentation concerned the German "Spiel", which is a broader term and arguably even harder to define.

Connolly et al. (2007) deny such inheritance and thus further expand criticism against the apparent lack of compositionality of prototypes. They investigated generic sentences that ascribe typical but non-analytic properties, for example, "Ravens are black" or "Rings are expensive".[2] Their subjects rated such sentences with unmodified and modified nouns. Connolly et al. (2007) discovered that humans tend to judge modified statements (e.g., "Feathered ravens are black") as less likely to be true than unmodified ones, especially if the modifier is atypical (as in "Jungle ravens are black"). This has been called the *modifier effect*. Apparently people do not "default to the stereotype", as Connolly et al. (2007, p. 5) call it. The work of Connolly et al. (2007) has inspired further experimental research. The upshot of the empirical work is that the modifier effect is extremely robust (Jönsson and Hampton, 2006, 2012; Gagné and Spalding, 2011, 2014; Hampton et al., 2011; Spalding and Gagné, 2015; Gagné et al., 2017; Spalding et al., 2019; Strößner and Schurz, 2020; Strößner et al., 2020). However, it has also been demonstrated that, even though modified statements are perceived as less plausible, the rated likelihood of an unmodified statement correlates with the rated likelihood of the modified one. This indicates that judgments about the modified concept are not independent of the original unmodified concept.[3]

Another debate revolves around the extent to which default inheritance is rationally expected. Connolly et al. (2007) deny that an inference from "Ravens are black" to "Jungle ravens are black" would be rationally justified. Indeed, the inference lacks logical certainty, unless it really means that *all* ravens are black. Statements that ascribe merely typical properties, however, allow for exceptions and the modified noun might refer to an exceptional subcategory. For example, birds can fly but Antarctic birds cannot fly. Nevertheless, the reasoning from categories to subcategories is often intuitively plausible. As a result, the formalization of such inferences gave rise to a whole branch of research on defeasible reasoning. Reiter (1980) started to develop formal logics of default-based logic for artificial intelligence and many other researchers from different disciplines followed (Pearl, 1988, 1990; Kraus et al., 1990; Gabbay et al., 1995; Veltman, 1996; Schurz, 2005). Though the rational justification of default inheritance is still researched, there is a consensus that at least typical subcategories should inherit typical properties. The corresponding inference scheme is called *cautious monotony* and allows inferring "S are typically P" from "C are typically P" and "C are typically S". Another famous inference pattern is *rational monotony*. It permits to reason from "C are typically P" and "C are not typically non-S" to "S are typically P". This rule corresponds to default inheritance to subcategories that are not atypical, for example, from a raven to a female raven. Strößner and Schurz (2020) argue that at least the inference to typical subcategories, that is, cautious monotony, is very reliable and entails almost no risk of deriving a false conclusion. Rational monotony is more risky but often still acceptable. Moreover, even the inference to exceptional categories can be quite reliable if the category members have a large overall similarity to each other (Thorn and Schurz, 2018; Strößner, 2020). For example, "Blind ravens are black" might be acceptable, even though blind ravens are atypical, because the blindness is unrelated to color. Very clearly, however, specific background knowledge should always dominate the judgment: No one should accept "Albino ravens are black". To sum up, default inheritance is often rationally justified. It is a useful reasoning pattern that allows to draw defeasible conclusions about properties of which one has no specific information.

While some forms of default inheritance should and actually do influence our understanding of modified nouns (i.e., subcategories), the details of this process are unclear. Is default inheritance really an inference in human cognition or rather the result of a prototype-induced bias? Is it a by-product of conceptual composition, that is, the result of forming the concept "jungle raven" from "raven"? Or is it detached from composition and only occurring after the meaning of the modifier noun compound has been processed? I present experimental evidence that shows that default inheritance is easily blocked by knowledge effects. This supports the view that default inheritance does not occur as a result of forming complex concepts and that it is rather an inference than a bias.

The paper proceeds with a presentation of empirical findings, starting with a re-analysis of the data from Connolly et al. (2007). I discuss several effects that were discovered in empirical research and how they can be interpreted on the theoretical level. The following part presents a series of experiments that test whether traces of default inheritance are still found when background knowledge intervenes. The largely negative results suggest that the inheritance is not a direct by-product of composition.

2. TYPICALITY IN UNKNOWN SUBCATEGORIES

2.1. Connolly et al.—A New Analysis

As noted above, research on the modifier effect goes back to an experimental study by Connolly et al. (2007). Their material consisted of 40 items, each with four sentences: an unmodified statement such as "Ravens are black" (Condition A), one with a typical modifier such as "Feathered ravens are black" (Condition B), one with an atypical modifier such as "Jungle ravens are black" (Condition C), and finally a double-modified statement such as "Young jungle ravens are black" (Condition D). Their 40 participants rated one version of each item on a scale from 1 (*very unlikely*) to 10 (*very likely*). **Figure 1** provides an overview of the mean ratings of the 40 items in the different conditions.

Connolly et al. (2007) reported the obvious decrease in rated likelihood from condition to condition. They establish its significance by analysis of variance (ANOVA) and pairwise comparisons with t-tests. I re-analyzed their data within a mixed-effect model approach, which became the standard method in psycholinguistic research during the last decade because

[2] Expressing typicality is one of the central functions of generics sentences (Krifka et al., 1995).

[3] Note that all experimental studies on this issue were undertaken with English or German material. The extent to which these findings are replicable for speakers of other languages, especially outside the Indo-European family, is not researched. Care should thus be taken when considering the results as representative for humans or languages in general.

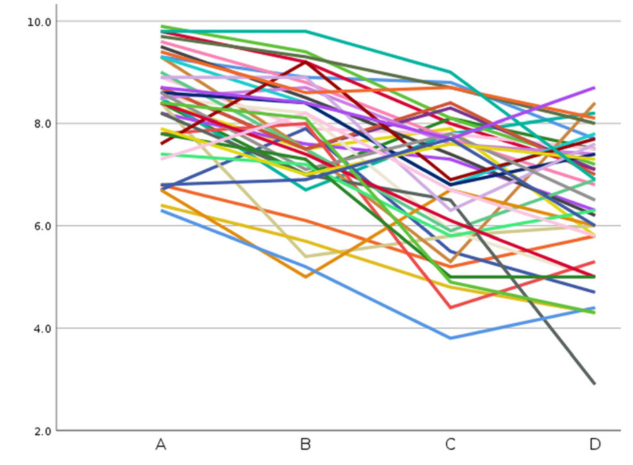


FIGURE 1 | Mean rating of 40 items in Connolly et al. (2007). Each graph corresponds to an item and displays its mean likelihood in the four Conditions A (unmodified), B (typical modifier), C (atypical modifier), and D (additional atypical modifier).

it accounts for the fact that subjects as well as the chosen material are random samples.[4] This model estimates the mean rating of unmodified conditions as 8.38 ($SE = 0.2$), of typical modifications as 7.72 ($SE = 0.2$), atypical ones as 6.88 ($SE = 0.2$), and of double-modified statements as 6.49 ($SE = 0.2$). All pairwise comparisons are significant (all $p < 0.001$, except atypical and double modification with $p = 0.003$). This conforms with the results reported by Connolly et al. (2007). Moreover, the calculation of model fit indicated a reasonably good model fit (conditional $R^2 = 0.332$) and a notable but not high effect size of the modifier (marginal $R^2 = 0.101$).

The decrease effect is quite obvious. However, this does not mean that no inheritance exists. In order to test for the influence of the unmodified statement, I further calculated correlations between the mean rating of the sentences in these different conditions.[5] Pearson's correlation test revealed that the rating of the unmodified statements was highly correlated with the rating of the typically modified sentences ($r = 0.71$, $p < 0.001$) as well as the atypically modified sentence ($r = 0.60$, $p < 0.001$). The same applies with regard to atypical and double-modified sentences ($r = 0.62$, $p < 0.001$). This speaks against the thesis that the rated likelihood of the unmodified sentences has no influence on the rating of the modified ones and makes it clear that there is not only decrease but also an inheritance effect.

When looking at the individual items, it becomes apparent that the general trend of gradually decreasing probability from A via B to C and D is violated severely by some items (see also **Figure 1**). A closer view shows that several items might have been affected by common knowledge of the participants. For example, against the general trend, the typically modified statements "Pet hamsters live in cages" and "Jazz Saxophones are made of brass" were judged as more likely (+1.2 and +1.6) compared to their unmodified counterparts. An obvious explanation is that most subjects know that pet rodents are held in cages and that they are acquainted with Jazz saxophones. An example of negative relevance is "flying" in "Flying yellow roosters live on farms" (−3.6). Flying is hardly compatible with being kept on a farm. Another item with potential knowledge effects is "Limousines are long". The atypical modifier "inexpensive" induced a more drastic loss in rated likelihood than the other items (−4), which points to subject's understanding that smaller cars are less expensive. The further modifier "old" led to an increase in the mean rating (+3.1) indicating that "old" moderates this relation. This search for knowledge effects may seem somewhat speculative, but the crucial point is that it is reasonable to assume that background knowledge influenced the ratings, although Connolly et al. (2007) tried to avoid this in the selection of the material. A thorough analysis of knowledge effects in Strößner et al. (2020) showed that items with potential knowledge effects had significantly greater deviations in the modified conditions.

2.2. Aspects of the Modifier Effect

In their discussion, Connolly et al. (2007) primarily focused on the decrease effect: For a concept C, prototypical property T and the modifier M, "MC are T" is usually rated as less likely than "C are T", especially if MC is an atypical subcategory. However, their data indicate three aspects:

- Decrease effect: The rated likelihood of "MC are T" is lower than for "C are T".
- Inheritance effect: The rated likelihood of "MC are T" depends on how likely "C are T" is.
- Knowledge effect: The rated likelihood of "MC are T" is strongly influenced by knowledge about M or MC.

Usually the term "modifier effect" is used to refer to the decrease effect. However, all three effects robustly influence the understanding of modified typicality statements. For example, Jönsson and Hampton (2012) repeated the experiment and reproduced these effects. The modifiers, especially atypical ones, lead to a reduction of the mean rated likelihood (A: 8.31, B: 7.51, C: 6.59, and D: 6.27), but there were also correlations between the judged likelihood of the unmodified and modified statements, which indicates inheritance. Potential influences from knowledge effects were indicated in self-reports by subjects. For example, "Edible catfish have whiskers" was rejected because the whiskers will be removed before eating the fish (c.f. Jönsson and Hampton, 2012, p. 103).

While knowledge may influence the rating of modified nouns, it needs to be stressed that neither the decrease effect nor the inheritance effect is explained by (factual) background knowledge. Gagné and Spalding (2011) replicated the modifier effect for artificial adjectives, that is, pronounceable but meaningless words. This design excludes factual knowledge.

[4] The reanalysis used R and the packages lme4, lmertest, and performance (Bates et al., 2015; Kuznetsova et al., 2017; R Core Team, 2017; Lüdecke et al., 2020). Subject and items were entered as factors with random intercepts. Modification condition was treated as fixed effect. I thank Andrew Connolly for providing the data from the original study.

[5] Such a test was not part of the analysis in Connolly et al. (2007) but has been carried out in later studies, for example by Jönsson and Hampton (2012) and Strößner and Schurz (2020).

In a study by Strößner and Schurz (2020), decrease effects appeared even when subjects mostly denied that the modifier was relevant. However, *if* background knowledge is available, it leads to very strong effects and *tends* to dominate the judgment.

Research has not only established that the modifier effect, especially the decrease effect, is very robust but also that it is more general than initially found by Connolly et al. (2007). It does not only occur for generic statements but also for universal statements such as "All (handmade) sofas have backrests", even if the universal quantification is emphasized as in "All (handmade) sofas always have a backrest", "Every single (handmade) sofa has a backrest", and "100% of (handmade) sofas have a backrest", as shown by Jönsson and Hampton (2006). Subjects often accept the unmodified universal statements but reject the modified statements, even though the latter are a *logical* consequence of the former. Notably, the effect was weaker in within subjects designs, that is, if the same subjects rated modified and unmodified statements. The effect was further moderated if the sentences were placed beneath each other. Moreover, Hampton et al. (2011) found that the modifier effect is not limited to merely typical properties but equally occurs for analytical properties, for example, in "(Jungle) ravens are birds". The statement "Ravens are birds" is rated as extremely likely, but the adding of a modifier "jungle" leads to the same amount of decrease as it does in a more contingent statement such as "(Jungle) ravens are black", where the property is less central (i.e., it is easy to imagine non-black ravens).

As mentioned above, Gagné and Spalding (2011) observed modifier effects even for meaningless words as modifiers. Besides a decrease in rated likelihood, they also noted a longer reaction time (1,406 ms compared to 1,172 ms). Moreover, Gagné and Spalding (2014) replicated these findings for relational sentences instead of modifiers (e.g., "kites that are made of silk" instead of "silk kites") and even for artificial nouns like "brinn", when subjects were told that "brinn" refers to a kind of bottle. In Gagné et al. (2017), the hedging words "normal" and "typical" produced a modifier effect. Subjects were told to assume that a generic is true (e.g., "Bottles are cold in annealing ovens"). They were then either asked how many bottles or how many normal bottles or how many typical bottles are cooled in annealing ovens. The mean judgment for the bare noun was 96%, while it was significantly lower for "normal bottles"/"typical bottles" (88%). Spalding and Gagné (2015) also showed that the modifier effect has a reverse sibling. Statements that attribute very *unlikely* properties (e.g., "Whales are small") are judged as less plausible than their modified counterparts (e.g., "Plary whales are small"). The modifier thus *increases* the judged likelihood of very atypical properties (see also Spalding et al., 2019).

2.3. The Role of (Rational) Reasoning

Christina Gagné and Thomas Spalding interpret their findings as evidence against the view that typical properties are directly inherited by subcategories. They deny to view concepts as "containers of properties" such that a modified noun automatically includes the properties as well. According to them, the inheritance is the result of a reasoning process: Participants reason by the meta-knowledge that a subcategory should be somewhat similar and somewhat different. This thesis has the advantage that it explains the inheritance (similarity) as well as the decrease (dissimilarity) as effects of a process that is more or less rationally justified.

However, the *decrease* effect occurs against rational intuitions. For example, rejecting "All handmade sofas have a backrest" but accepting "All sofas have a backrest" as done by subjects in Jönsson and Hampton (2006) is clearly fallacious. Also, it is not clear why central and even categorical properties like "is a bird" are subject to the same amount of decrease. One would expect that people more readily infer categorical properties (like being a bird) than accidental ones (being black).

Much of the apparently irrational effects have been attributed to the particular pragmatic aspects of the task. While logical factors (universal quantifier, essential properties) have little influence on the modification effect, the presentation of the material influences the extent of the decrease effect considerably. For example, placing statements beneath each other leads to a lower decrease effect (Jönsson and Hampton, 2006, 2012). Recently, Strößner and Schurz (2020) showed that the decrease effect was much smaller in a comparative task, where modified and unmodified statements were presented together, as well as in a story-based rating, in which single category members and modifying information were embedded in a story (e.g., about a girl who owns a lamb *Lamby*, a Norwegian lamb *Norwy*, and so on).[6] In some of their items, the modifier was relevant. Knowledge of positive relevance (e.g., in "Golden rings are expensive") had a strong effect in the story-based and comparative rating, but not in the normal likelihood rating. The authors conclude that there is still a decrease effect in the background: "In the normal likelihood rating, where not only sentences are evaluated separately, the negative pragmatic effect of the modifier and the positive effect of background knowledge cancel each other out" (Strößner and Schurz, 2020, p. 15). Positive relevance does not prevent a decrease effect but only superposes.

As explanation of the pragmatic effect, Strößner and Schurz (2020) name Gricean implicatures (Grice, 1989). Because people assume that a cooperative speech is as informative and relevant as necessary, the addition of the modifier is automatically perceived as potentially relevant. However, other pragmatic theories such as the relevance theory by Sperber and Wilson (1986) and the more recently developed Rational Speech Act theory (Goodman and Frank, 2016) support a similar prediction that additional information (e.g., a modifier) indicates a meaningful difference. Note that the modified statement is not only longer but takes additional effort in processing: it has a lower fluency. Reber and Unkelbach (2010, p. 568) note a relation between fluency and the relevance theory of Sperber and Wilson (1986), because a lack of fluency also might indicate relevance. A cooperative speaker should make her statements as simple to process as possible.[7]

[6] Note however that the ratings were generally low for the story-based task.

[7] Generally, the modifier effect seems to be related to fluency effects. However, this issue is under-researched since most studies are focused on what the modifier effect says about the (prototype) theory of concepts.

What appears to be a fallacy in reasoning might thus just be a side effect of otherwise useful cognitive mechanisms.

The pragmatic solution is not totally different from the reasoning approach by Gagné and Spalding. It is even similar to what Gagné and Spalding (2011, p. 189) call the meta-knowledge "that the purpose of using a combined concept is often to refer to a subcategory that is in some way distinct from other members of the head category". However, Strößner and Schurz (2020) emphasize the *unconscious* nature of the pragmatic component, stating that the *decrease* is not a result of reasoning but of a general relevance bias, which is evolutionarily adaptive but not rationally reflected and of which subjects are not even aware, while Gagné and Spalding leave the status of relevance assumptions open. Their central claim concerns the mechanism behind default inheritance. They criticize a container view of concepts according to which default inheritance is more or less an automatism of conceptual combination (c.f. Gagné et al., 2017, p. 225). Rather, they view inheritance as a result of reasoning.

In what follows, the paper addresses whether inheritance effects should be understood as a result of rational considerations or whether humans are biased toward inheritance just as they are biased toward relevance. To answer the question, I present two experiments that investigate inheritance in the presence of strong knowledge effects and for privative modifier noun combinations (e.g., "stone apple").

3. EXPERIMENTS

The following experiments aim to address default inheritance in a different way than studies on modification usually do. Most experiments avoid background knowledge. The following experiments do the reverse. I aim to look for inheritance effects when they are not rationally expected. I do this by introducing modifiers with strong negative knowledge effects that should prevent default inheritance. An example is the statement "Dirty pans are used for frying", where the modifier should prevent inheritance effects.

The experimental idea partly resembles earlier research by Springer and Murphy (1992). They compared modified sentences where a sentence's truth was either determined by the noun alone or was dependent on the modifier. For example, "Peeled apples are sweet" is generally true, while "Peeled apples are white" is true because of the relevant modifier "peeled". Analogously, the falsity of "Peeled apples are squared" has nothing to do with "peeled", while "Peeled apples are red" is false because of the modifier "peeled". It was found that true modified statements are easier and faster to verify if the modifier is relevant, as in "Peeled apples are white" (see also Gagné and Murphy, 1996). Regarding the false sentences, there were no significant differences between generally false statements and those with relevant modifications. The latter finding was cited by Connolly et al. (2007) as evidence against default inheritance. If typicality was inherited, they claim, then sentences such as "Peeled apples are red" should be more difficult to process because "red" would have to be inherited from "apple" and afterwards actively suppressed. However, the experimental design in Springer and Murphy (1992) did not intend to test default inheritance or the modifier effect, which had not been discovered at that time.

As argued above, multiple experiments have shown that the likelihood of "*C* are *T*" has a profound influence on "*MC* are *T*" in the absence of more specific knowledge about *MC*. The aim of the present experiment is to directly assess whether the influence of "*C* are *T*" on the acceptance of "*MC* are *T*" persists if *M* provides strong evidence *against T*. If default inheritance is the result of meta-knowledge or an inference pattern, its influence should be easily blocked if the modifier is sufficiently relevant. In this case, the more specific knowledge should determine the judgment. Thus, it would not be necessary to cognitively rely on usually uncertain default reasoning. If inheritance effects, however, come from a typicality bias or are a mere by-product of composition, their influence should persist.

In order to find these traces of irrational default inheritance, I investigate modified typicality statements with strongly relevant modifiers. However, instead of comparing them to unmodified statements, I compare them to statements with the same modifier but a noun for which no typicality association exists. For example, are there differences between the statement "Peeled apples are red" and "Peeled pears are red" that can be traced back to the fact that "Apples are red" is much more acceptable than "Pears are red"?

The following experimental study starts with a test of unmodified statements with and without typical properties. This is done in the preparatory experiment. An example is the pair of statements "Pans are used for frying" and "Pots are used for frying". The following two experiments use modifiers with negative knowledge constraints (e.g., "Dirty ____ are used for frying") and measure how the phrases are evaluated depending on whether the noun is prototypically associated (e.g., "pans") or unrelated (e.g., "pots"). Measured variables are acceptance (yes/no), reaction time, and a separate plausibility rating. Depending on how deeply people are entrenched to typicality inheritance, the modified sentence "Dirty pans are used for frying" should be still more acceptable than "Dirty pots are used for frying". An inference-based explanation of modification, on the other hand, predicts that there is no such influence of typicality and that people only rely on the prototype if more specific information is lacking. The effect I am thus mainly investigating is not the decrease effect but the persistence of inheritance effects even if they are not rationally expected.

3.1. Preparatory Experiment

My experiment required a set of adequate sentence pairs, consisting of a generic statement that expressed a typical property and a sentence which ascribed the same property to a noun concept for which it is not typical but possible. Apart from the different association to the property, the two nouns should be as similar as possible. I thus constructed 50 sentence pairs (in German) according to the following criteria:[8]

[8] The experiment was carried out with German native speakers.

TABLE 1 | Least square means of the preparatory experiment.

	Typical Est (SE) [0.95 CI]	Non-typical Est (SE) [0.95 CI]	R^2 Conditional / Marginal
Reaction time	1546 (100) [1344, 1746]	1992 (100) [1791, 2193]	0.45/0.06
Acceptance rate	0.96 (0.02) [0.92, 1.01]	0.25 (0.02) [0.21, 0.30]	0.58 / 0.53
Plausibility	87.5 (1.8) [83.9, 91.0]	31.4 (1.8) [27.8, 34.9]	0.64 / 0.58

Estimated means with standard error in round brackets, 0.95 confidence interval in square brackets and conditional as well as marginal R^2 in the last column.

- The noun concepts come from the same superordinate category and have a similar length.
- The typically true statement ascribes a property from the list of associated features by Cree and McRae (2003).
- The other statement ascribes the same property to a noun to which it is usually not associated but still possible.

An example of such a pair is "Rats carry diseases"/"Hamsters carry diseases" (original: "Ratten übertragen Krankheiten"/"Hamster übertragen Krankheiten"). The main purpose of the preparatory experiment was to choose appropriate sentences from the material. Forty subjects were recruited and received payment via the panel Prolific (app.prolific.co). The experiment was programmed and carried out on SoSciSurvey (www.soscisurvey.de).

The material was distributed over two surveys, each with 25 typical and 25 atypical generic statements. Typicality was a between subjects factor. Every participant saw either the true typicality statement or its counterpart. In the first part of the experiment, subjects were presented with the statements and had to decide whether they agree or disagree with the statement as fast and accurately as possible. Reaction time (including reading time) was recorded. In the second part, subjects were allowed to give a more fine-grained judgment on the plausibility of the same statements using a slider (0–100 scale) without any time pressure.

Among the 50 items, I selected 32 pairs that satisfied the following criteria:

- high acceptance of the typical statement, meaning at least 80% of subjects rated "I agree",
- a considerable difference of acceptability in the atypical and typical statements: acceptance rate of the atypical statement at least 30 points below the rate for the typical statement (e.g., at most 50% if the typical condition received 80% acceptance),
- contingency of the atypical statement, indicated by a plausibility with a mean of at least 10 and a median of at least 5 (on a scale from 0 to 100).

Table 1 displays the least mean squares of the experimental data for the 32 selected items estimated on the basis of a mixed-effect model.[9] As stipulated, acceptance and plausibility was high for typical generic statements and rather low but not extremely low for atypical generic statements. Moreover, reaction time was longer for the atypical generic statements. The fact that the reaction time of the true generic statements is faster is not unexpected. People are probably highly acquainted with generic statements like "Banana is yellow" and less exposed to statements like "Strawberries are yellow" and this might make them easier to verify and faster to process.[10]

3.2. Experiment 1

3.2.1. Methods

Material: The material consisted of the 32 sentence pairs from the preparatory experiment with an added modifier that conflicted the ascribed target property. An example is the sentence pair "Heated cellars are cold" and "Heated kitchens are cold" or the aforementioned "Dirty pans are used for frying" and "Dirty pots are used for frying". The full material is displayed in the **Appendix**. Additionally, I used 32 true modified sentences. About a half of them were true because of the modifier and the others were true independently of the modifier. Six further fillers were used as warm-up for the reaction time measurement.

Design: The 32 sentences with typical noun–property pairs were equally distributed over two questionnaires. Their non-typical counterparts appeared on the other questionnaire, respectively. Moreover, the 38 fillers were added. The experiment consisted of two major parts: a decision task in which participants had to decide as fast and accurately as possible whether they agree or disagree with the presented statements, and a plausibility rating of the same sentences.

Procedure: Eighty-two participants were recruited via Prolific and directed to SoSciSurvey, where they were randomly assigned to one of the two questionnaires. In the introductory texts, participants were told that the experiment tests the plausibility of generic sentences without explicitly referring to the notion of typicality. The structure of the experimental procedure was disclosed in the welcome text. That means, subjects were aware that they had to evaluate the same sentences during a decision and a rating task. They were explicitly told that some sentences concern objects of which they have no knowledge and that they should decide intuitively without much thought or research.

[9] Models were again calculated in R (R Core Team, 2017) with the packages lme4, lmerTest, and performance (Bates et al., 2015; Kuznetsova et al., 2017; Lüdecke et al., 2020). Subjects and items were entered with random intercepts. Estimation of degrees of freedom used Satterthwaite's method.

[10] It is a well-established fact that repetition tends to decrease processing time and increases perceived likelihood (c.f. Hasher et al., 1977; Dechêne et al., 2010; Unkelbach and Rom, 2017).

TABLE 2 | Least square means of Experiment 1.

	Typical Est (SE) [0.95 CI]	Non-typical Est (SE) [0.95 CI]	R^2 Conditional/Marginal
Reaction time	2576 (107) [2364, 2789]	2742 (107) [2530, 2955]	0.504/0.005
Acceptance rate	0.19 (0.02) [0.15, 0.24]	0.15 (0.02) [0.11, 0.20]	0.146/0.003
Plausibility	21.2 (2.2) [16.8, 25.6]	19.6 (2.2) [15.3, 24.0]	0.261/0.001

Estimated means with standard error are in round brackets, 0.95 confidence interval in square brackets, as well as conditional and marginal R^2 in the last column.

During the decision task, participants agreed or disagreed by pressing the buttons 0 or 1.[11] The next item was presented to them after pressing SPACE. This allowed participants to take self-paced breaks. The decision task was preceded by an instruction and a training run with 10 statements. The experimental block started with six filler questions to avoid warm-up effects. After that, the 32 target sentences and 32 fillers were presented in a random order. Similarly, the plausibility rating task started with a short instruction and a training block. After that, the target sentences and fillers were presented on one page in a random order. At this part of the experiment, subjects were allowed to take as much time as they needed. Other than in the decision task, the survey also allowed for correction of answers.

3.2.2. Results and Discussion

Prior to the analysis, extremely high reaction times (five data points over 15 s) were removed.[12] An overview of the results can be seen in **Table 2**. Sentences in which the noun was typically associated to the property were answered faster [$\beta_1 = -165$, $t_{(2505)} = -4.96$, $p < 0.001$]. They also had a slightly higher acceptance rate [$\beta_1 = 0.04$, $t_{(2510)} = 3.01$, $p = 0.003$]. However, the plausibility rating was only insignificantly higher [$\beta_1 = 1.5$, $t_{(2510)} = 1.80$, $p = 0.07$]. All effect sizes were very low, indicating that the typicality did barely influence variation in the data.

Let us now look how the reaction time changed in comparison to the preparatory experiment, where unmodified statements were evaluated. Generally, the reaction time was longer, which is expected, because the sentences were now longer and reaction time included reading time. However, the modifiers had a different influence on reaction time for the typical and atypical sentences. The increase on the median reaction time per item was on average 775 ms ($SD = 362$) for sentences without typicality and $1,007$ ms ($SD = 300$) for the sentences with typicality. A paired t-test confirmed that the mean difference of 232 ms is significant [$t_{(31)} = -3.20$, $p = 0.003$]. A cognitive mechanism that blocks default inheritance could in principle explain the larger increase in reaction time for sentences with typicality. However, the fact that modified typicality statements were still processed slightly faster than their counterparts speaks against such an interpretation. The more likely explanation is that the typicality statements had an initial processing advantage, which was lost by the added modifier. To check for a potential inheritance effect, I also calculated the correlations between the mean item plausibility rating for typical statements from the preparatory experiment and the ratings of the modified statements in this experiment: no significant correlation was found ($r = -0.11$, $p = 0.56$). The knowledge effects prevented default inheritance.

Another question worth exploring is whether typicality impacted the accuracy of the participants during the fast decision task. In order to address this questions, I detected cases in which the answer during the fast decision task did mismatch the answers in the plausibility rating, where the subjects answered without time pressure and had the option to correct answers. A case was considered to be inaccurate if the participant first accepted the sentence as true but rated its plausibility as lower than 20 or if a sentence was rejected but received a plausibility rate higher than 80. It turned out that the typicality of the noun property pair had no effect on such defined inaccuracy [atypical noun: $\beta_0 = 0.038$; difference for typical noun: $\beta_1 = +0.004$, $t_{(225510)} = 0.50$, $p = 0.61$].[13]

The fact that participants were equally consistent in handling negative relevant knowledge if a typical property noun combination was presented speaks against the thesis that a background inheritance needs to be actively blocked when confronted with relevant knowledge. On the other hand, there was a slightly but significantly higher acceptance rate for statements with typicality. This indicates a minor inheritance effect, even in view of the strongly negative background knowledge of the modifier. The somewhat higher—albeit only almost significant—plausibility values point in a similar direction. Is this the result of a prototype bias or was the negative relevance not perceived as sufficiently strong by the subjects?[14]

The second experiment explores this question by considering privative modifiers, where the modified nouns cannot be interpreted as referring to subcategories (e.g. "stuffed bear",

[11] 0 for "rather disagree" and 1 for "rather agree".

[12] Again, I used R with the packages lme4, lmerTest, and performance. Subjects and items were entered with random intercepts. Degrees of freedom were estimated by Satterthwaite's method. The exclusion of extremely long reaction time improved the model fit drastically from conditional $R^2 = 0.156$ to conditional $R^2 = 0.504$. Stricter exclusion rules did not further improve model fit.

[13] The mixed effect model was defined as above: item and subject with random intercepts. A more relaxed threshold (accepted, but rated as less than 50; or not accepted, but rated as more than 50 in plausibility) did not affect this general finding [atypical noun: $\beta_0 = 0.104$; difference for typical noun: $\beta_1 = +0.006$, $t_{(2510)} = 0.52$, $p = 0.60$]. Models that merely considered negative deviation (i.e., acceptance but low probability) lead to similar results.

[14] Especially one item in the experiment still received quite high acceptance "Daredevil tortoises are long-living".

TABLE 3 | Least square means of experiment 2.

	Typical	Non-typical	R^2
	Est (SE) [0.95 CI]	Est (SE) [0.95 CI]	Conditional/Marginal
Reaction time	2397 (85) [2227, 2567]	2441 (85) [2271, 2610]	0.384/0.000
Acceptance rate	0.17 (0.02) [0.12, 0.22]	0.16 (0.02) [0.11, 0.21]	0.172/0.000
Plausibility	15.0 (2.1) [10.9, 19.20]	16.0 (2.1) [11.9, 20.18]	0.236/0.000

Estimated means with standard error are in round brackets, 0.95 confidence interval in square brackets, as well as conditional and marginal R^2 in the last column.

"paper perl"). In this setting, biases from the noun could persist but a reasoning from categories to subcategories will not occur.

3.3. Experiment 2

This experiment investigates whether the effects from experiment 1 occur because the modified noun still refers to a subcategory or whether the noun just triggers an association to the property. If the noun concept's prototype biases participants to associate the property, a slight effect should persist for privative modification, which does not refer to a proper subcategory of the noun category.

3.3.1. Methods

Material: The sentence pairs were the same as in experiment 1. However, I now added modifiers that were not only negatively relevant but potentially privative. This means that the modified noun did not refer to a proper subcategory of the noun concepts, for example, "Paper pearls are expensive" and "Paper marbles are expensive". The full material is again presented in the **Appendix**.

Design: The design resembled that of experiment 1.

Procedure: The subjects were recruited and rewarded via Prolific. Overall, 82 persons participated in this part of the study.

3.3.2. Results and Discussion

As in experiment 1, I checked for undue long reaction times and removed one data point over 15 s. An overview of the outcome is given in **Table 3**, which presents the least square means of the dependent variables.[15] The noun's association to the property had no significant effect on reaction time [$\beta_1 = -43$, $t_{(2509)} = -1.38$, $p = 0.169$], acceptance [$\beta_1 = 0.02$, $t_{(2510)} = 1.21$, $p = 0.225$], or plausibility [$\beta_1 = -1.0$, $t_{(2510)} = -1.18$, $p = 0.238$]. As before, I checked for inconsistent answers, that is, cases in which a subject accepted a statement but judged its plausibility to be below 20 or rejected the statement but gave a plausibility score over 80. Again, typicality did not influence inconsistency [$\beta_0 = 0.050$; difference for typical nouns: $\beta_1 = +0.012$, $t_{(2541)} = 1.45$, $p = 0.146$].[16]

The correlation between the mean plausibility rating of the typical statements from the preparatory experiment and this experiment was not significant ($r = 0.25, p = 0.17$). Compared to the time measured for the unmodified sentences in the preparatory experiment, the effect of the modifier on the reaction time was different depending on whether the noun and property were associated. For typical nouns, the increase (887 ms, $SD = 277$) was higher than for atypical nouns (534 ms, $SD = 410$). The difference of 353 ms was highly significant [$t_{(31)} = 4.60$, $p < 0.001$]. In view of the other results, it seems unlikely that the additional time is needed to block a default inheritance. Rather, by adding the additional privative modifier, the sentence with a typical noun–property association lost its cognitive advantage and, thus, was processed just as a sentence without any involvement of typicality.

3.4. Discussion

Figure 2 provides a summary representation of the mean item trends over the different experiments. It is quite obvious that the typical statements were processed faster and rated as more plausible in the preparatory experiment, as seen on the left of **Figures 2A,C**. The adding of relevant (Exp. 1) or even privative (Exp. 2) modifiers lead to a profound increase in the reaction time and decrease in rated plausibility. This is just as one would expect in view of the strong knowledge influences that were introduced by these modifiers.

More interestingly and perhaps surprisingly, the typical noun–property association was *fully* canceled by the knowledge. In comparison to the statements with typicality involvement, the experiments revealed no strong effect of prototypical association between the noun and the target property. Though the acceptance rate was significantly higher for statements with a typical association in experiment 1, the effect was very small. For privative modifications, I found no effect of typicality at all. While it is to be expected that specific knowledge is much more influential than the prototype, the important result is that the prototype did barely influence the judgment at all. If understanding a noun like "raven" presupposes to process typical properties like blackness, it should have been harder to reject statements that mention these properties. However, there is no evidence that subjects were influenced by typicality and that they had to suppress typical properties in order to answer correctly. This becomes especially apparent by the fastness and accuracy of the answers. The results of my experiments thus support one key critique raised by Connolly et al. (2007) and also hold by Gagné and Spalding (2011). There is no evidence that the processing of typical features is necessary in order to understand the complex concepts.

[15] Subjects and items were again entered with random intercepts. Estimation of degrees of freedom is done using Satterthwaite's method. As before, models were calculated in R with the packages lme4, lmerTest, and performance.

[16] The mixed effect model was specified as above: random intercepts for items and subjects. Satterthwaite's method was used for estimation of degrees of freedom.

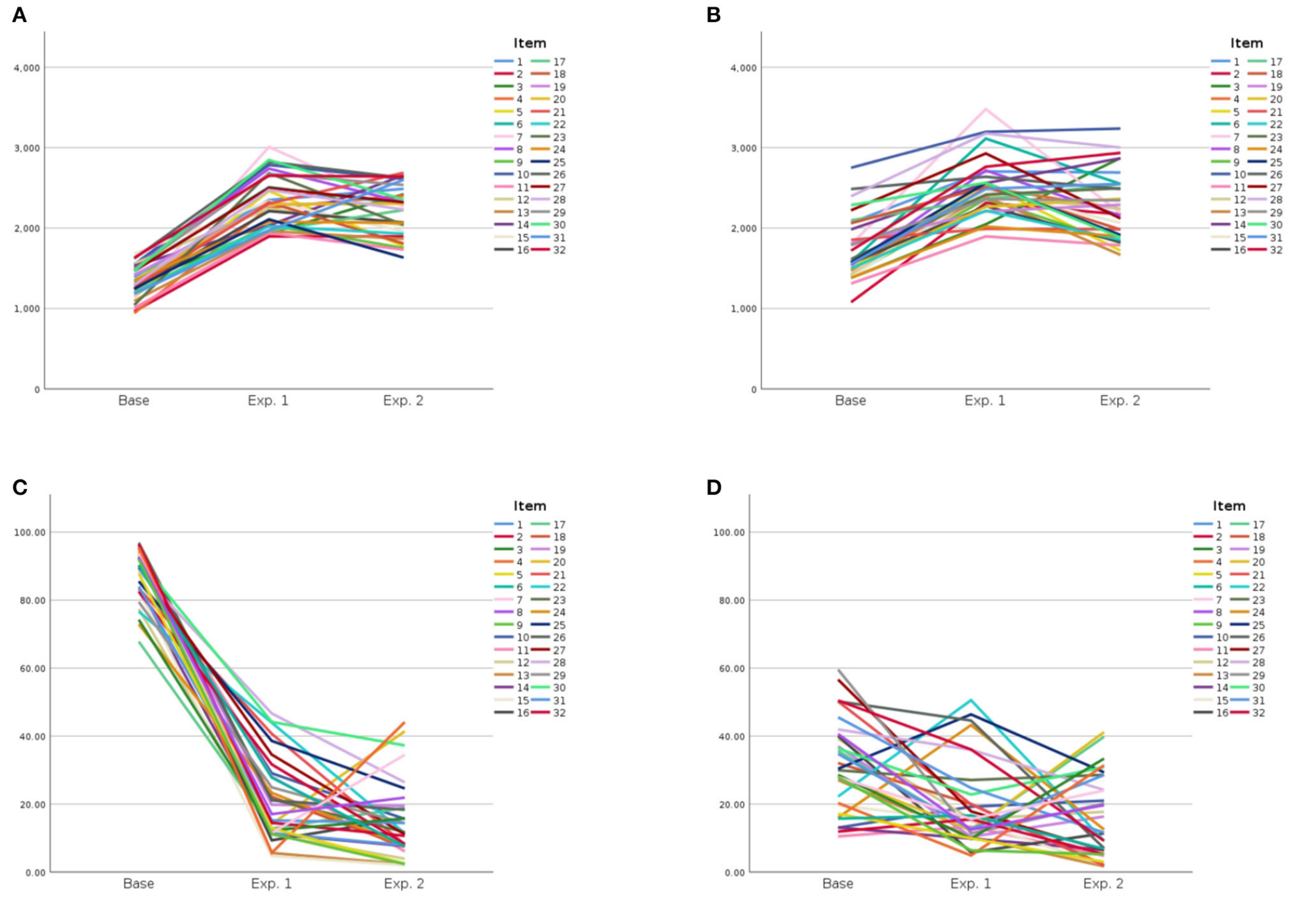


FIGURE 2 | Mean reaction time and mean plausibility of the items in the different experiment, where each line represents the trend of one item. **(A)** Reaction time: typical statements. **(B)** Reaction time: atypical statements. **(C)** Plausibility: typical statements. **(D)** Plausibility: typical statements.

A potential objection to this interpretation is that a lack of evidence of an effect is not equivalent to an evidence of a lacking effect. Indeed, the conclusion I am putting forward here should be viewed with some caution as it essentially rests on negative results. Note, however, that I do not draw the conclusions from the mere lack of statistical significance, which could be easily influenced by the numbers of participants and items. More importantly, the effect sizes in all relevant tests, even those that were significant, are negligibly small. In no way can they explain the considerable default inheritance effect that has been established in the research literature on the modification effect. This makes it very likely that a rational reasoning process—as studied in literature on default logic—lies behind the effect. The gathering of further and more direct evidence for this thesis is an open issue for further research.

4. CONCLUSION

As outlined above, three effects occur if humans are asked to rate the plausibility of a modified sentence: decrease, inheritance, and knowledge effects. Previous research has impressively shown that the decrease effect is extremely stable, even in cases where rational reasoning should block it, that is, for universal statements (Jönsson and Hampton, 2006) or analytic properties (Hampton et al., 2011). Even positively relevant knowledge does not fully block the decrease effect but rather superposes it (Strößner and Schurz, 2020).

The inheritance effect has been less intensively researched than the decrease effect even though it is central for understanding prototype theory to find the source of typicality inheritance. This paper aimed to investigate whether it occurs as a prototype-based bias. The experiments revealed that relevant modifiers tend to block inheritance effects. This result, I conclude, only makes sense if we assume that inheritance occurs as a reasoning process in the absence of knowledge, not as an automatic by-product of composing the meaning. In light of this finding, the reservations Gagné et al. (2017) expressed against a container model of concepts gain support. There is no evidence that we necessarily process concepts as a bundle of such features.

However, I do not reject that concepts are related to prototypes and that they evolve in a way which makes it possible to associate them to prototypes or typical properties (c.f. Jäger, 2007). Indeed, the whole idea of default inheritance, even as an inference, still presupposes concepts that are associated to typical properties (e.g., "cats" or "birds" rather than "non-cats" or "cat and birds"). One general idea of prototype theory is that concepts capture probabilistic covariances in the world (Rosch, 1978; Schurz, 2012) and this is not called into question by my experiments. With the experimental work of this article, I do not reject all ideas of prototype theory in general. The main point is rather that there is no evidence that the processing of a concept alone presupposes to process its prototype or typical features. In view of the many counter-rational findings concerning the decrease effect, this can be interpreted as an optimistic claim: we are easily fooled by our pragmatic biases, but we are not fooled by prototypes.

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

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 the Emmy-Noether programm of the German Research Foundation (From Perception to Belief and Back Again, BR 5210/1-1). I also acknowledge support by the DFG Open Access Publication Funds of the Ruhr-Universität Bochum.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.626023/full#supplementary-material

REFERENCES

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

Connolly, A. C., Fodor, J. A., Gleitman, L. R., and Gleitman, H. (2007). Why stereotypes don't even make good defaults. *Cognition* 103, 1–22. doi: 10.1016/j.cognition.2006.02.005

Cree, G. S., and McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *J. Exp. Psychol. Gen.* 132:163. doi: 10.1037/0096-3445.132.2.163

Dechêne, A., Stahl, C., Hansen, J., and Wänke, M. (2010). The truth about the truth: a meta-analytic review of the truth effect. *Pers. Soc. Psychol. Rev.* 14, 238–257. doi: 10.1177/1088868309352251

Fodor, J. (1998). *Concepts: Where Cognitive Science Went Wrong*. New York, NY: Oxford University Press.

Fodor, J., and Lepore, E. (1996). The red herring and the pet fish: why concepts still can't be prototypes. *Cognition* 58, 253–270.

Gabbay, D. M., Hogger, C. J., and Robinson, J. A. (1995). *Handbook of Logic in Artificial Intelligence and Logic Programming: Volume 3: Nonmonotonic Reasoning and Uncertain Reasoning*. Oxford: Clarendon Press.

Gagné, C. L., and Murphy, G. L. (1996). Influence of discourse context on feature availability in conceptual combination. *Discourse Process.* 22, 79–101. doi: 10.1080/01638539609544967

Gagné, C. L., and Spalding, T. L. (2011). Inferential processing and meta-knowledge as the bases for property inclusion in combined concepts. *J. Mem. Lang.* 65, 176–192. doi: 10.1016/j.jml.2011.03.005

Gagné, C. L., and Spalding, T. L. (2014). Subcategorisation, not uncertainty, drives the modification effect. *Lang. Cogn. Neurosci.* 29, 1283–1294. doi: 10.1080/23273798.2014.911924

Gagné, C. L., Spalding, T. L., and Kostelecky, M. (2017). "Conceptual combination, property inclusion, and the aristotelian-thomistic view of concepts," in *Compositionality and Concepts in Linguistics and Psychology*, eds J. A. Hampton and Y. Winter (Cham: Springer), 223–244.

Goodman, N. D., and Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends Cogn. Sci.* 20, 818–829. doi: 10.1016/j.tics.2016.08.005

Grice, P. (1989). *Studies in the Way of Words.*: Cambridge, MA: Harvard University Press.

Hampton, J. A. (1987). Inheritance of attributes in natural concept conjunctions. *Mem. Cogn.* 15, 55–71.

Hampton, J. A., and Jönsson, M. L. (2012). "Typicality and compositionality: the logic of combining vague concepts," in *The Oxford Handbook of Compositionality*, eds W. Hinzen and E. Machery (Oxford: Oxford University Press), 385–402.

Hampton, J. A., Passanisi, A., and Jönsson, M. L. (2011). The modifier effect and property mutability. *J. Mem. Lang.* 64, 233–248. doi: 10.1016/j.jml.2010.12.001

Hasher, L., Goldstein, D., and Toppino, T. (1977). Frequency and the conference of referential validity. *J. Verbal Learn. Verbal Behav.* 16, 107–112. doi: 10.1016/S0022-5371(77)80012-1

Jäger, G. (2007). The evolution of convex categories. *Linguist. Philos.* 30, 551–564. doi: 10.1007/s10988-008-9024-3

Jönsson, M. L., and Hampton, J. A. (2006). The inverse conjunction fallacy. *J. Mem. Lang.* 55, 317–334. doi: 10.1016/j.jml.2006.06.005

Jönsson, M. L., and Hampton, J. A. (2012). The modifier effect in within-category induction: default inheritance in complex noun phrases. *Lang. Cogn. Process.* 27, 90–116. doi: 10.1080/01690965.2010.544107

Kraus, S., Lehmann, D., and Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artif. Intell.* 44, 167–207. doi: 10.1016/0004-3702(90)90101-5

Krifka, M., Pelletier, F. J., Carlson, G., ter Meulen, A., Chierchia, G., and Link, G. (1995). "Genericity: an introduction," in *The Generic Book*, eds G. Carlson and F. J. Pelletier (Chicago, IL: The University of Chicago Press), 1–124.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13

Lüdecke, D., Makowski, D., Waggoner, P., and Patil, I. (2020). *Performance: Assessment of Regression Models Performance*. *CRAN*. R package.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Francisco, CA: Morgan Kaufmann.

Pearl, J. (1990). "System z: a natural ordering of defaults with tractable applications to nonmonotonic reasoning," in *Proceedings of the 3rd Conference on Theoretical Aspects of Reasoning About Knowledge* (San Francisco, CA: Morgan Kaufmann), 121–135.

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Reber, R., and Unkelbach, C. (2010). The epistemic status of processing fluency as source for judgments of truth. *Rev. Philos. Psychol.* 1, 563–581. doi: 10.1007/s13164-010-0039-7

Reiter, R. (1980). A logic for default reasoning. *Artif. Intell.* 13, 81–132.

Rosch, E. (1978). "Chapter 2: Principles of categorization," in *Cognition and Categorization*, eds E. Rosch and B. Lloyd (Hillsdale, NJ: Lawrence Erlbaum Associates), 27–48.

Rosch, E., and Mervis, C. B. (1975). Family resemblances: studies in the internal structure of categories. *Cogn. Psychol.* 7, 573–605.

Schurz, G. (2005). Non-monotonic reasoning from an evolutionary viewpoint: ontic, logical and cognitive foundations. *Synthese* 146, 37–51. doi: 10.1007/s11229-005-9067-8

Schurz, G. (2012). "Prototypes and their composition from an evolutionary point of view," in *The Oxford Handbook of Compositionality*, eds W. Hinzen and E. Machery (Oxford: Oxford University Press), 530–553.

Smith, E. E., Osherson, D. N., Rips, L. J., and Keane, M. (1988). Combining prototypes: a selective modification model. *Cogn. Sci.* 12, 485–527.

Spalding, T. L., and Gagné, C. L. (2015). Property attribution in combined concepts. *J. Exp. Psychol. Learn. Mem. Cogn.* 41, 693–707. doi: 10.1037/xlm0000085

Spalding, T. L., Gagné, C. L., Nisbet, K. A., Chamberlain, J. M., and Libben, G. (2019). If birds have sesamoid bones, do blackbirds have sesamoid bones? The modification effect with known compound words. *Front. Psychol.* 10:1570. doi: 10.3389/fpsyg.2019.01570

Sperber, D., and Wilson, D. (1986). *Relevance: Communication and Cognition.* Oxford: Blackwell.

Springer, K., and Murphy, G. L. (1992). Feature availability in conceptual combination. *Psychol. Sci.* 3, 111–117. doi: 10.1111/j.1467-9280.1992.tb00008.x

Strößner, C. (2020). Compositionality meets belief revision: a bayesian model of modification. *Rev. Philos. Psychol.* 11, 859–880. doi: 10.1007/s13164-020-00476-8

Strößner, C., and Schurz, G. (2020). The role of reasoning and pragmatics in the modifier effect. *Cogn. Sci.* 44:e12815. doi: 10.1111/cogs.12815

Strößner, C., Schuster, A., and Schurz, G. (2020). "Modification and default inheritance," in *Concepts, Frames and Cascades in Semantics, Cognition and Ontology*, Language, Cognition and Mind, eds T. Gamerschlag, T. Kalenscher, S. Löbner, M. Schrenk, and H. Zeevat (Cham: Springer).

Thorn, P. D.,and Schurz, G. (2018). *Inheritance Inference From an Ecological Perspective*. Mansucript. Available onnline at: https://cognitive-structures-cost18.phil.hhu.de/wp-content/uploads/2018/08/ThornInheritance-Reasoning-from-an-Ecological-Perspective.pdf

Unkelbach, C., and Rom, S. C. (2017). A referential theory of the repetition-induced truth effect. *Cognition* 160, 110–126. doi: 10.1016/j.cognition.2016.12.016

Veltman, F. (1996). Defaults in update semantics. *J. Philos. Logic* 25, 221–261. doi: 10.1007/bf00248150

Wittgenstein, L. (1953). *Philosophical Investigations, Translated by GEM Anscombe.* Oxford: Blackwell.

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

Table A1 | Target items of the experiments.

Typicality statement	Paired noun	Ex 1 modifier	Ex 2 modifier
Aschenbecher sind schmutzig.	Kafeebecher	Abgewaschen	Essbare
Bananen sind gelb.	Erdbeeren	Verfault	Unsichtbar
Keller sind kalt.	Küchen	Beheizt	Computeranimiert
Raben sind schwarz.	Spatzen	Albino	Marmor
Bären leben im Wald.	Enten	Eingefangen	Ausgestopft
Betten werden zum Schlafen genutzt.	Stühle	Ausgestellt	Zerlegt
Mais wächst auf Feldern.	Pilze	Gewächshaus	Synthetisch
Geschirrspüler stehen in der Küche.	Waschmaschinen	Unverkauft	Zerstört
Delfine leben im Meer.	Schwäne	Eingesperrt	Plastik
Garagen werden zum Parken genutzt.	Lauben	Abgesperrt	Eingerissen
Gorillas sind stark.	Mäuse	Krank	Porzellan
Trauben schmecken süß.	Rhabarber	Unreif	Eisern
Grashüpfer springen.	Marienkäfer	Beinlos	Gegrillt
Pistolen werden zum Töten genutzt.	Feuerzeuge	Leer	Unecht
Lämmer sind flauschig.	Schweinchen	Nackt	Schokoladen
Erdbeeren sind saftig.	Karotten	Getrocknet	Stein
Spiegel glänzen.	Wandgemälde	Verschmutzt	übermalt
Löwen leben in Afrika.	Füchse	Zoo	Versteinert
Pfannen nutzt man zum Braten.	Töpfe	Dreckig	Verrostet
Pinguine schwimmen.	Tauben	Betäubt	Geschnitzt
Ratten übertragen Krankheiten.	Hamster	Gesund	Zeichentrick
Perlen sind teuer.	Murmeln	Metall	Papier
Zitronen sind sauer.	Mandarinen	Geschmacklos	Gummi
U-Bahnen sind überfüllt.	Taxis	Nacht	Geister
Schwerter sind gefährlich.	Stöcke	Stumpf	Lego
Panzer werden von der Armee genutzt.	Züge	Ausrangiert	Papp
Krawatten sind formelle Kleidung.	Gürtel	Befleckt	Papier
Toiletten haben eine Spülung.	Waschbecken	Camping	Symbolisch
Tomaten isst man in Salat.	Kartoffeln	Ungewaschen	Pulverisiert
Schildkröten sind langlebig.	Salamander	Draufgängerisch	Elektrisch
Traktoren sind laut.	Kräne	Geparkt	Sandkasten
Dreiräder werden von Kindern benutzt.	Einräder	Riesig	Gläsern
Ashtrays are dirty.	Coffee mug	Washed up	Edibles
Bananas are yellow.	Strawberries	Rotten	Invisible
Cellars are cold.	Kitchens	Heated	Computer animated
Ravens are black.	Sparrows	Albino	Marble
Bears live in the forest.	Ducks	Captured	Stuffed
Beds are used for sleeping.	Chairs	Exhibited	Disassembled
Corn grows in fields.	Mushrooms	Greenhouse	Synthetic
Dishwashers are in the kitchen.	Washing machines	Unsold	Destroyed
Dolphins live in the sea.	Swans	Locked up	Plastic
Garages are used for parking.	Arbors	Locked	Torn down
Gorillas are strong.	Mice	Crane	Porcelain
Grapes taste sweet.	Rhubarb	Immature	Iron
Grasshoppers jump.	Ladybird	Legless	Grilled
Guns are used for killing.	Lighters	Empty	Fake
Lambs are fluffy.	Piggy	Nude	Chocolate
Strawberries are juicy.	Carrots	Dried	Stone
Mirrors are shiny.	Wall painting	Dirty	Painted over
Lions live in Africa.	Foxes	Zoo	Petrified

(Continued)

Table A1 |

Typicality statement	Paired Noun	Ex 1 Modifier	Ex 2 Modifier
Pans are used for roasting.	pots	Dirty	Rusty
Penguins swim.	Pigeons	Stunned	Carved
Rats carry diseases.	Hamster	Healthy	Cartoon
Pearls are expensive.	Marbles	Metal	Paper
Lemons are sour.	Tangerines	Tasteless	Rubber
Subways are crowded.	Taxis	Night	Ghosts
Swords are dangerous.	Sticks	Stump	Lego
Tanks are used by the army.	Trains	Discarded	Cardboard
Ties are formal wear.	Belt	Stained	Paper
Toilets have flushes.	Washbasin	Camping	Symbolic
Tomatoes are eaten in salads.	Potatoes	Unwashed	Pulverized
Turtles are long-lived.	Salamanders	Daredevil	Electrical
Tractors are loud.	Cranes	Parked	Sandbox
Tricycles are used by children.	Unicycles	Giant	Glass

Table A2 | Warm up fillers of experiment 1 and 2.

Talking animals can be found in fairy tales.
Gilded zebras are striped.
Lion kings have manes.
Glittering cushions are decorative.
Crumpled handkerchiefs are white.
Inflatable axes are sharp.

Table A3 | Plausible filler sentences of experiment 1 and 2.

Brown ants live in the ground.
Silver apples are round.
Perforated umbrellas have a handle.
Small blueberries are fruits.
Filterless cigarettes are unhealthy.
Beautiful crows have feathers.
New pens need ink.
Colorful tents are waterproof.
Clean benches are used for resting.
Unfurnished apartments have windows.
Prison beds are uncomfortable.
Successful actresses are rich.
Angry chimpanzees are loud.
Electric bikes are heavy.
Public pianos have many users.
Carving knives are used in the forest.
Fresh salad is green.
Paper boats are light.
Fake cops wear uniforms.
Artificial flowers are durable.
Fake certificates are rectangular.
Slaughtered calves are eaten.
Model trains are used by children.
Water pistols are toys.
Melted rings are hot.
Wooden horses have four legs.
Waving cats are colorful.
Miniature pyramids can be built by oneself.
Vegan sausages are edible.
Canned fish is edible.
Former US presidents are famous.
Candied nuts are sweet.

frontiers
in Communication

ORIGINAL RESEARCH
published: 19 May 2021
doi: 10.3389/fcomm.2021.662027

Rational Interpretation of Numerical Quantity in Argumentative Contexts

Chris Cummins[1]* *and Michael Franke*[2]

[1]*Linguistics and English Language, University of Edinburgh, Edinburgh, United Kingdom,* [2]*Institute for Cognitive Science, University of Osnabrück, Osnabrück, Germany*

Numerical descriptions furnish us with an apparently precise and objective way of summarising complex datasets. In practice, the issue is less clear-cut, partly because the use of numerical expressions in natural language invites inferences that go beyond their mathematical meaning, and consequently quantitative descriptions can be true but misleading. This raises important practical questions for the hearer: how should they interpret a quantitative description that is being used to further a particular argumentative agenda, and to what extent should they treat it as a good argument for a particular conclusion? In this paper, we discuss this issue with reference to notions of argumentative strength, and consider the strategy that a rational hearer should adopt in interpreting quantitative information that is being used argumentatively by the speaker. We exemplify this with reference to United Kingdom universities' reporting of their REF 2014 evaluations. We argue that this reporting is typical of argumentative discourse involving quantitative information in two important respects. Firstly, a hearer must take into account the speaker's agenda in order not to be misled by the information provided; but secondly, the speaker's choice of utterance is typically suboptimal in its argumentative strength, and this creates a considerable challenge for accurate interpretation.

Keywords: pragmatic inference, argumentative language use, non-cooperative dialogue, argument strength, information selection, quantity expressions

OPEN ACCESS

Edited by:
Matthew W. Crocker,
Saarland University, Germany

Reviewed by:
Christoph Hesse,
Leibniz Center for General Linguistics (ZAS), Germany
Anouschka Foltz,
University of Graz, Austria

****Correspondence:***
Chris Cummins
c.r.cummins@gmail.com

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

Received: *31 January 2021*
Accepted: *27 April 2021*
Published: *19 May 2021*

Citation:
Cummins C and Franke M (2021) Rational Interpretation of Numerical Quantity in Argumentative Contexts. Front. Commun. 6:662027. doi: 10.3389/fcomm.2021.662027

INTRODUCTION

How should a rational hearer interpret a statement of numerical quantity, such as 1)?

1) More than 30 states voted Democrat in the 1996 United States Presidential election.

Assuming that the speaker is accurate, the hearer can begin by deriving the semantic meaning of the quantity expression, and arrive at the interpretation that the cardinality of the set of Democrat-voting states in the 1996 election is greater than 30. If the hearer is willing to make additional assumptions about the speaker's cooperativity and knowledgeability, they can derive additional pragmatic inferences. Specifically, they can potentially infer that the speaker is unable to assert informationally stronger alternatives to 1), and hence either that these alternatives are false or that the speaker is ignorant as to their truth-value. In this case, informationally stronger alternatives potentially include those which give larger or more precise numbers (*more than 40*, *35*) or which describe wider date ranges (*in every Presidential election*).

But what if the speaker is strategic, in the sense that they wish to present information that will optimally support a particular argumentative agenda? For the rational hearer, this creates both a problem and an opportunity. On the one hand, the standard pragmatic inferences mentioned above may be unavailable, on the basis that the speaker may simply be declining to utter stronger alternatives that are known to be true, for purely strategic reasons. Thus, in 1), perhaps the speaker wishes to discuss the results of the 1996 election in isolation, in order to make a point about the relative strength of the candidates that particular year. On the other hand, if the speaker is known to be pursuing a particular argumentative agenda, this opens up the possibility of the hearer drawing inferences about the falsity of alternatives that would have been argumentatively stronger, whether or not these are informationally stronger in the usual pragmatic sense. For instance, a speaker who wished to argue that the Democrats can win a comfortable majority of states might choose to discuss the most recent example of them doing so, in which case in 1) they would have said *2008* rather than *1996* if the resulting sentence had still been true.

In this paper, we outline issues of rational use of language in argumentative discourse. Rational communication in non-cooperative contexts has been studied before, e.g., from the perspective of game theory (Franke et al.,2012; de Jaegher and van Rooij, 2014) and also via experimental methods (Franke et al.,2020). The argumentative dimension has been stressed as an important perspective on language use (Anscombre and Ducrot, 1983) that offers an alternative to purely information-based accounts of interaction. It has been used to explain a variety of natural language phenomena, such as the meaning and distribution of particles like *also* and *even* (Merin, 1999) or that of adversarial connectives such as *but* (Winterstein, 2012). Here, we focus specifically on argumentative language use in the domain of numerical quantity expressions. We first survey some of the relevant issues in current research on the semantics and pragmatics of numerical quantity, under standard assumptions about cooperativity in ***Standard Semantic and Pragmatic Meanings of Numerical Expressions***. We then discuss, in ***Argumentative Framing for a Single Numerical Quantity***, how argumentative motives affect a speaker's choice of utterance when describing a single numerical quantity. ***Argumentative Framing for Complex Information States With Complex Utterances*** extends these considerations to more complex cases where more than one numerical feature is potentially relevant for argumentative framing. ***Quantifying Argumentative Strength, and Allowing for Uncooperativity*** then introduces a notion of argumentative strength, following Merin (1999), which aims to subsume the considerations laid out in the foregoing discussion. ***A Case Study: Reporting the Research Excellence Framework*** subsequently derives some more concrete predictions of this approach and tests them with reference to a small corpus of argumentative usages of quantity expressions, drawn from the public statements made by United Kingdom universities concerning their rankings in the 2014 Research Excellence Framework (REF). We show that these usages can usefully be understood by appeal to the notion of argumentativity that we propose, but also that they present a particular interpretive challenge to the hearer as a consequence of their argumentative strength typically being suboptimal.

STANDARD SEMANTIC AND PRAGMATIC MEANINGS OF NUMERICAL EXPRESSIONS

It is tempting to assume that expressions of numerical quantity will be easy to formalise semantically. However, as the enduring debates in the semantics and pragmatics literature testify, many turn out on closer inspection to require sophisticated and subtle analyses. The question of how to formalise these meanings is important for semantic and pragmatic theory, but also for real-life communication, given the crucial role that number plays in conveying precise information that feeds into high-stakes decision-making.

As used in natural language, 'bare' (unmodified) numerals already admit multiple possible interpretations. Horn (1972) noted that bare numerals can express both exact readings, as in 2), and lower-bound ("at least") readings, as seems to be preferred for 3). In some cases, as pointed out by Carston (1998), bare numerals appear to contribute to upper-bound ("at most") readings, as in 4). And round numbers in particular can also convey approximate meanings, as discussed by Krifka (2009), as in 5), which is widely judged true, or at least true enough, if the number of people in the room is, for instance, 99 or 101 (Lasersohn, 1999).

2) I have three children.
3) People with three children are entitled to extra benefits.
4) You can have 2000 calories without putting on weight.
5) There are a hundred people in the room.

This ambiguity creates a potential challenge for the hearer: are they able to recover the speaker's intended meaning, given that this is not linguistically signalled? This is a widespread issue. Taking the case of approximate readings as in 5), speakers frequently round values before reporting them, and do not typically state that they have done so (for instance in telling the time, e.g. 7:30pm, cf. Van der Henst et al., 2002; and indeed in providing summary statistics for an experiment, e.g. "mean RT = 345 ms"). Hence, the way bare numerals are routinely interpreted in natural language gives rise to some pitfalls when we attempt to convey information with them at any given level of precision.

When speakers use modified numerals such as *more than/at most/up to 100*, a different set of issues arises. The ambiguity discussed above does not occur, as pointed out by Solt (2014): in this case, the semantic meaning contributed by the numeral is clearly exact. This imposes an additional constraint on the speaker. For instance, if there are 98 people in the room, a speaker can utter 5) and be judged to have told the truth, but if one further person then entered the room, a speaker who uttered 6) would still be judged to have spoken falsely, because *100* is interpreted as obligatorily exact in 6). That is to say, *more than 100* means *more than precisely 100* rather than merely *more than are present in a situation of which '100' could be truthfully asserted*.

6) There are more than 100 people in the room.

However, a different kind of ambiguity, at a pragmatic level, arises from utterances such as 6). In addition to conveying a (semantic) lower bound on the possible value under discussion, an expression such as *more than 100* appears to convey a (pragmatic) upper bound (Cummins et al., 2012). For instance, the utterance of 6) typically appears to convey the falsity of 7).

7) There are more than 200 people in the room.

Cummins et al. (2012) propose that these enriched meanings can be treated as quantity implicatures, and more specifically scalar implicatures: the use of *more than 100* implicates the falsity of the corresponding sentence with the stronger scalar alternative *more than 200*. But this analysis predicts further scope for misunderstanding between speaker and hearer, as it is not clear which stronger alternatives should be considered to have been negated. Should *more than 100* be taken to convey the falsity of *more than 110*, *more than 125*, *more than 150*, or none of these?

A partial solution to this problem, in the spirit of traditional approaches to scalar implicature, is to argue that the relevant stronger alternatives–which give rise to implicatures–involve numerals which are at least as salient as the original numeral. The notion of granularity, as discussed by Krifka (2009), offers one way of fleshing out this idea. The idea is that round numbers are scale points of scales with differing granularities–60 is at once a scale point in scales graduated by units, tens, perhaps twenties, and so on–and only numbers which are scale points on equally coarse-grained scales constitute scalar alternatives.

However, the limits of this approach are clear. As applied to round numbers in neutral contexts, the hearer still needs to understand which scale a speaker means to evoke–when they say *more than 100*, are they thinking of 100 as a scale point on a scale of tens, or 25s, or 100s? This will determine whether the scalar alternative is *more than 110*, *more than 125*, or *more than 200*. Various considerations might influence how hearers attempt to resolve this problem (see Hesse and Benz, 2020). And specific contexts may be associated with particular scales which supervene. For instance, salient milestones in the United Kingdom Singles Chart traditionally include Top 75 and Top 40, but not Top 50: a song that peaked at #48 could reasonably just be called a Top 75 hit, contrary to the predictions of a general granularity-driven account.

Both at a semantic and pragmatic level, then, the interpretation of numerical expressions creates challenges for the hearer, as the speaker is not obliged to signal the precise sense in which they intend a numeral to be interpreted. And so far we have assumed throughout that we are dealing with a cooperative discourse environment, in which the speaker intends their message to be perfectly reconstructed by the hearer.

What about discourses that are not fully cooperative in the sense of aiming for accurate, precise information transmission? Suppose, in particular, that the speaker wishes the hearer to get a false impression about a particular quantity. We have already seen how this situation might arise by accident–the hearer might take a precise numeral to be an approximation, a lower-bound numeral to be precise, or a modified numeral to give rise to an implicature that was not intended. Can an argumentative speaker exploit these natural possibilities for misunderstanding in order to mislead the hearer in a particular direction? And if so, how should a rational hearer respond in order not to be misled?

The following sections look in more detail at the interplay between, on the one hand, the pragmatic interpretation of quantity words as studied in the context of standard information-seeking cooperative discourse, and, on the other hand, a speaker's interest in presenting a known state of affairs in a particularly favourable light. ***Argumentative Framing for a Single Numerical Quantity*** looks at the arguably more basic case in which the relevant information is just a single numerical quality, and the speaker knows this precisely, but wishes the hearer to perceive it to be as high as possible. ***Argumentative Framing for Complex Information States With Complex Utterances*** extends this analysis to more complex situations where more than one feature matters for the speaker's argumentative framing.

ARGUMENTATIVE FRAMING FOR A SINGLE NUMERICAL QUANTITY

The goal of this section is to investigate how the pragmatic inferences discussed in the previous section, stemming from the usually assumed ideal of a cooperative information-conveying discourse, may be exploited by a speaker who knows the true value *N* of some numerical property but wishes to induce in the hearer an impression that this quantity is in fact higher than *N*. We refer to this situation as *high-framing* of a single quantity. We first look at possibilities of high-framing of a single quantity by using pragmatic slack, or pragmatic halos, associated with unmodified numerals in ***Exploiting Pragmatic Slack in Round Bare Numerals for High-Framing***. ***Exploiting the Imprecision of Round Modified Numerals for High-Framing*** then looks at roundness effects associated with modified numerals. Finally, ***The Potential Sub-Optimality of Non-Round Numbers***. explores the potential sub-optimality of using precise non-round number terms for high-framing.

Exploiting Pragmatic Slack in Round Bare Numerals for High-Framing

Suppose that a speaker, fully knowledgeable about a precise numerical quantity *N*, wishes to give a hearer a maximal impression of this quantity without speaking falsely[1]. What strategies might they adopt, given what we know about the interpretation of numerical quantity expressions?

One option is to make good use of imprecision and pragmatic slack. If *N* is just below a round number, the speaker might try

[1] We assume throughout that we are dealing with speakers who are disposed to be honest, in the minimal sense of not making assertions that could be judged semantically false. However, this leaves open the possibility that such speakers may choose to mislead their hearers pragmatically, appealing to plausible deniability (Pinker et al., 2008).

using that round number *M*: for instance, uttering 5) when there are in fact 98 people in the room. The hearer might interpret this as exact, or better yet from the speaker's point of view, as a lower-bound, i.e. as a commitment on the speaker's part to the existence of a set of 100 people who are in the room.

However, if the true attendance were 102, uttering 5) would risk the hearer getting a needlessly low impression of it, contrary to the speaker's interests; and if the attendance were not within the 'pragmatic halo' (Lasersohn, 1999) of 100, the speaker could not truthfully utter 5) at all.

In sum, we expect high-framing speakers who know true *N* to be able to use pragmatic slack to their advantage in the following way: use round number *M* > *N* to describe *N* if *N* is plausibly contained in a pragmatic halo around *M*.

Exploiting the Imprecision of Round Modified Numerals for High-Framing

A related but perhaps more powerful means of high-framing is to use round modified numerals which ensure a lower-bound interpretation in the semantics. For example, if *N* is the true known number, the high-framing speaker can use *more than M*, relating the quantity under discussion to some reference point *M*. Semantically, it would be natural to choose *M* to be as large as possible, thus ruling out as many (low) potential values as possible. However, pragmatically, as discussed above, the optimal choice of *M* is not straightforward, because *more than M* can implicate *not more than O* for various values *O* > *M*. Indeed, according to Cummins et al. (2012), the values that hearers associate with the description *more than 110* may be generally lower than those they associate with *more than 100* (although Hesse and Benz, 2020, have apparently conflicting data on this point). If this is so, a speaker wishing to emphasise the largeness of a crowd of 111 might be better off uttering 6), repeated below, rather than the semantically stronger 8).

8) There are more than 110 people in the room.

On a granularity-based account, this counterintuitive result arises because 8) effectively leaks information about the level of precision at which the speaker is operating–it seems highly likely that the speaker of 8) would have uttered 9) if they could do so. By contrast, it is not clear that the speaker of 6) is operating at such a fine-grained level, and they might not utter 9) even if they knew it to be true. Hence, the hearer may be more confident that 8) implicates the falsity of 9) than they could be that 6) implicates the falsity of 9).

9) There are more than 120 people in the room.

We conclude that speakers may choose to describe true known *N* for the purpose of high-framing by using a modified numeral like *more than M*, which semantically only contains a lower-bound. If they do so, they should select *M* in such a way that the expected pragmatic interpretation of *more than M* conveys higher values in information-seeking cooperative discourse than any other reference point or round number *M'* < *N* would in the phrase *more than M'*.

The Potential Sub-Optimality of Non-Round Numbers

So far, we have focused on round numbers and their potential usefulness for high-framing. Let us now consider whether high-framing might benefit from the use of non-round numbers.

We note first that, even with non-round numbers, the speaker can convey additional quantity information, such as in 10) where the non-round *19* is selected as the endpoint of a particular range.

10) If restored to operation, it would be one of the 19 largest telescopes existing today, all of which are in constant demand (https://www.nytimes.com/1988/11/15/science/volunteers-seek-revival-of-famed-telescope.html, retrieved 24/03/20).

Describing the telescope as *one of the 19 largest* rather than *one of the 20 largest* clearly makes a semantically stronger claim, which supports the speaker's apparent point that it would be an exceptionally large telescope. However, using *19* rather than *20* invites the hearer to draw inferences about the motivation for this precise choice–an available inference in this case being that the telescope would rank precisely 19th in size (unless there is some reason why we should care about precisely the 19 largest telescopes in particular). If the hearer infers this, the speaker has perhaps been less argumentatively effective than if the hearer had merely concluded that the telescope would be somewhere among the largest 20.

Similarly, in 11), the use of *top 19* strongly invites the inference that the salient stronger (given the entailment direction of the utterance) alternative *top 20* doesn't hold–i.e. that the team currently 20th in the CFP rankings, like Clemson, has not faced a team currently in the committee's top 25, which in turn suggests that Clemson's status is less special than the speaker seems to want to suggest.

11) Clemson is the only team among the top 19 in the CFP rankings that hasn't faced a team currently in the committee's top 25 (https://www.espn.co.uk/college-football/story/_/id/28196686/dabo-swinney-says-clemson-held-different-standard-cfp-voters, retrieved 24/03/20)

A similarly complex example occurs in 12).

12) Disappointingly, 10 of the world's 19 most unequal countries are in sub-Saharan Africa (https://www.un.org/africarenewal/magazine/december-2017-march-2018/closing-africa%E2%80%99s-wealth-gap, retrieved 24/03/20).

Here, by similar reasoning, the hearer can infer that *19* could not be replaced by *18*, as otherwise the speaker would have done so. It follows that the 19th most unequal country in the world is in sub-Saharan Africa, and thus only nine of the world's 18 most unequal countries are in that region. This is presumably considered to be a less compelling argument for the speaker's overall thesis than *10 of the world's 19*, as otherwise they would have uttered it in the first place.

In each of these cases, then, choosing the semantically strongest description invites pragmatic inferences which appear to push back against the speaker's argumentative goals (namely, in 11), that Clemson is distinguished by its lack of strong opposition so far, and in 12) that inequality is widespread in sub-Saharan Africa). Of course, the extent to which hearers draw these inferences is an empirical question, so it is not self-evident that these utterances constitute less effective arguments than informationally weaker alternatives would (for instance, *one of only two teams in the top 20*, or *10 of the world's 20*, respectively). However, it is equally unclear that they constitute better arguments than informationally weaker alternatives would.

In summary, then, the use of non-salient numbers in utterances such as 10)–12) invites inferences about the falsity of corresponding stronger statements involving more salient numbers. For this reason, we might expect non-salient numbers to be generally poor choices for high-framing.

ARGUMENTATIVE FRAMING FOR COMPLEX INFORMATION STATES WITH COMPLEX UTTERANCES

Examples 11) and 12) begin to show some of the complexity that is typical of argumentative language use. In these, unlike the previous examples, the speaker is not merely expressing one quantity as to make it sound large or small: rather, they have chosen two numbers with which to make a particular argument. In 12), the speaker has not only chosen the frame *X of the world's Y* but has made a deliberate choice about how to populate it, out of all the possible number pairs (*X*, *Y*) that would make the sentence true, and has presumably chosen numbers which they feel are rhetorically effective.

The broader point that this illustrates is that a speaker citing complex data in support of their argument can do so in many ways. An effective choice may invite the hearer to draw additional inferences that support the speaker's argument[2]. On the flip side, an ineffective choice may invite the hearer to draw inferences that undermine the speaker's argument. Bill Bryson (1998): 112f describes drawing just such inferences in response to a car advertisement:

"[The advert] says something like 'The new Dodge Backfire. Rated number one against the Chrysler Inert for handling. Rated number one against the Plymouth Repellent for mileage. Rated number one against the Ford Eczema for repair costs.' As you will notice . . . in each category the Dodge is rated against only one other competitor. . . .[I]f the Dodge were rated top against ten or twelve or fifteen competitors in any of those categories, then presumably the ad would have said so. Because it doesn't say so, one must naturally conclude that the Dodge performed worse than all its competitors except the one cited."

In this scenario, the sceptical hearer's inferences derive ultimately from the perception that a knowledgeable speaker, with a particular argumentative agenda, has chosen to present a very limited amount of information. The hearer infers that this reflects a strategic decision, motivated by the fact that presenting additional information (how the Dodge compares to the Chrysler in mileage, etc.) would undermine the speaker's broader communicative goal (presenting the Dodge as the most attractive choice).

From the standpoint of pragmatic analysis, we could formalise this idea by noting that the advert, as described, would give rise to a series of ad hoc implicatures to the effect that the Dodge is inferior to the Chrysler and Plymouth (and perhaps other competitors) in repair costs, inferior to the Chrysler and the Ford (and perhaps other competitors) in mileage, and inferior to the Plymouth and the Ford (and perhaps other competitors) in handling. These ad hoc implicatures are proposed to arise on the basis that entailment relations exist between sentence pairs such as 13) and 14), with 14) entailing 13); and given a context in which the stronger sentence 14) would be relevant, the utterance of the weaker sentence 13) is taken to implicate the stronger sentence's falsity each time.

13) The Dodge is rated higher than the Chrysler for handling.
14) The Dodge is rated higher than the Chrysler and the Plymouth for handling.

Given a sufficiently complex set of quantitative data, the set of true statements that could be made about the data will be very large. Under these circumstances, the speaker's decision to say whatever they decide to say, rather than any of the alternatives, could give rise to a rich array of inferences. As an example, consider a scenario in which 15) and 16) would each be plausible descriptions of a situation.

15) All of the students got some of the questions right.
16) Some of the students got all of the questions right.

In purely semantic terms, neither of these sentences is strictly more informative than the other, in the sense that no entailment relation obtains between them. However, a hearer might feel that one of them is more valuable than the other, as a conversational contribution, in a world where both are true. Suppose that such a hearer thinks that 16) is clearly the more valuable option. They should then take the utterance of 15) by a knowledgeable speaker to convey the negation of 16). An argumentative speaker who is aware of the hearer's preference can then potentially exploit it: they can cause the hearer to believe that 16) is false (perhaps incorrectly) by asserting 15).

In its effect, this would be much like a speaker asserting *some* in order to convey *not all* when they know that *all* is the case. But a speaker who asserts *some* when they know that *all* is the case could be argued to be dishonest, because there is a widespread

[2] As discussed earlier, whether or not those inferences are true may not be important to the argumentative speaker, although the speaker may wish them to be covered by plausible deniability.

understanding that *some* typically conveys *not all* in declarative contexts–a point discussed in more detail by Meibauer (2014) and Franke et al. (2020). By contrast, a speaker who asserts 15) in order to (misleadingly) convey the falsity of 16) might have some measure of plausible deniability against the claim of dishonesty, because speakers and hearers do not share contextually stable intuitions about the relative usefulness of these two possible utterances.

In summary, the above examples suggest that the effectiveness of a particular utterance, construed as an argument towards a particular goal, depends both on the semantic content of the utterance and the pragmatic inferences drawn by the hearer as a result of the utterance. Moreover, the eventual interpretation of a hearer who takes into account that the speaker has an argumentative agenda may diverge considerably from the pragmatic interpretation that they would be predicted to arrive at in cooperative contexts. Consequently, the usual tools with which we analyse the semantics and pragmatics of cooperative discourse are of limited use in helping us to systematise these ideas. In the following section, we explore how we can address this challenge by appeal to the notion of argumentative strength.

QUANTIFYING ARGUMENTATIVE STRENGTH, AND ALLOWING FOR UNCOOPERATIVITY

In the context of cooperative communication, we can use ideas around informativity and relevance to quantify the extent to which a candidate utterance would be a useful contribution to the discourse, in the sense of bringing about positive cognitive effects in the hearer, in Sperber and Wilson's (1986) terms. Somewhat analogously, given a (not necessarily cooperative) situation in which a speaker wishes to make a particular point, we can explore their choice of utterance by considering the extent to which candidate utterances would represent good arguments in support of that point. In the following we therefore explore a quantitative measure of argumentative strength of an utterance and consider the predictions that it makes about usage under various different assumptions. In ***Argumentative Strength for a Semantic Interpretation of an Utterance*** we consider argument strength in the case where hearers adopt a purely semantic interpretation of the speaker's utterance, and in ***Argumentative Strength for a Pragmatic Interpretation of an Utterance*** we expand this to the case where hearers are presumed to take into account the usual pragmatic inferences that would be available in a cooperative context. In ***Argumentative Strength for Complex Cases*** we exemplify how complex contexts invite the speaker to be more selective in their choice of utterance than standard pragmatic theories usually accommodate. Finally, in ***Rational Interpretation in an Argumentative Context***, we consider the perspective of a sceptical hearer confronted with a speaker who is selective in this way, and examine how argumentative strength can be evaluated in this kind of non-cooperative setting.

Argumentative Strength for a Semantic Interpretation of an Utterance

Working within the tradition of argumentative approaches to language (Anscombre and Ducrot, 1983), Merin (1999) proposes to model the argumentative strength (arg_str) of an utterance *u* with reference to the weight of evidence that it provides in support of the speaker's communicative goal hypothesis G. This notion of weight of evidence can be unpacked (following Good, 1950, and others) as a log-likelihood ratio as in **Eq. (17)** [3].

$$\text{arg\_str}\,(u, \mathrm{G}) = \log \frac{\mathrm{P}\,(u|\mathrm{G})}{\mathrm{P}\,(u|\neg\mathrm{G})} \tag{17}$$

Here, P(*u*|G) denotes the probability that utterance *u* is true if hypothesis G is true, and P(*u*|¬G) denotes the probability that utterance *u* is true if hypothesis G is false. The idea is that an utterance with a positive argumentative strength with respect to hypothesis G is, by definition, one that is more likely to be true if G is true than it is to be true if G is false[4].

For simple examples, it is easy to evaluate argumentative strength according to this measure. For instance, 18) has positive (indeed, infinitely large) argumentative strength in support of the contention G that the Poincaré conjecture holds, because $\mathrm{P}(u|\neg\mathrm{G}) = 0$ and $\mathrm{P}(u|\mathrm{G}) > 0$.

18) Grigori Perelman proved the Poincaré conjecture in 2006.

However, in more complex cases, it can be difficult to precisely calculate argumentative strength, while it is still possible to evaluate at least qualitative predictions based on intuition. To illustrate this, we can revisit 11), repeated here (omitting *disappointingly*) as 19). We might take it that the speaker's communicative goal in this context is something like 20).

19) 10 of the world's 19 most unequal countries are in sub-Saharan Africa.
20) Inequality is widespread in sub-Saharan Africa.

The argumentative strength of the utterance, as defined above, is calculated from the probability that 19) is true given 20) and the probability that 19) is true given the negation of 20). But the latter probability, in particular, is not readily calculable for speaker or hearer, because even if we define *widespread* crisply, *not widespread* clearly covers a range of values. However, the speaker and hearer may still have intuitions about the probabilistic relations between 19) and 20). For instance, we might say that 19) has positive argumentative

[3] The use of log-likelihood in Good's formalism ensures that the weight of evidence has desirable additive properties.

[4] Note that this definition of argumentative strength only makes reference to the truth conditions associated with the quantity expression. For the purposes of this paper, we do not explore the idea developed by Mira Ariel (2004) and subsequently that the use of particular quantity expressions is conventionally associated with particular argumentative effects, although this work is compatible with that idea: we could, for instance, take the nature of the quantity expression to inform the hearer's understanding of the identity of G.

strength with respect to 20) if 21) is judged more probable than 22), and negative argumentative strength if the reverse is true.

21) Inequality is widespread in sub-Saharan Africa, and 10 of the world's 19 most unequal countries are located there.
22) Inequality is not widespread in sub-Saharan Africa, and 10 of the world's 19 most unequal countries are located there.

By definition, an utterance with positive argumentative strength should constitute positive evidence in favour of the speaker's communicative goal G over its negation, and hence a rational hearer should respond to such an utterance by increasing the strength of their belief in G. However, as discussed in ***Argumentative Framing for a Single Numerical Quantity***, an utterance might also give rise to pragmatic enrichments that would tend to oppose the argument being made by the speaker. This possibility is not taken into account in **Eq. (17)**, which is concerned purely with the semantic content of the utterance.

Argumentative Strength for a Pragmatic Interpretation of an Utterance

To see how pragmatic inferences which are ordinarily associated with utterances of single-number descriptions (see ***Standard Semantic and Pragmatic Meanings of Numerical Expressions*** and ***Argumentative Framing for a Single Numerical Quantity***) might affect the notion of argumentative strength, let us return to a simpler example. Suppose that we, as hearers, believe that our conference will be a success if and only if more than 120 people attend. Let S be the event that more than 120 people attend the conference, and assume that it is common knowledge that people will attend if and only if they have registered. A speaker who (privately) has the argumentative goal of convincing us that the conference will be a success then utters either 23) or 24).

23) More than 100 people registered.
24) More than 110 people registered.

On semantic grounds, P(S|(24)) ≥ P(S|(23)): that is to say, the probability that S is true given that 24) is true is at least as great as the probability that S is true given that 23) is true. This holds because S is false in all worlds in which 23) is true and 24) is false. Therefore 24) should be a better argument for S than 23) is. However, as discussed earlier, 24) strongly invites the pragmatic inference that S is false, which is arguably not true of 23). If this pragmatic analysis is correct, taking that inference into account may change the picture and result in 23) being a better argument than 24) for the truth of S.

The general point here, once again, is that utterances which are effective arguments on their semantics may not be effective when pragmatic enrichments are included in the calculation. It would be helpful to have a notion of argumentative strength that takes this into account. More precisely, if we include pragmatic considerations, what is necessary for argument strength is not merely that the utterance u should be more likely true given G than given not-G, but rather that u should be more likely felicitously assertable–in the sense of both being true and not giving rise to false implicatures–given G than given not-G. Let A(u) stand for the fact that u is felicitously assertable. We could then propose a notion of pragmatic argument strength (prag_arg_str) as in **Eq. (25)**.

$$\text{prag\_arg\_str}(u, \mathrm{G}) = \log \frac{\mathrm{P}(A(u)|G)}{\mathrm{P}(A(u)|\neg G)} \tag{25}$$

To illustrate how this works, we can flesh out the example of 23) and 24) further with some additional assumptions: these are not intended to be realistic, but just serve to illustrate the calculation process. Suppose there is a 90% probability of an utterance being interpreted as conveying a pragmatic enrichment, and that for *more than 100* that enrichment is *not more than 150* while for *more than 110* it is *not more than 120*. For simplicity let us suppose that no other pragmatic interpretations are in play. Suppose further that the true value under discussion–the number of people who have registered for the conference–is uniformly distributed on the range [0, 200]. Recall that S is the event that more than 120 people will attend, and we are assuming that it is common knowledge that they will attend if and only if they have registered.

According to the measure in **Eq. (17)**, the argumentative strength of utterance u toward the goal G is the log of the ratio of P (u|G) and P (u|¬G). Here, G = S, and we consider first the utterance *more than 100*. The probability that *more than 100* is true given that *more than 120* is true equals 1; the probability that *more than 100* is true given that *more than 120* is false equals 1/6 here. Recall that we assume that the true value is uniformly distributed on [0, 200]–if *more than 120* is false, it must lie in the range [0, 120], again uniformly distributed. Hence the probability that it exceeds 100 is 20/120 = 1/6. So, according to **Eq. (17)**, the argumentative strength of *more than 100* is equal to log(1/(1/6)) = log 6 ≈ 0.78. Now we consider the utterance *more than 110*. Again, the probability that *more than 100* is true given that *more than 110* is true equals 1; the probability that *more than 100* is true given that *more than 110* is false equals 1/11 here. If *more than 110* is false, the true value is uniformly distributed on [0, 110] and has a 1/11 chance of exceeding 100. So, per **Eq. (17)**, the argumentative strength of *more than 110* is equal to log(1/(1/11)) = log 11 ≈ 1.04, which exceeds the argumentative strength of *more than 100*.

Now let us consider instead the measure in **Eq. (25)**, under which the argumentative strength of utterance u towards the goal G is the log of the ratio of P (A(u)|G) and P (A(u)|¬G). Again, G = S, and here we have adopted the assumptions that there is a 90% probability of the utterance being pragmatically interpreted, and that if it is, *more than 100* will be interpreted as *not more than 150* and *more than 110* will be interpreted as *not more than 120*. Consider first *more than 100*. This is assertable in two disjoint eventualities: i) it attracts a pragmatic interpretation and the true value lies in the range (100, 150][5], or ii) it does not attract a

[5] Here we use (100, 150] to refer to the half-open interval comprising values that are greater than 100 and less than or equal to 150, which are those values for which *more than 100* can be felicitously asserted if it implicates *not more than 150*.

pragmatic interpretation and the true value lies in the range (100, 200]. If S is true, then the probability that the true value lies in the range (100, 150] is 3/8 (because it is uniformly distributed on (120, 200]), and the probability that the true value lies in the range (100, 200] is 1. So the total probability that *more than 100* is assertable is (90% x 3/8 + 10% x 1) = 35/80. If S is false, then the probability that the true value lies in the range (100, 150] is 1/6 (because it is uniformly distributed on [0, 120]), and the probability that the true value lies in the range (100, 200] is also 1/6. So the total probability that *more than 100* is assertable is (90% x 1/6 + 10% x 1/6) = 1/6. Hence, under the measure in **Eq. (25)**, the argumentative strength of *more than 100* is log ((35/80)/(1/6)) = log (21/8) = 0.419.

Now consider *more than 110*. This is assertable in two disjoint eventualities: i) it attracts a pragmatic interpretation and the true value lies in the range (110, 120], or ii) it does not attract a pragmatic interpretation and the true value lies in the range (110, 200]. If S is true, then the probability that the true value lies in the range (110, 120] is zero, and the probability that the true value lies in the range (110, 200] is 1. So the total probability that *more than 110* is assertable is (90% x 0 + 10% x 1) = 1/10 (or, to put it another way, *more than 110* is only assertable if it attracts no pragmatic enrichment, and we are assuming this to happen with 1/10 probability in this illustration). If S is false, then the probability that the true value lies in the range (110, 120] is 1/12, and the probability that the true value lies in the range (110, 200] is also 1/12. So the total probability that *more than 110* is assertable is (90% x 1/12 + 10% x 1/12) = 1/12. Hence, under the measure in **Eq. (25)**, the argumentative strength of *more than 110* is log ((1/10)/(1/12)) = log (6/5) = 0.079, which is lower than for *more than 100*.

Hence, under these illustrative assumptions, *more than 110* is argumentatively stronger than *more than 100* by the purely semantic measure in **Eq. (17)**, but argumentatively weaker than *more than 100* by the pragmatic measure in **Eq. (25)**. A rational hearer in a world where these assumptions held should take either utterance as positive evidence for the goal S, but if they are sensitive to pragmatic considerations they should interpret *more than 100* as appreciably stronger evidence than the (very weak) *more than 110*.

Argumentative Strength for Complex Cases

In practice, we can think of complex quantitative data as inviting the speaker who summarises it to choose among a wide range of semantically true options, and even if we restrict the speaker to utterances that do not invite false pragmatic inferences, there may still be many possibilities in play. A striking example is provided by 26), which appeared as a newspaper sub-headline in 2018 on the subject of Oxford University's undergraduate admissions.

26) Figures show one in four of [sic] colleges failed to admit a single black British student each year between 2015 and 2017 (https://www.theguardian.com/education/2018/may/23/oxford-faces-anger-over-failure-to-improve-diversity-among-students, retrieved 25/03/20)

From the context (provided by the main headline) it is clear that the speaker's communicative goal here is to make the point that Oxford is failing in racial equality, as regards British students, through its admissions policy. The factual claim offered in the headline in support of this point clearly satisfies the criterion of having positive argumentative strength, by the definition in **Eq. (17)**. Moreover, although 26) does invite potential pragmatic inferences that weaken this effect (for instance, that three in four colleges succeeded in fulfilling this admissions criterion), it seems very likely that 26) also has positive argumentative strength by the pragmatic definition suggested in **Eq. (25)**.

At the same time, the utterance makes a strikingly complex quantitative claim, and it does so in a way that gives rise to several ambiguities, raising a number of potential questions in the mind of the hearer. Should the statement be interpreted as referring to the same colleges each year? Why are the years 2015–2017 focused on? Does *one in four (of) colleges* mean "a quarter of the colleges of the university" or "one out of the four colleges studied"? And is the scope ambiguity of *(they) failed [to do this] each year* to be resolved as meaning "each year, they failed to do this" or "in at least one year, they failed to do this"?[6]

We stress that, in discussing this and other examples, we do not aim to take a position on whether the speaker's argumentative goal in each specific case is ultimately supported by the data that the speaker summarises. Rather, we wish to consider how a rational hearer should adjust their belief about the speaker's argumentative goal, given the statement that the speaker chose to make on this occasion.

In the case of 26), it appears clear from the context that the speaker has a specific communicative goal in mind, and it would be reasonable to expect the speaker to choose an utterance which constitutes a good argument for that goal, when summarising the large and complex dataset under discussion. We take this to be a fairly standard argumentative context, distinguished only by the complexity of the utterance in 26), a complexity which suggests that the speaker is willing to entertain a wide variety of possible utterances with which to summarise their data. In effect, a rational hearer is entitled to note that such circumstances naturally seem to call for post hoc descriptions that involve some cherry-picking of the data. However, if a hearer believes that this kind of cherry-picking is occurring, this should make a difference to the interpretation that they place on the data that is ultimately reported, much like it does to our interpretation of post hoc statistical tests. We discuss the implications of this in the following subsection.

Rational Interpretation in an Argumentative Context

So far, we have only considered the perspective of an argumentative speaker who assumes that the hearer either

[6] The text of the full article suggests that the answers to these questions are: 26) does refer to the same colleges each year; the scope of the study being reported on was just the years 2015–2017; *one in four* means "a quarter of the colleges"; and *failed . . . each year* in fact means "in at least one year, they failed to do this". According to the article, "[t]he worst figures belonged to Corpus Christi College, which admitted a single black British student in those three years".

interprets utterances semantically (***Argumentative Strength for a Semantic Interpretation of an Utterance***) or pragmatically (***Argumentative Strength for a Pragmatic Interpretation of an Utterance***) in the usual non-argumentative manner. This is a simplifying assumption but arguably legitimate if the speaker can expect the hearer to be unaware or unsuspecting of a possibly misleading framing intention. However, we should also consider the perspective of a suspecting rational interpreter who is quite aware of the speaker's framing intentions.

So how should a rational hearer interpret an utterance made by an argumentative speaker? If the speaker merely produced a semantically truthful utterance that was drawn at random from the whole set of semantically truthful possibilities, it would appear rational for the hearer to increase their belief in G if the utterance had positive argumentative strength according to the definition in **Eq. (17)**. If the speaker produced a pragmatically felicitous truthful utterance that was drawn at random from the whole set of pragmatically felicitous truthful possibilities, it would appear rational for the hearer to increase their belief in G if the utterance had positive argumentative strength according to the definition in **Eq. (25)**. However, it would not be reasonable to suppose that an argumentative speaker should act in this way: we expect them to produce a true and felicitous statement which is selected to serve their argumentative goals. Consequently, the behaviour of a rational hearer should also be more nuanced.

If we consider the set of pragmatically felicitous and truthful utterances by which a complex data set can be summarised, post hoc, these will vary considerably in their argumentative strength. Indeed, for complex data, we might reasonably expect these utterances to range from having negative to positive argumentative strength, by either of the measures proposed above. An optimally argumentative speaker, according to such a metric, would be one who selected the utterance with the greatest positive argumentative strength with respect to their communicative goal G.

One way of characterising a rational hearer's expectation in such a case would be to assume that the speaker is optimally argumentative, taking pragmatic inference into account, and hence selects the maximally argumentatively positive utterance (of those that are true and pragmatically felicitous) according to the definition in **Eq. (25)**[7]. But the rational hearer should then not take this at face value: they should be aware that an utterance selected at random from the set of possible utterances would likely have had much less positive argumentative strength than the one that was in fact uttered.

In fact, if the speaker is argumentatively effective, the rational hearer should be interested in how likely G is under the assumption that *u* is the best thing that could be said in support of G (rather than just 'a thing that could be felicitously said in support of G'). From this perspective, when the hearer determines whether to concur with the speaker's argumentative goal G on the basis of 26), the hearer should not merely be asking whether the data presented in 26) are more compatible with a world in which Oxford's admissions policy is racist or one in which it is not. Rather, they should ask whether 26) exceeds in argumentative strength the most damning thing that could likely be asserted of Oxford's admissions policy in a world where it is not racist, and they should increase the strength of their belief in G only if that criterion is satisfied.[8]

To put it another way, if a rational hearer is aware that the speaker is trying to argue for G in an optimal way, and if *u* could likely be truthfully and felicitously asserted in a world where G was not the case (and the data under discussion reflected that G was not the case), the rational hearer should not take *u* as evidence in favour of G. Rather, as a criterion for increasing their belief in G, the rational hearer should adhere to a more stringent rule of interpretation, along the lines of 27).

27) Increase your belief in G on the basis of utterance *u* iff prag_arg_str (*u*, G) > prag_arg_str (*v*, G) for all *v* that are likely to be true and assertable given ¬G.

The point we wish to emphasise here is that, given a large dataset from a world in which G does not hold, it may well still be possible to summarise that dataset in a way that has positive argumentative strength with respect to G, according to the measures proposed in **Eqs (17** and **25)**–searching through the set of pragmatically assertable propositions that are true in the not-G world, we can find some that are (perhaps highly) suggestive of the truth of G. Given a large dataset from a world in which G does hold, an argumentatively effective speaker should be able to do better than this–they should be able to find pragmatically assertable propositions that constitute stronger evidence for G than any of those which would be available in a non-G world.

In practice, we cannot guarantee that this will be the case, because data from a not-G world may by chance be suggestive of the truth of G, just as data from a G world may by chance be suggestive of its falsity–hence the use of *likely* in 27) and the above argument. If, by chance, although G is in fact true, the data do not indicate it, then 27) predicts that no statement can be made about those data which should induce a sceptical rational hearer to increase the degree of their belief in G: we take this to be a reasonable corollary[9]

In practice, this approach appears to invite the hearer to be more sceptical than is warranted. For complex data, it is unlikely to be computationally tractable for the speaker to be able to find the argumentatively optimal utterance given their communicative

[7] Note that here we do not assume that the argumentative speaker is calibrating their choice of utterance to take into account the hearer's scepticism–although it is reasonable to think that an argumentative speaker may wish to do so. For ease of exposition we shall not attempt to address this case in this paper.

[8] Here we are assuming that the hearer is knowledgeable about which propositions are true in a world in which G is false. If the speaker takes the hearer to be less than perfectly knowledgeable, the picture becomes more complicated. We discuss this further in ***General Discussion***.

[9] A sceptical hearer might, of course, take it that even data that is extremely favourable for G might have arisen in a non-G world, just as, in the context of experimental science, even data that admit a very small *p*-value might have arisen under the null hypothesis. Consequently, they might hold that the condition in 27) is never satisfied, because any *u* might be true and assertable in a non-G world. However, beyond a point, scepticism of this kind will not be rational, in terms of leading to a correct understanding of the likely world state. Here we do not attempt to characterise the optimal degree of scepticism for the rational hearer under this idealisation.

goals. Allowing for this, an appropriate rule of interpretation for a rational speaker might instead be along the lines of 28).

28) Increase your belief in G iff prag_arg_str(u, G) > prag_arg_str(v, G) for all v that are likely assertable and accessible to the speaker given ¬G.

That is to say, the hearer should interpret an utterance as evidence for G if it has greater argumentative strength than any utterance that the speaker would, in practice, be able to produce in a world in which G was not true.

Interim Summary

The use of number in summarising data is associated with objectivity and precision, but these concepts are somewhat negotiable: number interpretation is pragmatically ambiguous in a number of ways, and the flexibility of numerical quantification makes it a particularly powerful domain in which a speaker can use language in the service of particular communicative goals that may not be shared by the hearer. If a speaker is argumentative in this sense, a rational hearer should strive to take this into account when determining whether to increase or decrease their belief in the proposition for which the speaker is ultimately arguing, based on the utterance(s) put forward in support of that proposition.

In the following section we exemplify some of these ideas with respect to a complex quantitative data set that is argumentatively described by a large number of distinct stakeholders with similar communicative goals, namely the results of REF 2014. Specifically, we will identify predictions that can be made about speaker behaviour in this context under the assumptions of the argumentative account, and examine the extent to which these are borne out.

A CASE STUDY: REPORTING THE RESEARCH EXCELLENCE FRAMEWORK

The approach outlined above allows us to make and test predictions about how speakers will use certain numerically quantified expressions in argumentative contexts. To do this, we wish to examine production data in a context in which speakers are summarising complex datasets with a clear argumentative goal in mind, and in order to evaluate the predictions we need to have access to the data as well as the speakers' productions. We would ideally be focusing on cases in which the speakers are expert users of argumentative language and are fully conversant with the details of the data they are summarising, as this is the scenario in which we expect speakers to produce argumentatively effective summaries of the data.

In all these respects, the public statements made by United Kingdom universities about their respective results in the REF 2014 assessment appear to constitute an appropriate object of study. In the following subsections, we briefly introduce the workings of the REF, consider the motivations and constraints that influence universities' public statements about the REF results, articulate a series of predictions about these statements that follow from our theory, and evaluate these predictions against the data. We will show that there are clear indications that the argumentative considerations we discuss are indeed influencing speakers' production choices; however, these productions are nevertheless suboptimal, as anticipated in the foregoing discussion, and this poses interpretative challenges for the rational hearer.

The Nature of the Research Excellence Framework 2014

REF 2014 (Research Excellence Framework) was an exercise designed to assess the quality of research in United Kingdom Higher Education Institutions. Its stated aims were to inform the allocation of research grant funds; to provide accountability for public investment in research; and to "provide benchmarking information and establish reputational yardsticks, for use within the higher education (HE) sector and for public information" (https://www.ref.ac.uk/2014/about/, retrieved 04/04/20).

For REF 2014, institutions made submissions consisting of research outputs, case studies of impact derived from research, and information about the research environment. These submissions were evaluated by 36 appointed sub-panels and awarded one of five possible grades, ranging from 4* to U/C (unclassified). In the case of research outputs, these grades corresponded to quality that was "world-leading", "internationally excellent", "recognised internationally", "recognised nationally", and which "falls below the standard of nationally recognised work" respectively (https://www.ref.ac.uk/2014/panels/assessmentcriteriaandleveldefinitions/, retrieved 04/04/20).

Institutions typically submitted to multiple sub-panels, and these distinct submissions were evaluated separately. In all, REF 2014 evaluated 1911 submissions from 154 different institutions: these submissions comprised 191,150 research outputs and 6975 impact case studies (and represented work by 52,061 academic staff).

The overall quality profile for each submission comprised a weighted average of the grades for outputs (65%), impact (20%) and environment (15%). Across all submissions, 30% were graded 4*, 46% 3*, 20% 2*, 3% 1* and 1% unclassified. However, this varied appreciably across the three sub-profile measures: only 22% of outputs achieved the 4* rating, whereas 44% of impact submissions and 45% of environment submissions did so.

When the REF 2014 results were published (December 18, 2014), several media outlets compiled 'league tables', perhaps the most influential being Times Higher Education (THE), who provided three rankings:

- **Grade point average (GPA).** 4 points were awarded for 4* grades, 3 points for 3*, and so on. The overall GPA measure for an institution was the weighted mean of the GPA for its individual panel submissions (weighted by the number of full-time equivalent (FTE) staff whose work was submitted to each panel).
- **Research power.** This was computed by multiplying the GPA by the number of FTE staff submitted by the institution.

- **Research intensity.** This was computed by multiplying the GPA by the proportion of REF-eligible staff whose work was submitted by the institution. The ranking based on this was published subsequently to the other two rankings.

The THE main league tables included only multi-subject institutions (those which submitted to more than one panel), with single-subject institutions listed separately; we focus on multi-subject institutions in what follows.

To exemplify the methodology, consider the results from the Institute for Cancer Research (ranked first on GPA), which submitted to two sub-panels, namely Clinical Medicine and Biological Sciences. Its submission for Clinical Medicine comprised 69 FTE staff and achieved a GPA of 3.33 (which itself was comprised of scores of 3.09 for outputs, 3.90 for impact and 3.63 for environment), while that for Biological Sciences comprised 34 FTE staff and achieved a GPA of 3.55 (3.44 outputs, 3.80 impact, 3.75 environment). The overall weighted mean GPA was 3.40, which, multiplied by 103 FTE staff, yielded a power score of 351. The Institute for Cancer Research had 108 FTE REF-eligible staff, so its research intensity measure was calculated by multiplying its overall GPA by 103/108: the resulting intensity-weighted GPA was 3.25, on which measure it again ranked first.

Additional statistics were computed by Research Fortnight (RF) and published by the Guardian: these prioritised research power, but added one further measure:

- **Research quality.** This was calculated as the proportion of 4* research plus one-third of the proportion of 3* research, based on the overall quality profile[10]. As an example, the Institute for Cancer Research achieved 50% 4* and 41.7% 3* outputs, and hence a quality index score of 63.9 (= 50 + (41.7/3)).

The average GPA scores for the whole REF were 3.01 for outputs, 3.24 for impact, and 3.28 for environment. This represented an appreciable increase in scores from the previous assessment, the 2008 Research Assessment Exercise (RAE). Although the official REF results did not report GPA, they noted an increase in the percentage of outputs judged world-leading (22% against 14%) and internationally excellent (50% against 37%). The official summary further noted that "three-quarters of the universities had at least 10% of their submitted work graded as world-leading (4*). The top quarter had at least 30% graded as world-leading (4*)" (REF Brief Guide 2014, https://www.ref.ac.uk/2014/media/ref/content/pub/REF%20Brief%20Guide%202014.pdf, retrieved 30/01/21).

Reporting the Research Excellence Framework

Many institutions issued press releases summarising their results, in keeping with the REF's stated goal to "establish reputational yardsticks". However, the REF team did not articulate an official line as to how the results should be interpreted as evidence of reputational strength. Consequently, institutions were largely free to interpret and present the results as they saw fit. This therefore represents a case in which expert communicators (the institutional press officers), with full access to a complex dataset, have the opportunity to select what information to present and how to present it, in the service of a clearly motivated argumentative agenda (advancing the perceived research reputation of their institution).

Against this, of course, it might be argued that–again in the absence of national policy as to what should be considered prima facie evidence of reputational strength–institutions were free to pursue different objectives, and their reportage of the results might merely reflect that. For example, if an institution had pursued a strategy of boosting research power at the expense of GPA, and this was successful, it would be reasonable for them to present research power data as evidence of their success. Thus, we cannot exclude an optimistic interpretation under which the selective reporting of results actually corresponds to the prior goals of the institutions. Even so, such reporting could mislead the (non-sceptical) hearer, who might interpret a press release focusing only on one metric as evidence that the institution in question could–if challenged–offer similarly strong evidence of its high reputation across a broader range of metrics, whereas this might in fact not be the case.

Hypotheses

Our overarching question is whether institutions use argumentatively effective strategies in the way our theoretical account predicts, when selectively reporting REF outcome data. From the rational hearer's point of view, the corresponding question is whether it is necessary to take the institutions' likely argumentative agenda into account when interpreting the data that they present. Here we aim to unpack this into specific testable predictions concerning how speakers will act under the assumption that they are argumentatively effective, judged by the standard that we proposed in ***Quantifying Argumentative Strength, and Allowing for Uncooperativity***. That is to say, we aim to test whether the speakers in this study–the authors of the institutional reports about their REF results–are optimising the argumentative strength of their utterances.

Firstly, we expect argumentatively effective speakers to avoid presenting information that gives rise to inferences that run counter to their communicative goals. One potential source of such information is quantity implicature. We discussed how numerical expressions of the form *top M* might give rise to implicatures of this kind: not only do they convey that *top O* is not the case for salient $O < M$, but, particularly in the case of non-round *M*, they potentially convey that *top M-1* is not the case. Consequently, we expect argumentatively effective speakers to use *top M* formulations only when they can do so while avoiding argumentatively disadvantageous quantity implicatures.

Secondly, we expect argumentatively effective speakers to avoid presenting contextual information when doing so would promote inferences that run counter to their communicative

[10] A motivation for the use of this measure was the expectation that funding allocations would be based on the proportions of 4* and 3* research, with 4* weighted three times as heavily as 3* research.

goals. In the REF context, multiple rankings are available for discussion, most notably the GPA and power rankings, and this is evident to the speaker but not necessarily evident to the hearer. A rational hearer, aware of the existence of multiple rankings, might expect the speaker to quote the most favourable one and could infer that other unmentioned rankings were less favourable to that institution. We might therefore expect an argumentatively effective speaker to avoid indicating to the hearer that multiple rankings exist, in order to preserve the hearer's ignorance on this point and thus prevent the hearer from drawing an unfavourable inference.

Thirdly, we expect argumentatively effective speakers to avoid presenting information that fails to support their communicative goal more clearly than it supports the negation of that goal. In the context of the REF, we assume that the press releases issued are intended to bolster the reputation of the institution in question with reference to its competitors. Presenting statements in support of the institution's quality that would also be true of its competitors would therefore be an ineffective strategy in terms of argumentative force. Moreover, in a sceptical hearer (of the kind discussed in ***Rational Interpretation in an Argumentative Context***), it would invite the inference that nothing more favourable could be said about the institution in question than that which could be said of its competitors. Thus, such statements would be ineffective (given a rational hearer unaware of the speaker's argumentative agenda) or actively counterproductive (given a sceptical hearer who takes the speaker's argumentative agenda into account), when considered as arguments for the institution's quality. Note that we assume, in making this prediction, that the speaker takes the hearer to be knowledgeable as regards what could be truthfully said of the institution's competitors: we return to the implications of this assumption in ***General Discussion***.

Hence, in summary, we make the following predictions about the reporting of REF results:

H1: Speakers will use argumentatively appropriate reference points: an institution will be described as "top *M*" only if its ranking is near *M*, and speakers will avoid using non-round *M*.

H2: Speakers will prioritise favourable rankings and suppress unfavourable rankings: if the GPA and power rankings differ in how highly they place an institution, the more favourable ranking will be reported and the report will not convey the existence of an alternative ranking scheme.

H3: Speakers will avoid argumentatively unhelpful statements: they should not attempt to argue for the reputational strength of their institution on the basis of statements that would also be true of lower-ranked institutions.

Procedure

We collated data from the top 40 institutions, according to the GPA rankings, focusing in each case on descriptions of institution-wide accomplishments rather than those of individual faculties or departments. We first searched for press releases that had been issued at an institutional level on December 18, 2014 in connection with REF 2014 results, as archived on institutions' websites: these were available for 29 of the 40 institutions. Where these were not available we looked for summary pages detailing REF 2014 results as part of the institutions' general profiles: these were available for 10 of the remaining 11 institutions. In this way we obtained information from all institutions in the top 40 except the London School of Economics and Political Science (ranked third by GPA), which is hence excluded from the following analysis.

Results

H1: Use Best Available Reference Points

We predicted that expressions such as *top M*, used argumentatively, will be uttered only in connection with institutions that are ranked just above the relevant threshold, and only with round *n*, in order to avoid argumentatively unfavourable implicatures.

29)–38) represent all the uses of *top M* in the REF reports we examined that make reference to the overall institutional ranking. We indicate in square brackets the precise ranking that these quotes allude to.

29) Cardiff in top five for research excellence . . . The quality and impact of Cardiff's research has led to a meteoric rise in league tables, pushing it into the UK's top 5 universities [5th]
30) [King's College London is] Top 10 nationally for research 'power' and 'quality' [6th, 7th].
31) Warwick repeats top 10 success in UK research ranking exercise. [8th equal]
32) The [London School of Hygiene and Tropical Medicine] is ranked in the top 10 of all universities in the UK. [10th]
33) The results demonstrate excellence across research, putting Sheffield in the top 10 per cent of all UK universities. [16th equal = 10th percentile[11]].
34) University of Leeds in top 10 for research and impact power [10th]
35) Royal Holloway is within the top 25 per cent of UK universities for research rated 'world-leading' or 'internationally excellent' [26th equal on unweighted measure = 21st percentile]
36) Swansea research breaks into UK top 30. [26th equal]
37) Essex has re-confirmed its position as one of the UK's top 20 research universities. [20th]
38) [Strathclyde is] Top 20 in the UK for Research Intensity. [18th]

As predicted, each of these descriptions uses round values of *M* in the *top M* formulation, and in each case no comparably salient $O < M$ exists for which the *top O* claim would be true. Hence we can see these examples as demonstrating a preference on the part

[11] Among multi-subject institutions, Sheffield ranks equal 14th out of 128 on the GPA measure; including single-subject institutions, it ranks equal 16th out of 154, hence on the cusp of the top decile. We assume this is the metric that the authors of the press release have in mind.

of the speaker to choose *top M* descriptions that are argumentatively effective, by the measures we discuss.

There are also indications in these data that the possibility of describing the institution as *top M* for some relatively small round value of *M* has motivated the choice of ranking criteria. Essex, in 37), and Strathclyde, in 38), both appeal to the research intensity measure, on which they are ranked considerably higher than on either of the measures initially published. Strikingly, Essex places 22nd on this measure, but 20th among universities-that is to say 37) is true if we do not consider the Institute for Cancer Research or the London School of Hygiene and Tropical Medicine to be universities (notwithstanding 32)). Similarly, Cardiff places 6th on research excellence as measured by GPA, but improves to 5th if we exclude the Institute for Cancer Research from consideration. Thus, their rhetorical move of focusing on universities may be motivated by the argumentative advantage of being able to make the *top five* claim rather than merely *top six*, which is semantically weaker but also gives rise to an argumentatively disadvantageous implicature *exactly 6th*.

Other uses of *top M* in these data involve generalisations over faculties or subject areas, and are sometimes combined with appeal to non-obvious ranking choices, as for example in 39) and (perhaps most extremely) 40). However, as we are restricting our attention here to descriptions of the institutions as a whole, we will not discuss these cases further, other than to note that they represent an alternative way to present the data for argumentative effect.

39) On actual research outputs 19 of Warwick's departments were ranked in the top 10 in the UK.
40) More than 25 per cent of the Durham University subjects entered for REF 2014 were in the top 5 subjects [sic] nationally for grade point average (overall score).

H2: Prioritise Favourable Rankings, Suppress Unfavourable Rankings

Taking the GPA and power rankings to be the most salient, we hypothesise that institutions will prefer to report the measure on which they rank more highly, as this constitutes better evidence of their high reputation. We also hypothesise that institutions will decline to mention the existence of the alternative measure, as this would invite inferences about their relative performance on that measure that would be detrimental to their reputational claim.

Of the 39 institutions for which we have data, 19 are ranked higher on GPA than power and 19 are ranked higher on power than GPA (the University of Durham places 20th on both rankings). Of the former group, nine mention GPA in their report and none mention power (a significant difference: $p < 0.01$, sign test), while ten do not make explicit reference to either measure. Of the latter group, 11 prioritise reference to power over GPA (eight of which do not mention the GPA measure at all) and two prioritise reference to GPA and do not mention power (again a significant difference: $p < 0.05$, sign test), while six do not make explicit reference to either measure. There is thus a significant interaction ($p < 0.001$, Fisher's exact test), showing a clear preference for institutions to prefer the measure on which they rank higher and most commonly not to acknowledge the existence of the less favourable measure.

Outside of these two major statistics, the most popular measure for first mention was the combined proportion of research attaining a particular quality threshold, which was cited first by 14 institutions. 11 institutions focused on their proportion of 4* and 3* research: of these 11, 8 rank more highly on this measure than on either GPA or power. However, although this is compatible with a view in which the choice of this measure has been generally motivated by the wish to report a high ranking, in fact only two of these institutions comment on their rankings by this measure: Royal Holloway, in 40), which makes a claim that it could also make with reference to the GPA measure, and Queen Mary University of London, in 41), although the data from the summary table appears to place it 8th on this measure.

41) Royal Holloway is within the top 25 per cent of UK universities for research rated 'world-leading' or 'internationally excellent'.
42) Overall QMUL is ranked 5th in the UK [among multi-faculty institutions] for the percentage of its 3* and 4* research outputs.

Alongside the reference to the combined proportion of 4* and 3* research, two of these 11 institutions also make reference to GPA, one to power, and eight to neither. Thus, the general pattern is once again one in which institutions do not acknowledge the existence of alternative rankings which would describe them less favourably.

As we discussed earlier, the extent to which institutions acknowledge alternative measures could reasonably be expected to bear heavily on hearers' interpretations of the information provided. 43), from King's College London, represents a particularly transparent presentation of the alternative measures (the 'quality' measure here referring to GPA): the institution's preferred measure is complemented immediately by reference to the salient alternative. 44), from the University of East Anglia (UEA), is somewhat more opaque in this respect: the institution's preferred measure (focusing wholly on outputs, rather than the combined measure) is not one that is usually tabulated in its own right, and neither the overall GPA nor the power rating are alluded to in the following text. The hearer of 44) might reasonably be surprised to find UEA ranked 23rd by the THE for research quality.

43) King's has risen to 6th position nationally in the 'power' ranking–up from 11th in the Research Assessment Exercise (RAE) 2008. 'Power' takes into account both the quality and the quantity of research activity. King's has also risen to 7th position for quality–up from 22nd in 2008.
44) UEA is 10th in the UK for quality of research outputs. Over 82% of UEA research is rated as 'world-leading' or 'internationally excellent'.

H3: Avoid Argumentatively Unhelpful Statements

Our third prediction was that speakers would tend to avoid arguing for their institutions' reputational strength on the basis of statements that could also be truthfully asserted of lower-ranked institutions, on the basis that such statements would be argumentatively at best ineffective and at worst (given a sceptical hearer) counterproductive. However, there are a striking number of apparent counterexamples to this among the data, as exemplified by 45)–50), which include several article headlines.

45) The REF 2014 showed that the vast majority of Newcastle University's research was placed in the top two categories of 4*(world leading) or 3* (internationally excellent).
46) The University of Nottingham is a leading international institution carrying out world-class research, according to the Research Excellence Framework (REF) 2014.
47) [The University of Exeter has] world-leading research in all the units we submitted to REF. . .
48) Liverpool research ranked in UK top 10.
49) REF highlights world leading research at Aston.
50) Research at the University of Dundee has been ranked among the very best in the United Kingdom.
51) University of Sussex research is 'world-leading', major review finds.

Of these examples, 45) makes a quantitative claim, but the strength of this depends on the interpretation of *the vast majority*. The relevant figure in this case is 79%, which places Newcastle 34th on this metric. Were the claim merely *a majority*, Newcastle would share this distinction with all the top 88 institutions in the GPA ranking; if we interpret the threshold for *vast majority* at, for instance, 66%, then 63 institutions still meet this criterion. We note that the rest of the press release does not encourage the reader to contextualise the claim in this way, and does not present any information that would be helpful to them in doing so.

The headline 48) also makes a quantitative claim, but this turns out, on closer inspection, to be existential in character: the body text clarifies that seven subjects at Liverpool were ranked in the top 10 nationally (by the measure of "research excellence"). As there are 36 sub-panels in play, and given the possibility of appealing to multiple distinct measures, the claim of having "research ranked in UK top 10" is argumentatively a relatively weak one, although it is impossible to verify precisely how weak without detailed examination of the overall distribution of outcomes by sub-panel.

The subsequent examples here all focus on the existence of world-leading research at the respective institutions. In the context of the REF results, this is a surprisingly weak claim from an argumentative perspective. As noted earlier, three-quarters of the universities submitting to REF 2014 had at least 10% of their work graded as 4*. Indeed, only 72 of the 1911 submissions failed to have any work at all graded at 4*, so the claim made by Exeter as 47) is one that could be made by the majority of institutions submitting to REF, while the existential claims of 49)–51) could be made by 151 of the 154 institutions. Thus, to the extent that these claims are to be understood as arguments to the effect that Aston, Dundee and Sussex are above-average institutions (which they are, according to the GPA measure), they appear to have very little argumentative strength, according to the measures proposed in this paper.

GENERAL DISCUSSION

Individual institutions' reporting of the results of REF 2014 represents a scenario in which speakers can be expected to summarise complex data in an argumentatively effective way, in the service of a generally clear communicative goal, namely to emphasise the high quality of the institution's research. Based on the approach to argumentation discussed in this paper, we were able to articulate three predictions as to how speakers would behave in this case. Two of these were borne out. Given a choice of rankings to report, institutions have broadly behaved in accordance with a strategy of selecting the ranking that is most favourable, and presenting little information to hint at the existence of other, less favourable, data. This would accord with a strategy of presenting argumentatively strong information while dissuading the hearer from drawing ad hoc inferences that undermine its argumentative point. The use of the formulation *top M* also adheres to the predicted principles: the formulation is used only when the precise ranking is close to *M* and *M* is a round number. Again, the effect is not to invite the hearer to draw inferences that would be deleterious to the argument being advanced.

Speaker behaviour in this case study, however, deviated strikingly from our third prediction: argumentatively weak information was frequently presented, as seen in 45)–51), where assertions are made that could equally truthfully be made of institutions which had performed much less well. This represents a challenge for the explanatory utility of the approach we suggest–how can we explain this choice of communicative strategy?

Recall that in ***Quantifying Argumentative Strength, and Allowing for Uncooperativity*** we raised the question of how sceptical a rational hearer should be about the use of simple descriptions of complex data, when evaluating the argumentative strength of these descriptions and using that to update beliefs. A minimally sceptical approach would be to increase one's belief in some proposition G given an utterance *u* if the probability that u is true given G exceeds the probability that *u* is true given the negation of G (and to decrease one's belief in G if the reverse is true). For example, if we consider 47) as *u* and assume G is the proposition that Exeter is an outstanding research university, this condition is clearly satisfied, and we should increase our belief in G on hearing 47). A maximally sceptical approach would be to increase one's belief in G given *u* only if *u* is argumentatively better than any of the things that could be said given that G were false (and to decrease one's belief in G otherwise). In this case, taking the same values of *u* and G as before, this condition is not satisfied: 47) would likely be true even if Exeter were not an outstanding research university (and its REF results reflected that), so we should decrease our belief in G given 47).

In practice, 47) illustrates an intermediate case: it represents a relatively weak argumentative claim, but this could be for distinct reasons. One possibility is that the speaker of 47) thinks that the hearer will not be sceptical in the way suggested by the above account in how they update their beliefs, and therefore expects this argument to convey positive argumentative strength: we could think of this as the speaker being optimistic about the receptiveness of the audience to their argument. Another possibility is that the speaker has simply not considered that 47) is an objectively weak argument, given that it is something that tens of institutions ranked below Exeter on the standard metrics could also say[12]. In this case, we could regard the speaker as being incompetent at maximising argumentative strength–and, to the extent that speakers behave this way, we could conclude that the model is inadequate for capturing speaker behaviour.

It is also worth considering a third possibility. Perhaps the speaker of 47) thinks that the hearer is not aware that this assertion would also be true for lower-ranked institutions, and consequently believes that the hearer will perceive the utterance to have positive argumentative strength, even if the speaker knows this not to be the case. This is somewhat analogous to the case of Hypothesis 2, in which the speaker exploits the hearer's ignorance about alternative measures: it is reasonable to expect the hearer to be less than fully informed about the REF results for competing institutions, and this would license the speaker to exploit the argumentative potential of utterances that would not be predicted to be argumentatively effective with fully knowledgeable hearers[13]. In general, we feel that this is a plausible explanation for argumentative speakers' divergence from the theoretically optimal strategy. However, in order to evaluate this explanation empirically, we would need to establish the hearers' knowledgeability (and specifically how this is perceived by the speakers), which cannot be read off the data we examine in this paper.

In summary, then, the picture presented by the REF reports is (perhaps characteristically) mixed. The authors of these reports are, collectively, not entirely consistent in maximising argument strength, by the measures proposed in ***Quantifying Argumentative Strength, and Allowing for Uncooperativity***. However, at the same time, they are clearly not neutral in their treatment of the data. Consequently, these texts place considerable demands on the rational hearer who wishes to interpret the claims being made. Given a hearer who accepts their reports at face value, perhaps 20 or more institutions might be able to convince that hearer that they belong in the top 10; however, given a maximally sceptical hearer, perhaps only about 10 institutions might be able to convince that hearer that they belong in the top 20.

Thus, as far as these press releases are concerned, the hearer cannot arrive at any close approximation of an objectively accurate interpretation of the results by adopting any of the first three strategies canvassed earlier in this paper. Adopting the straightforward semantic or pragmatic approaches to argumentative strength, the hearer will generally infer that the universities' research has been evaluated more favourably by REF than is in fact the case. Adopting the more demanding stance of expecting the best possible descriptions, the hearer will overcompensate and infer that the evaluations are in fact worse, in most instances, than was actually the case. To decipher the descriptions accurately, from the standpoint of argumentative strength, the hearer has to be aware that the speakers are systematically making efforts in the direction of maximising argumentative strength, but also that they are inconsistent in how effectively they achieve this.

These data exemplify a much more widespread problem, concerning both how complex information should be summarised in order not to mislead the hearer, and how the hearer should interpret summary information in order to reconstruct the best possible approximation to the underlying reality. The problem is clearly accentuated when a speaker has a particular argumentative agenda, even when they are determined to advance that agenda only through the presentation of true and accurate (albeit carefully selected) facts. It is perhaps rather unfortunate, although not entirely surprising, that this challenge is so strongly in evidence in the context of the reporting of REF 2014 results, in which some of the United Kingdom's most esteemed institutions participate in an exercise designed to determine their "reputational strength".

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

CC collected data; CC and MF analysed the data and co-wrote the paper.

FUNDING

The authors gratefully acknowledge financial support via the project "Probabilistic models of content choice in manipulative language use" funded by the Lower Saxony Ministry of Science and Culture (15-50115-Schottland-13/19).

[12] This is rather analogous to the problem of chance capitalisation in statistical analysis: the speaker may not have considered that 47) is likely to be true irrespective of whether or not G holds.

[13] Note that a cooperative speaker might equally be aware that the hearer's understanding of the situation is limited and that the hearer may miscalculate the argumentative strength of utterances as a result. For example, addressed to a non-specialised audience, *Daniel Kahneman has won the Nobel Prize* would constitute an effective argument as to that person's academic credentials, but *Tim Gowers has won the Fields Medal* might not, and the speaker might need to add more context to explain how the argumentative strength of this should be evaluated.

REFERENCES

Anscombre, J.-C., and Ducrot, O. (1983). *L'argumentation dans la langue*. Brussels: Mardaga.

Ariel, M. (2004). Most. *Language*. 80 (4), 658–706. doi:10.1353/lan.2004.0162

Bryson, B. (1998). *Notes from a Big Country*. New York: Doubleday.

Carston, R. (1998). "Informativeness, Relevance and Scalar Implicature," in *Relevance Theory: Applications and Implications*. Editors R. Carston and S. Uchida (Amsterdam: Benjamins), 179–236. doi:10.1075/pbns.37.11car

Cummins, C., Sauerland, U., and Solt, S. (2012). Granularity and Scalar Implicature in Numerical Expressions. *Linguist Philos.* 35, 135–169. doi:10.1007/s10988-012-9114-0

de Jaegher, K., and van Rooij, R. (2014). Game-theoretic Pragmatics under Conflicting and Common Interests. *Erkenntnis* 79, 769–820. doi:10.1007/s10670-013-9465-0

Franke, M., de Jager, T., and van Rooij, R. (2012). Relevance in Cooperation and Conflict. *J. Logic Comput.* 22 (1), 23–54. doi:10.1093/logcom/exp070

Franke, M., Dulcinati, G., and Pouscoulous, N. (2020). Strategies of Deception: Under-Informativity, Uninformativity, and Lies-Misleading with Different Kinds of Implicature. *Top. Cogn. Sci.* 12, 583–607. doi:10.1111/tops.12456

Good, I. J. (1950). *Probability and the Weighing of Evidence*. London: Griffin.

Hesse, C., and Benz, A. (2020). Scalar Bounds and Expected Values of Comparatively Modified Numerals. *J. Mem. Lang.* 111, 104068. doi:10.1016/j.jml.2019.104068

Horn, L. R. (1972). On the Semantic Properties of Logical Operators in English. Ph.D. thesis. Los Angeles: University of CaliforniaDistributed by Indiana University Linguistics Club.

Krifka, M. (2009). "Approximate Interpretations of Number Words: A Case for Strategic Communication," in *Theory and Evidence in Semantics*. Editors E. Hinrichs and J. Nerbonne (Stanford, CA: CSLI Publications), 109–132.

Lasersohn, P. (1999). Pragmatic Halos. *Language*. 75, 522–551. doi:10.2307/417059

Meibauer, J. (2014). *Lying at the Semantics-Pragmatics Interface*. Berlin: Mouton de Gruyter.

Merin, A. (1999). "Information, Relevance, and Social Decision-Making: Some Principles and Results of Decision-Theoretic Semantics," in *Logic, Language, and Computation*. Editors L. S. Moss, J. Ginzburg, and M. de Rijke (Stanford, CA: CSLI Publications), Vol. 2, 179–221.

Pinker, S., Nowak, M. A., and Lee, J. J. (2008). The Logic of Indirect Speech. *Proc. Natl. Acad. Sci. U.S.A.* 105 (3), 883–888. doi:10.1073/pnas.0707192105

Solt, S. (2014). "An Alternative Theory of Imprecision," in *Proceedings of the 24th Semantics and Linguistic Theory Conference (SALT 24)*. Editors T. Snider, S. D'Antonio, and M. Weigand (Washington DC: Linguistics Society of America), 514–533.

Sperber, D., and Wilson, D. (1986). *Relevance: Communication and Cognition* (2nd Edn 1995). , Harvard University Press, Cambridge, MA: Harvard University Press.

Van der Henst, J.-B., Carles, L., and Sperber, D. (2002). Truthfulness and Relevance in Telling the Time. *Mind Lang.* 17, 457–466. doi:10.1111/1468-0017.00207

Winterstein, G. (2012). What *But*-Sentences Argue for: An Argumentative Analysis of *but*. *Lingua* 122 (5), 1864–1885. doi:10.1016/j.lingua.2012.09.014

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: 26 May 2021
doi: 10.3389/fpsyg.2021.661087

The Role of UID for the Usage of Verb Phrase Ellipsis: Psycholinguistic Evidence From Length and Context Effects

Lisa Schäfer[1,2]*, *Robin Lemke*[1,2], *Heiner Drenhaus*[1,3] *and Ingo Reich*[1,2]

[1] Collaborative Research Center 1102, Saarland University, Saarbrücken, Germany, [2] Department of Modern German Linguistics, Saarland University, Saarbrücken, Germany, [3] Department of Language Science and Technology, Saarland University, Saarbrücken, Germany

We investigate the underexplored question of when speakers make use of the omission phenomenon verb phrase ellipsis (VPE) in English given that the full form is also available to them. We base the interpretation of our results on the well-established information-theoretic Uniform Information Density (UID) hypothesis: Speakers tend to distribute processing effort uniformly across utterances and avoid regions of low information by omitting redundant material through, e.g., VPE. We investigate the length of the omittable VP and its predictability in context as sources of redundancy which lead to larger or deeper regions of low information and an increased pressure to use ellipsis. We use both naturalness rating and self-paced reading studies in order to link naturalness patterns to potential processing difficulties. For the length effects our rating and reading results support a UID account. Surprisingly, we do not find an effect of the context on the naturalness and the processing of VPE. We suggest that our manipulation might have been too weak or not effective to evidence such an effect.

Keywords: ellipsis, VP ellipsis, information theory, uniform information density, rating study, self-paced reading study

OPEN ACCESS

Edited by:
Hannah Rohde,
University of Edinburgh,
United Kingdom

Reviewed by:
Dan Parker,
College of William & Mary,
United States
Jason Merchant,
University of Chicago, United States
Christina S. Kim,
University of Kent, United Kingdom

****Correspondence:***
Lisa Schäfer
lisa.schaefer@uni-saarland.de

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

Received: *30 January 2021*
Accepted: *15 April 2021*
Published: *26 May 2021*

Citation:
Schäfer L, Lemke R, Drenhaus H and Reich I (2021) The Role of UID for the Usage of Verb Phrase Ellipsis: Psycholinguistic Evidence From Length and Context Effects. Front. Psychol. 12:661087. doi: 10.3389/fpsyg.2021.661087

1. INTRODUCTION

When speakers want to get a message across, they often have the choice between ellipsis and the corresponding full form (1) and it is not always obvious which form to use. The underexplored question of why speakers sometimes prefer the ellipsis over the full form and sometimes do not is the topic of this paper, which we explore at the example of VP ellipsis.

VP ellipsis (Sag, 1976; Williams, 1977) is one of the most extensively studied omission phenomena in linguistics. The term refers to a kind of constituent ellipsis where the omitted element, i.e., the target of ellipsis, is a complete verb phrase. Only a corresponding auxiliary is left in the position of the omitted verb phrase (1).

(1) Sam played football
  a. and Dean played football too.
  b. and Dean did ⟨play football⟩ too.
  c. and Dean should ⟨play football⟩ too.

The extensive literature on this phenomenon has focused on systemic questions like the modeling of the ellipsis site, the relation between the ellipsis site and its antecedent (or postcedent) and the licensing conditions of VP ellipsis (see e.g., Merchant, 2018; Reich, 2018, for recent overviews). Analogously, the psycholinguistic literature mainly addressed procedural aspects of the relation between antecedent and target such as complexity effects (see e.g., Frazier et al., 2000; Frazier and Clifton, 2001; Apel et al., 2007; Martin and McElree, 2008; Paape et al., 2017). However, to the best of our knowledge, the question of when and why speakers actually make use of VP ellipsis given that the corresponding full form is also available to them has not yet been investigated in the literature.

We pursue the hypothesis that VP ellipsis is preferred more strongly the more redundant the omitted material is, because this makes the most efficient use of the hearer's processing resources[1]. We base our account on the well-established information-theoretic Uniform Information Density (UID) hypothesis (Levy and Jaeger, 2007). According to UID, speakers tend to distribute information uniformly across utterances avoiding information minima caused by redundant material. We focus on two sources of redundancy that could impact the preference for VP ellipsis: the length of the redundant VP which leads to a longer redundant region and its predictability in context which causes a deeper redundant region. To test the predictions of UID with respect to length and predictability in context we first manipulate either the length of the redundant VP or its predictability in context and determine the naturalness of VP ellipsis in comparison to the corresponding full form. Second, we focus on the full forms and use a self-paced reading experiment to measure the processing effort associated with the redundant VP. This allows us to correlate differences in naturalness with potential processing difficulties caused by information minima.

This paper is structured as follows: In section 2, we present our information-theoretic account to the usage of VP ellipsis based on UID and discuss its predictions with respect to length and context effects. In section 3, we discuss length effects and present a naturalness rating study and a self-paced reading study on length effects. Section 4 is dedicated to effects of predictability in context and presents a pre-test, a rating study and a self-paced reading experiment. Section 5 summarizes our central findings and contributions.

2. INFORMATION-THEORETIC ACCOUNT TO VP ELLIPSIS

The Uniform Information Density (UID) hypothesis (Levy and Jaeger, 2007) has been successfully applied to account for a variety of omission phenomena from acoustic reduction (Aylett and Turk, 2004; see Jaeger and Buz, 2017 for an overview), to the omission of functional elements such as relativizers (Levy and Jaeger, 2007), complementizers (Jaeger, 2010) and discourse markers (Asr and Demberg, 2015) in English, case markers in Japanese (Kurumada and Jaeger, 2015) and articles in German newspaper articles (Lemke et al., 2017), to the omission of content words, for instance the deletion of parts of the utterance in German fragments (Lemke et al., 2020) and the omission of preverbal subjects in Russian (Kravtchenko, 2014). In a recent study, Lemke et al.[2] found that UID also constrains other elliptical phenomena such as sluicing. This makes UID a promising approach for describing the omission process of VP ellipsis where the ellipsis targets a whole VP with both function and content words.

In the information theoretic framework, the *information* of an expression is defined as the negative binary logarithm of its conditional probability given context, i.e., $-log_2\ p(word|context)$ (Shannon, 1948). Psycholinguistic research has established the synonymous term *surprisal* and has shown that information or surprisal indexes processing effort (Hale, 2001; Demberg and Keller, 2008; Levy, 2008). The central idea of the UID hypothesis is that communication is successful when surprisal or processing effort is distributed as uniformly as possible across an utterance. Such a uniform distribution avoids suprisal minima (*troughs*) and maxima above channel capacity (*peaks*) in the information density profile, i.e., it prevents that the processing capacities of the hearer are underutilized or exceeded. As a consequence, there are two ways in which an utterance can be optimized with respect to UID: First, speakers can omit predictable words which have low surprisal and would cause troughs in the information density profile. Second, speakers can smooth peaks by inserting a word before a very unpredictable word that is hard to process. If this insertion increases the predictability of the word that is hard to process, this reduces the processing effort on this word. With respect to VP ellipsis, the important point is the fact that surprisal minima are caused by redundant material. In full forms like (1-a), the repeated VP *played football* is redundant and we would in principle expect that a repetition of redundant material causes a surprisal minimum in the information density profile. In contrast, the ellipsis in (1-b) avoids such a minimum and thus smooths the information density profile. This results in a more uniform distribution and a more efficient use of the hearer's processing resources. This idea is illustrated in **Figure 1**[3] using hypothetical surprisal values for example(1).

We investigate two potential sources of redundancy: the length of a VP and its predictability in context. Firstly, following UID we expect that the redundancy of a VP increases as a function of its length: Longer repeated VPs create longer regions of low information in the information density profile as shown in **Figure 2**. In this example the repeated VP is longer and hence causes a longer trough in the information density profile. Such longer regions make the utterance less efficient and we expect the pressure put on the speaker to omit the redundant part and to use VP ellipsis to be stronger in this case.

Secondly, in line with UID also the predictability of the VP in context should impact its redundancy. Hence, exactly the same VP should create a deeper trough in the information density profile when it occurs in a predictive context compared to a neutral context. When the example in (1) is uttered in a predictive context like (2-a) compared to a neutral context like (2-b), the repeated VP *played football* becomes even more redundant because the context makes Dean more likely to play

[1] We use the term "hearer" to refer to the recipient of the communication, regardless of whether this communication is auditory or written.

[2] Lemke, R., Schäfer, L., and Reich, L. (under review). *Can identity conditions on ellipsis be explained by processing principles?*

[3] All figures in this paper were created with the package ggplot2 (Wickham, 2016) in R (R Core Team, 2020).

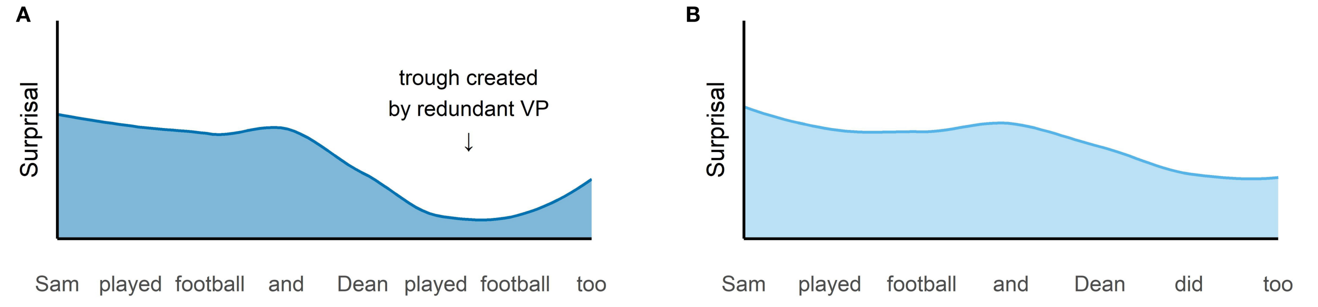


FIGURE 1 | Hypothetical information density profiles for example (1): The surprisal values for the words of the full form **(A)** and for the words of the ellipsis **(B)** are plotted.

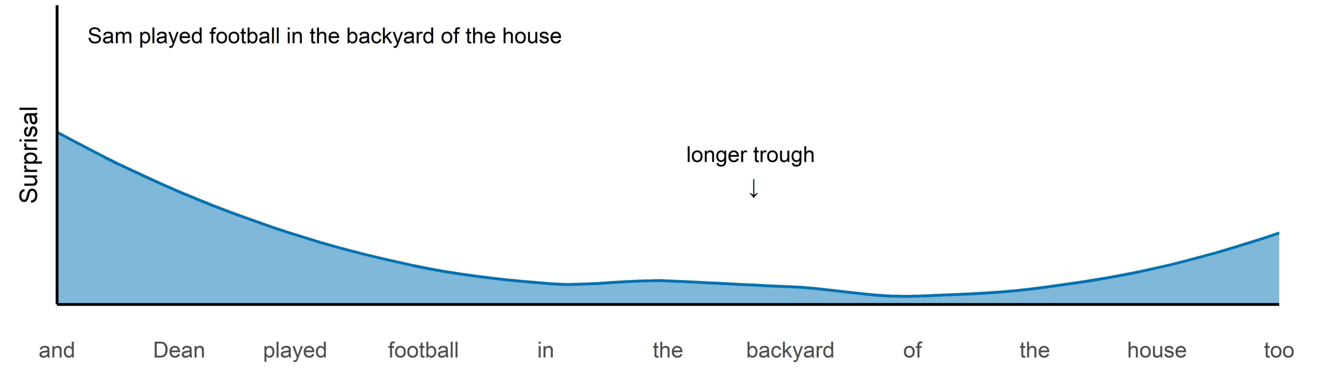


FIGURE 2 | Hypothetical information density profile for the second conjunct of a longer version of example (1).

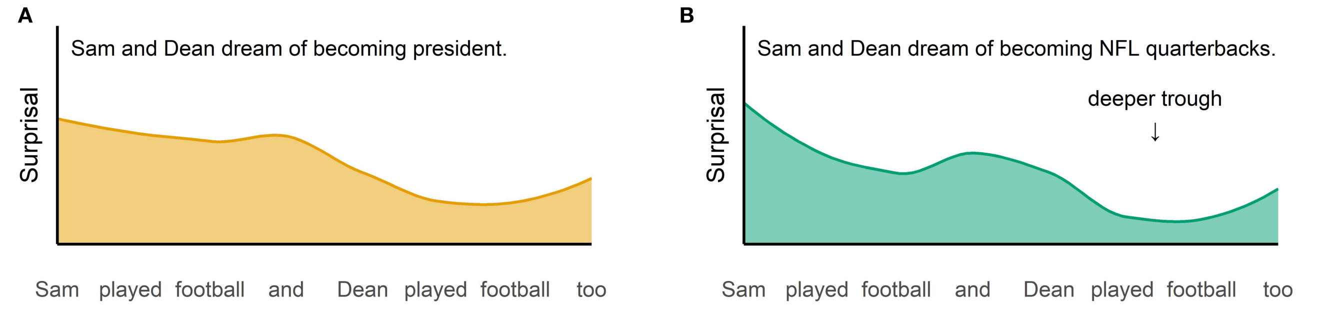


FIGURE 3 | Hypothetical information density for profiles for example (1) in a neutral context like (2-b) **(A)** and a predictive context like (2-a) **(B)**.

football (**Figure 3**). It thus conveys fewer information in this case and leads to a deeper trough in the information density profile. And such a deeper trough is equivalent to a less efficient use of the hearer's processing capacities. To avoid this, a speakers should have a stronger preference to use VP ellipsis in such predictive contexts.

(2) a. Sam and Dean dream of becoming NFL quarterbacks some day. (**predictive**)
b. Sam and Dean dream of becoming President some day. (**neutral**)

UID explains the production of utterances from the perspective of a speaker who performs audience design (Bell, 1984): She or he adapts her or his utterances as to facilitate comprehension for the hearer. We can assess the success of this audience design with naturalness rating and self-paced reading experiments which allows us to link the relative naturalness of ellipsis to the processing effort associated with the competing full forms.

Note that the UID predictions of avoiding redundancy are partially shared by accounts from research on anaphora[4].

[4] We would like to thank two anonymous reviewers for pointing this out.

First, Williams (1997, p. 603) postulates the principle *Don't Overlook Anaphoric Possibilities* (*DOAP*), according to which any opportunity to anaphorize text must be seized and a repeated phrase must be destressed (Williams, 1997, p. 595). Since Williams (1997) interprets deleted material as an instance of anphora, DOAP should also apply to VP ellipsis. Whenever deletion as extreme form of destressing is possible, speakers should make use of it and hearers should expect it. Realizing redundant material can in turn lead hearers to assume that there is a reason for this explicitness, e.g., in the form of a contrast. Consequently, if no such reason exists, hearers should reject the more redundant forms. A possible account based on the DOAP principle would hence predict that the repetition of redundant material is penalized, i.e., that it leads to degraded ratings. Conversely, the use of reduced forms such as VP ellipsis should be beneficial in that case and lead to better ratings.

Second, previous research has evidenced the so called *repeated-name penalty* (Gordon et al., 1993; Gordon and Hendrick, 1998; Almor, 1999) and the similar *overt pronoun penalty* in languages with null pronouns (Almor et al., 2017; Shoji et al., 2017): Participants read sentences more slowly when they contain a repeated name instead of a pronoun or an overt pronoun instead of a null pronoun. Gordon and Hendrick (1998, p. 390) argue that pronouns are primarily used to establish coreference, while names introduce entities into the discourse. Hence, coreference with names instead of pronouns requires additional processing effort resulting in increased reading times. Kertz (2010) adapts the concept of repetition penalties to VP ellipsis and rating data (see also Kim et al., 2011). She observes degraded ratings in contexts where a matched repeated VP was introduced by a parallel connective, calling this a *repeated verb phrase penalty*. A potential account based on the repetition penalties would consequently predict that processing difficulties caused by redundant material result in degraded acceptability.

The predictions of a possible DOAP approach and a potential repetition penalties account are partially consistent with those of the information-theoretic UID hypothesis: DOAP and the repetition penalties both predict degraded ratings through redundant material, which the latter account explains with processing difficulties. UID, however, explicitly makes gradual predictions: According to UID, a repeated VP is expected to be worse or more difficult to process, the longer it is or the more predictable it is in context. Possible accounts based on DOAP and the penalties would predict that any repetition of redundant material should be degraded and would not straightforwardly account for gradual or categorical effects of length or predictability. Hence, these predictions allow us to distinguish our UID account from the potential DOAP and repetition penalty accounts.

3. LENGTH EFFECTS

As outlined above, we expect, following UID, that the length of redundant material impacts the preference of a speaker to omit this material. More specifically, a longer redundant repeated verb phrase should be more likely to be omitted than a corresponding short repeated redundant verb phrase. We test this hypothesis first with a naturalness rating study which investigates the perception of long and short redundant verb phrases compared to their elliptical counterparts. This tells us whether the usage of ellipsis is motivated by a form of audience design: When VP ellipsis is preferred over full forms by hearers, speakers in turn should be more likely to use them to increase the efficiency of communication. Assessing whether repeated redundant verb phrases indeed lead to less efficient communication is the goal of the self-paced reading study on only the full forms. With respect to length we test whether the information minimum caused by redundancy is more severe when the repeated part is longer.

3.1. Experiment 1 – Naturalness Rating Study

In a 2 × 2 (LENGTH: short vs. long × FORM: full form vs. VPE) naturalness rating study we test the prediction that a long redundant verb phrase is more dispreferred than a short redundant verb phrase compared to the corresponding VP ellipsis.

3.1.1. Materials

We constructed 32 items[5] like (3) which consist in two coordinated main clauses with SVO word order respectively. The basic verb phrase is always a verb object pair like *play football* with the object being a DP without an overt determiner like *football*. We varied the LENGTH of this verb phrase between short and long. In the short conditions we presented only the basic verb phrase, in the long conditions we expanded the verb phrase by a complex locative adverbial consisting of two nested prepositional phrases that defines more closely where the event described by the verb is happening. The verb phrase in the second conjunct was varied in its FORM between the full form and VP ellipsis.

(3) a. Sam played football and Dean played football, too. (**short, full form**)
b. Sam played football and Dean did, too. (**short, VPE**)
c. Sam played football in the backyard of the house and Dean played football in the backyard of the house, too. (**long, full form**)
d. Sam played football in the backyard of the house and Dean did, too. (**long, VPE**)

We mixed the items with 72 fillers, among which were 24 gapping constructions (4) and 24 constructions with a subject lacking (5), half of which were elliptical, half syntactically complete. We included these ellipses to ensure that our items did not stand out as being the only syntactically incomplete utterances and balanced ellipses and full forms across the experiment. Sixteen of the fillers were followed by polar comprehension questions that served as attention checks.

(4) Mary hates broccoli and John (hates) cauliflower.

(5) Cass entered the theatre after the start of the movie and (he) looked for his seat but it was already taken.

[5] The items of all experiments can be found in the **Supplementary Materials**.

3.1.2. Procedure

We recruited 48 self-reported native speakers of British English from the crowdsourcing platform Prolific Academic who received a compensation of £2. The survey was conducted over the Internet using the LimeSurvey survey presentation software[6]. Subjects were asked to rate the naturalness of the stimuli on a 7-point Likert scale where 1 was *completely unnatural* and 7 *completely natural*. Materials were distributed across four lists with a 2 × 2 Latin square design. Each subject saw each token set once and only in one condition. The 32 items were mixed with the 72 fillers and presented in pseudo-randomized order.

3.1.3. Results

Before the main analysis we excluded 7 participants who failed our attention checks by answering more than the beforehand set threshold of 4 comprehension questions incorrectly. This threshold was established because at this point there is no significant difference to a purely random answering as evidenced by a chi-square goodness of fit test. We analyzed the remaining data in R (R Core Team, 2020) with cumulative link mixed models for ordinal data (Christensen, 2019). In all analyses in this paper we used a backward model selection procedure to find the final model: By performing likelihood ratio tests with the `anova` function we compared a model with and without an effect in question and continued with the simpler model if this did not significantly improve model fit. In our full model[7] we model the ratings as a function of the two binary predictors LENGTH and FORM, the scaled and centered POSITION of the item in the experiment and all two way interactions between them. We used deviation coding for the two categorical variables with −0.5 and 0.5 as levels. We included the full random effects structure justified by the data (Barr et al., 2013), i.e., random intercepts for subjects and items and by-subject and by-item random slopes for LENGTH, FORM, POSITION and their two-way interactions. The final model (**Table 1**) contains a significant main effect of LENGTH ($\chi^2 = 29.45, p < 0.001$) which shows that participants in general preferred utterances with short verb phrases over utterances with long verb phrases. The final model also revealed a significant main effect of FORM ($\chi^2 = 17.7, p < 0.001$): The ratings for VP ellipsis were generally better than the ratings for the full forms. We found a significant interaction between FORM and LENGTH ($\chi^2 = 11.85, p < 0.001$) (see **Figure 4**): Full forms with a long repeated verb phrase are rated significantly worse than full forms with a short verb phrase as compared to utterances with VP ellipsis. A significant interaction between FORM and POSITION ($\chi^2 = 5.8, p < 0.05$) and a significant main effect of POSITION ($\chi^2 = 4.42, p < 0.05$) show that in general the ratings became better in the course of the experiment and that they improved in particular for VP ellipsis which might indicate a familiarization effect.

TABLE 1 | Fixed effects in the final clmm for experiment 1.

Predictor	Estimate	SE	χ^2	*p*-value	
FORM	−1.24	0.27	17.7	< 0.001	***
LENGTH	0.88	0.14	29.45	< 0.001	***
POSITION	0.28	0.13	4.42	< 0.05	*
FORM:LENGTH	1.1	0.3	11.85	< 0.001	***
FORM:POSITION	−0.3	0.12	5.8	< 0.05	*

**p < 0.05, **p < 0.01, ***p < 0.001.*

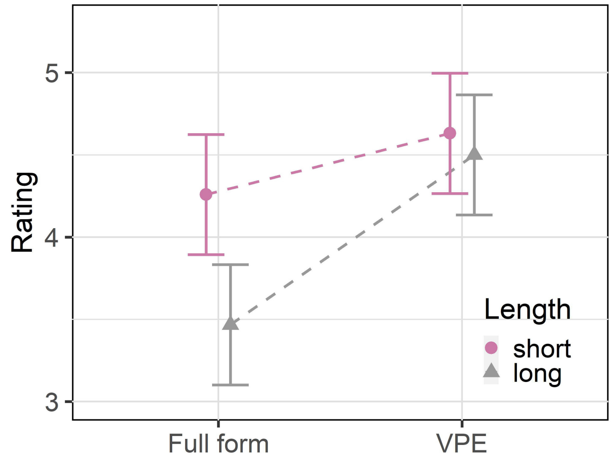


FIGURE 4 | Mean ratings and 95% confidence intervals per conditions for experiment 1.

3.1.4. Discussion

Our naturalness rating study confirms the prediction of the UID hypothesis on length effects: The results show that while participants overall prefer utterances with short repeated verb phrases and with VP ellipsis, long redundant full forms are particularly dispreferred as compared to the corresponding VP ellipsis conditions.[8] This is in line with the prediction that from a hearer perspective VP ellipsis is particularly preferred in the long conditions where the full form would create a long surprisal minimum. If a speaker performs audience design, she or he

[6] https://www.limesurvey.org/

[7] `Ratings ~ (FORM + LENGTH + POSITION)^2 + (1 + (FORM + LENGTH + POSITION)^2 | Subjects) + (1 + (FORM + LENGTH + POSITION)^2 | Items).`

[8] We argue that the degraded ratings are caused by redundancy and hence expect a gradual effect, i.e., the ratings get worse the more redundant the target utterance is. A reviewer suggests to test this prediction with partially redundant utterances, such as (i), where the PP is repeated but the core VP is new.

(i) a. Sam played football in the backyard of the house and Dean flew a kite in the backyard of the house.
b. Sam played football in the backyard of the house and Dean flew a kite there.

We must leave a systematic investigation of such cases to future research, but we have tentative data from experiment 4, where a part of our fillers had a similar structure (see section 4.2.1): In this experiment the fully redundant long full forms received a mean rating of 3.82 ($\sigma = 1.92$) and the corresponding ellipses got 5.1 ($\sigma = 1.67$). With a mean rating of 4.24 ($\sigma = 1.82$), the partly redundant full forms with a new second VP and a repeated PP (i-a) lie between these two, which could be a hint toward a gradual effect of redundancy on naturalness ratings. However, this result is questioned by the fact that the corresponding partly reduced forms (i-b) are rated best (5.32, $\sigma = 1.56$), even better than the completely reduced forms.

should take the hearer perspective into account and there should be a stronger pressure to omit the redundant material.

The main effect of form that shows a general preference for VP ellipsis over full forms is also expected by UID: Participants favor the more reduced form of an ellipsis over the redundant repetition of identical material in the full form. The repeated verb phrase is redundant in both length conditions because it is completely identical to the verb phrase in the first conjunct. This means that even in the short conditions two words are used to communicate what in the ellipsis conditions can be said with a single *did*. Ellipsis hence avoids a trough in the ID profile that would be caused by the redundant repetition of the identical verb phrase. The result that redundant repetitions are generally dispreferred is also in line with the DOAP principle of Williams (1997) and with the repetition penalties (e.g., Gordon et al., 1993; Kertz, 2010), but these approaches cannot account for the observed interaction, i.e., they do not straightforwardly predict the gradual nature of the length effect.

Participants seem to generally prefer shorter utterances which might be related to the fact that the locative adverbials consisting of two PPs are more demanding than the very simple plain VPs. In sum, experiment 1 is in line with the UID predictions: Speakers prefer VP ellipsis especially when it avoids the redundant repetition of a long verb phrase.

3.2. Experiment 2 – Self-paced Reading Study

While experiment 1 showed the expected naturalness pattern, we need to complement it with an on-line self-paced-reading study to test the UID predictions about processing effort. According to our UID account the degraded ratings for the long redundant full forms are caused by an information minimum that underutilizes the hearer's processing capacities. To test this prediction we use a 1 × 2 (Length: short vs. long) self-paced reading paradigm. We measure the reading times for the redundant verb phrase to see whether participants indeed speed up on this region. Our UID account predicts that a redundant verb phrase is read relatively faster when it is longer than when it is shorter.

3.2.1. Materials

We used only the full forms of the same 32 items and 72 fillers that were tested in experiment 1 including the 16 comprehension questions that served again as attention checks. We measured reading times on the first and the second verb phrase as illustrated in (6). The items were expanded by a spillover region always consisting in a clause introduced by *whereas* or *while* which described a different action performed by a third person. This prevents a wrap-up effect on the final word of the second verb phrase and makes the two verb phrases more comparable.

(6) a. Sam played football$_{\text{1st VP}}$ and Dean played football$_{\text{2nd VP}}$ too whereas Jack studied for university. (**short**)

b. Sam played football in the backyard of the house$_{\text{1st VP}}$ and Dean played football in the backyard of the house$_{\text{2nd VP}}$ too whereas Jack studied for university. (**long**).

3.2.2. Procedure

We recruited 96 self-reported native speakers of British English from the crowdsourcing platform Prolific Academic who were paid £2. None of the participants had taken part in experiment 1. The experiment was conducted over the Internet using IBEX[9]. Subjects read the stimuli in a centered self-paced reading paradigm. Materials were presented word by word on the screen. The experiment was preceded by a practice phase with 7 sentences and 2 comprehension questions to familiarize subjects with the procedure. Materials were distributed across two lists with a Latin square design. Each subject saw 32 items (16 per condition) which were mixed with the 72 fillers and presented in fully randomized order. Sixteen of the fillers were followed by attention checks in the form of polar comprehension questions.

3.2.3. Pre-processing

The dependent variable that we use in our analysis are residualized cumulated reading times (RCRT in what follows) which we compare between the first and the second verb phrase. To obtain these reading times we first excluded all by-word reading times that were faster than 90 ms and slower than 3,000 ms. Since we compare the reading times of a whole region of interest, i.e., the whole verb phrase as underlined in (6), we excluded all regions that had become incomplete due to the by-word exclusions. These exclusions resulted in a loss of approximately 2% of the regions of interest. For each region of interest we summed up the plain by-word reading times. These cumulated reading times were then residualized based on the item data of all participants. That means that the cumulated reading times were normalized for length per participant by using the residuals of a linear model computed on the items of all participants with reading times as a function of number of characters (see Gibson and Levy, 2016).[10] This allows us to compare the speed-up on the second verb phrase between short and long verb phrases despite the varying number of characters.

3.2.4. Results

We excluded the data of 26 participants who had answered more than 4 of our 16 comprehension questions incorrectly.[11] We analyzed the remaining data with linear mixed effects models (Bates et al., 2015) in R. Our full model contained the RCRT as dependent variable and the binary predictors Length (short vs. long VP) and VP (first vs. second VP), the scaled and centered Position of the trial in the experiment and all two-way interactions between the predictors. We coded the two categorical variables with −0.5 and 0.5 respectively using deviation coding. We included a random intercept for items, a by-item random slope for Length and a by-subject random slope for VP.[12] Given that we use a dependent variable that is

[9] https://spellout.net/ibexfarm/

[10] We adapted the code provided by Gibson and Levy (2016) at https://osf.io/swyux/.

[11] The higher number of exclusions as compared to experiment 1 might be the result of the more demanding reading task. In particular the long repeated VPs might have led to fatigue and consequently to stronger inattention.

[12] `RCRT ~(Length + VP + Position)^2 + (0 + VP || Subjects) + (1 + Length | Items)`.

TABLE 2 | Fixed effects in the final lmer for experiment 2.

Predictor	Estimate	SE	df	χ^2	*p*-value	
LENGTH	–179.85	45.48	30.99	13.07	< 0.001	***
VP	206.38	28.61	68.60	39.22	< 0.001	***
POSITION	–211.09	7.48	4291.06	730.55	< 0.001	***
LENGTH:VP	–289.88	29.46	4265.28	95.82	< 0.001	***
LENGTH:POSITION	219.23	14.97	4287.63	209.46	< 0.001	***
VP:POSITION	–33.16	14.75	4275.01	5.05	< 0.05	*

*p < 0.05, **p < 0.01, ***p < 0.001.

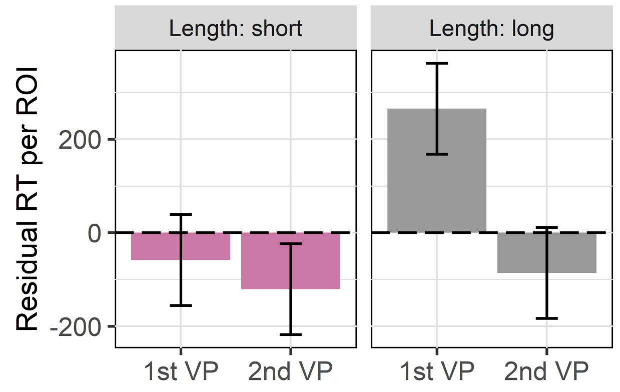


FIGURE 5 | Mean residual cumulated reading times per region of interest and 95% confidence intervals for experiment 2.

already normalized for subject and length effects and given that the two verb phrases are always identical for each item we used this informed random effects structure.

The final model (**Table 2**) revealed a significant main effect of VP ($\chi^2 = 39.22, p < 0.001$): Participants read the second (redundant) verb phrase faster than the first (non-redundant) verb phrase. The model also revealed a significant main effect of LENGTH ($\chi^2 = 13.07, p < 0.001$): Participants were overall faster on the short verb phrases. The model contained a significant interaction between LENGTH and VP ($\chi^2 = 95.82, p < 0.001$) (see **Figure 5**): The speed-up on the second verb phrase as compared to the first was especially fast for the long verb phrases. Furthermore, the final model contained a significant main effect of POSITION ($\chi^2 = 730.55, p < 0.001$) and significant interactions of POSITION with LENGTH ($\chi^2 = 209.46, p < 0.001$) and with VP ($\chi^2 = 5.05, p < 0.05$). Participants became notably faster during the experiment which indicates an increased familiarity with the task, in particular they speeded up on the first verb phrase and on the long verb phrases.

3.2.5. Discussion

The result of the self-paced reading study is in line with the UID prediction: The speed-up on the second verb phrase is bigger for the long conditions than for the short conditions. A long redundant verb phrase should thus create a longer region of low surprisal and result in a more severe underutilizing of the hearer's processing resources. This is exactly what is reflected in the degraded naturalness ratings for the long full form in the rating study in section 3.1. Hence, the reading study shows that the degraded ratings can be traced back to a non-optimal information density profile.

The reading study furthermore showed that participants were faster on the short verb phrases even after normalizing for the differing number of characters. This might be due to the fact that there is less material to be integrated when processing shorter utterances. Additionally there was a general speed-up between the first and the second verb phrase. Since participants already know the verb phrase when they encounter it for the second time, they may consequently read it faster. The massive position effects observed in the analysis indicate that participants became more and more familiar with the experimental design and the structures. It might be the case that the long redundant verb phrases are particularly marked and that participants are slow when they first encounter them, but become faster in the course of the experiment as a familiarization effect.

In total, the results of this reading study are in line with UID: They suggest that the degraded ratings from experiment 1 are indeed caused by a non-optimal information density profile with a long trough.

4. CONTEXT EFFECTS

Experiments 1 and 2 showed that redundant structures are dispreferred and harder to process as predicted by UID. In what follows we explore a second source of redundancy that in contrast to length allows us to keep the target verb phrase constant across conditions: the predictability through context. The central idea is that a verb phrase is the more redundant, i.e., the less informative, the more predictable it is based on the previous linguistic context (2), repeated here as (7). For instance, in (8), Dean should be more likely to also play football if he wants to become a NFL quarterback (7-a) than if he wants to become President (7-b).

(7) a. Sam and Dean dream of becoming NFL quarterbacks some day. (**predictive**)
b. Sam and Dean dream of becoming President some day. (**neutral**)

(8) Sam played football in the backyard of the house and Dean played football in the backyard of the house too.

Just as with the length effects, we test this prediction with a naturalness rating experiment and a self-paced reading study. Again, we want to measure the naturalness of ellipsis as compared to the corresponding full forms and to trace back possible differences to processing as indexed by reading times. Before our actual experiments, we conducted a pre-test to test whether our contexts were indeed either predictive or neutral.

4.1. Experiment 3 – Pre-test

Up to now we have only assumed that the context (7-a) is more predictive than the context (7-b). We verify this assumption with a pre-test in which we obtain estimates for the likelihood of the second conjunct in context, independent of ellipsis. This pre-test should evidence that our verb phrases are likely in the predictive contexts and significantly less likely in the neutral contexts. Based on the results we select those items for the subsequent rating and reading study for which we find a significant difference in likelihood between the predictive and neutral context condition. Additionally, it is crucial to avoid that our neutral contexts are not only less predictive but implausible. Implausible contexts could be problematic for at least two reasons: First, if participants cannot make sense of the respective items, this might lead to an overall rejection of the neutral conditions. This would mask any fine-grained UID effects. Second, being confronted with too many implausible contexts could lead participants to abandon predictive processing during the rating study (see e.g., Fine et al., 2013; Brothers et al., 2017, who show that participants rapidly adapt their predictions during sentence comprehension) and this could override the predictability manipulation altogether. Therefore, we needed to assure that our neutral contexts make the critical verb phrases significantly more likely than implausible controls.

4.1.1. Materials

We constructed a presumably predictive and a presumably neutral context sentence respectively for each of the 32 items from experiments 1 and 2 which were slightly adapted to better fit to the contexts. We tried to keep both context conditions as parallel as possible by either varying only the object of the VP or in some cases an embedded VP.[13]

Instead of presenting the coordinated structures to participants we used only the second conjunct, i.e., the one that will be targeted by VP ellipsis in the actual experiment (9). This way we ensured that we only test the predictability of the target verb phrase in the given context. In order to have more material on which we could measure reading times in the planned reading experiment, we used the long variants from experiments 1 and 2.

(9) a. Sam and Dean dream of becoming NFL quarterbacks some day. Dean played football in the backyard of the house$_{\text{target sentence}}$. (**predictive**)
b. Sam and Dean dream of becoming President some day. Dean played football in the backyard of the house$_{\text{target sentence}}$. (**neutral**)

We mixed our items with 90 fillers including 32 similar items for another experiment with two context sentences and 34 script-based (Schank and Abelson, 1977) fillers with one context sentence. For half of these fillers the context made the target sentence predictable, for half not. The remaining 24 fillers were pre-tested stimuli, 12 with 1 context sentence, 12 with 2 context sentences, of which half were implausible because they contained severe script violations as exemplified in (10).[14] We included them as controls to verify that our neutral contexts were not implausible, i.e., that the ratings for items with neutral contexts are significantly higher than the ratings for items with implausible contexts.

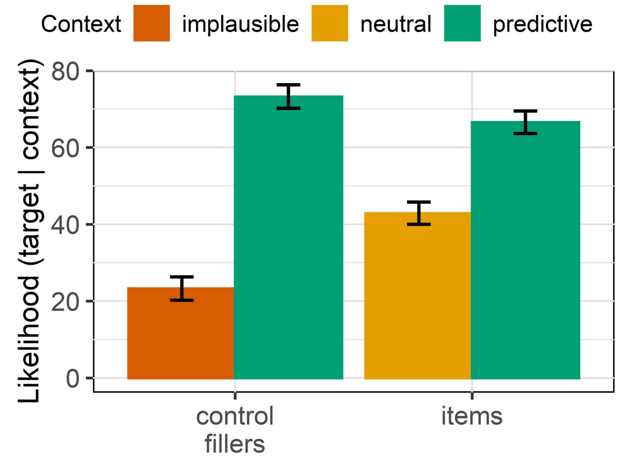


FIGURE 6 | Mean likelihood ratings and 95% confidence intervals for items and control fillers in experiment 3. The implausible conditions of the control fillers were rated as significantly less likely than the neutral items.

(10) Rowena was hungry. She called the delivery service. She greeted the employee warmly and ordered a blouse in extra large$_{\text{target sentence}}$.
(**control filler, implausible**).

4.1.2. Procedure

We recruited 48 self-reported native speakers of American English from Prolific Academic who had not participated in experiments 1 and 2 and compensated them with £2.50. They had to rate how likely it is that the event described by the target sentence, which was presented in bold face, happens in the given context using a slider scale from 0 (*cannot happen*) to 100 (*must happen*). The items were distributed across two lists with a Latin square design. Each subject rated 32 items (16 with a predictive, 16 with a neutral context) which were mixed with the fillers and presented in fully randomized order.

4.1.3. Results

Figure 6 shows the mean likelihood ratings and 95% confidence intervals for our items and the implausible (and corresponding predictive) controls. The implausible context fillers had a mean likelihood rating of 23.08 points ($\sigma = 24.73$) whereas the neutral context conditions of our items were rated with an average of 42.82 points ($\sigma = 25.19$) This indicates that our items are not implausible, but only less probable. This is confirmed by

[13] An example of the latter is given in (i).

(i) a. Jodie and Donna were eager to see the new season of their favorite show. Donna watched television on the sofa in the living room.
b. Jodie and Donna were eager to go for a jog in the park. Donna watched television on the sofa in the living room.

[14] We thank Elisabeth Rabs for providing us with the original German materials as used in Rabs et al. (under review): *Situational Expectancy or Word Association? The Influence of Event Knowledge on the N400.*

the results of a linear mixed effects model (Bates et al., 2015) on a subset of the data consisting of the control fillers and the items. For the analysis we collapsed the implausible and the neutral context conditions which are jointly contrasted with the predictive conditions. We model the likelihood score as a function of stimulus type and context and find a significant interaction between both predictors in the expected direction ($\chi^2 = 29.58, p < 0.001$): The implausible fillers received significantly lower likelihood ratings than the neutral items. This indicates that our neutral contexts should be plausible and we should receive valid ratings for them.

In order to select the items for the rating and the reading experiment, we assessed for each item whether the likelihood rating for the predictive context was significantly higher than for the neutral context. We compared the mean rating for the neutral context condition to the mean rating for the predictive context condition for each token set separately with one-sided Wilcoxon-tests in R. For 24 of 32 items the rating for the predictive context was significantly higher than for the neutral context, so we selected them for our main experiments.

4.2. Experiment 4 – Naturalness Rating Study

Our UID account predicts that a redundant verb phrase is more likely to be omitted. While experiment 1 and 2 showed that this redundancy increases as a function of the verb phrase's length, a second source of redundancy could be predictability in context. A repeated verb phrase should also be more redundant if it is likely given the previous context. We expect that this additional redundancy creates a deeper information minimum in the full forms which leads to degraded naturalness ratings. We test this with a 2 × 2 (CONTEXT: predictive vs. neutral × FORM: full form vs. VPE) naturalness rating study.

4.2.1. Materials

We used the 24 items which we had selected with the pre-test including predictive and neutral contexts. We reinserted the first conjunct to the target sentence (11) so that the target sentences were basically identical to the long conditions of experiments 1 and 2 and added a sentence-initial adverbial.

(11) a. Sam and Dean dream of becoming NFL quarterbacks some day. Wednesday afternoon Sam played football in the backyard of the house and Dean played football in the backyard of the house too. (**predictive, full form**)

b. Sam and Dean dream of becoming NFL quarterbacks some day. Wednesday afternoon Sam played football in the backyard of the house and Dean did too. (**predictive, ellipsis**)

c. Sam and Dean dream of becoming President some day. Wednesday afternoon Sam played football in the backyard of the house and Dean played football in the backyard of the house too. (**neutral, full form**)

d. Sam and Dean dream of becoming President some day. Wednesday afternoon Sam played football in the backyard of the house and Dean did too. (**neutral, ellipsis**)

The items were mixed with 36 fillers which resembled the items in consisting of a context sentence and a target sentence with two coordinated verb phrases. Their purpose was to avoid a habituation effect caused by the structure of our items. Since the structure of our items was relatively constant, subjects could anticipate a redundant verb phrase as soon as they encounter a verb phrase followed by an *and*. This could overwrite or weaken the predictability manipulation of the verb phrase that we intended through the context sentence. Therefore we created 12 filler sentences where a completely different conjunct followed the coordination (12), 12 fillers where we changed the prepositional phrase but maintained the basic verb phrase (13) and 12 fillers where the prepositional phrase was kept constant but the verb phrase changed (14). For half of the sentences with a repeated phrase ($n = 12$) we substituted this phrase with an ellipsis (13) or a pro-form such as *there* in (14). This way, participants could not anticipate an identical second verb phrase when encountering *and*.

(12) Gabriel and Michael had taken leave. In the morning Gabriel packed provisions at the table in the kitchen and Michael loaded the car in the street before the house. (**filler, different conjunct**)

(13) Claire and Alex have a green thumb. Last year Claire grew tomatoes in flowerpots on the terrace and Alex (grew tomatoes) in patches in the garden. (**filler, same VP, different PP**)

(14) Bobby and Gordon enjoy life to the full. Last Saturday Bobby lost money in a casino in Reno and Gordon saw a performance (in a casino in Reno | there). (**filler, different VP, same PP**)

We further included 24 items from another experiment and 24 fillers which both had a structure similar to our items and each of which were half elliptical. This again was intended to ensure that our items did not stand out as the only syntactically incomplete utterances. Sixteen of the fillers were followed by polar comprehension questions asking either for information from the context or the target sentence that served as attention checks.

4.2.2. Procedure

We recruited 96 self-reported native speakers of American English on Prolific Academic who had not taken part in any of the previous experiments. They were compensated with £2. We presented the survey over the Internet using IBEX. Subjects rated the naturalness of the critical utterance which was set in italics on a 7-point Likert scale (7 was *completely natural*). Materials were distributed across four lists with a Latin square design. Each subject saw each token set once and only in one condition. The FORM of the items was varied between subjects, i.e., 48 subjects saw only ellipses, 48 subjects only full forms in order to avoid floor effects for the marked redundant full forms.

TABLE 3 | Fixed effects in the final clmm for experiment 4.

Predictor	Estimate	SE	χ^2	*p*-value	
Predictability	4.61	0.76	27.58	< 0.001	***
Form	–1.74	0.54	9.75	< 0.01	**
Position	0.16	0.13	1.6	> 0.2	
Form:Position	–0.53	0.24	4.89	< 0.05	*

**p < 0.05, **p < 0.01, and ***p < 0.001.*

4.2.3. Results

Before the analysis we excluded 13 subjects who had not passed our attention checks by answering more than 4 of 16 comprehension questions incorrectly. The threshold was set analogously to experiment 1 in section 3.1.3. The data of the remaining 83 subjects was analyzed using cumulative link mixed models (Christensen, 2019) in R following the procedure described for experiment 1 in section 3.1.3. The full model contained the ratings as an ordinal dependent variable and as independent variables the binary Form predictor, the numerical mean pre-test score by item and condition indicating Predictability, the scaled Position of the trial in the experiment and all two-way interactions between them. The categorical variable Form variable was transformed to -0.5 and 0.5 respectively using deviation coding. We included random intercepts for subjects and items and by-subject random slopes for Predictability and Position, as well as by-item random slopes for all three predictors and a by-item random slope for the interaction between Predictability and Form.[15]

The final model (**Table 3**) contains a significant main effect of Form ($\chi^2 = 9.75, p < 0.01$) which indicates a preference for VP ellipses over full forms. We also find a significant main effect of the Predictability score ($\chi^2 = 27.58, p < 0.001$): Utterances that are predictable given the previous context received better ratings. The interaction between Form and Predictability is marginal ($\chi^2 = 3.19, p = 0.07$) and therefore not part of the final model. There is a trend toward better ratings for VP ellipsis in predictive contexts as illustrated in **Figure 7**.

4.2.4. Discussion

In this rating study, we investigated predictability in context as a source of redundancy for a repeated verb phrase. Our UID account predicts that VP ellipsis should be more strongly preferred when the omitted verb phrase is more predictable in context. In the data, we do not find this predicted interaction between the predictability and the form of the redundant verb phrase. There is only a marginal effect in the expected direction. While the pre-test evidenced a clear difference in likelihood between the two context conditions, this does not result in a stronger preference for VP ellipsis. We find however that our predictability manipulation works: Participants preferred utterances in predictive contexts over such in neutral contexts.

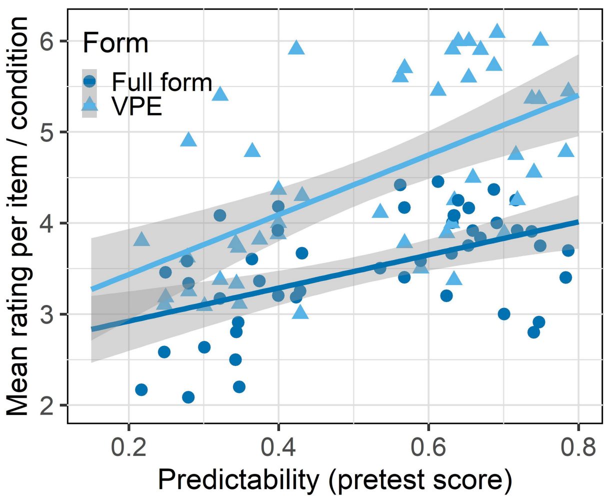


FIGURE 7 | Mean rating per item and condition as a function of the numerical pretest score indicating Predictability for experiment 4.

Similar to the length rating study, there was also a general preference for the more compact VP ellipsis over the long redundant full forms which is also predicted by the DOAP principle and the repetition penalty account.

So why is there only a marginal preference for VP ellipsis in the predictive conditions? A possible explanation might be that our context manipulation did not affect VP ellipsis because the verb phrase is still too predictable even in our neutral conditions and therefore VP ellipsis is also preferred in these conditions according to UID. Regardless of whether the VP ellipsis follows a predictive or a neutral context, there is always a parallel first verb phrase available which is straightforwardly accessible as antecedent for the ellipsis. Thus, VP ellipsis can be easily processed even in the neutral condition and there is no need to use the redundant full form. This is supported by the overall preference for VP ellipsis over the full form, which we did find in both naturalness rating studies presented in this article.

We further need to consider that the set of possible encodings for the message that Sam played football and that Dean played football does not consist only of the full form and the corresponding VP ellipsis. An alternative encoding is a simple sentence with a coordinated subject like (15) which might be a competitor to the full form but which cannot be readily compared to the other two forms with UID.

(15) Sam and Dean played football in the backyard of the house.

We will turn back to these potential issues in section 5.

4.3. Experiment 5 – Self-paced Reading Study

In a 1 × 2 (Context: predictive × neutral) self-paced reading study we investigate whether the context impacts the processing effort on the redundant verb phrase. Our UID based account

[15] Ratings ~(Form + Predictability + Position)ˆ2 + (1 + Predictability + Position | Subjects) + (1 + (Form + Predictability)ˆ2 + Position | Items).

predicts that the redundant repeated VP is read faster in a predictive compared to a neutral context. This speed-up would evidence deeper regions of low information, i.e., the underutilization of the hearer's processing resources. For the length effects, we found both degraded ratings and a longer trough for the more redundant full forms. For the context effects, we want to test whether a predictable verb phrase leads to a deeper trough in the information density profile indexed by faster reading times. If we did not find such an effect, i.e., if there was no speed-up in the predictive condition, this would explain why we did not find the expected interaction in the rating study, i.e., why VP ellipsis was not more strongly preferred in the predictive contexts.

4.3.1. Materials

We used the same materials as in experiment 4, but tested only the full forms, both of the items (16) and the fillers. The method is similar to experiment 2, but instead of comparing the reading times between the first and the second verb phrase we compare the reading times on only the second verb phrase between both CONTEXT conditions.

(16) a. Sam and Dean dream of becoming NFL quarterbacks some day. Last Saturday Sam played football in the backyard of the house and Dean played football in the backyard of the house too. **(predictive)**

b. Sam and Dean dream of becoming President some day. Last Saturday Sam played football in the backyard of the house and Dean played football in the backyard of the house too. **(neutral)**.

4.3.2. Procedure

49 self-reported native speakers of American English who had not participated in any of the previous experiments were recruited over Prolific Academic to take part in the study.[16] They received a compensation of £2. We conducted the self-paced reading experiment over the Internet using IBEX. In each trial, subjects first saw the context sentence as a whole and then read the target utterance word-by-word[17] in a centered self-paced reading paradigm. Before the actual experiments subjects passed a practice phase consisting of 7 sentences and 3 comprehension questions. Materials were distributed across two lists with a Latin square design. In the main experiment each participant read 24 items (12 in each condition) and 84 fillers presented in fully randomized order. Sixteen fillers had a subsequent polar comprehension question that served as attention checks.

In our analysis, we compared the residualized cumulated reading times (RCRT) calculated as described in section 3.2.3 for the identical second VP between the predictive and the neutral condition. We excluded by-word reading times faster than 90 ms and slower than 3,000 ms and all regions of interest that have become incomplete due to these by-word exclusions. This resulted in a loss of about 1% of all regions of interest.

TABLE 4 | Fixed effects in the final lmer for experiment 5.

Predictor	Estimate	SE	df	χ^2	*p*-value	
POSITION	−158.91	16.26	486.81	87.55	< 0.001	***

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

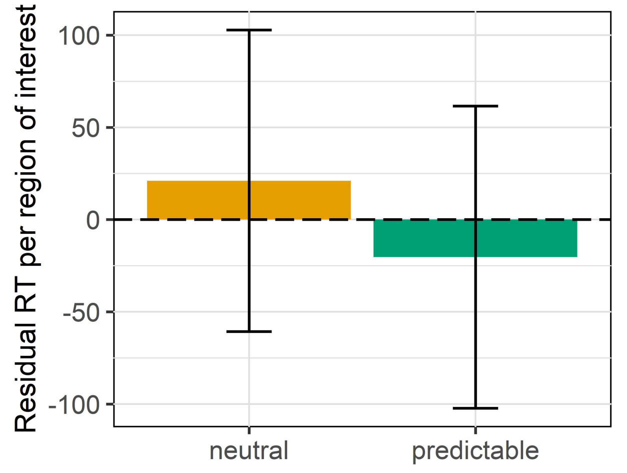


FIGURE 8 | Mean residual reading times and 95% confidence intervals per region of interest per condition for experiment 5.

4.3.3. Results

Before the analysis we excluded 6 participants who had failed our attention checks in having answered more than 4 of 16 comprehension questions incorrectly. We analyzed the data of the remaining 43 participants in R using linear mixed effects models (Bates et al., 2015) and the same procedure of backward model section described in experiment 1 in section 3.1.3. Our full model contained the RCRTs as dependent variable and as independent variables the numerical pre-test score indicating PREDICTABILITY, the scaled and centered POSITION of the item in the experiment and their interaction. We only included a random intercept for items because the reading times are already normalized per subject and more complex random effect structures resulted in singular fit.[18] The final model (**Table 4**) contained only a significant main effect of POSITION ($\chi^2 = 87.55, p < 0.001$) indicating that participants became faster in the course of the experiment. The main effect of predictability was not significant ($\chi^2 = 0.63, p = 0.43$). The redundant VP did not differ in reading times between the predictive and the neutral conditions (**Figure 8**).

4.3.4. Discussion

We investigated the processing of a redundant verb phrase in a predictive vs. a neutral context and found no difference in reading times of the second redundant verb phrase between context conditions. Specifically, participants did not show a speed-up on the repeated verb phrase after a predictive compared to a

[16] Due to internal processes of the crowd sourcing platform Prolific, we had the complete data of 49 instead of the planned 48 participants.

[17] An anonymous reviewer suggested that a phrase-by-phrase presentation could help to isolate effects in a clearer way. We will consider this for future studies.

[18] `RCRT ~(PREDICTABILITY + POSITION)^2 + (1 | Items).`

neutral context. This way, the results of the self-paced reading study pattern with the results of the rating study in section 4.2. This suggests that the repeated VP is equally redundant in both context conditions. The predictive context does not lead to a deeper information minimum in the information density profile than the neutral context. In section 4.2.4, we already presented a possible explanation for why we do not find the context effects that a UID account would predict. For the self-paced reading study, we add that we presented full forms that are highly unnatural in both conditions given that the second verb phrase is completely identical to the first verb phrase and that a simpler alternative in the form of a sentence with a coordinated subject would be available. This intuition is confirmed by the results of both rating studies in this paper where the long redundant full forms received degraded ratings. We hypothesize that during the reading task this unnaturalness masked the effect of the more subtle context manipulation or even led to severe processing difficulties that resulted in an equally strong slow down for both context conditions.

5. GENERAL DISCUSSION

We present a novel information-theoretic account to the underexplored question of when VP ellipsis is used. According to the UID hypothesis an increased redundancy leads to information minima which speakers tend to avoid when producing utterances. VP ellipsis or ellipsis in general is a possible strategy to avoid such troughs: The redundant material is omitted or at least drastically reduced. We investigated length and predictability in context as two sources of redundancy of the repeated verb phrase. A longer repeated verb phrase should cause a longer information minimum, while a repeated verb phrase in a predictive compared to a neutral context should result in a deeper information minimum. In both cases, these minima underutilize the hearer's processing resources and we expect that this is reflected in degraded naturalness ratings and faster reading times.

For the length effects manipulation, our results are in line with the predictions of our UID account. In the rating study we found that VP ellipsis is especially preferred over the full form when the redundant verb phrase is longer. In this case also the corresponding information minimum is longer which is equivalent to the underutilizing of the hearer's processing resources for a longer time. In a self-paced reading study we could evidence that the naturalness pattern is caused by processing: The redundant second verb phrase was read relatively faster compared to the first verb phrase when it was longer which indicates a longer information mimimum. The length of the redundant material seems to be indeed a factor that affects the information density profile and hence the usage of VP ellipsis. It is an advantage of our UID account over the DOAP principle (Williams, 1997) and the repetition penalties accounts (e.g., Gordon et al., 1993; Kertz, 2010) that it does not only predict a general categorical penalty for the repetition of redundant material, but a gradual effect of length.

We could not evidence an effect of predictability in context on the redundant verb phrase. In the naturalness rating study we found a non-significant trend toward a preference for VP ellipsis in predictive contexts. In the self-paced reading study, the reading times of the redundant verb phrases did not differ regardless of whether the verb phrase followed a predictive or a neutral context. We identified two possible explanations for this result: (i) The unnaturalness of the long redundant verb phrases could mask more subtle effects. The rating study on length effects evidenced that the long redundant full forms received particularly bad ratings. However, we had to use these full forms in the context studies in order to have enough material to measure on in the self-paced reading study. Since the context manipulation is more subtle than the length manipulation, the effect of the context might be overridden by the penalty caused by the long redundant full form. (ii) It might be the case that our context manipulation itself is too subtle. From a UID perspective there is no need for the speaker to use the full forms in any of the conditions that we tested. The form of our items entails that the first verb phrase is always immediately available as an antecedent for ellipsis. Hence, the ellipsis can be straightforwardly resolved even in the neutral context conditions. VP ellipsis as the shorter form always has an advantage over the less well-formed full form. This, in a future study, it may be promising to find a way to make the VP ellipsis less redundant. That is, the verb phrase should not be highly predictable through a given identical first verb phrase and the discourse connective *and*. A starting point might be to look at cases where the antecedent of the VP ellipsis differs in its morphosyntactic properties from the reconstruction of the ellipsis site. Arregui et al. (2006) tested structures like (17) where the antecedent is not a verb phrase but a gerund or a nominalization. In such cases a UID account could argue that an increased mismatch in form results in decreased redundancy of the repeated verb phrase. A full form as more explicit form could reduce the processing effort here because the effort associated with the more difficult resolving of ellipsis is canceled.

(17) a. *Singing the arias* tomorrow night will be difficult but Maria will.
b. Tomorrow night's *singing of the arias* will be difficult but Maria will.
(Arregui et al., 2006, p. 238).

In sum, we find partial support for our information-theoretic account to the usage of VP ellipsis. While the results on length effects are in line with our account based on UID, the results on context effects are not. The context reading study suggests that for structural reasons the redundant verb phrase is still too predictable even in the neutral contexts. This does not provide evidence against UID, but further studies in which VP ellipsis is made less redundant are needed to strengthen our account.

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 Ethics committee of the Deutsche Gesellschaft für Sprachwissenschaft (German society for language science). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

LS was responsible for preparing and conducting the experiments, for analyzing and visualizing the resulting data and writing the initial draft of this paper. RL supported LS in preparing and conducting the experiments and in the analysis of the data, and critically commented on the initial draft of this paper. HD and IR developed and formulated the overarching research goals, managed and supervised the research activities, and critically reviewed the analysis of the data and the initial draft of this paper. All authors contributed to the article and approved the submitted version.

FUNDING

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG)–Projektnummer 232722074–SFB 1102/Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–Project-ID 232722074-SFB 1102.

ACKNOWLEDGMENTS

We thank Elisabeth Rabs for providing the control fillers used in experiment 3, Matthew Kuhn for revising the experimental stimuli, the audience of the RAILS 2019 conference for their valuable feedback and the DFG for making this research possible.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.661087/full#supplementary-material

REFERENCES

Almor, A. (1999). Noun-phrase anaphora and focus: the informational load hypothesis. *Psychol. Rev*. 106, 748–765. doi: 10.1037/0033-295X.106.4.748

Almor, A., de Carvalho Maia, J., Cunha Lima, M. L., Vernice, M., and Gelormini-Lezama, C. (2017). Language processing, acceptability, and statistical distribution: A study of null and overt subjects in Brazilian Portuguese. *J. Mem. Lang*. 92, 98–113. doi: 10.1016/j.jml.2016.06.001

Apel, J., Knoeferle, P., and Crocker, M. W. (2007). "Rocessing parallel structure: evidence from eye-tracking and a computational model," in *Proceedings of the Second European Cognitive Science Society Conference* (Delphy), 125–131.

Arregui, A., Clifton, C., Frazier, L., and Moulton, K. (2006). Processing elided verb phrases with flawed antecedents: the recycling hypothesis. *J. Mem. Lang*. 55, 232–246. doi: 10.1016/j.jml.2006.02.005

Asr, F. T., and Demberg, V. (2015). "Uniform information density at the level of discourse relations: negation markers and discourse connective omission," in *Proceedings of the 11th International Conference on Computational Semantics* (London), 118–128.

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

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. (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

Bell, A. (1984). Language style as audience design. *Lang. Soc*. 13, 145–204. doi: 10.1017/S004740450001037X

Brothers, T., Swaab, T. Y., and Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *J. Mem. Lang*. 93, 203–216. doi: 10.1016/j.jml.2016.10.002

Christensen, R. H. B. (2019). *Ordinal—Regression Models for Ordinal Data*. R package version 2019.12-10. Available online at: https://CRAN.R-project.org/package=ordinal

Demberg, V., and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109, 193–210. doi: 10.1016/j.cognition.2008.07.008

Fine, A. B., Jaeger, T. F., Farmer, T. A., and Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE* 8:e77661. doi: 10.1371/journal.pone.0077661

Frazier, L., and Clifton, C. (2001). Parsing coordinates and ellipsis: copy α. *Syntax* 4, 1–22. doi: 10.1111/1467-9612.00034

Frazier, L., Munn, A., and Clifton, C. (2000). Processing coordinate structures. *J. Psycholinguist. Res*. 29, 343–370. doi: 10.1023/A:1005156427600

Gibson, E., and Levy, R. (2016). An attempted replication of Hackl, Koster-Hale, Varvoutis (2012). *arXiv*:1605.00178v001.

Gordon, P. C., Grosz, B. J., and Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cogn. Sci*. 17, 311–347. doi: 10.1207/s15516709cog1703_1

Gordon, P. C., and Hendrick, R. (1998). The representation and processing of coreference in discourse. *Cogn. Sci*. 22, 389–424. doi: 10.1207/s15516709cog2204_1

Hale, J. (2001). "A probabilistic earley parser as a psycholinguistic model," in *Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (Pittsburgh, PA), 159–166. doi: 10.3115/1073336.1073357

Jaeger, T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cogn. Psychol*. 61, 23–62. doi: 10.1016/j.cogpsych.2010.02.002

Jaeger, T. F., and Buz, E. (2017). "Signal reduction and linguistic encoding," in *The Handbook of Psycholinguistics*, eds E. M. Fernández and H. S. Cairns (Hoboken, NJ: John Wiley & Sons, Ltd.), 38–81. doi: 10.1002/9781118829516.ch3

Kertz, L. (2010). *Ellipsis reconsidered* (Ph.D. thesis). UC San Diego, San Diego, CA, United States.

Kim, C. S., Kobele, G. M., Runner, J. T., and Hale, J. T. (2011). The acceptability cline in VP ellipsis. *Syntax* 14, 318–354. doi: 10.1111/j.1467-9612.2011.00160.x

Kravtchenko, E. (2014). "Predictability and syntactic production: evidence from subject omission in Russian," in *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (Quebec City, QC), 785–790.

Kurumada, C., and Jaeger, T. F. (2015). Communicative efficiency in language production: optional case-marking in Japanese. *J. Mem. Lang*. 83, 152–178. doi: 10.1016/j.jml.2015.03.003

Lemke, R., Horch, E., and Reich, I. (2017). "Optimal encoding!-information theory constrains article omission in newspaper headlines," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (Valencia, Spain), 131–135. doi: 10.18653/v1/E17-2021

Lemke, R., Schäfer, L., Drenhaus, H., and Reich, I. (2020). "Script knowledge constrains ellipses in fragments-evidence from production data and language modeling," in *Proceedings of the Society for Computation in Linguistics 2020* (New York, NY), 441–444.

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

Levy, R., and Jaeger, T. F. (2007). "Speakers optimize information density through syntactic reduction," in *Advances in Neural Information Processing Systems 19*, eds B. Schlökopf, J. Platt, and T. Hofmann (Cambridge, MA: The MIT Press), 849–856.

Martin, A. E., and McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *J. Mem. Lang.* 58, 879–906. doi: 10.1016/j.jml.2007.06.010

Merchant, J. (2018). "Ellipsis: a survey of analytical approaches," in *The Oxford Handbook of Ellipsis, Oxford Handbooks*, eds J. V. Craenenbroeck and T. Temmerman (New York, NY: Oxford University Press), 18–45. doi: 10.1093/oxfordhb/9780198712398.013.2

Paape, D., Nicenboim, B., and Vasishth, S. (2017). Does antecedent complexity affect ellipsis processing? An empirical investigation. *Glossa J. Gen. Linguist.* 2, 1–29. doi: 10.5334/gjgl,.290

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reich, I. (2018). "Ellipsen," in *Handbuch Pragmatik*, eds F. Liedtke and A. Tuchen (Stuttgart: J.B. Metzler), 240–251. doi: 10.1007/978-3-476-04624-6_24

Sag, I. A. (1976). *Deletion and logical form* (thesis). Massachusetts Institute of Technology, Cambridge, MA, United States.

Schank, R. C., and Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* 27, 623–656. doi: 10.1002/j.1538-7305.1948.tb00917.x

Shoji, S., Dubinsky, S., and Almor, A. (2017). The repeated name penalty, the overt pronoun penalty, and topic in Japanese. *J. Psycholinguist. Res.* 46, 89–106. doi: 10.1007/s10936-016-9424-4

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag. doi: 10.1007/978-3-319-24277-4_9

Williams, E. (1997). Blocking and anaphora. *Linguist. Inq.* 28, 577–628.

Williams, E. S. (1977). Discourse and logical form. *Linguist. Inq.* 8, 101–139.

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 Psychology

ORIGINAL RESEARCH
published: 28 May 2021
doi: 10.3389/fpsyg.2021.660730

Limits to the Rational Production of Discourse Connectives

Frances Yung[1]*, *Jana Jungbluth*[1] *and Vera Demberg*[1,2]

[1] *Department of Language Science and Technology, Saarland University, Saarbrücken, Germany,* [2] *Department of Computer Science, Saarland University, Saarbrücken, Germany*

Rational accounts of language use such as the uniform information density hypothesis, which asserts that speakers distribute information uniformly across their utterances, and the rational speech act (RSA) model, which suggests that speakers optimize the formulation of their message by reasoning about what the comprehender would understand, have been hypothesized to account for a wide range of language use phenomena. We here specifically focus on the production of discourse connectives. While there is some prior work indicating that discourse connective production may be governed by RSA, that work uses a strongly gamified experimental setting. In this study, we aim to explore whether speakers reason about the interpretation of their conversational partner also in more realistic settings. We thereby systematically vary the task setup to tease apart effects of task instructions and effects of the speaker explicitly seeing the interpretation alternatives for the listener. Our results show that the RSA-predicted effect of connective choice based on reasoning about the listener is only found in the original setting where explicit interpretation alternatives of the listener are available for the speaker. The effect disappears when the speaker has to reason about listener interpretations. We furthermore find that rational effects are amplified by the gamified task setting, indicating that meta-reasoning about the specific task may play an important role and potentially limit the generalizability of the found effects to more naturalistic every-day language use.

Keywords: rational speech act model, discourse processing, discourse connectives, production, experimental pragmatics, crowdsourcing experiment, gamification

OPEN ACCESS

Edited by:
Gerhard Jäger,
University of Tübingen, Germany

Reviewed by:
Melissa Michaud Baese-Berk,
University of Oregon, United States
Ciyang Qing,
University of Amsterdam, Netherlands

****Correspondence:***
Frances Yung
frances@coli.uni-saarland.de

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

Received: *29 January 2021*
Accepted: *19 April 2021*
Published: *28 May 2021*

Citation:
Yung F, Jungbluth J and Demberg V (2021) Limits to the Rational Production of Discourse Connectives. Front. Psychol. 12:660730. doi: 10.3389/fpsyg.2021.660730

1. INTRODUCTION

A speaker faces a number of choices when encoding a discourse relation: they can choose whether to leave it implicit, or mark the relation explicitly using a discourse connective. Discourse connectives (DC) are linguistic devices that signal coherence relations. Discourse theories such as the Rhetorical Structure Theory (RST; Mann and Thompson, 1988) distinguish between a large number of coherence relations and corresponding DCs; however, there is no one-to-one correspondence between them. One discourse relation can be signaled by multiple DCs, and one DC can signal a variety of different discourse relations. For example, a *causal* relation can be marked by *because* or *since*. In turn, *since* can signal a *causal* relation or a *temporal* relation about the starting point of an event.

The speaker thus often also needs to decide between several lexical alternatives for marking a specific discourse relation. The resulting variation in discourse connective choice is to date largely unexplained. We therefore here set out to test whether rational accounts of language processing, such as the uniform information density theory (Levy and Jaeger, 2007; Jaeger, 2010) or the rational speech act theory (Frank and Goodman, 2012) can account for this production choice.

These theories have been proposed to account for a wide range of phenomena in language production including speech articulation and the inclusion of optional syntactic markers (Jaeger and Buz, 2018), as well as referring expression production (Degen et al., 2013; Graf et al., 2016; Degen et al., 2020), omission of pronouns (Chen et al., 2018), ordering of adjectives (Hahn et al., 2018), and expression of exhaustivity (Wilcox and Spector, 2019). While the uniform information density hypothesis is most suitable for studying phenomena where the production variants are meaning-equivalent, the rational speech act theory also involves reasoning about alternative meanings of an utterance, and hence seems best suited for studying the production of discourse connectives. In fact, the RSA model has already been used to account for the distribution of explicit and implicit discourse connectives, and found it to be in line with the qualitative prediction of the RSA model (Yung et al., 2016, 2017). They found that, in the Penn Discourse Treebank (Rashmi et al., 2008), an explicit connective is more often omitted when the it is not informative enough to offset its production cost, or if there are enough other discourse signals in the arguments. The rational speech act theory (RSA) (Frank and Goodman, 2012; Goodman and Frank, 2016) is a formalization of Gricean pragmatics (Grice, 2000). It models and makes quantitative predictions on language production and comprehension in terms of a rational process by which speakers and listeners iteratively reason about each other. According to the RSA, a rational speaker aims at successful communication by calculating how the hearer would understand the speakers' utterance and choosing their utterance by trading off the likelihood that the utterance will successfully communicate the intended meaning against the speaker-effort of producing that utterance. Several variants of the RSA account have been proposed, including the *incremental* RSA (Cohn-Gordon et al., 2019), which allows the model to operate not only on the level of a sentence as a unit for defining successful communication, but holds that speakers may even aim to avoid temporary misunderstandings.

The current work thus seeks to find out whether the choice for a specific discourse connective is the result of a rational choice process in the speaker, who reasons about what discourse inferences the listener might make when hearing a specific discourse connective.

For example, a speaker might prefer the connective *whereas*, which signals *contrast*, over the connective *while*, which can signal both *contrast* and *temporal synchrony*, in a situation where the listener might be expecting a temporal relation, in order to direct listener expectations in the intended direction and avoid the risk of later misunderstandings. On the other hand, a speaker may well choose an ambiguous connective, if the intended coherence relation is easily predictable, and hence easy to disambiguate, by the listener.

Yung and Demberg (2018) set out to test whether connective choice in language production is a rational process as predicted by the RSA account, by setting up a language game experiment. In this experiment, the speaker is asked to express a target discourse relation to a listener by uttering a discourse connective, which either signals the relation unambiguously, or is ambiguous in that it can also signal other relations. The communicative utility of the connectives is determined by the set of possible interpretations that the listener might infer. These are shown explicitly to the speaker in the gamified setting used in Yung and Demberg (2018). In one case, both the ambiguous and unambiguous connective can safely be chosen to signal the relation, as no alternative interpretation that fits these connectives is part of the set of interpretations for the listener. In the other condition, the set of listener interpretations contains two relations that both fit the ambiguous connective. In this case, choosing the unambiguous connective is communicatively most useful, as it uniquely picks out the intended interpretation.

Yung and Demberg (2018) found that speakers in their experiment did choose the unambiguous DC option more often when the ambiguous option could fit with another given interpretation rather than the intended meaning, suggesting that people do reason about the comprehension of the listener. The results of that study were thus in line with the quantitative predictions of the RSA theory. However, there is an obvious gap between this gamified experimental design and naturalistic language use in communication: most importantly, the possible interpretations of a listener are normally not directly available to the speaker, but would have to be inferred. The prior study of Yung and Demberg (2018) thus only allows us to conclude that speakers CAN choose connectives rationally when they have the chance to reason about what the listener may understand, but does not show whether people actually DO make these rather complex inferences during normal language production. The question left unanswered is whether the explicit restriction on the valid interpretations, which only occurs in a gamified setup, is a critical factor that allows the speaker to reason about the listener's mind and make a rational choice, or whether the behavior found in Yung and Demberg (2018) also plays out in naturalistic language production. This is a concern that has been voiced also previously in the context of the RSA model: while the rational account allows to calculate what a perfectly rational speaker should do, there are concerns regarding the cognitive plausibility of the model (Borg, 2012; Carston, 2017; Borg, 2017): it is not always clear whether speakers actually make all those computations in real time every day language use. A specific contribution of this article from the point of view of rational models is that it does not approach this question by manipulating the necessary depth of reasoning or number of alternatives that need to be considered in reasoning, but investigates a case where reasoning needs to be done about an abstract object, namely coherence relations expected by the listener.

The current study aims to fill this gap by assessing the rational account of DC production under more realistic settings. In particular, we test people's discourse connective production choice in a setting where the possible interpretations of the listener are not limited to a specific set and are not explicitly available to the speaker. Instead, we manipulate discourse expectations of the listener, since people are sensitive to various signals in the context and build up expectation about the upcoming coherence relation (Lascarides et al., 1992; Kehler et al., 2008; Rohde et al., 2011; Rohde and Horton, 2014; Scholman et al., 2017, 2020; Schwab and Liu, 2020). In this more natural setting, showing that people prefer unambiguous DCs when there is a higher risk of misinterpretation by the listener (because the expectation is not in line with the actual ending) would substantially strengthen the empirical evidence for the rational account.

We here report on a series of experiments conducted via crowd-sourcing, where we manipulate what information regarding comprehender interpretation options is visible to the speaker. We replicate the effect found in Yung and Demberg (2018) when using the same game-like setup with explicitly given alternative continuations (section 2.4.3), but do not find any effect of discourse-related predictability on connective choice in our experimental settings where these continuations are not shown explicitly (section 3).

In section 4, we report on a follow-up experiment which tests whether the failure to find an effect of contextual constraint on connective choice is due to the lack of showing these alternatives, or whether it could be related to feedback during the experiment or other factors in the experimental setup which might encourage explicit reasoning about speaker interpretation.

Our results indicate that experimental design has a sizeable effect on connective choice—the game-like setting leads to more unambiguous connectives being chosen than more naturalistic designs. This brings up the question to what extent the results from gamified language tasks generalize to every-day language comprehension and production, or whether they are constrained to tasks that involve more explicit meta-reasoning.

This case study on DC production indicates that an easily calculable or explicitly available set of alternative interpretations is crucial for speakers to perform RSA-style reasoning. Overall, this study shows that accounts of rational language production might not be able to account for connective choice in everyday language communication.

1.1. Background

1.1.1. The Rational Speech Act Theory

The rational speech act model (RSA) is a Bayesian computational framework based on Gricean pragmatic principles, which state that speakers try to be informative based on the knowledge shared with the listeners. Formally, given an intended meaning m to be conveyed, the pragmatic speaker in the RSA model chooses a particular utterance u' from a number of alternative utterances that are compatible with meaning m. The probability that the speaker chooses u' is proportional to the *utility* of u' with respect to m and the shared background C (Equation 1). The *utility* of an utterance depends on the *cost* for the speaker to produce it and its *informativity*, which is quantified as the log probability of the listener inferring meaning m when they hear the utterance (see Equation 2). In the basic RSA model, the speaker reasons about a *literal* listener, who chooses an interpretation that is compatible with the utterance in context (Equation 3).

$$S_{prag}(u'|m,C) \propto e^{\alpha\, utility(u';m,C)} \tag{1}$$

$$utility(u;m,C) = \log L_{lit}(m|u,C) - cost(u) \tag{2}$$

$$L_{lit}(m'|u,C) \propto P(m',C) \tag{3}$$

The *utility*, in the basic RSA model, is based on the comprehension of the listener after the complete utterance is processed.

The *incremental* RSA (Cohn-Gordon et al., 2019) further considers the informativeness of the incomplete utterance; it optimizes the utility of the next unit of production (e.g., word) where the context C is defined as the partial sentence uttered so far. Based on this modified version, speakers should choose their words such that temporary misunderstandings on the part of the listener are also avoided.

The RSA and other Bayesian rational accounts of language processing are supported by a set of experimental data on human communication, spanning a wide range of language phenomena (see Goodman and Frank, 2016 for an overview). A set of empirical study results can speak to the consideration of alternative interpretations and alternative utterances during language processing: it has been shown that the existence of alternative interpretations for an utterance affects the listener's processing of the actual utterance (Beun and Cremers, 1998; Bergen et al., 2012; Degen et al., 2013; Degen, 2013; Degen and Tanenhaus, 2015, 2016) and that speakers are sensitive to the informativity of referring expressions given the choices of objects in context (Olson, 1970; Brennan and Clark, 1996; Brown-Schmidt and Tanenhaus, 2006, 2008; Yoon and Brown-Schmidt, 2013). For example, while *"trousers"* is specific enough for the listener in a context containing a pair of jeans and a shirt, it would be ambiguous if there is also a pair of sweatpants. In turn, the speaker would avoid using the generic term *"trousers"* and prefer the more specific *"jeans"* to refer to the pair of jeans in the latter context. The availability of alternative utterances also matters. For example, Degen and Tanenhaus (2016) demonstrated that the processing of the scalar *"some"* is delayed in a context where the speaker is allowed to use exact numbers compared to when that option is not available.

The iterative reasoning between the speaker and listener proposed by RSA is in line with the literature on perspective-taking in the formulation and interpretation of utterances, which states that people generally take into account the knowledge and perspectives of their interlocutors (Stalnaker, 1978; Sperber and Wilson, 1986; Clark and Brennan, 1991; Clark, 1992, 1996; Pickering and Garrod, 2004; Barr and Keysar, 2005, 2006; Bard et al., 2000; Galati and Brennan, 2010; Pickering and Garrod, 2013; Ryskin et al., 2015).

For instance, a number of studies on referring expression production report that speakers generally adapt their production preferences to the knowledge of the listeners (Isaacs and Clark, 1987; Wilkes-Gibbs and Clark, 1992; Nadig and Sedivy, 2002; Yoon et al., 2012). Similar rational production behavior has also been found regarding the omission of pronouns when the referent is clearly understood (Chen et al., 2018) and the preference order of subjective adjectives (Hahn et al., 2018). One characteristic of the production scenarios that were examined is that the *intended meaning* is a concrete object with certain clearly distinguishable properties; the representation of its meaning does not require a high level of abstraction. It is as of yet unclear whether speakers can also reason about the informativity of an utterance when the meaning to convey is an abstract one, such as a coherence relation between segments of texts. We will discuss the evidence for comprehenders forming discourse expectations during comprehension below.

1.1.2. Language Game Experiments

Many studies on RSA and perspective-taking make use of referential language games to test people's interpretation or production of referring expressions (e.g., Frank and Goodman, 2012; Degen et al., 2013; Vogel et al., 2014; Franke and Degen, 2016; Mozuraitis et al., 2018; Ryskin et al., 2015; Kreiss and Degen, 2020; Ryskin et al., 2020). Typically, in these game-styled experiments, a limited set of objects are presented to the participants, who are asked to interpret which object a particular referring expression refers to, or are asked to utter an expression to refer to a particular object. These studies typically show large RSA-consistent effects. This experimental paradigm defines toy worlds where the possible interpretations are limited to a controlled set of objects and allows the researchers to precisely manipulate the knowledge accessible to the speaker and the listener. For example, Ryskin et al. (2015) design privileged perspective where one interlocutor sees certain alternative objects in the context while the other does not. Despite these advantages related to experimental control, it has to be noted that these artificial settings are a simplification of the situation in the real world, where the possible interpretations are usually not limited, at least not as explicitly. In section 1.1.3, we will discuss the game-styled experiment presented in Yung and Demberg (2018), which captures the speaker's preference of DC choices when misinterpretation is (im)possible.

A more open-ended experimental environment is explored by a series of studies in Sulik and Lupyan (2016, 2018a,b). They use a signaling task where participants are asked to provide a single cue word to give a hint for the partner to guess a target word. In their setup, no alternative options of hints or target words are given, and the results show that the director uses salience information from their own perspective rather than that of the guesser. Participants' performance in choosing a cue word following the guesser's perspective can be improved with added contextual constraints and repeated interactions with feedback provided by the guesser, but further studies find that the improvement is based on other heuristics rather than better reasoning of their partner's perspective (Nedergaard and Smith, 2020). These findings suggest that people do not seem to reason about their listener's perspective in situations where the alternatives are completely unconstrained or unknown. However, the signaling game is a highly demanding task: it is not straightforward for the participant to come up with a cue from the guesser's perspective even if they actually try to do so.

These findings are consistent with studies in more complex perspective-taking settings, which suggest that there may be limitations to rational processing, showing that people do in fact often not behave optimally from the perspective of rational models. In tasks that require perspective-taking, i.e., when the speaker is aware that the information available to the listener is different from their own information, speakers tend to prioritize their own perspective when they are under time pressure (Horton and Keysar, 1996), or in situations where the information on their perspective is more salient (Lane and Ferreira, 2008). The recent study of Vogels et al. (2020) also found that speakers do not adapt their production to the cognitive load of the listeners on a fine-grained level, but rather adopt a very coarse strategy that they then follow: when the driver in a simulated driving task was under cognitive load, the speaker only made their utterances more redundant (easier to understand) if they had previously experienced the difficulty of the driver task themselves, and didn't adapt their strategy for trials where the cognitive load on the driver was lower.

In addition to the limitation on the alternative interpretations, the settings of game-styled experiments might entice people to engage in more extensive reasoning than usual, in order to guess the correct answer of the "riddle." In particular, Sikos et al. (2019) found that, comparing with one-shot web-based experiments, increasing the participant's engagement in the task leads the participants to follow more closely to reasoning based on RSA, while the results of one-shot games are in some cases better fitted by simpler models based on literal interpretation (Qing and Franke, 2015; Frank et al., 2016; Sikos et al., 2019).

Taken together, results from referential language games show that people *can* reason about the reasoning of their interlocutors, but it is not clear if they would actually perform the same reasoning in everyday language use. Furthermore, prior findings indicate that speakers may not always have the capacity to behave optimally, even if they may strive to do so, and that they are happy to follow coarse heuristics for successful communication instead of reasoning on an utterance-by-utterance basis. A reason for this observation could be that maintaining a detailed mental model of the addressee's needs may be cognitively costly (Koolen et al., 2011; Horton and Keysar, 1996; Roßnagel, 2000).

1.1.3. Language Game for DC Production (Yung and Demberg, 2018)

A gamified experimental design, similar to other RSA studies, is used in Yung and Demberg (2018) to compare the qualitative prediction of RSA against the choice of human subjects. The design adapts the language games of referring expression production for DC production. An example of the stimuli used is shown below.

Example item from Yung and Demberg (2018):

That tennis player has been losing his matches...
Options: *since / as / but*

A. **(Target production)... we know he is still recovering from the injury.**
B1. ...the season started.. /B2....he was close in every match.
C. ...his coach believes that he still has chance.

In this experiment, the subjects act as speakers. They are given the first half of a sentence (*That tennis player has been losing his matches*) and a continuation which represents the speaker's communicative goal (continuation A: .. *we know he is still recovering from the injury.*) They are asked to choose one connective from one of the three given options (e.g., *since*, *as*, and *but*) to provide a "hint" to the other player regarding the intended target relation (*continuation A* as the target continuation out of three given continuations: A, B1 or B2, and C.

Furthermore, the speaker in the game can also see a set of alternative discourse continuations which the listener could possibly infer (continuation C and either continuation B1 or B2, depending on the condition). The subjects are told that the sets of connectives and continuations are also visible to the listener player, except that continuation A is the target.

The set of connectives and the set of discourse continuations are manipulated. Two of the connective options (*since* and *as*) can be used to mark the target relation (continuation A), but one of them is ambiguous (*since*, which fits both continuations A and B1) and the other is unambiguous (*because*, which fits only with continuation A).

The set of alternative continuations is set up such that it does or does not contain a continuation that is compatible with the other reading of the ambiguous connective (continuation B1, which matches the temporal reading of *since*). This manipulation of continuations thus creates a toy situation where mis-guessing is possible (including B1 in the alternative set), or not (including B2 instead of B1). Under this gamified setting, it was found that speakers do choose and unambiguous DC significantly more often in the former situation. In the current study, we are set to find out if the result still holds under a more naturalistic setting, where misinterpretation is manipulated by discourse expectation.

1.1.4. Discourse Expectations

A variety of studies have shown that comprehenders use a range of cues to anticipate the continuation of the discourse (Sanders and Noordman, 2000; Rohde et al., 2011; Canestrelli et al., 2013; Köhne and Demberg, 2013; Rohde and Horton, 2014; Drenhaus et al., 2014; Xiang and Kuperberg, 2015; Scholman et al., 2017; Van Bergen and Bosker, 2018). Relevant cues include discourse connectives, as well as more subtle signals such as implicit causality verbs and negation. Köhne and Demberg (2013), for instance, found that people have different expectation about the upcoming discourse after reading a *causal* connective (e.g., *therefore*) vs. a*concession* connective (e.g., *however*). Similar results were also found in related studies such as Drenhaus et al. (2014), Xiang and Kuperberg (2015). People are also sensitive to more implicit signals apart from explicit connectives. For example, comprehenders anticipate a causal relation after encountering implicit causality verbs, such as *blame* (Rohde and Horton, 2014).

Apart from lexical signals, context is also considered to be important for the interpretation of discourse (Sanders et al., 1992; Lascarides et al., 1992; Cornish, 2009; Spooren and Degand, 2010; Song, 2010). Contextual signals that influence the expectation of a particular coherence relation are not limited to specific words that occur locally in the segments of texts joined by relation, but could locate in the more global context. For example, Scholman et al. (2020) showed that, in a story continuation task, people generate more *list* relations following a context where several similar events occurred, e.g., *"the woman experienced several unfortunate events last night. She got wine thrown at her by her dining companion..."*. However, the sensitivity to such contextual signal was shown to vary between different people: while some showed very high sensitivity, others seemed to ignore the signal, or not be able to take it into account Scholman et al. (2020). Furthermore, Schwab and Liu (2020) found that contrasting information in the context, e.g., *"he likes to run outdoors. He has a treadmill in the living room..."* facilitates the processing of a *concession* relation.

These works point to the fact that comprehenders generate expectation about the upcoming discourse continuation based on lexical and contextual cues in the preceding contexts. The current study aims at investigating the effect of contextual discourse expectation in combination with a rational account of connective production.

2. MATERIAL CONSTRUCTION AND METHODS

The objective of this study is to find out whether speakers choose discourse connectives rationally in a more naturalistic setting. Specifically, when the possible interpretations are not restricted and are not explicitly shown to the speaker, do speakers still reason about the listeners' difficulty in interpretation and prefer a disambiguating connective if the comprehender is likely to make the wrong inference?

In our experimental design, it is thus necessary to manipulate how easily the intended discourse relation can be inferred, without explicitly listing the possible interpretations. Our materials are constructed based on the strategy used in Yung and Demberg (2018). However, instead of limiting the possible interpretations allowed in the game, we propose to manipulate the interpretation difficulty by means of *contextual expectation*. We hypothesize that the target discourse relation is expected to be more difficult to infer in a context where a different coherence relation is expected, compared to a situation where the target discourse relation itself is highly expected. For instance, referring to the example presented in section 1.1.3, we create a contextual situation where the listener is expecting a reason (e.g., *The drop in performance of the tennis player was not coincidental. He has been losing his matches BECAUSE...*) or a specification of time (e.g., *Let me tell you how long that tennis player has been disappointing his fans. He has been losing his matches SINCE...*). The alternative interpretations, on the other hand, are not limited nor visible to the subjects. Following the prediction of the RSA model, speakers should use a more specific DC to express an unexpected discourse relation, while they may safely use an ambiguous connective if the target relation is already expected anyway. The construction of the stimuli will be explained in more details in the following subsection.

It is worth noting that the stimuli might not work properly if both meanings of the ambiguous connective are compatible with the target continuation. For example, in the sentence *"That tennis player has been losing his matches ____________ he changed his coach,"* the second clause can be read as a reason or a specification

TABLE 1 | Material construction pattern illustrated with a concrete example.

1a	Context for TA	*Chris is a professional artist and so is his wife. However, his talent is very different from hers:*
1b	Context for CA	*I am going to the music festival with my friends next week. I look forward to a particular performance by a musician who can play two instruments at the same time:*
1c	Neutral context	*I had a very nice lunch with my old friend Chris today. I haven't seen him in a long time. Chris loves music:*
2	Arg1	*he plays the saxophone*
3	Connective choice	*while* (TA / CA), *whereas* (TA), *specifically* (other)
4a	TA (target Arg2)	*his wife is a ballet dancer*
4b	CA (competitor Arg2)	*he accompanies himself on the drums.*

TA stands in a contrastive relation to Arg1 in this example, while CA stands in a temporal-synchronous relationship with Arg1.

of time. The choice of *since* (ambiguous DC) instead of *ever since* (unambiguous DC for the temporal relation) may not be due to the adjustment in ambiguity level that we would like to test, but rather because the reason reading is preferred by the subject. In other words, the alternative relation senses of the ambiguous DCs elicited by the stimuli have to be distinctive enough. To verify this, we conducted a pretest on the stimuli on another group of subjects. The details will be explained in section 2.4.1.

Another necessary verification of the experimental materials is to test whether the situational contexts of the items do increase the expectation of a particular discourse relation as we expect in the design. Section 2.4.2 describes the pretest we carried out to verify this. Finally, the newly constructed stimuli should also work in the gamified setting. We thus try to replicate the results of Yung and Demberg (2018) with the new set of stimuli in our third pretest (section 2.4.3).

2.1. Stimuli Construction

The pattern of our experimental stimuli is as follows: We determine two alternative discourse relations, the target relation (TR) and the competitor relation (CR). Next, we select a pair of connectives such that one of the connectives is ambiguous and can signal both TR and CR, while the other connective is unambiguous and can only signal TR. We then need to design a discourse relational argument Arg1 which is compatible with either relation, and two continuations, one conveying the target relation, and the other conveying the competitor relation. We denote these relational arguments as TA and CA respectively. Finally, in order to manipulate which of the coherence relations is expected, we construct two different contexts that raise discourse expectations for each of these relations. As a baseline, we also add a neutral context, and a third unrelated connective which marks neither TR nor CR. An example of an item is given in **Table 1**.

In this example, the target discourse relation to be produced is a CONTRAST relation, between *"he plays the saxophone"* and *"his wife is a ballet dancer."* We call *"he plays the saxophone"*' the *first argument*, abbreviated Arg1 of the discourse relation, and *"his wife is a ballet dancer"* the *target second argument* (TA). The TA is a specific instantiation of the abstract relation type to be produced by the speaker, and connectives that mark the relation type are provided as options for the speaker to choose from. Among the provided options, both *while* and *whereas* mark a contrast relation, but *while* is more ambiguous because it can also mark the temporal relation between two events happening at the same time. On the other hand, *and specifically* does not fit the *target continuation*. We call *while*, *whereas* and *and specifically* the *ambiguous*, *unambiguous* and *incompatible* DC respectively. The *incompatible* DC is chosen such that the relation it signals is considerably different from any of the relations signaled by the *ambiguous* and *unambiguous* DCs.

In our experiment, the speaker will see one of the contexts (1a, 1b, or 1c) and the first argument (2), and will be asked to choose among the three connectives (3). The speaker will also see the intended second argument (4a). The competitor second argument (4b) will never be shown, it thus remains implicit. We constructed a total of 62 items following this pattern.

According to the RSA, it is rational to prefer the unambiguous connective *whereas* over the ambiguous connective *while*, especially in a context where the competitor argument (CA) is contextually expected: selecting the ambiguous connective which is compatible with CA would leave the comprehender on the wrong track and lead to difficulty in inferring the correct continuation TA.

In order for the stimuli to work in the intended way, it is important for the two coherence relations that are marked by the ambiguous connective to be distinct from one another, such that the unambiguous connective intended to mark only the target relation TR is not compatible with the competitor relation CR. We therefore selected three connectives, *since*, *as*, and *while* as the ambiguous connectives in our experiment, as they each signal two relations that are distinct from one another. **Table 2** summarizes the target discourse relations and the DC options covered by the stimuli. The intended mismatch between the unambiguous connective and the competitor second argument is tested in our first pretest, see section 2.4.1.

2.2. Participants

All pretests and experiments reported in this article were conducted online via the crowd-sourcing platform Prolific. Participants were restricted to English native speakers currently residing in English-speaking countries. Also, only participants with past approval rates of 99% or more were selected. Details on the participants will be reported in each specific experiment.

2.3. Procedure

In the beginning of the experiment, the participants were informed that collected data will be used for research purposes and that all data will be anonymized prior to analysis. They were also informed that there are no risks or benefits to participating in the study and their contribution is voluntary, and thus they might decline further participation, at any time, without adverse consequences. The participants' consent and confirmation of

TABLE 2 | Summary of the stimuli.

Ambiguous DC	Target discourse relation	Unambiguous DC options	Stimulus count
since	CAUSAL	*because*	10
since	PRECEDENCE	*ever since*	10
as	CAUSAL	*because*	10
as	SYNCHRONOUS	*when, while, at the same time as, etc.*	10
while	CONTRAST	*whereas, but*	11
while	SYNCHRONOUS	*when, as, during the time when, etc.*	11

The incompatible DCs used in the stimuli include if, unless, in other words, for example, so that etc.

being at least 18 years old were obtained before the start of the experiment.

2.4. Norming and Pretests

We conducted three pretests to make sure that our stimuli work as intended. The first pretest was run to validate whether the unambiguous connective is indeed incompatible with the competitor relation. This pretest is reported in section 2.4.1 below. The second pretest aims at testing whether the biasing contexts 1a vs. 1b indeed raise different discourse expectations (4a vs. 4b), and is reported in section 2.4.2. Finally, we repeated the experimental setup described in Yung and Demberg (2018) with our new materials, in order to check whether we can replicate their results (section 2.4.3).

2.4.1. Pretest 1: Validation Relation Interpretations

One difficulty in stimulus design is that the relations themselves can sometimes be ambiguous. In those cases, a participant might infer both readings, or the participant may only infer one reading, but we don't know ahead of time which one. Both of these cases are problematic.

For an example of a case where the participant may infer both readings, consider the sentence *John started to clean his flat regularly* **since** *his girlfriend moved in*. In this example, *his girlfriend moved in* could be the reason, or just the marker of the specific time. Both of the unambiguous markers (*because* for the causal reading and *ever since* for the temporal reading) would in that case be compatible with the continuation, and hence there would be no rational advantage to choosing the unambiguous connective over the ambiguous one.

If, on the other hand, a participant only infers one of these relations, we also have a problem because we don't know ahead of time which one it will be and what connective we should hence provide as the unambiguous alternative. For instance, if the participant interprets *his girlfriend moved in* as the continuation of a temporal relation (the CR), then *because* is no longer a valid marker for the TR in the stimulus. Hence, we do not want to include sentences where both the TR and CR are possible.

The objective of this pretest is thus to confirm that the target continuation of each stimulus represents a discourse relation that is highly distinguishable from the competitor discourse relation. Accordingly, the acceptability of each connective option is tested with respect to the intended discourse relation.

2.4.1.1. Materials and Procedure

The pretest was carried out in the form of a coherence rating task. We created two sentences for each experimental item by inserting the unambiguous connective and an unambiguous connective expressing only the competitor relation between the first argument and target second argument as shown in the following example.

Stimulus item:

James has been studying very hard ____________ he entered secondary school 2 years ago.
(Options: *since*, *ever since*, *instead*)

Pretest items:

1. connective compatible only with TA:
James has been studying very hard **ever since** he entered secondary school 2 years ago.
2. connective compatible only with CA:
James has been studying very hard ***because** he entered secondary school 2 years ago.
3. ambiguous connective:
James has been studying very hard **since** he entered secondary school 2 years ago.
4. incompatible connective:
James has been studying very hard ***instead** he entered secondary school 2 years ago.

Participants were asked to rate the coherence of each pretest item on a scale of 1 (least acceptable) to 4 (most acceptable). They could also optionally suggest a word or phrase to replace the bold DC to improve the acceptability of the sentence. This additional feedback provided suggestions for the improvement of the stimuli. Since the focus of this pretest is the discourse relation between the first argument and the target continuation, preceding contexts are not included in the pretest items.

For items that work as intended, variant (1) with the unambiguous connective from the original item should be judged to be substantially better than variant (2). Furthermore, variant (3) verifies if the ambiguous connective fits the original item and variant (4) confirms the incompatible connective is not acceptable. Hence, variant (3) should be judged with high ratings while variant (4) should be rated worse.

2.4.1.2. Participants

The items were distributed evenly across 16 lists, and each list was completed by 15 participants. Each participant only saw one version of an item. They also did not see items sharing the same first arguments. A total of 411 participants (age range: 20–75, mean age: 36, 257 females) took part in several rounds of the pretest. They were recruited via the crowdsourcing platform Prolific according to the criteria described in section 2.2.

2.4.1.3. Analysis

We define the *semantic gap between alternative discourse relations* of a stimulus based on the difference in the average rating of the intended and unintended version of the pretest item. For example, the average ratings of pretest items 1 and 2 shown above were 3.87 and 2.27, respectively. The *semantic gap* is thus $3.87 - 2.27 = 1.60$, which can be normalized to 53% based on the maximally possible difference of 3. Stimuli with semantic gap below 5% were replaced or revised. The revised stimuli underwent another round of pretest. Several rounds of pretests were conducted on several subsets of the items until the semantic gaps of all items were above 5%. The results of the final version of the items were collected from a total of 360 participants.

2.4.1.4. Results

The average coherence rating for the final items was 3.47 for the variant with the unambiguous connective fitting with target second argument TA, and 1.59 for the unambiguous connective that fits the competitor second argument CA. Ambiguous DCs and incompatible DCs received average ratings of 3.15 and 1.25, respectively. The semantic gap between final versions of the stimuli ranged between 5 and 96%, with an average of 62%.

2.4.2. Pretest 2: Validation of Target- and Competitor-Predicting Contexts

The second pretest is performed to confirm the contextual conditions of the stimuli. Referring to the example shown in **Table 1**, we want to make sure that the target-predicting contextual condition (1a) raises the prediction for a contrastive relation and fits together with the target relational argument (4a). On the other hand, the competitor-predictive context (1b) should be predictive of a temporal synchronous relation and should fit with competitor continuation (4b), but not vice versa.

2.4.2.1. Materials and Procedure

The pretest was formulated as a forced choice task in which participants were asked to select the discourse continuation that best fit the context, see the following example:

Pretest items:

(1a) Context A, here CONTRAST-predicting context:
Chris is a professional artist and so is his wife. However, his talent is very different from hers: he plays the saxophone

(1b) Context B, here SYNCHRONOUS-predicting context:
I am going to the music festival with my friends next week. I look forward to the particular performance by a musician who can play two instruments at the same time: he plays the saxophone

(2a) Continuation fitting Context A, here CONTRAST:
...whereas his wife is a ballet dancer.

(2b) Continuation fitting Context B, here SYNCHRONOUS:
...at the same time as he accompanies himself on the drums.

The order of the two options was randomized in the study.

2.4.2.2. Participants

The items were distributed evenly among 9 lists, such that each item was responded to by 15 participants. Like in the previous pretest, each participant only saw one condition of each item. Across several rounds of pretests, we recruited a total of 263 participants (age range: 22–74, mean age: 36, 188 females) via Prolific, based on the same criteria as mentioned above, and excluding participants who had taken part in the previous pretest.

2.4.2.3. Analysis

We define the *contextual gap* between target- and competitor-predicting contexts based on the difference in the number of participants choosing the matching vs. non-matching continuations. For example, 14 participants chose continuation (2a) when given context (1a), and 0 participants chose continuation (2a) when given context (1b). The score of *contextual gap* of this stimulus pair is thus $14 - 0 = 14$, which can be normalized to 93% based on the possible range of $0 - 15$. Stimulus pairs with a contextual gap below 25% were replaced or revised.

Several rounds of pretests were conducted such that the final version of the items all have a contextual gap larger than 25%. The results of the final version were collected from 135 participants.

2.4.2.4. Results

The mean number of votes of the expected and unexpected relations are 12.48 (SD=2.51) and 2.52 (SD=2.51) respectively, showing that the situational contexts used in the stimuli do trigger the expectation of one discourse relation in comparison to the alternative relation signaled by the ambiguous DC. The average contextual gap for the final stimuli was 68%, ranging from 27 to 93%.

2.4.3. Pretest 3: Replication of Yung and Demberg (2018)

The final pretest aims at verifying whether the created stimuli can elicit pragmatic inference under setting used in Yung and Demberg (2018), where the alternative Arg2 continuations are shown to the speaker explicitly.

2.4.3.1. Materials and Procedure

As contextual prediction is less relevant when the alternative continuations are presented explicitly, we performed this pretest using the *neutral* contexts. The speaker is shown the context, a choice of three connectives, and a set of three alternative second arguments. The speaker is told that the listener will have to guess which argument is the correct continuation, based on the connective that the speaker provides as a cue.

The three alternative second arguments consist of the target argument TA (continuation A in the below example), the competitor argument CA (continuation B1 below) and an unrelated completion C which is linked to the first argument via a different coherence relation. The target argument TA is indicated to the speaker by bold font. The experimental condition displaying options A, B1, and C corresponds to the CA-predictive context, for which a rational speaker should prefer the unambiguous marker *whereas* to mark relation A. A second condition in this experiment consists of continuations A, B2, and C. This condition corresponds to the TA-predictive context; here, both *while* and *whereas* signal relation TA unambiguously, therefore, the choice between them doesn't matter in this condition.	Here is an example of the items used in the pretest.

Pretest 3 item:

> I had a very nice lunch with my old friend Chris today. I haven't seen him in a long time. Chris loves music: he plays the saxophone...

Options: *while / whereas / specifically*

A. ... his wife is a ballet dancer.

B1. (with competitor) ... he accompanies himself on the drums.

B2. (no competitor) ... he plays it every evening after dinner.

C. ... he is good at playing jazz.

We also constructed filler items, which had the same format as the test items, except that the target was continuation B or C. In the fillers, only one of the provided DCs thus fit the target continuation. In the example, only *while* fits continuation B1 and only *specifically* fits continuations B2 and C.

The second player was programmed to be a rational listener, i.e., the simulated hearer would choose continuation A if the speaker selected the unambiguous connective, and continuation B if the speaker selected the ambiguous connective. In the unambiguous condition, the simulated player was programmed to choose option A for both connectives. The participant received one point if the guess of the hearer was correct. At the end of the experiment, bonuses were issued based on the total points. The bonus system encourages participants to engage more in the communicative task.

The items were evenly distributed into 12 lists. Each list contained 10-11 items and fillers. The conditions, discourse connectives and relation types were fully counterbalanced. The target continuation was always presented to the speaker as continuation A, while the other two alternative continuation were randomly assigned to B and C. The order of the three discourse connectives was also randomized per participant.

2.4.3.2. Participants

We recruited 180 participants (age range: 19–71; mean age: 34; 99 females) via Prolific under the same criteria as the other studies, excluding participants who had taken part in the previous pretest. Participants who chose 4 or more non-matching connectives were replaced.

The participants were assigned evenly to the 12 lists; each participant saw 10-11 experimental items and 10-11 fillers.

2.4.3.3. Analysis

We analyzed the data using a Binomial Liner Mixed-Effects Regression Model (*lme4* implementation in R, Bates et al., 2015), with connective choice as a response variable and continuation set as a predictor. The unambiguous DC was coded as 1, and the ambiguous connective as coded as 0. The models reported below include random intercepts by participant, as well as random intercepts and slopes for continuation set by item. Random slopes by participant had to be removed since they couldn't be effectively estimated by the model (their random effects correlation was 1.0).

2.4.3.4. Results

The linear mixed effects analysis reveals a significant effect of condition (what options are shown as possible continuations) on connective choice $\beta = 0.560$; $SE = 0.138$ $z = 4.049, p < 0.001$. This finding is consistent with the results by Yung and Demberg (2018) and indicates that the presence of a competitor relation in the alternative options increases the preference for the unambiguous connective.

Figure 1 compares the results of this pretest with those from Yung and Demberg (2018). While Yung and Demberg (2018) found that speakers did not have a preference between the ambiguous and unambiguous DCs in the *no competitor* condition, the results for our new items show a general preference for the unambiguous DC, even when there is no ambiguity.

2.4.3.5. Discussion

We believe that this discrepancy in results can be attributed to the differences in the stimuli we use: our stimuli include a different distribution of ambiguous DCs and their unambiguous alternatives compared to Yung and Demberg (2018). For example, the unambiguous DC *because*, which is a very frequent marker, is used in our stimuli as the unambiguous option for a *causal* relation, but it is not included as an option in Yung and Demberg (2018). On the other hand, the ambiguous DC *while* is frequently used to mark the synchronicity of two continuous events, while the unambiguous version *at the same time as* is much rarer. This highlights the importance of experimental control over other factors of DC production, such as frequency.

We therefore also included connective identity as a predictor in the model, and found significant differences between the connective pairs with respect to how likely the unambiguous connective was to be chosen by the participants (*since*: $\beta = 0.905$; $SE = 0.347$ $z = 2.609, p < 0.01$; *while*: $\beta = -0.949$; $SE = 0.320$ $z = -2.963, p < 0.01$). These differences did however not change the overall effect of the presence of a competitor second argument on connective choice.

Overall, the pretest results confirm that this set of stimuli can elicit RSA-like rational DC production, in a language game setup where the alternative interpretations are restricted. We next proceed to examine whether similar results can be produced in a more natural setup, i.e., when the possible interpretations are not restricted.

3. EXPERIMENT 1: SPEAKER'S CHOICE OF DCS WHEN THE INTERPRETATIONS ARE UNRESTRICTED

The objective of this work is to examine whether the explicit availability of the comprehensible discourse relations—an artificial situation presented in a language game experiment—is a crucial factor for speakers to rationally choose a connective. To this end, in our first experiment, we replace this experimental design choice by creating an "invisible" set of alternatives based on the contextual predictions which should lead the comprehenders to expect a specific discourse relation, even if it is not explicitly shown. Our goal is to test whether the speaker's

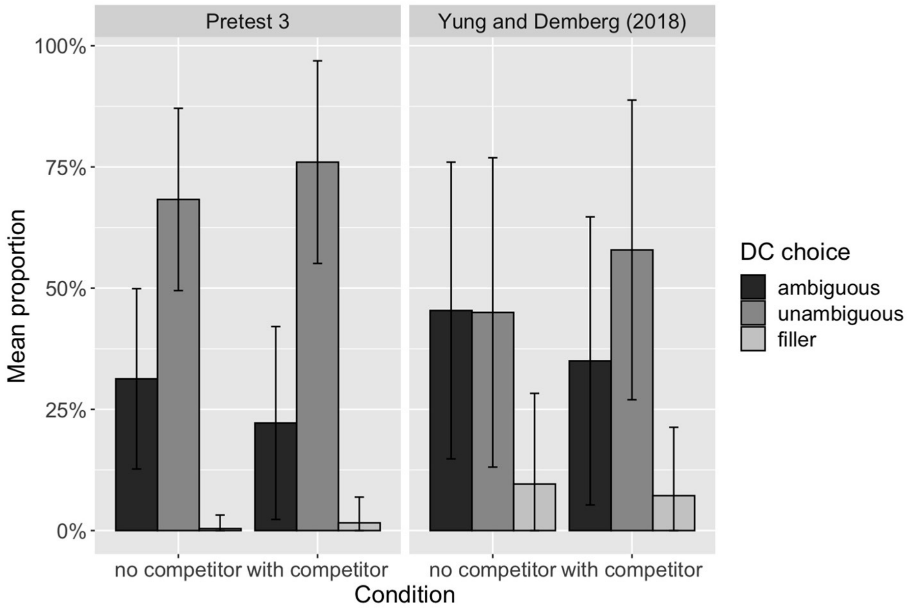


FIGURE 1 | Distribution of the proportion of DC choices made by each participant in the language game experiment of Pretest 3 and Yung and Demberg (2018).

preference between an ambiguous and unambiguous DC shows the same tendency as in the language game experiment, is observed.

In this experiment, the possible continuations of the discourse are neither restricted nor explicitly defined—a situation that resembles natural communication more closely. To assess the rational account of connective production, it is however necessary to create a condition where mis-interpretation is predicted, or not, by the speaker. Such manipulation is achieved in the language game design by including a competitor or not in the available interpretations.

Here, we create two contrasting conditions that correspond to the *with* and *without competitor conditions* by manipulating the preceding contexts without restricting the interpretations, as described in section 2.1. A context where the *target discourse relation* is expected corresponds to the *without competitor condition*, as mis-comprehension is less likely. In contrast, the listener may fail to interpret the target relation when *the competitor relation* is contextually expected, and this condition corresponds to the *with competitor condition*. Following the qualitative prediction of RSA, we expect that speakers will choose the unambiguous connective more often when the preceding context elicit the expectation of the competitor relation.

3.1. Procedure and Materials

The 62 stimuli described in section 2.1 were split into 15 lists, each containing 12 stimuli, such that stimuli sharing the same first argument were never included in the same list. The types of ambiguous DCs, target discourse relations and experimental conditions were counterbalanced. Each list also contained 12 filler items which were taken from a total pool of 18 unique fillers. The items and options were presented to each participant in random order. The fillers have the same structure as the actual stimuli, but are always unambiguous. The purpose of the fillers is to avoid expectation from the participants that there are always two correct options per question. The fillers also help us in screening spammers who answer randomly.

The participants were instructed to imagine that they were reading the sentences to a friend over the phone, but one of the words was blurred and illegible, and they should choose a word from the options to replace it.

3.2. Participants

Two hundred and twenty-five native English speakers (age range: 19–70, mean age: 38, 125 females) were recruited via Prolific.ac. 144 of them reside in the U.K, 54 in the U.S. and the rest in Australia or Canada. They did not take part in any of the pretests. They took an average of 10 min to finish the task and were awarded 1.34 GBP for their contribution. 16 workers who had more than 10% wrong answers (choosing a DC that does not match the target continuation) were removed and replaced.

3.3. Analysis

We used a binomial linear mixed effects regression model to analyze the effect of the three contextual conditions on DC choice. Again, the unambiguous connective was coded as 1 and the ambiguous connective as 0. Context type was dummy coded, with the competitor predicting context as the base level. Random by-participant and by-item intercepts as well as by-item slopes for the contextual condition were included. We furthermore included semantic gap and contextual gap, which were estimated as part of pretests 1 and 2, as covariates in the model, to account for differences between the items. Responses choosing the *incompatible* DCs were not taken into account. Additionally,

TABLE 3 | Regression coefficients of the binomial linear mixed effects model for Experiment 1.

Variable	β	SE	z	p
Intercept	−0.006	0.765	−0.008	0.994
Target-predicting context	−0.022	0.118	−0.184	0.854
Neutral context	−0.123	0.128	−0.958	0.338
Semantic gap	1.755	0.694	2.529	0.011*
Contextual gap	−0.395	0.786	−0.502	0.616

*$p < 0.05$.

we also performed a Bayes Factor analysis using full Bayesian multilevel models. The Bayesian inferences were done using Markov Chain Monte Carlo (MCMC) sampling with 4 chains, each with iter = 6,000; warmup = 1,000; thin = 1; post-warmup = 20,000. The models were implemented using the BRMS package in R (Bürkner, 2017). We here report results for the default prior, which is an improper flat prior over the reals. For the effects that were not found significant in the linear mixed effect model, we report the Bayes factor expressed as BF_{01}, indicating the odds for the null hypothesis H0 compared to the H1 based on the data.

3.4. Results

The binomial linear mixed effects regression model showed no difference between the competitor-biasing context condition and the target-biasing context condition ($\beta = -0.022; z = -0.184, p > 0.05$), and also no significant difference between the competitor-biasing context and the neutral context ($\beta = -0.123, z = -0.958, p > 0.05$), see also **Table 3**.

We therefore also ran Bayesian multilevel models. Their results were consistent with the results of the linear mixed effects models, and showed no effect of context (target-biasing context: $t = -0.03$; 95% CI $[-0.26, 0.21]$, neutral context: $t = -0.13$; 95% CI $[-0.39, 0.13]$). The Bayes Factor (BF_{01}) comparing the reduced model without context as predictor (H0) to the model including context as predictor (H1) is 32.88, indicating very strong evidence in support of H0.

The lack of effect is also visualized in **Figure 2**, which displays the proportion of connective choices in the three conditions. Excluding the small number of choices of the incompatible DCs, which can be interpreted as the cases where the participants were not producing the intended target discourse relation, the proportion of unambiguous DC choices are similar, namely 65, 66, and 65% under the *target-predicting*, *competitor-predicting*, and *neutral* conditions, respectively. In contrast to our hypothesis, the *competitor-predicting* condition does not increase the speaker's preference to use an unambiguous DC.

We furthermore find a statistically significant effect of semantic gap on connective choice, see **Table 3**. Items with a large semantic gap between the alternative discourse relations result in a larger proportion of unambiguous DC production compared to items with a smaller semantic gap.

This effect indicates that the unambiguous connective was preferred when the unambiguous connective could clearly not mark the competitor continuation. There was no effect of contextual gap (this is an expected outcome given that the contextual conditions do not affect the DC choice). The interactions between contextual gap and context type [$\chi^2(2) = 2{,}105, p > 0.05$], or between semantic gap and context type [$\chi^2(2) = 2.194, p > 0.05$] did not improve model fit.

3.5. Discussion

The experiment results suggest that the expectation of the forthcoming discourse relation to be produced does not affect the speaker's choice of discourse connective. This means that contextual expectation of the competitor discourse relation does not specifically trigger speakers to use an unambiguous DC to encode the target relation, while explicitly displaying the competitor continuation does, as shown in Pretest 3. A possible explanation would be that people perform RSA-style reasoning only in a game setting, where (i) meta-reasoning about what the listener will choose as a coherence relation is encouraged, and where (ii) reasoning about listener interpretation is facilitated by explicitly showing the alternative interpretations, i.e., this inference does not have to be performed by the speaker, and by rewarding the speaker if the listener would guess correctly. It is thus possible that this setup encouraged deeper reasoning about the task, or facilitated learning: when the rational listener gave a non-target response, the speaker may have used this feedback to adapt their strategy and subsequently avoid ambiguous connectives.

We therefore conducted a follow-up experiment in which we still do not show the alternative possible interpretations by the listener, but try to encourage meta-reasoning to a similar extent as in pretest 3.

4. EXPERIMENT 2: INVESTIGATING THE EFFECT OF EXPERIMENTAL DESIGN

Experiment 1 and pretest 3 yielded different results (pretest 3 was consistent with the RSA hypothesis, while experiment 1 was not). These experiments however differ in two ways: Firstly, pretest 3 explicitly lists the different listener interpretations, while experiment 1 manipulates discourse expectations, without explicitly showing what the discourse expectations are; secondly, the experiments also differ in terms of setup and instructions, specifically, the instructions of pretest 3, which ask the participant to provide the connective cue in order for the listener to guess the correct second argument, might entice participants more strongly to perform meta-level reasoning to gain points in the game, while experiment 1 uses a more naturalistic situational setting.

Experiment 2 thus aims at teasing apart these two factors. We do this by designing the instructions to match the instructions of pretest 3, while still not showing the alternative listener interpretations to the speaker. A comparison between experiments 1 and 2 will then allow us to investigate whether the lack of effect in experiment 1 can be attributed to the difference in study instructions. To this end, we run the first half of the experiment just like pretest 3, thus providing the participants with training and the mindset of pretest 3. We then add a novel condition in the second half of the experiment. In this novel condition, the speaker sees three alternative listener

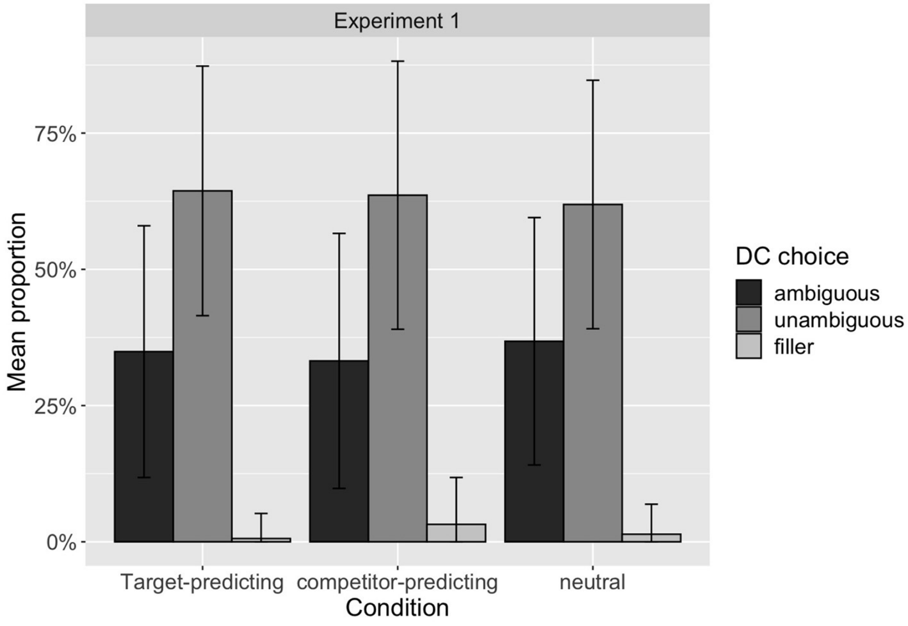


FIGURE 2 | Distribution of proportion of the DC choices made each participant in Experiment 1.

TABLE 4 | An example of a stimulus in various conditions.

1. Preceding context:

Target-predicting condition^{1,2n,2w}	**Competitor-predicting condition^{1,2n,2w}**	**Neutral condition^{p3,1,2n,2w}**
Chris is a professional artist and so is his wife. However, his talent is very different from hers:	I am going to the music festival with my friends next week. I look forward to the particular performance by a musician who can play two instruments at the same time:	I had a very nice lunch with my old friend Chris today. I haven't seen him in a long time. Chris loves music:

2. Core stimulus: first argument and connective choices ^{p3,1,2n,2w}:

he plays the saxophone _______________________ (while / whereas / and specifically,) ...

(*while*=ambiguous DC, *whereas*=unambiguous DC, *and specifically*=incompatible DC)

3. Target and alternative continuations:

No competitor condition^{p3,2n,2w}	**With competitor condition^{p3,2w}**	**Blinded condition^{1,2n,2w}**
between semantic A. his wife is a ballet dancer...	A. his wife is a ballet dancer...	A. his wife is a ballet dancer...
B2. he plays it every evening after dinner...	B1. he accompanies himself on the drums...	B. ■■■■■■■■■■■■■
C. he is good at playing jazz...	C. he is good at playing jazz ...	C. ■■■■■■■■■■■■■

For comparison, experiments including the corresponding conditions are indicated: p3, Pretest 3; 1, Experiment 1; 2n, Experiment 2 (no pragmatic exposure); 2w, Experiment 2 (with pragmatic exposure).

interpretations, but only one of them (the target) is readable, while the other two are blinded, see bottom right cell in **Table 4**. In experiment 2, we again use the three contexts (target-predicting, competitor-predicting and neutral), and balance them across all conditions. Note though that we do not expect an effect of context in the first half of the experiment—here the listener interpretations are shown explicitly and hence overrule any expectations about listener inferences. We do however expect an effect of context in the second half of the experiment, where the alternative continuations are not readable and hence need to be "instantiated" by the speaker based on the predictions derived from the context.

In summary, the most interesting part of the second experiment is its second half: here, the participants have all the instructions and experience just like in pretest 3, but cannot see the alternative continuations, like in experiment 1.

In addition, we want to evaluate how the language game experience in the first half would affect people's performance under the blinded-continuations-condition. Specifically, do people consider the potential risk of ambiguity in the connective, if they haven't seen any effects of ambiguity earlier? And do people adapt their choice based on feedback during the first half of the experiment, such that an incorrect guess by our rational listener may entice the speaker to subsequently prefer the unambiguous connective. To test whether there is such an effect of language game experience, we therefore introduce two training conditions in the first half of the experiment: in the one condition, the alternative continuations explicitly shown to the participants include competitor continuations, and the feedback is from the rational listener, just like in pretest 3, while in the other condition, the participants never see any competitor continuations in the first half, and feedback comes from a literal listener, thus avoiding to give feedback that may specifically encourage rational behavior.

This study will help to shed light on the effect of task formulation on rational reasoning effects in experimental studies.

4.1. Materials

We again use the materials as described in section 2, but add a *blinded* condition. The blinded condition is designed to resemble the situation where the possible discourse continuations are unlimited, because it does not provide any information of the alternative continuations. **Table 4** provides an overview of the conditions of all experiments.

4.2. Procedure

The experiment is based on the language game design used in Pretest 3 (section 2.4.3) that manipulates the alternative continuations, except for the following modifications:

1. Instead of using the *neutral* context in all items, *target-predicting* and *competitor-predicting* contexts are also included as experiment conditions, and are counterbalanced with the *no* and *with competitor* conditions.
2. Each task to be finished by one participant is divided into two halves. In the first half of the task, the alternative continuations are always shown to the participants. For half of the participants, the setup is the same as in pretest 3, for the other half, only the unambiguous condition is included, in which there is never a competitor second argument. In the second half of the task, however, only the target continuation is shown, and the alternative continuations are *blinded*, i.e., NOT readable to the participants. An example is shown in the bottom right corner of **Table 4**.
3. Each task is implemented in two different versions, which we call the *with* and *without pragmatic exposure* versions respectively. The two versions differ in whether the *with competitor* condition is included or not. In the first half of the *with pragmatic exposure* version, half of the stimulus items have a competitor in the given alternatives, just as in Pretest 3, while in the *without pragmatic exposure* version, there are never competitors in the alternative continuations.

To summarize, the first half of the *with pragmatic exposure* version is a 3 × 2 design (*target-predicting/competitor-predicting/neutral* by *with competitor/no competitor*), while the first half of the *without pragmatic exposure* version is a 3 × 1 design (3 contextual conditions by *no competitor*). The second halves of both versions also have 3 conditions (3 contextual conditions, with *blinded* continuations).

Note that the feedback provided by "Player 2" (the listener) is also programmed differently in the two versions. "Player 2" of the *with pragmatic exposure* version reasons about Player 1's choices and answers rationally, while "Player 2" of the *without pragmatic exposure* version will correctly guess the target as long as it's compatible with the chosen connective. Although the alternative continuations are blinded, the guesses made by "Player 2" are shown to the participants as feedback. In the *without pragmatic exposure* version, "Player 2" never guesses a competitor while in the *with pragmatic exposure* version, a competitor is always returned as a feedback whenever an *ambiguous* DC is chosen. An overview of the experimental design is provided in **Table 5**.

With the restriction that each participant does not see the same first argument more than once, the 62 stimuli with counterbalanced contextual and alternative conditions were divided into 60 lists of 31 items each, following the task structure described above. Each half of the task contained 12-13 active stimuli and 2-3 fillers, which were randomly shuffled for each participant within each half of the task. The rest of the procedure is similar to the setup of Pretest 3. The participants were given the same instructions, except that they were also informed that the alternative continuations would be blinded in the second half of the task.

4.3. Participants

Nine hundred native English speakers (age range: 19–85, mean age: 35, 536 females) were recruited on Prolific.ac, and were randomly assigned to the *with* and *without pragmatic exposure* groups. Six hundred and eighty eight of them reside in the U.K, 156 in the U.S. and the rest in Canada, Ireland, Australia or New Zealand. They did not take part in any of the pretests nor Experiment 1. They took an average of 21 min to finish the task and were awarded 1.8 GBP plus and average of 1 GBP bonus for their contribution. Workers who had more than 15 wrong answers were removed and replaced[1].

4.4. Analysis

The experimental design of experiment 2 allows us to address several questions.

1. Is there an effect of contextual constraint on connective choice in the setting with blinded continuation alternatives? For

[1] A wrong answer refers to a DC that does not match the target continuation, so both the ambiguous and unambiguous DCs are considered correct, even in cases where it results in a wrong guess.

TABLE 5 | Task structure of Experiment 2.

	With pragmatic exposure **version**	***Without pragmatic exposure*** **version**
1st half	**12-13 stimuli**: the target continuation matches **both the ambiguous and unambiguous DCs**	
	Contextual condition: target-predicting / competitor-predicting / neutral (counterbalanced)	
	Alternatives: no / with competitor (counterbalanced)	Alternatives: **no competitor only**
	Rational feedback: "Player 2" guesses the competitor continuation if the *ambiguous* DC is chosen under *with competitor* condition, otherwise literal feedback.	**Literal feedback**: "Player 2" guesses the target continuation if either the *ambiguous* or *unambiguous* DC is chosen.
	2-3 fillers: the target continuation matches **the incompatible DC only**	
	Contextual condition: randomly assigned per item; alternatives: no competitor only	
	Literal feedback: "Player 2" (correctly) guesses the target if the "incompatible" DC is chosen.	
2nd half	**12 stimuli**: the target continuation matches **both the ambiguous and unambiguous DCs**	
	Contextual condition: target-prediting/ competitor-predicting/neutral (counterbalanced)	
	Feedback biasing the unambiguous DC: "Player 2" guesses the competitor continuation whenever the ambiguous DC is chosen. The guesses are **unblinded** and displayed.	**Literal feedback**: "Player 2" guesses the target continuation if either the *ambiguous* or *unambiguous* DC is chosen. The guesses are **unblinded** and displayed.
	2-3 fillers: the target continuation matches **the incompatible DC only**	
	Contextual condition: randomly assigned per item; alternatives: blinded	
	Literal feedback: the guesses are unblinded and displayed	

this, we analyse the data from the second half of the second experiment.

2. Does the result from the pretest 3 replicate? We can test this based on the first half of the experiment.
3. Does the experimental task formulation play a major role in connective choice? For this, we will analyse the rate of unambiguous connectives inserted in the first vs. second half of experiment 2, and vs. experiment 1.
4. Finally, we can investigate the effect of pragmatic experience on connective choice: comparing the with pragmatic exposure vs. without pragmatic exposure settings from experiment 2 will allow us to quantify the effect of the language game experience, such as feedback, on communicative success.

For each of these questions, we will analyse different subsets of the data using linear mixed effects regression models in R, as described above. The full random effects structure is used whenever convergence is achieved. When a smaller random effects structure had to be chosen, this will be reported with the specific model. In all analyses, we only consider instances where the participant chose the ambiguous or the unambiguous connective. Cases where the incompatible DC was chosen are ignored in the analysis (this happened only in 3% of cases).

4.5. Results

4.5.1. Effect of Context in Blinded Condition

Our first analysis tests for the main effect of interest: whether discourse connective choice is affected by whether the context is target-predicting or competitor-predicting, when the alternative continuations are not explicitly shown. According to the RSA hypothesis, a rational speaker should prefer the unambiguous connective more strongly in the competitor-predicting condition. The information that the speaker has in this setting is identical to the information available in experiment 1, but this time, the task formulation and instructions are comparable to pretest 3, and participants have already experienced the task with visible alternative continuations during the first part of the experiment. We thus here analyse the second half of the experiment, where the alternative continuation are blinded, and collapse across exposure type (with vs. without pragmatic exposure). Random slopes by participant had to be removed since they couldn't be effectively estimated by the model (their random effects correlation was 1.0). A binomial mixed effects analysis with connective choice as a response variable and context type as a predictor variable shows no significant effect of context type (target-predicting context: $\beta = 0.050$, $z = 0.644$, $p > 0.05$; neutral context: $\beta = 0.061$, $z = 0.840$, $p > 0.05$). We therefore also performed a Bayes Factor analysis with Bayesian multilevel models. The same settings as Experiment 1 were used, except that the number of iterations was increased (4 chains x iter = 10,000; warmup = 1,000; thin = 1; post-warmup = 36,000) due to increased data size, such that the Bayes Factor analysis could converge. In line with the glmer model, the Bayesian multilevel model also shows no effect of context type (target-biasing context: $t = 0.03$; 95% CI $[-0.13, 0.19]$, neutral context: $t = 0.04$; 95% CI $[-0.11, 0.19]$). The Bayes Factor (BF_{01}) comparing the reduced model without context as predictor (H0) over the model including context as predictor (H1) is 368, indicating very strong evidence for H0. The value of BF_{01} is thus about 10 times larger than the BF_{01} we obtained in Experiment 1. We think that this can be explained by the much larger number of observations in experiment 2 ($10,834$ observations from 900 workers vs. $2,741$ observations from 225 workers).

The mean rate of unambiguous connectives is at 74% both in the target-predicting context condition and the competitor-predicting context condition. Again, we find a statistically

TABLE 6 | Regression coefficients of the logistic linear mixed effects model including the responses from the blinded conditions (second half) of Experiment 2.

Variable	β	SE	z	p
Intercept	0.731	0.715	1.023	0.306
Target-predicting context	0.050	0.078	0.644	0.520
Neutral context	0.061	0.073	0.840	0.401
Semantic gap	1.958	0.658	2.974	$< 0.003^{**}$
Contextual gap	−0.978	0.738	−1.325	0.185

$^{**}p < 0.01$.

TABLE 7 | Regression coefficients of the logistic linear mixed effects model including the responses from the first half of Experiment 2.

Variable	β	*SE*	*z*	*p*
Intercept	0.904	0.143	6.303	$< 0.001^{***}$
With competitor	0.446	0.075	5.974	$< 0.001^{***}$
Target-predicting context	0.001	0.068	0.010	0.992
Neutral context	−0.061	0.065	−0.940	0.347

$^{***}p < 0.001$.

significant effect of semantic gap on connective choice, and no effect for contextual gap (see **Table 6**). These findings are consistent with experiment 1, but inconsistent with highly rational connective choice. We note that the overall rate of unambiguous connectives in this experiment is substantially higher than in experiment 1; we will analyse the effects of experimental design in more detail in section 4.5.3.

4.5.2. Replication of Pretest 3

We next analyse the data from the first half of the experiment. The setup here is identical to pretest 3, except that all three different contexts are included, not just the neutral context. We do however not expect any difference between the context conditions, as the possible alternative interpretations of the hearer are shown explicitly. If the contextually predicted alternative is not presented among the alternative continuations, we do not expect this alternative to affect connective choice. However, we do expect to replicate the effect of competitor presence among the explicitly shown alternatives on connective choice.

A binomial linear mixed effects model (see also **Table 7**) showed a statistically significant effect of competitor presence among the explicitly shown alternatives ($\beta = 0.446$, $z = 5.974$, p <0.001), in line with pretest 3. As expected, we do not find a significant effect of either context condition ($\beta = 0.001$, $z = 0.010$, $p > 0.05$) for the target-predicting context compared to the competitor-predicting context, and ($\beta = 0.061$, $z = -0.940$, $p > 0.05$) for the neutral context compared to the competitor-predicting context when alternative completions are shown.

4.5.3. Comparison Across Experimental Designs

Table 8 provides an overview of the proportion of instances where participants chose an *unambiguous* DC instead of an *ambiguous* DC across the different experimental designs in this study. (Cases were participants selected an incompatible DC are not counted in the table.)

We ran a binomial linear mixed effects model with connective type as the response variable and experimental design (with vs. no competitor vs. expt1 vs. blinded with pragmatic exposure vs. blinded without pragmatic exposure) as the predictor variable; the blinded without pragmatic exposure condition was used as the baseline condition, to test whether results are significantly different from the pretest 3 setting or the setting from experiment 1.

First, we found that there are significantly more insertions of unambiguous connectives in the blinded condition with pragmatic exposure, compared to no pragmatic exposure ($\beta = 0.178$, $z = 2.798$, $p < 0.01$). This means that experience with ambiguity in the first half of the experiment does affect participants' connective choices, such that they are more likely to choose unambiguous connectives subsequently. It is possible that this effect is the result of *learning* from unsuccessful communication during the experiment (i.e., where the comprehender chose a competitor completion)

As expected, there is an even stronger effect for the with-competitor condition, where the competitor interpretations are shown explicitly ($\beta = 0.226$, $z = 2.789$, $p < 0.01$), compared to the blinded no pragmatic exposure baseline.

We also find a graded effect in the other direction: there are significantly fewer insertions of unambiguous connectives in the no-competitor condition, where the alternatives are explicitly limited to non-confusable options ($\beta = -0.180$, $z = -3.063$, $p < 0.01$, see also **Table 9**); people choose unambiguous connectives less often when they know that the alternatives don't include any instances which would lead to misunderstandings.

Furthermore, we also see that there is an even lower rate of unambiguous connectives in the experiment 1 design ($\beta = -0.472$, $z = -5.030$, $p < 0.001$, compared to the blinded condition). This indicates that, even though the same information is available to the speaker in both cases, there is an influence of experimental task: participants are more aware of the existence of interpretation alternatives on the side of the hearer in the blinded setting, and therefore are also aware of the risk of misunderstanding, which leads them to prefer the unambiguous connective.

These results hence reveal a graded effect of restriction on alternative interpretations: when the possible interpretations are not limited, the speaker will use more precise DCs than in situations where the alternatives are limited to non-confusable relations; the speaker is sure that the listener won't misunderstand. When a confusable interpretation is explicitly included in alternatives, the speaker will, in turn, be more aware that a misunderstanding is possible, compared to when interpretations aren't explicitly provided. As the difference between the with and without pragmatic exposure conditions shows, participants' previous experience in the gamified task affects their choice. They are more aware of the chance of mis-interpretation if they have previously seen confusable alternatives in the "training phase," or even received some corrective feedback.

TABLE 8 | Mean unambiguous DC proportion per participant under various conditions in Experiment 1 and 2.

	Exp. 2: with pragmatic exposure			Exp. 2: without pragm. exp.		Exp. 1
	No competitor (first half)	With competitor (first half)	Blinded (second half)	No competitor (first half)	Blinded (second half)	Unknown
Overall	70%	77%	75%	70%	72%	64%
Target -predicting	71%	78%	75%	70%	72%	65%
Competitor -predicting	71%	77%	75%	68%	73%	65%
Neutral	69%	76%	74%	71%	72%	63%

TABLE 9 | Regression coefficients of the logistic linear mixed effects model including the responses from Experiment 1 as well as both halves of Experiment 2.

Variable	β	*SE*	*z*	*p*
Intercept	1.212	0.155	7.834	< 0.001***
Expt-1	−0.472	0.094	−5.030	< 0.001***
Blinded (with prag. exposure)	0.178	0.064	2.798	< 0.01**
No competitor	−0.180	0.059	−3.063	< 0.01**
With competitor	0.226	0.081	2.789	< 0.01**

The base level of the predictor variable is the blinded condition of the without pragmatic exposure version of Experiment 2. **$p < 0.01$, ***$p < 0.001$.

4.6. Summary and Discussion

Summarizing the results of Experiment 2, we found that contextual expectation of a competitor discourse relation does not have the same effect as presenting it explicitly as a possible continuation. These findings replicate the results of experiment 1. We therefore conclude that the lack of effect in experiment 1 cannot be attributed to the instructions of the task, but rather to the not explicitly listing the alternative listener interpretation options. It is possible that it is too difficult for the speaker to reason about the discourse expectations that the context raises for the listener.

The empirical results we found here are thus not consistent with our expectations based on the rational speech act model: while we had expected to find an effect of discourse relation expectation on connective choice, similar to the effect found in pretest 3, we were not able to detect any such effect, and in fact, our Bayesian Factor analysis indicates that the data strongly support the null hypothesis.

The argument we made here is based on a qualitative prediction of the RSA theory, and qualitative results from our empirical data. As the RSA framework is capable of making quantitative probabilistic predictions, it would be possible to also test more exact quantitative predictions. The required ingredients include the prior distribution of the salience of a relation based on the biasing contexts, which serves as the literal listener model (Equation 3) and a function that defines the production cost of a given DC (Equation 2). Both measures could be obtained empirically in separate experiments.

We assume that the production costs do not vary across experimental conditions, and the main driving factor of the effect would be the discourse relation inferences of the listener after having perceived the connective. Based on our prestest 2, we believe that our experimental manipulation was effective in changing comprehender interpretations, and that there would thus be a substantial difference between context conditions also in an experiment that collects this prior probability more directly. However, given the lack of even a qualitative effect in our data, even when we used a very large number of participants in experiment 2, we think that it is not very promising to proceed to a more quantitative comparison at this point.

The comparison of experimental designs provided evidence that gamified elements such as the explicit listing of alternatives, and experience with the task induce participants to choose an unambiguous connective more often. These results thus indicate that gamification of the task affects rational reasoning and thereby the results of the RSA study.

5. OVERALL DISCUSSION AND CONCLUSION

The RSA model states that speakers reason about the interpretation of the listener and weigh the cost of an utterance against its utility in avoiding misunderstandings. According to this theory, it is predicted that when the listener is likely to confuse an intended discourse relation with another relation, the speaker should avoid the (albeit temporary) misunderstanding by using a more informative utterance, by using a DC that signals the target relation more exclusively. Following the success of predicting human behavior in a variety of language processing tasks, such as the production of referring expressions, the RSA account had also been shown to make correct qualitative predictions on the speaker's choice of DCs in an language game experiment by Yung and Demberg (2018). Language games of this kind are widely used to explore pragmatic inferences in contexts because they allow precise manipulation by explicitly displaying the a set of listener interpretations to the speaker.

The current study set out to test RSA's prediction on discourse relation production using a methodology of improved ecological validity by removing the explicit statement of what the interpretations of the listener might be. Instead, manipulation on the preceding context is used to elicit discourse expectations, which either match or do not match with the target discourse relation to be conveyed by the speaker. We hypothesized that a situation where a confusable discourse relation is highly expected in context will lead to similar increased demand in choosing a suitable connective to avoid temporary misinterpretations by the listener.

Experimental results show that the context manipulation is successful in that the connective marking the target relation is inconsistent with the expected relation and can hence help to correct listener expectations early on.

However, our experiment 1, which did not explicitly show the discourse expectations, reveals that, contrary to our hypothesis, the preference to produce a particular discourse relation with a specific, unambiguous DC does not depend on whether the target relation or other competing relations are expected. Further experiments, using a modified language game design, confirms that the contextual expectation of a competitor discourse relation does not affect the production of the DC.

We however did find that participants use more informative utterances when the listener's interpretation is unrestricted than when the interpretations are restricted to non-confusable alternatives. That is, when the listener can see that there is no risk of misinterpretation, they do not use unambiguous connectives as much, but also use ambiguous ones (which in this context, in fact are also unambiguous). These results indicate that people do reason about the listener's interpretation, consistent with earlier findings, but only when the interpretation alternatives are easily accessible. Our results are consistent with an account according to which speakers adopt general strategies, instead of reasoning about each case, in line with earlier results by Vogels et al. (2020). One such strategy that we observed here was to more often choose unambiguous connectives, if ambiguity had been experienced earlier in the study.

How can the absence of any effect of contextual expectation be explained? We see two possible options:

a) inferring discourse expectations and reasoning about them is too difficult, therefore participants don't do it (i.e., they only engage in reasoning when the discourse expectation inference step is done for them by the experimental design).
b) they do infer discourse expectations and reason about listener interpretations, but feel that it's not necessary to disambiguate the relation as the content of the second argument of the relation will eventually lead to full disambiguation anyway.

Regarding option (a), let's first take a step back: the RSA crucially states that speakers reason about the interpretation of the listeners in order to maximize the informativeness of the utterance. An underlying assumption is that they are equipped with the necessary resources, such as computational resources and background knowledge, to do so. Communication of concrete meanings, such as reference to particular objects or numerical quantities, have been extensively studied in existing work. Oftentimes, the alternative interpretations from the point of view of the listener were also directly available to the speaker in those studies, consider for instance referring expression generation. In the main experiment, the subjects had to do another level of inference: to rationally select an informative connective, they would have to reason about what the listener would expect. These discourse expectations are not only abstract concepts (which may be more difficult to juggle in memory), but they also are not present in the visible context of the interaction.

The results from Yung and Demberg (2018) and pretest 3 demonstrate that speakers *can* choose connectives in order to avoid misinterpretations on the side of the listener, pretest 2 further demonstrated that the stimuli do give rise to expectations, and earlier work has provided ample evidence that listeners generate discourse expectations during comprehension (Sanders and Noordman, 2000; Rohde et al., 2011; Canestrelli et al., 2013; Rohde and Horton, 2014; Xiang and Kuperberg, 2015; Scholman et al., 2017; Van Bergen and Bosker, 2018; Schwab and Liu, 2020; Köhne-Fuetterer et al., 2021). However, there is no direct evidence that speakers also simulate the discourse expectations that listeners would generate.

The RSA theory does not provide explicit limits or definitions as to when a speaker reasons about a listener, and for what linguistic phenomena or under which situational circumstances this reasoning would be too effortful. In fact, a common criticism of RSA (and Gricean pragmatics) is that it falls short in explaining speaker productions: utterances are sometimes longer than they need to be, underinformative or ambiguous (Engelhardt et al., 2006; Gatt et al., 2013; Baumann et al., 2014; McMahan and Stone, 2015), and speakers also sometimes fail to take listener perspective into account when generating referring expressions (Horton and Keysar, 1996; Lane and Ferreira, 2008; Yoon et al., 2012).

These findings have lead to discussions as to whether speakers really always behave rationally, and more specifically, whether speakers reason about listeners in all cases, and how many levels of recursion in reasoning should be considered (Degen and Franke, 2012; Franke and Degen, 2016) (in most previous models, the default is set to 2 levels of recursion). Yuan et al. (2018) explored these questions in the context of reference games and found that pragmatic listeners and speakers always outperform their literal counterparts and that model performance becomes more accurate as more levels of recursion are assumed.

Yuan et al. (2018) also explored the effect of limiting the number of considered alternatives, and found that this does not detrimentally affect model results (in fact, it improves model fit). Note that the results found in the present experiment do not require a large number of alternatives: strictly speaking, even reasoning about the top-1 alternative from the point of view of the listener would be sufficient to elicit an effect of context on discourse connective choice. Also, the number of levels of recursion depth required in the reasoning in our experiment is not large—default 2-level recursion would be sufficient. Therefore, those prior concerns do not explain why we fail to find an effect here.

In summary, while it is well-known that there can be differences between individuals as to how deeply they engage in the reasoning process, there has previously been little discussion with respect to the potential differences in cognitive difficulty of making a single reasoning step. Our study hence sheds light on a potential additional source of limitation with respect to reasoning about the interlocutor, outside of recursion depth of the number of alternatives that need to be considered.

Option (b) is a possible criticism for the design in experiment 1, in particular if dropping the assumption of *incremental* RSA.

However, if this was the determining factor of why there was no difference between the context conditions, we should have seen an effect of context in the blinded conditions in the second part of experiment 2: in this setting, people were very aware that the task of their listener was to guess the second argument, so not providing a helpful hint for that guess would be pretty non-cooperative. Therefore, the outcome of experiment 2 doesn't seem compatible with this explanation.

In terms of methodological contribution, we found that the widely used gamified experimental design substantially affects results, even in a setting where the experimental items are the same. We designed an experimental structure that tests the speaker's DC production in relation to different levels of guidance of pragmatic reasoning. Minimum guidance was provided in Experiment 1: the speaker did not know if the communication was successful or not, as no feedback was given. Maximum guidance is provided in the previous work by Yung and Demberg (2018) and the first half of the *with pragmatic exposure* version of Experiment 2: the speaker learns that the listener might guess the competitor relation if the utterance is not specific enough. An interesting condition between these two extremes was the blinded condition in the second half of experiment 2. Here, the task is identical to the maximum guidance setting, but the knowledge available to the speaker is identical to the minimum guidance setting. Here, we see a generic increase in the use of unambiguous connectives compared to the minimal guidance setting, but no condition-specific increase as would have been expected according to the RSA.

We therefore conclude that gamification of the task, which encourages reasoning about alternatives, boosts RSA-consistent behavior. However, it remains unclear to what extent the findings from a language game actually represent people's normal language production. One direction for future work is thus to validate RSA-styled production and interpretation outside the assumption of a toy world for other phenomena.

We think that alternative experimental designs should be explored, which seek for free production of utterances given a manipulated prompt or situation, such as a story generation task given a sequence of pictures. While such a free production task is much closer to naturalistic language use, it is not trivial to elicit specific discourse relations and closely control the experimental conditions in such a design. However, given enough data, it is still possible to collect a distribution of intended discourse relations and the corresponding connectives. Crowd-sourcing could be an effective way to collect such a database, as the additional noise introduced through the less controlled experimental design might be counter-weighed by a larger number of participants.

Furthermore, the results of this experiment lead us to the question of what it takes for the speaker to engage in reasoning about the listener interpretation. When do speakers consider listener misinterpretation risk, and under which circumstances is this calculation too effortful? One possibility might be that reasoning of the speaker about the listener might be triggered only when a small set of explicit alternatives is available and in full view, such that it doesn't need to be held in memory. This question thus addresses the generality with which the RSA account holds: in principle, it is formulated such that it covers reasoning about listener interpretations independent of the form in which they are available to the speaker. However, experiments conducted so far have only addressed a small area of the possible production phenomena that RSA could be applied to, and have addressed these questions in settings where the alternative listener interpretations are in a shared visible space.

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://osf.io/q32vj/?view_only=61f5a72585c74690ac2bbc6222403c76.

ETHICS STATEMENT

The studies involving human participants were approved by Deutsche Gesellschaft für Sprache (DGfS). The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

FY and VD conceived the idea, designed the experiments, and wrote the manuscript. FY carried out the experiments and analyzed the data. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) "Project-ID 232722074" SFB 1102.

ACKNOWLEDGMENTS

We thank Jana Jungbluth for her help in constructing the stimuli, as well as her help in conducting the experiment. We would also like to thank the reviewers for their valuable comments which helped to improve this article.

REFERENCES

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., and Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *J. Mem. Lang.* 42, 1–22. doi: 10.1006/jmla.1999.2667

Barr, D. J., and Keysar, B. (2005). "Making sense of how we make sense: the paradox of egocentrism in language use," in *Figurative Language Comprehension: Social and Cultural Influences*, eds H. L. Colston and A. N. Katz (Lawrence Erlbaum Associates Publishers), 21–41.

Barr, D. J., and Keysar, B. (2006). "Chapter 23 - perspective taking and the coordination of meaning in language use," in *Handbook of Psycholinguistics (Second Edition)*, eds M. J. Traxler and M. A. Gernsbacher (London: Academic Press), 901–938. doi: 10.1016/B978-012369374-7/50024-9

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

Baumann, P., Clark, B., and Kaufmann, S. (2014). "Overspecification and the cost of pragmatic reasoning about referring expressions," in *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, eds P. Bello, M. Guarini,

M. McShane, and B. Scassellati (Austin, TX: Cognitive Science Society), 1898–1903.

Bergen, L., Goodman, N., and Levy, R. (2012). "That's what she (could have) said: how alternative utterances affect language use," in *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, eds N. Miyake, D. Peebles, and R. P. Cooper (Austin, TX: Cognitive Science Society), 120–125.

Beun, R.-J., and Cremers, A. H. (1998). Object reference in a shared domain of conversation. *Pragmat. Cogn.* 6, 121–152. doi: 10.1075/pc.6.1-2.08beu

Borg, E. (2012). *Pursuing Meaning*. Oxford: Oxford University Press.

Borg, E. (2017). Local vs. global pragmatics. *Inquiry* 60, 509–516. doi: 10.1080/0020174X.2016.1246862

Brennan, S. E., and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol. Learn. Mem. Cogn.* 22:1482. doi: 10.1037/0278-7393.22.6.1482

Brown-Schmidt, S., and Tanenhaus, M. K. (2006). Watching the eyes when talking about size: an investigation of message formulation and utterance planning. *J. Mem. Lang.* 54, 592–609. doi: 10.1016/j.jml.2005.12.008

Brown-Schmidt, S., and Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: a targeted language game approach. *Cogn. Sci.* 32, 643–684. doi: 10.1080/03640210802066816

Bürkner, P.-C. (2017). Advanced bayesian multilevel modeling with the r package brms. *arXiv [Preprint]. arXiv:1705.11123.* doi: 10.32614/RJ-2018-017

Canestrelli, A. R., Mak, W. M., and Sanders, T. J. (2013). Causal connectives in discourse processing: how differences in subjectivity are reflected in eye movements. *Lang. Cogn. Process.* 28, 1394–1413. doi: 10.1080/01690965.2012.685885

Carston, R. (2017). Pragmatic enrichment: beyond gricean rational reconstruction–a response to mandy simons. *Inquiry* 60, 517–538. doi: 10.1080/0020174X.2016.1246863

Chen, G., van Deemter, C., and Lin, C. (2018). "Modelling pro-drop with the rational speech acts model," in *Proceedings of the 11th International Conference on Natural Language Generation (Association for Computational Linguistics (ACL))*, eds E. Krahmer, A. Gatt, and M. Goudbeek (Tilburg: Tilburg University), 159–164. doi: 10.18653/v1/W18-6519

Clark, H. H. (1992). *Arenas of Language Use*. Chicago, IL: University of Chicago Press.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Clark, H. H., and Brennan, S. E. (1991). "Grounding in communication," in *Perspectives on Socially Shared Cognition (American Psychological Association)*, eds L. B. Resnick, J. M. Levine, and S. D. Teasley (Washington, DC: American Psychological Association), 127–149. doi: 10.1037/10096-006

Cohn-Gordon, R., Goodman, N., and Potts, C. (2019). "An incremental iterated response model of pragmatics," in *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, eds G. Jarosz, M. Nelson, B. O'Connor, and J. Pater, 81–90. doi: 10.7275/cprc-8x17

Cornish, F. (2009). ""Text" and "discourse" as "context": discourse anaphora and the fdg contextual component," in *Working Papers in Functional Discourse Grammar (WP-FDG-82): The London Papers I*, Vol. 1, 97–115.

Degen, J. (2013). *Alternatives in pragmatic reasoning* (dissertation), University of Rochester, Rochester, NY, United States.

Degen, J., and Franke, M. (2012). "Optimal reasoning about referential expressions," in *Proceedings of SemDial 2012 (SeineDial)*, eds S. Brown-Schmidt, J. Ginzburg, and S. Larsson (Paris: Université Paris Diderot - Paris 7), 2–11.

Degen, J., Franke, M., and Jager, G. (2013). "Cost-based pragmatic inference about referential expressions," in *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (Austin, TX: Cognitive Science Society), 376–381.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., and Goodman, N. D. (2020). When redundancy is useful: a Bayesian approach to "overinformative" referring expressions. *Psychol. Rev*. 127, 591–621. doi: 10.1037/rev0000186

Degen, J., and Tanenhaus, M. K. (2015). Processing scalar implicature: a constraint-based approach. *Cogn. Sci.* 39, 667–710. doi: 10.1111/cogs.12171

Degen, J., and Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: a visual world eye-tracking study. *Cogn. Sci.* 40, 172–201. doi: 10.1111/cogs.12227

Drenhaus, H., Demberg, V., Köhne, J., and Delogu, F. (2014). "Incremental and predictive discourse processing based on causal and concessive discourse markers: erp studies on German and English," in *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (Austin, TX: Cognitive Science Society), 403–408.

Engelhardt, P. E., Bailey, K. G., and Ferreira, F. (2006). Do speakers and listeners observe the gricean maxim of quantity? *J. Mem. Lang.* 54, 554–573. doi: 10.1016/j.jml.2005.12.009

Frank, M. C., Emilsson, A. G., Peloquin, B., Goodman, N. D., and Potts, C. (2016). Rational speech act models of pragmatic reasoning in reference games. *psyarxiv*. doi: 10.31234/osf.io/f9y6b

Frank, M. C., and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science* 336:998. doi: 10.1126/science.1218633

Franke, M., and Degen, J. (2016). Reasoning in reference games: individual- vs. population-level probabilistic modeling. *PLoS ONE* 11:e0154854. doi: 10.1371/journal.pone.0154854

Galati, A., and Brennan, S. E. (2010). Attenuating information in spoken communication: for the speaker, or for the addressee? *J. Mem. Lang.* 62, 35–51. doi: 10.1016/j.jml.2009.09.002

Gatt, A., van Gompel, R. P., van Deemter, K., and Krahmer, E. (2013). "Are we bayesian referring expression generators?," in *Proceedings of 35th Annual Conference of the Cognitive Science Society*, eds M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth (Austin, TX: Cognitive Science Society), 1228–1233.

Goodman, N. D., and Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends Cogn. Sci.* 20, 818–829. doi: 10.1016/j.tics.2016.08.005

Graf, C., Degen, J., Hawkins, R. X., and Goodman, N. D. (2016). "Animal, dog, or dalmatian? level of abstraction in nominal referring expressions," in *Proceedings for the 38th Annual Meeting of the Cognitive Science Society*, eds A. Papafragou, D. Grodner, D. Mirman, and J. C. Trueswell (Austin, TX: Cognitive Science Society). 2261–2266.

Grice, H. P. (2000). "Logic and Conversation," in *Perspectives in the Philosophy of Language: A Concise Anthology*, 271.

Hahn, M., Degen, J., Goodman, N. D., Jurafsky, D., and Futrell, R. (2018). "An information-theoretic explanation of adjective ordering preferences," in *Proceedings for the 40th Annual Meeting of the Cognitive Science Society*, eds C. Kalish, M. A. Rau, X. Zhu, and T. T. Rogers (Austin, TX: Cognitive Science Society), 1766–1771.

Horton, W. S., and Keysar, B. (1996). When do speakers take into account common ground? *Cognition* 59, 91–117.

Isaacs, E. A., and Clark, H. H. (1987). References in conversation between experts and novices. *J. Exp. Psychol. Gen.* 116:26.

Jaeger, T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cogn. Psychol.* 61, 23–62. doi: 10.1016/j.cogpsych.2010.02.002

Jaeger, T. F., and Buz, E. (2018). "Signal reduction and linguistic encoding," in *Blackwell handbooks in linguistics. The handbook of psycholinguistics*, eds E. M. Fernández, and H. S. Cairns (Hoboken, NJ: Wiley Blackwell), 38–81.

Kehler, A., Kertz, L., Rohde, H., and Elman, J. L. (2008). Coherence and coreference revisited. *J. Semant.* 25, 1–44. doi: 10.1093/jos/ffm018

Köhne, J., and Demberg, V. (2013). "The time-course of processing discourse connectives," in *Proceedings of the Annual 35th Meeting of the Cognitive Science Society*, ed. M. Knauff (Austin, TX: Cognitive Science Society), 1760–1765.

Köhne-Fuetterer, J., Drenhaus, H., Delogu, F., and Demberg, V. (2021). The online processing of causal and concessive discourse connectives. *Linguistics* 59, 417–448. doi: 10.1515/ling-2021-0011

Koolen, R., Gatt, A., Goudbeek, M., and Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *J. Pragmat.* 43, 3231–3250. doi: 10.1016/j.pragma.2011.06.008

Kreiss, E., and Degen, J. (2020). "Production expectations modulate contrastive inference," in *Proceedings for the 42nd Annual Meeting of the Cognitive Science Society*, eds S. Denison., M. Mack, Y. Xu, and B. C. Armstrong (Austin, TX: Cognitive Science Society), 259–265. Available online at: https://cogsci.mindmodeling.org/2020/

Lane, L. W., and Ferreira, V. S. (2008). Speaker-external versus speaker-internal forces on utterance form: do cognitive demands override threats to referential success? *J. Exp. Psychol. Learn. Mem. Cogn.* 34:1466. doi: 10.1037/a0013353

Lascarides, A., Asher, N., and Oberlander, J. (1992). "Inferring discourse relations in context," in *30th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, PA: Association for Computational Linguistics), 1–8. doi: 10.3115/981967.981968

Levy, R., and Jaeger, T. F. (2007). "Speakers optimize information density through syntactic reduction," in *Advances in Neural Information Processing Systems 19*, eds B. Schölkopf, J. Platt, and T. Hofmann (Cambridge, MA: The MIT Press). 19:849. doi: 10.7551/mitpress/7503.003.0111

Mann, W. C., and Thompson, S. A. (1988). Rhetorical structure theory: toward a functional theory of text organization. *Text* 8, 243–281.

McMahan, B., and Stone, M. (2015). A bayesian model of grounded color semantics. *Trans. Assoc. Comput. Linguist.* 3, 103–115. doi: 10.1162/tacl_a_00126

Mozuraitis, M., Stevenson, S., and Heller, D. (2018). Modeling reference production as the probabilistic combination of multiple perspectives. *Cogn. Sci.* 42, 974–1008. doi: 10.1111/cogs.12582

Nadig, A. S., and Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychol. Sci.* 13, 329–336. doi: 10.1111/j.0956-7976.2002.00460.x

Nedergaard, J., and Smith, K. (2020). "Are you thinking what i'm thinking? Perspective-taking in a language game," in *Proceedings for the 42nd Annual Meeting of the Cognitive Science Society*, eds S. Denison, M. Mack, Y. Xu, and B. C. Armstrong (Austin, TX: Cognitive Science Society), 1001–1007.

Olson, D. R. (1970). Language and thought: aspects of a cognitive theory of semantics. *Psychol. Rev.* 77:257.

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

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

Qing, C., and Franke M. (2015). "Variations on a Bayesian theme: comparing bayesian models of referential reasoning," in *Bayesian Natural Language Semantics and Pragmatics. Language, Cognition, and Mind*, Vol. 2, eds H. Zeevat and H. C. Schmitz (Cham: Springer). doi: 10.1007/978-3-319-17064-0_9

Rashmi, P., Nikhil, D., Alan, L., Eleni, M., Robaldo, L., Aravind, J., et al. (2008). *The Penn Discourse Treebank 2.0*. Marrakech: European Language Resources Association (ELRA).

Rohde, H., and Horton, W. S. (2014). Anticipatory looks reveal expectations about discourse relations. *Cognition* 133, 667–691. doi: 10.1016/j.cognition.2014.08.012

Rohde, H., Levy, R., and Kehler, A. (2011). Anticipating explanations in relative clause processing. *Cognition* 118, 339–358. doi: 10.1016/j.cognition.2010.10.016

Roßnagel, C. (2000). Cognitive load and perspective-taking: applying the automatic-controlled distinction to verbal communication. *Eur. J. Soc. Psychol.* 30, 429–445. doi: 10.1002/(SICI)1099-0992(200005/06)30:3<429::AID-EJSP3>3.0.CO;2-V

Ryskin, R., Stevenson, S., and Heller, D. (2020). "Probabilistic weighting of perspectives in dyadic communication," in *Proceedings for the 42nd Annual Meeting of the Cognitive Science Society*, eds S. Denison., M. Mack, Y. Xu, and B. C. Armstrong (Austin, TX: Cognitive Science Society), 252–258.

Ryskin, R. A., Benjamin, A. S., Tullis, J., and Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: an individual differences approach. *J. Exp. Psychol. Gen.* 144:898. doi: 10.1037/xge0000093

Sanders, T. J., and Noordman, L. G. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Process.* 29, 37–60. doi: 10.1207/S15326950dp2901_3

Sanders, T. J., Spooren, W. P., and Noordman, L. G. (1992). Toward a taxonomy of coherence relations. *Discourse Process.* 15, 1–35.

Scholman, M. C., Demberg, V., and Sanders, T. J. (2020). Individual differences in expecting coherence relations: exploring the variability in sensitivity to contextual signals in discourse. *Discourse Process.* 57, 844–861. doi: 10.1080/0163853X.2020.1813492

Scholman, M. C., Rohde, H., and Demberg, V. (2017). on the one hand as a cue to anticipate upcoming discourse structure. *J. Mem. Lang.* 97, 47–60. doi: 10.1016/j.jml.2017.07.010

Schwab, J., and Liu, M. (2020). Lexical and contextual cue effects in discourse expectations: experimenting with german'zwar... aber'and english'true/sure... but'. *Dialogue Discourse* 11, 74–109. doi: 10.5087/dad.2020.203

Sikos, L., Venhuizen, N., Drenhaus, H., and Crocker, M. W. (2019). "Reevaluating pragmatic reasoning in web-based language games," in *Poster presented at: CUNY Conference on Human Sentence Processing* (Boulder, CO: University of Colorado).

Song, L. (2010). The role of context in discourse analysis. *J. Lang. Teach. Res.* 1:876. doi: 10.4304/jltr.1.6.876-879

Sperber, D., and Wilson, D. (1986). *Relevance: Communication and Cognition*, Vol. 142. Cambridge, MA: Harvard University Press.

Spooren, W., and Degand, L. (2010). Coding coherence relations: reliability and validity. *Corpus Linguist. Linguist. Theor.* 6, 241–266. doi: 10.1515/cllt.2010.009

Stalnaker, R. C. (1978). "Assertion," in *Pragmatics* (Leiden: Brill), 315–332. doi: 10.1163/9789004368873_013

Sulik, J., and Lupyan, G. (2016). "Failures of perspective taking in an open-ended signaling task," in *The Evolution of Language: Proceedings of the 11th International Conference (evolangx11)*. doi: 10.12775/3991-1.100

Sulik, J., and Lupyan, G. (2018a). Perspective taking in a novel signaling task: effects of world knowledge and contextual constraint. *J. Exp. Psychol. Gen.* 147:1619. doi: 10.1037/xge0000475

Sulik, J., and Lupyan, G. (2018b). "Success in signaling: the effect of feedback to signaler and receiver," in *The Evolution of Language: Proceedings of the 12th International Conference*. Available online at: http://evolang.org/torun/proceedings/papertemplate

Van Bergen, G., and Bosker, H. R. (2018). Linguistic expectation management in online discourse processing: an investigation of dutch inderdaad'indeed'and eigenlijk'actually'. *J. Mem. Lang.* 103, 191–209. doi: 10.1016/j.jml.2018.08.004

Vogel, A., Gomez Emilsson, A., Frank, M. C., Jurafsky, D., and Potts, C. (2014). "Learning to reason pragmatically with cognitive limitations," in *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, eds P. Bello, M. Guarini, M. McShane, and B. Scassellati (Austin, TX: Cognitive Science Society), 3055–3060.

Vogels, J., Howcroft, D. M., Tourtouri, E., and Demberg, V. (2020). How speakers adapt object descriptions to listeners under load. *Lang. Cogn. Neurosci.* 35, 78–92. doi: 10.1080/23273798.2019.1648839

Wilcox, E., and Spector, B. (2019). "The role of prior beliefs in the rational speech act model of pragmatics: exhaustivity as a case study," in *Proceedings for 41st the Annual Meeting of the Cognitive Science Society*, eds A. K. Goel, C. M. Seifert, and C. Freksa (Austin, TX: Cognitive Science Society), 3099–3105.

Wilkes-Gibbs, D., and Clark, H. H. (1992). Coordinating beliefs in conversation. *J. Mem. Lang.* 31, 183–194.

Xiang, M., and Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Lang. Cogn. Neurosci.* 30, 648–672. doi: 10.1080/23273798.2014.995679

Yoon, S. O., and Brown-Schmidt, S. (2013). Lexical differentiation in language production and comprehension. *J. Mem. Lang.* 69, 397–416. doi: 10.1016/j.jml.2013.05.005

Yoon, S. O., Koh, S., and Brown-Schmidt, S. (2012). Influence of perspective and goals on reference production in conversation. *Psychon. Bull. Rev.* 19, 699–707. doi: 10.3758/s13423-012-0262-6

Yuan, A., Monroe, W., Bai, Y., and Kushman, N. (2018). "Understanding the rational speech act model," in *Proceedings for the 40th Annual Meeting of the Cognitive Science Society*, eds C. Kalish, M. A. Rau, X. Zhu, and T. T. Rogers (Austin, TX: Cognitive Science Society), 2759–2764.

Yung, F., and Demberg, V. (2018). "Do speakers produce discourse connectives rationally?," in *Proceedings of the Eight Workshop on Cognitive Aspects of Computational Language Learning and Processing* (Stroudsburg, PA: ACL). 6–16.

Yung, F., Duh, K., Komura, T., and Matsumoto, Y. (2016). "Modelling the usage of discourse connectives as rational speech acts," in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning* (Berlin: ACL), 302–313.

Yung, F., Duh, K., Komura, T., and Matsumoto, Y. (2017). A psycholinguistic model for the marking of discourse relations. *Dialogue Discourse* 8, 106–131. doi: 10.5087/dad.2017.104

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: 28 May 2021
doi: 10.3389/fpsyg.2021.661898

A Rose by Any Other Verb: The Effect of Expectations and Word Category on Processing Effort in Situated Sentence Comprehension

Les Sikos, Katharina Stein and Maria Staudte*

Language Science and Technology, Saarland University, Saarbrücken, Germany

OPEN ACCESS

Edited by:
Hannah Rohde,
University of Edinburgh,
United Kingdom

Reviewed by:
Aya Meltzer-Asscher,
Tel Aviv University, Israel
Aine Ito,
Humboldt University of Berlin,
Germany
Shayne Sloggett,
University of York, United Kingdom

***Correspondence:**
Les Sikos
sikos@coli.uni-saarland.de

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

Received: *31 January 2021*
Accepted: *12 April 2021*
Published: *28 May 2021*

Citation:
Sikos L, Stein K and Staudte M (2021)
A Rose by Any Other Verb: The Effect
of Expectations and Word Category
on Processing Effort in Situated
Sentence Comprehension.
Front. Psychol. 12:661898.
doi: 10.3389/fpsyg.2021.661898

Recent work has shown that linguistic and visual contexts jointly modulate linguistic expectancy and, thus, the processing effort for a (more or less) expected critical word. According to these findings, uncertainty about the upcoming referent in a visually-situated sentence can be reduced by exploiting the selectional restrictions of a preceding word (e.g., a verb or an adjective), which then reduces processing effort on the critical word (e.g., a referential noun). Interestingly, however, no such modulation was observed in these studies on the expectation-generating word itself. The goal of the current study is to investigate whether the reduction of uncertainty (i.e., the generation of expectations) simply does not modulate processing effort-or whether the particular subject-verb-object (SVO) sentence structure used in these studies (which emphasizes the referential nature of the noun as direct pointer to visually co-present objects) accounts for the observed pattern. To test these questions, the current design reverses the functional roles of nouns and verbs by using sentence constructions in which the noun reduces uncertainty about upcoming verbs, and the verb provides the disambiguating and reference-resolving piece of information. Experiment 1 (a Visual World Paradigm study) and Experiment 2 (a Grammaticality Maze study) both replicate the effect found in previous work (i.e., the effect of visually-situated context on the word which uniquely identifies the referent), albeit on the verb in the current study. Results on the noun, where uncertainty is reduced and expectations are generated in the current design, were mixed and were most likely influenced by design decisions specific to each experiment. These results show that processing of the reference-resolving word—whether it be a noun or a verb—reliably benefits from the prior linguistic and visual information that lead to the generation of concrete expectations.

Keywords: processing effort, expectations, situated surprisal, visual world paradigm, language comprehension, referential uncertainty, grammaticality maze, pupillometry

1. INTRODUCTION

Recent language processing literature converges on establishing a predictive mechanism in which expectations about upcoming words can be determined by both linguistic and visual contexts. On the one hand, word expectancy, as derived from the linguistic context, has been shown to reliably correlate with processing effort, i.e., more predictable words are easier to process (e.g., Kutas and Hillyard, 1980; Federmeier et al., 2007; Van Berkum et al., 2007; Demberg and Keller, 2008; Smith and Levy, 2013). On the other hand, more recent work has shown that the visual context can also influence linguistic expectancy and, for instance, reduce the processing effort for a word when the co-present scene enables very clear and concrete predictions for that word (Ankener et al., 2018; Tourtouri et al., 2019; Staudte et al., 2021).

For example, Ankener and colleagues examined two critical regions in German stimulus sentences in order to investigate the chain of processing from generating an expectation, to the downstream effects of that expectation. A sentence such as *Die Frau verschüttet jetzt das Wasser* ("The woman spills now the water") was accompanied by a visual scene depicting four objects, some of which were "spillable." The two regions of interest within the sentence were: (a) the verb (e.g., *verschüttet*, "spills"), where linguistic expectations for upcoming spillable object nouns were generated, and (b) the sentence-final noun (e.g., *Wasser*, "water") whose expectancy varied depending on whether one, three, four, or none of the depicted objects could be spilled. To analyze eye movements, new inspections of the target object were extracted during the verb region (i.e., before the target was mentioned). Eye movements indicated that participants were more likely to shift their attention to the target when it was the only spillable object in the display, than when there were three or four spillable objects. Although these results did not distinguish anticipation strength between three and four potential target objects, they do provide evidence for listeners' strong(est) anticipation of the target when it was the only object that matched the verb's selectional restrictions. This suggests that uncertainty about the upcoming referent was reduced by exploiting linguistic knowledge about the verbal restrictions. Results further showed that processing effort at the object noun, as measured by the pupillometric Index of Cognitive Activity (ICA, Marshall, 2002; Ankener et al., 2018) and electrophysiological measures (Staudte et al., 2021), was higher when more spillable objects were co-present. In contrast, the object noun was easiest to process when no other spillable competitors were co-present, and thus the object noun was most predictable. These results demonstrate that processing effort is directly influenced by both visual and linguistic contexts, which together modulate *visually-situated* expectations.

Other work further suggests that words which reduce uncertainty about upcoming linguistic continuations require greater processing effort than words that do not reduce uncertainty (Frank, 2013; Hale, 2016; Linzen and Jaeger, 2016; Maess et al., 2016). Linzen and Jaeger (2016) revealed, for instance, that a word which reduces the uncertainty about possible continuations elicits longer reading times. Maess et al. (2016) measured magnetoencephalography (MEG) while participants listened to simple German sentences in which the verbs either constrained expectations for a particular noun or not (e.g., *Er dirigiert/leitet das Orchester*, "He conducts/leads the orchestra") and found that more constraining verbs (e.g., *dirigiert*, "conducts") elicited greater processing difficulty, as reflected in larger N400 amplitudes, than unconstraining verbs (e.g., *leitet*, "leads"). Moreover, when a noun (e.g., "orchestra") followed a constraining verb, the noun elicited a reduced N400 relative to the same noun following an unconstraining verb, indicating that it was easier to process. This pattern of effects was interpreted as "trade-off" in processing effort between the moment at which a prediction is made and a later point in time when the prediction is cashed out. Although Maess et al. (2016) attribute this difference in processing cost to the constraining word preactivating semantic features of the upcoming predictable noun, the effect is also consistent with the reduction of uncertainty. Lastly, similar trade-off effects, but in the P600 component, were found by Ness and Meltzer-Asscher (2018) and attributed to pre-updating, a mechanism thought to reflect an early integration of the predicted upcoming verb argument.

Interestingly, however, neither measure of processing difficulty in Ankener et al. (2018) or Staudte et al. (2021) indicated a modulation of processing effort at the verb itself, despite the fact that the verb reduced uncertainty about upcoming referents to a greater or lesser extent depending on the visual context. This is somewhat surprising given the results of previous work indicating that more constraining words/verbs elicit greater processing effort than unconstraining words/verbs (Frank, 2013; Hale, 2016; Linzen and Jaeger, 2016; Maess et al., 2016; Ness and Meltzer-Asscher, 2018). In contrast, Ankener and colleagues interpreted their findings as an indication that processing effort at the predictive stage (i.e., at the verb) was simply not affected by the amount of (referential) uncertainty that can be reduced at the verb and that this might be specific to the *situated* and *referential* nature of expectations.

An alternative explanation for the findings of Ankener and colleagues is that the particular word categories used in their stimuli contributed to the pattern of null effects found at the verb and significant effects found at the subsequent noun. More specifically, the linear order of words in Ankener et al. (2018) and Staudte et al. (2021)—in which participants first encountered the verb and then the noun—may have emphasized the referential aspects of the object noun. That is, while nouns in general can be thought of as direct pointers to objects in the world, this function receives particular emphasis when the noun is used to uniquely disambiguate a reference and, consequently, to decode the entire sentence proposition. The verb, in contrast, does not index the displayed objects as directly, and therefore may not strictly exclude objects that do not fit the verb's selectional restrictions. This difference could potentially explain the lack of effects found at the verb.

Thus, the goal of the current study is to disentangle two potential explanations for these previous findings: (1) Is it the case that the generation of expectations—and the resulting reduction of referential uncertainty—simply does not modulate processing effort, as suggested by Ankener and colleagues? Or (2) can the lack of effects found at the verb in these previous studies be better explained by differences in the referential function of nouns and verbs and their linear order of occurrence? We address these questions in two visually-situated experiments that each employ a common German construction in passive voice wherein the mention of the object noun is followed by a past participle form of the verb. This construction allowed us to reverse the linear order of the object noun and the verb, such that the noun now serves to reduce (some) uncertainty while the subsequent participle provides the necessary information for uniquely identifying the target scene/object in the display.

Experiment 1 is similar to Ankener et al. (2018) in that it employs a visual world paradigm design and assesses pupillary measures of processing difficulty: auditory sentences are presented while listeners view scenes depicting actions being performed on objects. Experiment 2 uses a more exploratory design in which participants preview the same scenes as in Experiment 1, but processing difficulty is assessed via word-by-word reading times in the Grammaticality Maze (G-Maze) task (Forster et al., 2009).

Crucially, both experiments find a similarly graded effect of visually-situated context on the word which uniquely identifies the referent (i.e., the verb in the current design). These findings replicate the effect found at sentence-final nouns in Ankener et al. (2018) and Staudte et al. (2021). The results of Experiment 1 are also consistent with Ankener and colleagues in that we find no modulation of processing effort (as indexed by ICA) on the word where expectations are first generated (i.e., the noun in the current design). In contrast, Experiment 2 also reveals a significant effect at the noun. This combined pattern of effects in Experiment 2 is consistent with Maess et al. (2016) and may, among other things, reflect a trade-off in processing effort between expectation generation and reference resolution.

2. EXPERIMENT 1: PUPILLOMETRIC MEASURE OF NOUN AND VERB PROCESSING

The primary goal of Experiment 1 was to investigate whether the expectedness of a verb, as modulated by a co-present visual referential context, predicts processing effort at that verb. In addition, we also examined whether processing effort at the prediction-generating object noun was modulated by the degree to which the noun reduced uncertainty about the upcoming verb. Following Ankener et al. (2018), we assessed processing effort using ICA, a pupillometric measure of cognitive load which is robust to eye movements and changes in ambient lighting (Marshall, 2000, 2002).

2.1. Methods

2.1.1. Participants

Thirty-two native speakers of German (22 female; 19–40 years old, $M = 25.3$, $SD = 4.9$) were recruited from Saarland University community and were compensated 7.50€ for their participation. All participants reported normal or corrected-to-normal vision. Due to a technical error, the data from one participant could not be used for analyses.

2.1.2. Materials

Participants listened and responded to pre-recorded sentences (in German) while viewing visual displays. Forty experimental sentences were constructed using the following template: *Sag mir, ob* [ARTIKEL OBJEKT], *die von der Figur* [ge-VERB-n] *wird,* [POSITION] *ist* ("Tell me if [ARTICLE OBJECT] that by the figure is being [VERB-ed] is [POSITION]"). For instance, *Sag mir, ob die Rose, die von der Figur gegossen wird, oben ist* ("Tell me if the rose that the figure is watering is on the top"). Queried positions rotated through five possibilities: *oben/unten/links/rechts/fehlt* ("on the top/bottom/left/right/ missing"). Auditory stimuli were recorded with a natural speaking rate and intonation with Audacity 2.2.14 and annotated with Praat 6.0.37 (Boersma, 2001).

The expectedness of the target verb was manipulated by pairing each auditory stimulus with four visual displays in a 1 X 4 design (**Figure 1**). Each display consisted of four scenes, wherein each scene depicted a different action being performed on an object. Displays differed in the number of scenes (1, 3, 4, or 0) that contained the mentioned object (e.g., *die Rose*; "the rose")[1]. In the *1-match* condition, the mentioned object was depicted in only one of the four scenes. Thus, upon hearing the object, the target verb (e.g., *gegossen*; "watering") becomes highly expected. A distractor object (e.g., pizza) was depicted in the remaining three scenes. In the *3-match* condition, the mentioned object was depicted in three of the four scenes, but only one of these scenes was consistent with the target verb. This manipulation decreases the expectedness of the target verb relative to the 1-match condition because upon hearing the object three action verbs were still equally likely. The other two scenes containing the mentioned object served as competitors. The distractor object appeared in the remaining scene. In the *4-match* condition, the mentioned object was depicted in all four scenes, further decreasing the expectedness of the target verb because upon hearing the object four verbs are still possible. Again, however, only one scene was consistent with the target verb. Finally, in the *0-match* condition, the mentioned object did not appear in any of the scenes. Thus, at the point when the object is mentioned, it becomes clear that the visual display cannot provide any information about the target verb. Visual displays were counterbalanced across items such that the mentioned object from one item served as the distractor for another item. For instance, the displays in **Figure 1** were also paired with the sentence, *Sag mir, ob die Pizza, die von der Figur belegt wird, oben ist.* ("Tell me if the pizza that the figure is making is on the top."). Scenes were composed in Paint S (version 5.6.9 (312)5) by arranging images from open source clipart websites (https://openclipart.org; http://clipart-library.com). The position of targets, competitors, and distractors were rotated across items.

In order to disguise the critical manipulation, 40 filler sentences were constructed using three different question

[1] The 1-, 3-, 4-, and 0-match conditions were chosen to facilitate a comparison of the current results with Ankener et al. (2018), Experiment 4. The authors of that study chose the 1- and 4-match conditions in order to maximize the difference in expectation across conditions while simultaneously keeping the overall visual complexity of the displays low. The 3-match condition was chosen in order to allow visual displays to be counterbalanced across items.

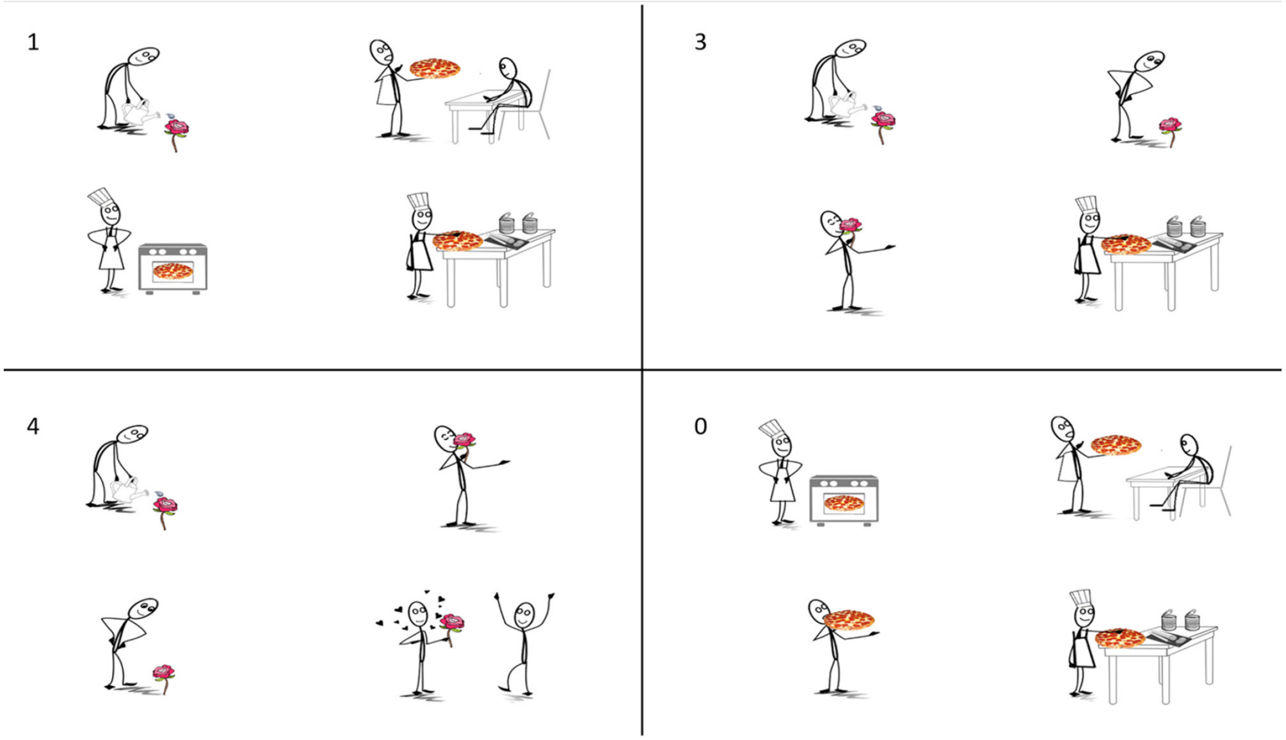


FIGURE 1 | Experiment 1: Example visual display in all four conditions. From **left** to **right** and **top** to **bottom**: *1-match*, *3-match*, *4-match*, and *0-match* conditions, given the sentence: "Tell me if the rose that the figure is watering is on the bottom".

structures: one resembled experimental items, one used a verb-subject-object (VSO) construction (e.g., *Verstaut die Figur die Bluse auf der linken Seite?*; "Does the figure package the blouse on the left?"), and one used a relative clause construction (e.g., *Ist der Kugelschreiber, der von der Figur benutzt wird, links?*; "Is the pen that the figure is using on the left?"). Each filler sentence was paired with one visual display consisting of four scenes. Displays differed across filler items in the number of scenes that contained the mentioned object (7 filler displays contained the object in one scene, 12 contained the object in two scenes[2], 7 contained the object in three scenes, 7 contained the object in four scenes, and 7 contained the object in zero scenes).

Four stimulus lists were created from the above materials according to a Latin square design. Experimental items were counterbalanced across lists such that each participant observed ten items in each condition but no participant observed any item in more than one condition. All participants saw the same fillers. Presentation order was pseudorandomly mixed such that no more than two items of the same condition occurred in sequence. No objects or verbs were repeated across experimental or filler items.

2.1.3. Procedure

Participants were randomly assigned to a stimulus list (8 per list). Following informed consent participants were seated approximately 60 cm from a computer monitor and an Eye-Link 1000+ (SR Research, Ltd.; Mississauga, Ont., Canada). Participants completed a brief, self-paced familiarization session that introduced all the actions that would later appear in the visual displays, but with different objects than in the experimental trials. Each action appeared one at a time while an auditory recording of the corresponding verb was played via external loudspeakers. Participants were then fitted into a chin rest and the eye tracker was calibrated. Each trial began with the presentation of a visual display. Participants were allowed to freely view the display for 1,000 ms, after which the auditory stimulus began. The display remained on screen during the auditory stimulus and for an additional 1,000 ms thereafter. The participants' task was to give the correct answer by pressing a button as quickly and as accurately as possible. Answers were balanced so that *Richtig* ("True") was the correct answer on half of the trials. Feedback was given to participants after each response by displaying (*Korrekt/Inkorrekt*, "Correct/Incorrect"). Participants initiated each new trial by button press. The experiment was implemented in Experiment Builder (SR Research, v 2.1.140) and began with three practice trials. The entire session lasted approximately 45 min.

[2] Because the experimental items did not contain a 2-match condition, we included more filler items in which two scenes contained the mentioned object in order to approach an equal balance of 1-, 2-, 3-, 4-, and 0-match items across the entire study.

2.2. Results

Analyses were conducted using the *lmer* package *lme4* library, version 1.1-10; Bates et al. (2015) in the statistics software package R (version 3.4.2; R Development Core Team, 2017). Fixed effects were contrast-coded and evaluated via likelihood ratio tests implemented in *lmerTest* (Kuznetsova et al., 2017), where denominator *df* was estimated using the Satterthwaite method. Participants and items were entered as crossed, independent, random effects. All models included maximal random effects structures (Barr et al., 2013). We report estimates, standard errors, *t* and *p*-values associated with likelihood ratio tests.

2.2.1. Eye Movements

For presentation purposes only, **Figure 2A** shows the overall proportion of fixations across an averaged trial in all conditions. Visual inspection suggests that when the visual scene allowed for the anticipation of potential target verbs (i.e., 1-match and 3-match conditions), fixations on the scenes containing the mentioned object began to increase at the onset of the noun phrase (left-most dashed vertical line). In contrast, no discrimination is possible in the noun region for the 0-match and 4-match conditions: in the 0-match condition, none of the scenes are relevant, while in the 4-match condition, all of the scenes are equally relevant.

These observations were assessed statistically by comparing whether new inspections to the target scene were detected across conditions within the *noun region* (i.e., noun phrase onset to offset: $M = 658, SD = 103$). The presence/absence of new inspections to each scene were encoded as a binary dependent variable and were analyzed using generalized mixed-effects regression models (GLMER) with a binomial distribution. Orthogonal comparisons between conditions were contrast coded and entered into each model as fixed effects (C1C3: 1-match vs. 3-match, C3C4: 3-match vs. 4-match) with crossed random effects for subjects and items: glmer(number_of_target_inspections ∼ C1C3 + C3C4 + (1 + C1C3 + C3C4 || Subject) + (1 + C1C3 + C3C4 || Item), data, family = "binomial"). Comparisons with the 0-match condition were omitted here as there was no target scene. Results confirmed a significant increase in new inspections of the target scene during the noun region in the 1-match condition ($M = 0.20, SD = 0.40$) compared to the 3-match condition ($M = 0.12, SD = 0.33$) [$\beta = 0.59, SE = 0.16, z = 3.77, p < 0.01$]. In contrast, new target-scene inspections in 3-match where not significantly higher than in the 4-match ($M = 0.09, SD = 0.29$) condition [$\beta = 0.32, SE = 0.19, z = 1.70, p = 0.09$].

2.2.2. Index of Cognitive Activity

ICA reflects fluctuations in the pupil signal that are due to effortful cognitive activity (Marshall, 2000). It is computed as the number of times per second that an abrupt discontinuity (i.e., an ICA event) in the pupil signal is detected, after controlling for any effects due to eye movements and the light reflex (Marshall, 2007). Low values of ICA indicate lower cognitive effort, while higher values reflect greater effort. Importantly, ICA maintains both time and frequency information and can therefore provide a fine-grained analysis of changes in cognitive effort over time.

To assess the effects of visual context on processing effort, we compared the number of ICA events across conditions for two critical regions, namely a noun and a verb region, defined as follows: Consistent with previously established methods (Ankener et al., 2018; Sekicki and Staudte, 2018), analyses for each region were conducted on non-overlapping time windows spanning 600 ms and beginning from the middle of each critical word's duration. ICA values that were 2.5 standard deviations or greater than an individual subject's mean were considered outliers and were excluded from analyses (0.02%).

Figure 2B presents the ICA results for all conditions in the critical noun and verb regions. For presentation purposes only, a baseline region ("Tell me if") is also included. No differences can be seen in either the baseline or noun regions. However, clear differences emerge in the verb region. To assess these observations statistically, the ICA events obtained within the two critical time windows were treated as count variables and analyzed as dependent variables in separate GLMER models with Poisson distributions. The assessed contrasts were C0C1 (0-match vs. 1-match), C1C3 (1-match vs. 3-match), and C3C4 (3-match vs. 4-match). The following model was used to analyze both the noun and the verb region: glmer(ICA ∼ C0C1 + C1C3 + C3C4 + (1 + C0C1 + C1C3 + C3C4 || Subject) + (1 + C0C1 + C1C3 + C3C4 || Item), data, family = poisson (link = "log")). In the noun region, no significant differences between conditions were found ($ps > .16$). In contrast, results for the verb region revealed significantly fewer ICA events in the 0-match condition ($M = 38.55, SD = 15.40$) than the 1-match condition ($M = 42.94, SD = 13.15$) [$\beta = -0.13, SE = 0.03, z = -3.76, p < 0.01$], and significantly fewer ICA events in the 1-match condition than the 3-match condition ($M = 47.12, SD = 11.22$) [$\beta = -0.10, SE = 0.03, z = -3.70, p < 0.01$]. No reliable differences were found between the 3-match and 4-match ($M = 47.08, SD = 12.25$) conditions [$\beta = 0.003, SE = 0.02, z = 0.16, p = 0.87$].

Taken together, these results indicate that there is an effect of multimodal information on the expectedness of the disambiguating verb, and consequently on the effort required to process that verb.

2.3. Discussion

Results from Experiment 1 revealed that processing effort at the target verb was modulated by the number of actions in the display that were consistent with the verb (**Figure 2B**). More specifically, the verb was easier to process when only one verb-consistent action was displayed (1-match condition) than when three or four verb-consistent actions were shown (3- and 4-match conditions), as reflected in lower mean ICA values during the verb region. Somewhat surprisingly, the 0-match condition yielded the lowest ICA values. This finding differs from results in Ankener et al. (2018), where the equivalent condition yielded the highest values. However, in the 0-match condition in the current experiment, participants could already determine at the noun that the correct answer to the question (e.g., "Tell me if the rose...") could only be "Yes" if the question ended with

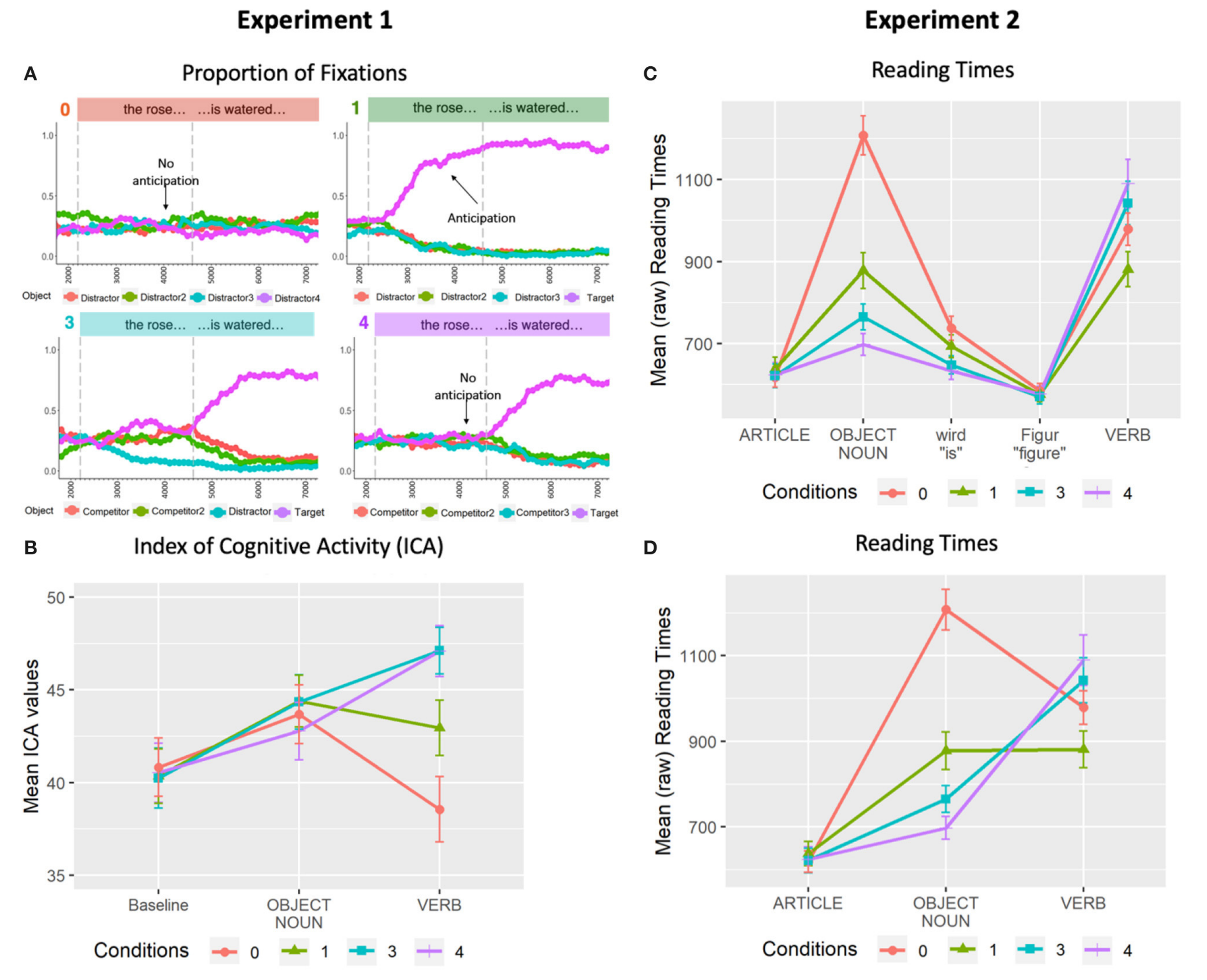


FIGURE 2 | Results from Experiment 1 (left) and Experiment 2 (right). **(A)** Proportion of fixations across averaged trial length in 100 ms bins in the 0-match, 1-match, 3-match, and 4-match conditions. **(B)** Mean ICA values for all four conditions. **(C)** Mean word-by-word (raw) reading times by condition across entire sentence. **(D)** Mean word-by-word (raw) reading times by condition for critical regions only. All error bars indicate 95% confidence intervals.

"...is missing." Thus, listening to the verb was not required in this case, thereby making the verb in the 0-match condition the easiest to process. Finally, statistical analyses further revealed typical anticipatory eye movements during the noun region (i.e., looks only to likely upcoming actions/verbs) even though the difference in new target-inspections between the 3- and 4-match conditions did not reach significance in this study. However, as in Ankener et al. (2018), the distinct allocation of attention (1-match vs. other) did not appear to modulate processing effort at the expectation-generating word.

Taken together, these results indicate that visual context can similarly affect the predictability of both verbs and nouns. We also replicate the lack of an effect on processing effort for the word that provides the constraining information (i.e., the word that reduces referential uncertainty). Thus, regardless of word class, processing effort seems to correlate with visually-situated expectancy but not with the reduction of referential uncertainty.

3. EXPERIMENT 2: G-MAZE READING TIMES AS A MEASURE OF NOUN AND VERB PROCESSING

The aim of Experiment 2 was to replicate the pattern of effects on processing effort from Experiment 1 (i.e., the influence of expectations on the processing of the reference-resolving word, and the lack of an effect on processing of the expectation-generating word) using a different dependent measure. To this end, we collected self-paced reading times using a novel combination of the visual world paradigm and the

Grammaticality Maze (G-Maze) task (Forster et al., 2009). The G-Maze task is a variation of self-paced reading, which has been shown to have better precision (i.e., is less susceptible to spill-over effects) than standard forms of self-paced reading (Witzel et al., 2012), and therefore can more accurately identify the point at which processing time differences emerge during online comprehension (Sikos et al., 2017). Sentences are presented word by word as a sequence of forced choices between two alternatives, only one of which continues the sentence grammatically. If the participant successfully navigates the "maze" by choosing the correct word from each pair, the selected words form a coherent sentence (**Figure 3**). Specifically, we predicted that more predictable verbs would elicit less processing effort, reflected in shorter reading times. Because Experiment 1 and previous work found no impact of the reduction of referential uncertainty on processing effort, we expected to find no differences in reading time on the object noun in the current study. If, however, uncertainty reduction does modulate processing effort in the current design, then the 1-match condition could elicit longer reading times than the 3- and 4-match conditions, because it allows for greater reduction of uncertainty.

3.1. Methods

3.1.1. Participants

Thirty-two native speakers of German (19 female; 18–28 years old, $M = 22.3$, $SD = 2.5$) who had not participated in Experiment 1 were recruited from Saarland University community and were compensated with 10€ for their participation. All participants reported normal or corrected-to-normal vision. Participants who did not successfully complete at least 70% of experimental trials, including both the G-Maze task and the subsequent truth-value judgement task, were excluded ($n = 1$). Two additional participants were excluded due to data corruption issues, resulting in a total of 29 participants.

3.1.2. Materials

On each trial, participants (a) viewed a visual display consisting of four scenes, then (b) completed a G-Maze task presenting a sentence which either did or did not refer to one of the scenes in the display, and finally (c) decided whether the sentence correctly described one of the scenes or not (*Richtig/Falsch*; "True/False"). As in Experiment 1, the expectedness of the target verb was manipulated by pairing each sentence with four visual displays in a 1 X 4 design (**Figure 1**). The same visual displays and conditions were used as in Experiment 1. Linguistic stimuli were adapted from the materials in Experiment 1 by using an alternate template so as to be compatible with a True/False response: [ARTIKEL OBJEKT] *wird von der Figur* [ge-VERB-n] ("[ARTICLE OBJECT] is by the figure [VERB-ed]"). For instance, *Die Rose wird von der Figur gegossen* ("The rose is by the figure watered"). This construction also ensured that the sentence-final word (the verb) was the locus of both sentence-level integration and visual scene identification in the 1-, 3-, and 4-match conditions. Note, however, that in the 0-match condition participants could already recognize upon encountering the noun that the sentence would not refer to any of the depicted scenes. Thus, the correct response to one-quarter of the experimental items was *Falsch* ("False").

To disguise the critical manipulation, the same 40 filler items were used as in Experiment 1, with the following modifications. First, the sentence structures were adapted so as to be compatible with the truth-value judgment task (e.g, *Zum Zeichnen benutzt die Figur den Kugelschreiber* ("The figure uses the ballpoint pen to draw"). Second, sentences varied in length from 5 to 12 words. Finally, half of the filler sentences did not correctly describe a scene in the corresponding display, either because the mentioned action or the mentioned object (as in the 0-match condition) was not present in the display, or because the mentioned object and action (which were both depicted) did not appear together in any of the scenes. The goal of these fillers was to discourage participants from basing their response only upon the presence or absence of the mentioned action and object in any of the scenes. Thus, the correct response for half of the filler items was *Falsch* ("False"). Four stimulus lists were created from the above materials using the same constraints as in Experiment 1.

3.1.3. Procedure

The same general procedure was used as in Experiment 1, with the following modifications. During the familiarization session, the verb corresponding to each scene was presented visually rather than auditorily. During the experimental session each trial began with the presentation of a visual display that participants were allowed to freely view for as long as they wished. Upon pressing a button the display was replaced with the G-Maze task, which began with two crosses (+) that remained on screen for 1,000 ms, indicating where subsequent word pairs would appear. Each word in the sentence (except the first word) was then presented together with a foil word, which was not a grammatical continuation of the sentence[3]. The first word in every sentence was paired with ellipses ("..."). Presentation side (left, right) was randomized such that the correct word appeared equally often on each side. Any punctuation (i.e., comma, period) that appeared with a word also appeared with its foil. Participants were instructed to choose as quickly and as accurately as possible the word that best continued the sentence. Participants indicated their selection by pressing the left or right button on a button box, and the amount of time required for selecting the grammatical continuation was recorded as the reading time for that word. If the correct word was chosen, the next pair of words appeared automatically. However, if a foil word was selected, or if no response was given within 8 s, negative feedback (*Inkorrekt*, "Incorrect") was displayed and the trial was aborted. Once the end of a sentence was reached, participants were asked for a truth value judgment. They used a button box to indicate whether the sentence contained a correct descriptive statement or not. For 62.5% of the trials the correct answer was *Richtig* ("True"). Feedback was given after each response (*Korrekt/Inkorrekt*,

[3] Foils were created in a two-stage process following Sikos et al. (2017). First, a custom Python script randomly selected a foil candidate for each word in each experimental and filler item. Foil candidates were constrained such that they did not appear in bigrams with the correct word at the previous position in the sentence within a large German corpus. Second, each foil was then hand checked by a trained native-German linguist to ensure that it was not a grammatical continuation of the sentence. The same foil was used for identical words across conditions.

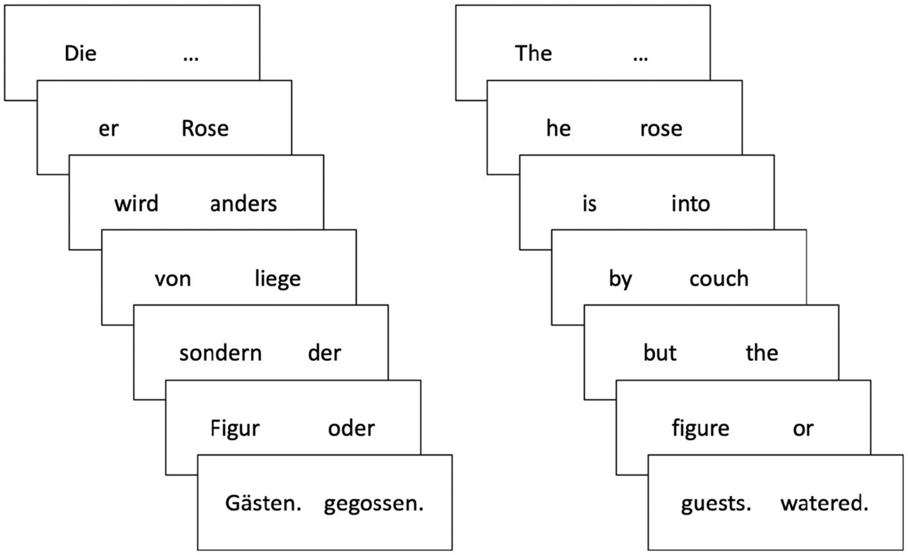


FIGURE 3 | Experiment 2: Structure of the G-Maze task for the example sentence, *Die Rose wird von der Figur gegossen* **(left)**, and its rough English translation "The rose is by the figure watered" **(right)**. Sentences were presented word by word as a sequence of forced choices between two alternatives, only one of which continued the sentence grammatically.

"Correct/Incorrect"). Participants initiated each new trial by button press. After half of the trials were completed, participants were given the opportunity for a short break. The experiment was also implemented in Experiment Builder and began with three practice trials. The entire session lasted approximately 60 min.

3.2. Results

3.2.1. Accuracy

Overall performance on the G-Maze task was near ceiling. Participants successfully completed 96.0% ($SD = 0.20$) of all experimental and filler mazes. Performance on the subsequent truth-value judgment task was also high ($M = 94.3\%$, $SD = 0.23$), confirming that participants were reading the sentences for meaning during the G-Maze task. Only experimental trials for which both the G-Maze task and the truth-value judgment task were completed successfully (92.2%) were included in the analyses reported below.

3.2.2. Reading Times

Noun and verb reading times exceeding 2.5 standard deviations by participant were trimmed, excluding 1.9% (noun) and 2.1% (verb) of the data. The remaining noun and verb reading times were log-transformed and analyzed separately using linear mixed effects models. Orthogonal comparisons between conditions were again contrast coded and entered as fixed effects (C1C3: 1-match vs. 3-match; C3C4: 3-match vs. 4-match; C0C1: 0-match vs. 1-match). The following model was used to analyze both the noun and the verb region: lmer(log(RT) ∼ C0C1 + C1C3 + C3C4 + (1 + C0C1 + C1C3 + C3C4 || Subject) + (1 + C0C1 + C1C3 + C3C4 || Item), data).

For presentation purposes only, **Figure 2C** presents the mean word-by-word (raw) reading times by condition. To visualize changes in processing difficulty across the entire sentence, regions are also included for the article, *wird*, and *Figure*. In order to facilitate a comparison of these results to the ICA results from Experiment 1, **Figure 2D** presents only the key regions. Counter to our predictions, differences between conditions first emerged at the object noun: reading times were faster for the 4-match condition than the 3-match condition ($\beta = 0.10$, $SE = 0.02$, $t = 4.30$, $p < 0.001$); object nouns in the 3-match condition were read more quickly than in the 1-match condition ($\beta = 0.12$, $SE = 0.03$, $t = 4.12$, $p < 0.001$); and object nouns in the 1-match condition were read more quickly than in the 0-match condition ($\beta = 0.30$, $SE = 0.02$, $t = 13.39$, $p < 0.001$). As predicted, verbs were read more quickly in the 1-match condition than the 3-match condition ($\beta = -0.14$, $SE = 0.03$, $t = -5.27$, $p < 0.001$). Verbs in the 3-match condition were read more quickly than the 4-match condition, although this difference did not reach significance ($\beta = -0.02$, $SE = 0.02$, $t = -0.83$, $p = 0.41$). In addition, verbs in the 1-match condition were read more quickly than the 0-match condition ($\beta = 0.12$, $SE = 0.02$, $t = 4.79$, $p < 0.001$).

3.3. Discussion

Figures 2B,D present the key results from both experiments side-by-side for comparison. Reading times from Experiment 2 revealed a graded effect of visual context on processing effort at the object noun. These results showed that nouns were easiest to process in the 4-match condition and became parametrically more difficult as fewer and fewer objects in the display matched the mentioned noun. Processing of the noun was most difficult in the 0-match condition. This pattern may indicate that scenes primed/preactivated the mentioned nouns in Experiment 2: Participants were asked to carefully view and remember the scenes, which were then removed before the G-Maze task began. Thus, the noun (e.g., rose) may have been more prominent—and thus potentially remembered better—when it was depicted in multiple scenes.

Reading times in the verb region largely replicated the ICA results from Experiment 1. That is, verbs were read most quickly when the noun reduced all uncertainty about which scene was being referred to, and consequently made the upcoming verb highly predictable (1-match condition). In contrast, when referential uncertainty remained, participants took longer to process the verb. However, the reading time difference between 3-match and 4-match conditions did not reach significance. This result also replicates previous findings (Ankener et al., 2018; Staudte et al., 2021) and suggests that discrimination between three and four potential target objects/scenes has relatively little impact on processing effort. Finally, when no expectations for the verb are generated because the mentioned object is not depicted in any of the scenes (0-match condition), processing time increases relative to the 1-match condition. Interestingly, however, reading times indicate that the verb in the 0-match condition is still easier to process than when there is some referential uncertainty (3- and 4-match conditions). We attribute this intermediate level of processing effort to a combination of two effects: a facilitation effect due to recognizing that the verb is not relevant for the answer (i.e., recognizing that the answer will be "False") and an inhibition effect due to not being able to anticipate the verb, despite still having to fully process the verb in order to complete the G-Maze task (but see also the General Discussion).

4. GENERAL DISCUSSION

The aim of the current studies was to investigate whether the generation of expectations for upcoming words in visually-situated language comprehension—and the resulting reduction of referential uncertainty—simply does not modulate processing effort, as suggested by Ankener et al. (2018) and Staudte et al. (2021), or alternatively, whether the lack of effects found at the expectation-generating word in these previous studies can be better explained by word category differences in the referential function of nouns and verbs in a reversed word order.

To address these questions, we conducted two visually-situated comprehension experiments, which essentially reverse the functional roles of nouns and verbs. Experiment 1 employed the visual world paradigm and ICA as a measure of processing effort. Experiment 2 sought to validate those findings using a more exploratory method in which processing difficulty was assessed via reading times in the G-Maze task.

4.1. Reference Resolution

In the current design, comprehenders were only able to uniquely resolve the referent upon encountering the verb. Crucially, both experiments and dependent measures revealed a similarly graded influence of situated context on processing effort at the verb: results of both studies showed an increase in processing effort as the number of depicted actions matching the verb increased. These findings largely replicate the effect found at sentence-final nouns in Ankener et al. (2018), where the noun served the role of uniquely identifying the referent. The results therefore indicate that verbs as well as nouns can be used to resolve referents in a visual scene—despite the inherent functional differences due to word category—and thereby allow the reader to recover the intended proposition of the sentence. In addition, processing of the reference-resolving word (whether it be a noun or a verb) reliably benefits when prior linguistic and visual information combine to generate concrete expectations for that word.

One obvious difference in results across Experiments 1, 2, and Ankener et al. (2018) is the pattern of effects found at the reference-resolving word in the 0-match condition. Whereas Ankener et al. (2018) found that the 0-match condition elicited the highest processing effort in this region, ICA results from Experiment 1 revealed that this condition elicited the lowest processing effort. Moreover, reading time results from Experiment 2 indicate that processing effort for the reference-resolving word in the 0-match condition was intermediate between the 1-match and 3-match conditions. However, these differences can be readily explained as a consequence of the different tasks used in each study. In contrast to Ankener et al. (2018), participants in Experiment 1 did not need to fully process the reference-resolving word in the 0-match condition in order to successfully respond to the query (e.g., Tell me if the rose that is by the figure watered is on the top/bottom/left/right/is missing). This is because the mentioned object did not appear in any of the scenes, thus it became immediately clear upon encountering the noun that the correct response could only be "is missing." In Experiment 2, however, while processing of the verb was not strictly necessary to successfully complete the subsequent truth-value judgment task (e.g., The rose is by the figure watered; True/False), the G-Maze task forces each word in the sentence to be accessed and integrated into the unfolding utterance representation—only then can the comprehender select the correct word instead of a foil and successfully navigate the maze to the end of the sentence. In the 0-match condition, it might be obvious that the correct response will eventually be "False," however the verb must still be fully processed and selected beforehand. Moreover, in contrast to the 1-match condition, in which the verb can be anticipated, comprehenders in the 0-match condition do not have the benefit of visual preactivation of the verb. Thus, the combination of these two processes (i.e., facilitation in the truth judgment task and lack of preactivation) may explain why reading times for the 0-match condition in Experiment 2 fall between the 1-match and 3-match conditions.

4.2. Generation of Expectations

In both experiments of the current study, expectations for upcoming verbs were generated at the object noun. Consistent with Ankener et al. (2018), ICA results from Experiment 1 revealed no modulation of processing effort at the expectation-generating word. In contrast, however, reading time results from the same expectation-generating noun in Experiment 2 showed a reliable modulation of effort: processing difficulty was greatest in the 0-match condition (when the mentioned object did not appear in the display) and parametrically decreased as more potential target objects were depicted in the visual context. In fact, this pattern of effects at the expectation-generating noun appears to be the inverse of the subsequent pattern found at the reference-resolving verb, where processing difficulty was greatest in the 4-match condition and decreased as fewer potential target objects were depicted. Although this combined pattern of effects in Experiment 2 is consistent with the notion of uncertainty reduction, it is also consistent with the results of Maess et al. (2016), which argues for a direct "trade-off" in processing effort between preactivation and a later point in time when that preactivated word is encountered. On this account, the effort expended at the noun reflects preactivation of semantic features of the expected, upcoming scene description, which is then offset by a complementary facilitation in processing the subsequent (expected) verb (Maess et al., 2016). Similarly, these results are in line with the findings on pre-updating (during verb processing) and the processing trade-off with the predicted word (noun) in Ness and Meltzer-Asscher (2018).

Yet another explanation for the parametric effects found at the expectation-generating noun in Experiment 2 is that the difference in reading times may have been driven by task-based effects rather than uncertainty reduction *per se*. In adapting the materials of Experiment 1 to Experiment 2, we chose to remove the visual display during the G-Maze portion of the task. This decision was driven by two related concerns: (1) that a co-present visual scene could potentially draw the participant's gaze away from the G-Maze task, and (2) that this effect would vary systematically by condition. Accordingly, participants were instructed to view each display carefully so that they could later respond as to whether or not the sentence corresponded to one of the displayed scenes. One unintended consequence of this decision is that participants may have utilized a non-trivial amount of working memory to accomplish this goal, which may then have influenced the processing of the object noun. For instance, it is possible that some proportion of participants consciously or unconsciously labeled the objects or actions depicted in each display in order to better remember the key information. Exit survey results provide some support for this account. When asked whether they used any particular strategies in order to successfully perform the task, 15 participants reported that they tried to memorize either the depicted objects, actions, or both. If this was indeed the case, explicitly labeling objects during the preview phase of the task would presumably preactivate[4] the mentioned noun such that its subsequent processing would be facilitated. This could therefore explain the reading time advantage for the noun in those conditions in which the mentioned object was present in the visual context (1-, 3-, and 4-match conditions) relative to when it was absent (0-match condition). Moreover, preactivation of the noun is likely to be greater when more of the scenes in the display contain the mentioned object. Thus, this account is also consistent with the parametric effects found at the expectation-generating noun.

If this final explanation is correct—and the observed noun effects in Experiment 2 are therefore specific to the procedure used in our G-maze design, wherein the visual context was removed before sentence processing began—then one of our original research questions would remain unresolved: Why were no effects of uncertainty reduction or preactivation/pre-updating observed during the processing of the constraining word, neither on the noun in Experiment 1 (noun), nor on the verb in Ankener et al. (2018) and Staudte et al. (2021)? Here we speculate that the co-present visual scene used in these latter experiments may have played a role in why processing effort was not affected in such cases, and we offer several explanations as to why this might be the case. Firstly, participants in those experiments did not necessarily need to maintain (one or more) predictions in working memory. Instead, they could simply rely on the external representations (Spivey et al., 2004) visible in the co-present visual display to mentally flag objects with regard to match vs. no-match, rather then computing and maintaining representations of all matches. Thus, processing effort might not have been affected by whether or not one or more objects/actions in the visual display served as potential verb (arguments). Secondly, the amount of referential uncertainty that is reduced when going from four to one potential objects/actions is relatively small, at least when compared to the difference between high and low constraining words in purely linguistic contexts with no co-present visual scene, as in Maess et al. (2016), Ness and Meltzer-Asscher (2018). In such cases, low constraining words allow for dozens or even hundreds of continuations (comparable to the 0-match conditions in the current studies), while high constraining words typically license only a few concrete predictions. Thirdly, the reduction of uncertainty and the maintenance of multiple predictions could each elicit processing effort, which could then cancel each other out. That is, while reducing uncertainty from four to one option might require increased effort, less effort would then be required to maintain that single object/action representation in working memory than four representations. In contrast, the comparison of processing effort across conditions in Maess et al. (2016) was not among different numbers of preactivated representations, but was instead between preactivation and the lack thereof.

All of these alternative explanations are grounded in the specifics of simultaneously perceiving linguistic and visual information. This makes Experiment 2 particularly interesting—although exploratory—because no objects were co-present and instead had to be mentally-represented, predicted, and maintained in working memory. However, further research is needed to tease apart whether the effects during noun processing in Experiment 2 do indeed index any of the above mentioned "forward-looking" mechanisms to predict upcoming content, or whether they are instead a result of preactivation based on the previously shown scenes.

[4] Note that preactivation here refers to the noun being preactivated through viewing and memorizing the preceding visual display as opposed to the noun preactivating the upcoming verb (al features).

5. CONCLUSION

In sum, the results of the current study indicate that verbs as well as nouns can be used to resolve referents in a visual scene, and thus to reconstruct the speaker's intended proposition. Moreover, processing of the reference-resolving word—whether it be a noun or a verb—reliably benefits from the prior linguistic and visual information that leads to the generation of concrete expectations for that word.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories: https://doi.org/10.7910/DVN/ST41P8.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Deutsche Gesellschaft für Sprachwissenshaft (DGfS). Participants provided their written informed consent prior to participation.

AUTHOR CONTRIBUTIONS

LS, KS, and MS conceived and designed the studies and analyses and contributed to the manuscript. KS collected the data. All authors contributed to the article and approved the submitted version.

FUNDING

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 232722074—SFB 1102/Gefördert durch die Deutsche Forschungsgemeinschaft (DFG)—Projektnummer 232722074—SFB 1102. Open Access publication costs were granted by SFB 1102.

ACKNOWLEDGMENTS

We would like to thank Christine Muljadi for composing the visual stimuli and collecting data for Experiment 1. We also thank Christine Ankener and Linda Sommerfeld for valuable discussions and assistance with data collection.

REFERENCES

Ankener, C. S., Sekicki, M., and Staudte, M. (2018). The influence of visual uncertainty on word surprisal and processing effort. *Front. Psychol*. 9:2387. doi: 10.3389/fpsyg.2018.02387

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, 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

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot. Int*. 5, 341–345.

Demberg, V., and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109, 193–210. doi: 10.1016/j.cognition.2008.07.008

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., and Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Res*. 1146, 75–84. doi: 10.1016/j.brainres.2006.06.101

Forster, K. I., Guerrera, C., and Elliot, L. (2009). The maze task: measuring forced incremental sentence processing time. *Behav. Res. Methods* 41, 163–171. doi: 10.3758/BRM.41.1.163

Frank, S. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Top. Cogn. Sci*. 5, 475–494. doi: 10.1111/tops.12025

Hale, J. (2016). Information-theoretical complexity metrics. *Lang. Linguist. Compass* 10, 397–412. doi: 10.1111/lnc3.12196

Kutas, M., and Hillyard, S. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* 4427, 203–205. doi: 10.1126/science.7350657

Kuznetsova A., Brockhoff P. B., and Christensen R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Softw*. 82, 1–26. doi: 10.18637/jss.v082.i13

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

Maess, B., Mamashli, F., Obleser, J., Helle, L., and Friederici, A. D. (2016). Prediction signatures in the brain: semantic pre-activation during language comprehension. *Front. Hum. Neurosci*. 10:591. doi: 10.3389/fnhum.2016.00591

Marshall, S. P. (2000). U.S. Patent No. 6,090,051. Washington, DC: U.S. Patent and Trademark Office.

Marshall, S. P. (2002). "The index of cognitive activity: measuring cognitive workload," in *Proceedings of the IEEE 7th Conference on Human Factors and Power Plants* (Scottsdale, AZ: IEEE), 7. doi: 10.1109/HFPP.2002.1042860

Marshall, S. P. (2007). Identifying cognitive state from eye metrics. *Aviat. Space Environ. Med*. 78, B165–B175.

Ness, T., and Meltzer-Asscher, A. (2018). Predictive pre-updating and working memory capacity: evidence from event-related potentials. *J. Cogn. Neurosci*. 30, 1916–1938. doi: 10.1162/jocn_a_01322

Sekicki, M., and Staudte, M. (2018). Eye'll help you out! How the gaze cue reduces the cognitive load required for reference processing. *Cogn. Sci*. 42, 2418–2458. doi: 10.1111/cogs.12682

Sikos, L., Greenberg, C., Drenhaus, H., and Crocker, M. W. (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)* (London), 3168–3173.

Smith, N. J., and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition* 128, 302–319. doi: 10.1016/j.cognition.2013.02.013

Spivey, M. J., Richardson, D. C., and Fitneva, S. A. (2004). "Thinking outside the brain: spatial indices to visual and linguistic information," in *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*, eds J. M. Henderson and F. Ferreira (New York, NY: Psychology Press), 161–189.

Staudte, M., Ankener, C., Drenhaus, H., and Crocker, M. W. (2021). Graded expectations in visually situated comprehension: costs and benefits as indexed by the N400. *Psychon. Bull. Rev*. 28, 624–631. doi: 10.3758/s13423-020-01827-3

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

Van Berkum, J. J., Koornneef, A. W., Otten, M., and Nieuwland, M. S. (2007). Establishing reference in language comprehension: an electrophysiological perspective. *Brain Res*. 1146, 158–171. doi: 10.1016/j.brainres.2006.06.091

Witzel, N., Witzel, J., and Forster, K. (2012). Comparisons of online reading paradigms: eye tracking, moving-window, and maze. *J. Psycholinguist. Res.* 41, 105–128. doi: 10.1007/s10936-011-9179-x

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 Psychology

HYPOTHESIS AND THEORY
published: 31 May 2021
doi: 10.3389/fpsyg.2021.672408

An Information-Theoretic Account of Semantic Interference in Word Production

*Richard Futrell**

Department of Language Science, University of California, Irvine, Irvine, CA, United States

I present a computational-level model of semantic interference effects in online word production within a rate–distortion framework. I consider a bounded-rational agent trying to produce words. The agent's action policy is determined by maximizing accuracy in production subject to computational constraints. These computational constraints are formalized using mutual information. I show that semantic similarity-based interference among words falls out naturally from this setup, and I present a series of simulations showing that the model captures some of the key empirical patterns observed in Stroop and Picture–Word Interference paradigms, including comparisons to human data from previous experiments.

Keywords: language production, information theory, bounded rationality, semantic interference effect, Stroop, rate-distortion

1. INTRODUCTION

In cognitive science and related fields, **bounded rationality** is the idea that our cognitive systems are designed to take actions that are approximately optimal, given that only limited computational resources are available for calculating the optimal action (Simon, 1955, 1972; Kahneman, 2003; Howes et al., 2009; Lewis et al., 2014; Gershman et al., 2015; Lieder and Griffiths, 2019). The idea is appealing because it maintains the mathematical precision of theories based on rationality, while avoiding the paradoxes and empirical shortcomings that come from claiming that human beings act in ways that are entirely rational. There has been recent interest in formalizing bounded rationality within the mathematical framework of rate–distortion theory (Berger, 1971; Cover and Thomas, 2006) with applications to cognitive science (Sims, 2016, 2018; Zaslavsky et al., 2018; Gershman, 2020).

In this paper, I apply rate–distortion theory to derive a model of online word production. The goal is to model the difficulty of online word production, as measured using psychometric dependent variables, such as reaction time and rates and patterns of errors. The main contribution of this paper is to show that rate–distortion theory generically predicts the well-documented **semantic interference effects** that a subject experiences when trying to produce a target word in the presence of a semantically related distractor. For example, the Stroop task famously exhibits interference (Stroop, 1935): given a stimulus, such as the word **BLUE** printed in red ink, and an instruction to name the color of the ink, it is hard to produce "red" because of interference from the similar word "blue." A similar kind of interference is present in the Picture–Word Interference task, where a drawing must be named in the presence of a superimposed distractor word (Lupker, 1979; Starreveld and La Heij, 2017). Beyond the basic interference effect, I show that rate–distortion theory predicts a number of key phenomena observed in such tasks.

OPEN ACCESS

Edited by:
Matthew W. Crocker,
Saarland University, Germany

Reviewed by:
Chiara Gambi,
Cardiff University, United Kingdom
Flavia De Simone,
Suor Orsola Benincasa University, Italy
Ardi Roelofs,
Radboud University Nijmegen, Netherlands

****Correspondence:***
Richard Futrell
rfutrell@uci.edu

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

Received: *25 February 2021*
Accepted: *27 April 2021*
Published: *31 May 2021*

Citation:
Futrell R (2021) An Information-Theoretic Account of Semantic Interference in Word Production. Front. Psychol. 12:672408. doi: 10.3389/fpsyg.2021.672408

2. BACKGROUND: RATE–DISTORTION THEORY OF CONTROL

2.1. Bounded Rationality

Ultimately, our cognitive systems implement an **action policy**: a function from sensory inputs to motor outputs. For example, an animal might see another animal and decide among a large set of possible actions, including attacking, approaching, ambushing, fleeing, etc. In general, we can conceive of an action policy as a stochastic function mapping states S (including perceptual, physiological, and memory information) to motor actions A:

$$q : S \to A.$$

We can also think of the policy as a *probability distribution* on actions given states, where $q(a|s)$ denotes the probability of taking action a in state s.

A **bounded-rational action policy** is a policy that chooses an action to maximize some measure of reward, or equivalently, to minimize the cost of the *consequences* of taking a certain action in the world, subject to a constraint on the computational resources used in finding and implementing this action. These resources include factors, such as time—in many circumstances, it may be more important to act quickly than to take the time to compute the best action—as well as physiological resources, such as the energy required to perform computations. Formally, letting $D(s, a)$ represent the **action cost** or the cost of the consequences of taking action a in state s, and letting $C(s, a)$ denote the **computation cost** required to compute the action a given state s, then the overall cost for a policy q can be written as

$$\mathcal{L}(q) = \left\langle D(s,a) + \frac{1}{\gamma} C(s,a) \right\rangle, \tag{1}$$

where $\langle \cdot \rangle$ denotes an average over the joint probability distribution on states and actions given those states $p(s)q(a|s)$, and $\frac{1}{\gamma}$ is a scalar value which indicates how much a unit of computation cost C should be weighed against a unit of action cost D. The scalar γ can also be viewed as a parameter giving the amount of resources available for computation: high γ means that the agent is willing to perform a lot of computation in order to minimize the action cost D.

The expression $\mathcal{L}(q)$ in (1) is called the **control objective**, and a bounded optimal action policy is derived by minimizing it:

$$q_{\text{bounded rational}} = \arg\min_q \mathcal{L}(q),$$

where the minimization is over the set of all possible policies. The bounded-rational policy reduces to the fully rational policy in the case when computation costs have negligible importance, i.e., $\frac{1}{\gamma} \to 0$ in Equation (1).

Without further specifications, the theory of bounded rationality goes no farther than the formalization above. Given a set of cost functions, the bounded rational action policy is derived as the solution to a multi-objective minimization problem involving those cost functions. The theory only makes precise predictions when the cost functions and their relative weights are further specified. Below, we will see how we can do this in a principled way using tools from information theory.

2.2. Rate–Distortion Theory

Rate-distortion theory is the mathematical theory of lossy communication and compression, a subfield of information theory. It provides mathematical tools to answer questions like: if I want to transmit a picture of a zebra to you, and I do not have the capacity to send it to you perfectly, how can I encode the image such that your received picture looks approximately like what I sent? This problem involves two constraints: (1) my capacity to transmit information (called **rate**), and (2) a measure of how much your received picture differs from my picture (this measure is called **distortion**). Rate–distortion theory describes the problem of finding a data encoding which minimizes the distortion subject to a constraint on the rate.

The link between rate–distortion theory and bounded rational action policies was not immediately clear, although the original paper on rate–distortion theory did note a connection with control theory (Shannon, 1959, p. 350). The key insight that has enabled researchers to link these two theories is that rate–distortion theory can be applied to constrain the perception–action loop. The idea is to treat an action policy as a communication channel from sensory input to motor output. Then the action cost D in Equation (1) is the distortion, and the computation cost C in Equation (1) is the rate. This connection was introduced first in the economics literature by Sims (2003, 2005, 2010) under the name **rational inattention**: the idea being that an agent might decide not to attend to certain information because the computational resources required to sustain that attention are not worth the investment. The idea was then picked up in the robotics, cybernetics, machine learning, and psychology literature (van Dijk et al., 2009; Tishby and Polani, 2011; Rubin et al., 2012; Ortega and Braun, 2013; Genewein et al., 2015; Sims, 2016, 2018; Gershman and Bhui, 2020, among others).

In the **rate–distortion theory of control** (RDC), a bounded-rational action policy is derived by minimizing the following control objective:

$$\mathcal{L}(q) = \left\langle D(s,a) \right\rangle + \frac{1}{\gamma} I[S : A], \tag{2}$$

where $D(s, a)$ is the distortion or action cost for taking action a in state s, and $I[S : A]$ denotes the **mutual information** between the random variables S representing the state and A representing the action policy:

$$I[S : A] = \left\langle \log \frac{q(a|s)}{q(a)} \right\rangle,$$

where the probability $q(a)$ is the marginal probability of taking action a under the policy q, averaging over all states:

$$q(a) = \sum_s p(s) q(a|s).$$

The substantive claim of the RDC is that computation costs should be modeled as the mutual information between states and actions $I[S:A]$. This quantity can be interpreted as the amount of information that must be extracted from S in order to specify A (Sims, 2003), or as the information throughput of a controller implementing the policy $q(a|s)$ (Fan, 2014). I will argue below that this is a natural measure of computation cost, and that it subsumes many other measures.

I summarize four converging motivations for the use of the mutual information between states and actions $I[S:A]$ (and related measures, such as relative entropy) as a measure of computation cost. I provide pointers into the literature for the full forms of these arguments. See also Zénon et al. (2019, section 4) for a comprehensive discussion and review.

1. **Computation time.** The mutual information reflects the *search time* taken to find the action A given state S by a rejection sampling algorithm. When the mutual information $I[S:A]$ is lower, the correct action can be found using fewer samples from $q(a)$ (Braun and Ortega, 2014, section 2).
2. **Algorithmic complexity.** The mutual information reflects how many bits of information an agent must store to remember the policy, or how many bits of information an agent needs to observe to learn the policy. This argument is presented in a PAC-Bayes framework by Rubin et al. (2012), who also show that action policies with a mutual information penalty are less prone to overfitting to their immediate environment.
3. **Free energy.** The RDC objective in Equation (1) is technically a **free energy** functional (Ortega and Braun, 2013), bringing the theory in line with neuroscientific theories of brain function formulated in terms of minimizing free energy (Friston, 2010).
4. **Congruence with empirically-observed laws of behavior.** Information-theoretic models of cognitive control have proposed that the time taken to initiate an action should be proportional to the amount of information required to specify that action (Fan, 2014). We can derive well-validated empirical laws of behavior under this assumption. For example, Hick's Law is the observation that the time taken to decide among a set of actions A is directly proportional to the logarithm of the number of possible actions $\log|A|$ (Hick, 1952; Hyman, 1953). The RDC computation cost $I[S:A]$ reduces to $\log|A|$, yielding Hick's Law, in the case where (1) an agent is deciding among a set of actions A, (2) the default policy $q(a)$ is uninformative about which action to take, and (3) the state-dependent policy $q(a|s)$ specifies the desired action deterministically.

In summary, there is a convergence among a number of previous intuitive notions of computation cost, all of which point toward $I[S:A]$ as a reasonable measure. In addition to these theoretical arguments, a growing neuroscience literature has linked information measures, such as $I[S:A]$ to brain activity in the prefrontal cortex (Koechlin and Summerfield, 2007; Fan, 2014).

The form of the RDC objective in Equation (2) is only the simplest member of a family of possible control objectives. In reality, a cognitive agent must integrate information from many different inputs and produce motor output on many different actuators. Each input and each motor output can be associated with its own channel, with its own information-based penalty. Multiple input channels can be modeled by adding further weighted mutual information terms to Equation (2) (for example, see van Dijk and Polani, 2011, 2013; Genewein et al., 2015). In fact, we will see that our model of Picture–Word Interference requires at least two input channels: a top-down goal signal and a bottom-up perceptual signal.

2.3. Solutions to the RDC Objective

The policies admitted under the rate–distortion theory of control have a common mathematical form. The minima of Equation (2) obey the following equations:

$$\begin{aligned} q(a|s) &= \frac{1}{Z(s)} q(a) \exp\{-\gamma D(s,a)\} \\ q(a) &= \sum_s p(s) q(a|s) \\ Z(s) &= \sum_a q(a) \exp\{-\gamma D(s,a)\}. \end{aligned} \tag{3}$$

Note that the Equation (3) do not specify a policy uniquely. The equations are called self-consistent, meaning that any $q(a|s)$, $q(a)$, and $Z(s)$ jointly constitute a minimum of the control objective as long as they satisfy the three equations simultaneously. In general, multiple solutions can exist. A numerical solution to the equations can be found by starting with a random value of $q(a|s)$, then evaluating the equations iteratively until a fixed point is reached.

One generalization that we can deduce immediately from this system of equations is that RDC policies favor re-use of common actions. We can see this because the factor $q(a)$ in Equation (3) will be high for actions that are taken frequently across all states. Therefore, these actions will be preferred, sometimes in lieu of the action that would be more appropriate in a particular state s. Intuitively, the factor $q(a)$ represents a "habit": a propensity to take a certain action regardless of the present context (van Dijk and Polani, 2013; Wood and Rünger, 2016; Gershman, 2020).

2.4. Link to Behavioral Measures

The RDC describes the derivation of bounded-rational action policies, but does not immediately make predictions about the timing of these actions nor other behavioral and neural dependent measures that are commonly deployed in the study of cognitive control and language production. A linking hypothesis is required from the mathematical policy $q(a|s)$ to predictions about dependent measures, such as reaction time, the usual measure of difficulty in word production studies.

There are a number of perspectives in the psychological literature on the relationship between reaction times (RTs) and information-theoretic measures of complexity (Laming, 1968, 2003; Luce, 2003; Ortega and Braun, 2013; Fan, 2014; Zénon et al., 2019; Lynn et al., 2020). The simplest possible hypothesis is that the time required to initiate an action is linearly proportional to the amount of computation that needs to be done to select the action. For example, Fan (2014) conceptualizes cognitive control as the means by which uncertainty about the output action is reduced at a constant rate in terms of bits per millisecond. I adopt this linking hypothesis here, with a modification to account for the fact that the computation required to select an action breaks into multiple parts, which I call computation cost and decision cost:

1. **Computation cost**. The computation required to produce the action policy $q(a|s)$. This is equal to the cost term in the control objective $\mathcal{L}$ that generates $q(a|s)$. For example, given the control objective in Equation (2), the average computation cost is the mutual information $I[S:A] = \left\langle \log \frac{q(a|s)}{q(a)} \right\rangle$. For a particular action a in state s, the cost is the pointwise mutual information $\log \frac{q(a|s)}{q(a)}$. This notion of computation cost combines Zénon et al. (2019)'s notions of "perceptual cost" and "automatic cost." For human behavioral work relating this notion of computation cost to computation time, see Ortega and Stocker (2016) and Schach et al. (2018).
2. **Decision cost**. A policy $q(a|s)$ is a probability distribution on actions, but in any given state, an agent must take a single action. Decision cost is the cost associated with selecting a single action a^* from a distribution $q(a|s)$; it represents a decision that still needs to be made (perhaps randomly) after considering state information. I take decision cost to be equal to the KL divergence from $q(a|s)$ to a delta distribution specifying a single action a^*:

$$\begin{aligned} D_{\mathrm{KL}}[\delta_{aa^*}||q(a|s)] &= \left\langle \log \frac{\delta_{aa^*}}{q(a|s)} \right\rangle \\ &= -\log q(a^*|s), \end{aligned}$$

   where δ_{aa^*} is a Kronecker delta function (equal to 1 when $a = a^*$ and 0 otherwise). Thus, decision cost comes out to be the surprisal (negative log probability) of the action a^* given the state s under the action policy.

It stands to reason that both computation cost and decision cost make contributions to dependent measures, such as reaction time, although perhaps not according to a simple function. In this work I will present computation and decision cost in terms of bits of information, and where appropriate I will discuss their possible translation into observable dependent measures.

There have been other, more complex proposals about the link between RDC policies and observable measures, such as reaction time. For example, Ortega and Braun (2013, p. 10–11) link RDC policies to drift–diffusion models of choice behavior (Bogacz et al., 2006). While I do not pursue these other linking hypotheses here, they could provide different perspectives or more precise predictions in future work.

2.5. Level of Analysis

RDC as applied to word production is a computational-level theory in Marr's sense (Marr, 1982), meaning that it attempts to model the problem that is being solved in language production. Because it is stated at this level of abstraction, it is not necessarily in conflict with existing more mechanistic models of word production. RDC states simply that the cognitive cost of taking certain actions is determined by a trade-off of minimizing action cost while also minimizing information-processing costs, measured using mutual information. This trade-off might be implemented in terms of spreading activation in networks with constrained topology, production rules, etc. Nevertheless, it will be interesting to see where the predictions of more mechanistic theories diverge from those of the more abstract RDC.

To sum up this section, I have presented the rate–distortion theory of control (RDC) as a model of bounded-rational action. Below, I will present a new application of this model to model human word production, which exhibits a property of the model which has not previously been explored. In particular, I will show that similarity-based interference effects, which are common in word production as well as other aspects of cognition, arise as a generic prediction of RDC models.

3. INTERFERENCE IN THE RATE–DISTORTION THEORY OF CONTROL

In this section I will demonstrate the basic mechanism by which RDC predicts similarity-based interference effects.

3.1. The Empirical Phenomena

The term **similarity-based interference** encompasses a large number of phenomena in human perception, action, and memory. It refers to the idea that percepts, actions, or memories are confused for each other when they are "similar" according to some metric (Shepard, 1987), that is, when they share features or associated cues. Furthermore, there may be increased latency in identifying a percept, retrieving information from memory (Jäger et al., 2017), or initiating in action (Stroop, 1935) in the presence of some "similar" distractor. Capturing similarity-based interference is a key goal of cognitive models, including those based on cue-based retrieval, spreading activation, and production rules (Watkins and Watkins, 1975; Ratcliff, 1978; Anderson and Lebiere, 1998; Roelofs, 2003).

3.2. RDC Account

Similarity-based interference arises generically in RDC models because the action cost $D(s, a)$ naturally defines a similarity metric among actions, an insight used by Sims (2018) in his model of generalization in absolute identification tasks. The function $D(s, a)$ gives the cost of taking action a in state s. Two actions are similar when they have similar cost, that is, when there is low cost for failing to distinguish them. Accordingly, we can define a distance metric between two actions. In state s, let a_s be the action with minimal cost, and a_d be any other action. The state-dependent distance metric among actions can be defined as a function

$$d(a_s, a_d) = D(s, a_d) - D(s, a_s).$$

This distance metric[1] will play the role of the distortion metric in rate–distortion theory.

[1] The function $d(a_s, a_d)$ is technically a pre-metric. It satisfies $d(a, a) = 0$ for all actions a, and it is always non-negative. It is non-negative because a_s is defined as the action with minimal cost in state s. The function is only a pre-metric, not a full metric, because it is not generally symmetrical. That is, $d(a_s, a_d) \neq d(a_d, a_s)$ in general.

Now that we have a distance metric among actions, we can see that interference effects arise even in the simplest formulation of the RDC. Suppose the control system is attempting to solve the following problem: in a state s (for example, seeing a picture of an apple), there is a single unique target action a_s corresponding to that state (for example, saying the word "apple"). The agent is attempting to generate the right target action in state s. In this setting, RDC predicts generally that the probability that any two actions (e.g., words) a_s and a_d are confused will increase as they get closer in the distance metric $d(a_s, a_d)$—thus predicting similarity-based interference among competitors.

More formally, let the control objective be

$$\mathcal{L}(q) = \Big\langle d(a_s, a) \Big\rangle + \frac{1}{\gamma} I[S:A]. \tag{4}$$

This equation expresses that the agent will try to minimize the average distance between the selected action a and the target action a_s, subject to a computation cost of $\frac{1}{\gamma}$ units per bit of information from the states S used to specify actions A. Then following the logic in Equation (3), the bounded-rational policy has the form

$$\begin{aligned} q(a|s) &= \frac{1}{Z(s)} q(a) \exp\{-\gamma d(a_s, a)\} \qquad (5) \\ q(a) &= \sum_s p(s) q(a|s) \\ Z(s) &= \sum_a q(a) \exp\{-\gamma d(a_s, a)\}. \end{aligned}$$

This policy exhibits exponentially-decaying interference effects as a function of the distance $d(a_s, a)$. To see this, let's simplify the setting, considering a scenario where there are only two possible actions given a state s: the target action a_s and a single distractor a_d. Plugging in to Equation (5), we find that the probability of the target action a_s in state s is given by a logistic curve[2]:

$$q(a_s|s) = \frac{1}{1 + \frac{q(a_d)}{q(a_s)} \exp\{-\gamma d(a_s, a_d)\}}. \tag{6}$$

The curve is illustrated in **Figure 1**. The important part of Equation (6) is the second term in the denominator, which represents the effect of interference between the target action a_s and the distractor action a_d. As this interference term gets larger, the probability of the target action $q(a_s|s)$ gets smaller. This interference term is large when (1) the distractor action a_d is a priori likely, and (2) the distractor action a_d is close to the target action a_s.

An agent with a control objective as in Equation (4) will therefore show similarity-based interference in terms of errors in the action taken. This interference also manifests in decision cost for action a_s:

$$\begin{aligned} \text{Decision cost} &= -\log q(a_s|s) \\ &= \log\Big(1 + \frac{q(a_d)}{q(a_s)} \exp\{-\gamma d(a_s, a_d)\}\Big), \end{aligned}$$

visualized in **Figure 1**. This function decreases as $d(a_s, a_d)$ increases. The computation cost, on the other hand, decreases when $d(a_s, a_d)$ decreases, reflecting the main mechanism by which similarity-based interference arises in this model: at small distances $d(a_s, a_d)$, the policy achieves lower computation cost at the expense of decreased accuracy in the action selected.

Applying this logic to word production, we predict interference effects among semantically similar production targets when both are likely actions given the agent's state. Consider a state where a person sees a picture of an apple, and the words "apple" and "pear" are both a priori likely for some reason. This corresponds to target action $a_s = \texttt{say "apple"}$ and distractor action $a_d = \texttt{say "pear"}$, with $q(a_s)$ and $q(a_d)$ both high, and $d(a_s, a_d)$ low. A bounded-rational agent will erroneously say "pear" in this state more often than if the distractor were something less similar, such as $a'_d = \texttt{say "car"}$; furthermore, the action $a_s = \texttt{say "apple"}$ can only be produced at higher decision cost due to the presence of the distractor. The reason is that when the distractor is "car," the relevant distance is $d(a_s, a'_d) \gg d(a_s, a_d)$, leading to a lower probability of confusion in the action policy.

This example embodies the core logic of the RDC account of interference. Below, I will demonstrate this logic in a more thoroughly worked out model of the Stroop/Picture–Word Interference Task including fits to human behavioral data. That simulation will require a more involved control model, but the underlying cause of similarity-based interference remains the same as in this example.

4. MODEL OF PICTURE–WORD INTERFERENCE

Here, I show that RDC can capture some of the major characteristics of semantic interference in the Picture–Word Interference task.

4.1. Phenomena

Picture–Word Interference (PWI) is one of the most well-studied phenomena in language production and cognitive control (Schriefers et al., 1990; Damian and Martin, 1999; Bürki

[2] The probability of the target action $q(a_s|s)$ is calculated as follows:

$$\begin{aligned} q(a_s|s) &= \frac{q(a_s) \exp\{-\gamma d(a_s, a_s)\}}{q(a) \exp\{-\gamma d(a_s, a_d)\} + q(a_d) \exp\{-\gamma d(a_s, a_d)\}} \\ &= \frac{q(a_s) \exp\{0\}}{q(a_s) \exp\{0\} + q(a_d) \exp\{-\gamma d(a_s, a_d)\}} \\ &= \frac{q(a_s)}{q(a_s) + q(a_d) \exp\{-\gamma d(a_s, a_d)\}} \\ &= \frac{1}{1 + \frac{q(a_d)}{q(a_s)} \exp\{-\gamma d(a_s, a_d)\}}. \end{aligned}$$

This is an instance of the general logistic curve

$$f(x) = \frac{1}{1 + \exp\{-k(x - x_0)\}}$$

with slope parameter $k = \gamma$ and initial condition $x_0 = \frac{1}{\gamma} \log \frac{q(a_d)}{q(a_s)}$. More generally, given a set of distractors $a_d \neq a_s$, the probability of the correct action a_s is

$$q(a_s|s) = \frac{1}{1 + \sum_{a_d \neq a_s} \frac{q(a_d)}{q(a_s)} \exp\{-\gamma d(a_d, a_s)\}}.$$

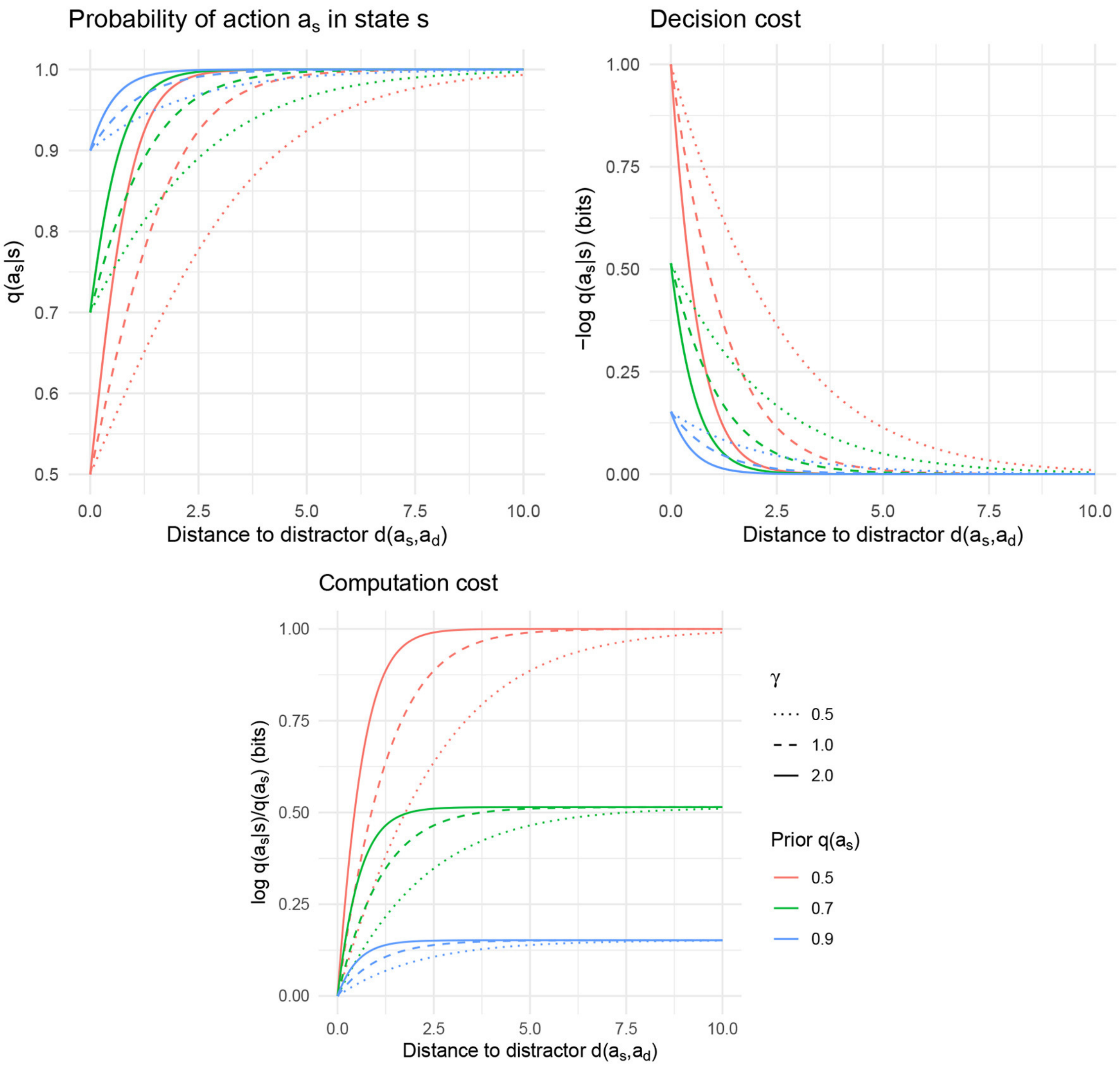


FIGURE 1 | Interference between a target action a_s and a distractor a_d as a function of the distance $d(a_s, a_d)$, for varying values of resource parameter γ and the a priori probability $q(a_s)$. **(Top left)** The probability $q(a_s|s)$ of taking the appropriate action a_s in state s. **(Top right)** The decision cost $-\log q(a_s|s)$, which is high when a_s and a_d have low semantic distance. **(Bottom)** The computation cost $\log \frac{q(a_s|s)}{q(a_s)}$.

et al., 2020). The task evokes similarity-based interference in picture naming by superimposing a text word over an image, and asking a subject to name the image (Lupker, 1979). Examples are shown in **Figure 2**. The **Stroop task** is closely related (Stroop, 1935; MacLeod, 1991; van Maanen et al., 2009; Starreveld and La Heij, 2017): in this task, a word, such as **BLUE** is presented in red ink, and subjects are asked to name the color of the ink.

The hallmark PWI effect is that subjects are slower to name the image in the presence of a superimposed word which is semantically categorically related to the image (the **semantic** condition in **Figure 2**), as compared to their reaction times when the superimposed text is a neutral string, such as XXXXX (the **neutral** condition in **Figure 2**). Furthermore, reaction times are fastest when the superimposed word is the same as the name of the image (the **congruent** condition), and if the superimposed text is a semantically unrelated word (the **unrelated** condition), reaction times are somewhere between the neutral and semantic conditions. "Semantic interference" in the PWI task refers to this additional slowdown and increased probability of error for the semantic condition relative to the unrelated condition.

Many PWI and Stroop experiments include only a neutral or an unrelated condition, rather than all four of these conditions, which has resulted in some variance in terms of the size of the reported interference effect (MacLeod, 1991). The neutral and unrelated conditions are referred to together as the **baseline** conditions, and the semantic and unrelated conditions are referred to together as the **incongruent** conditions.

4.2. Related Work

Because of its empirical robustness and (apparent) conceptual simplicity, PWI and Stroop tasks have been the target of many computational cognitive models throughout the past

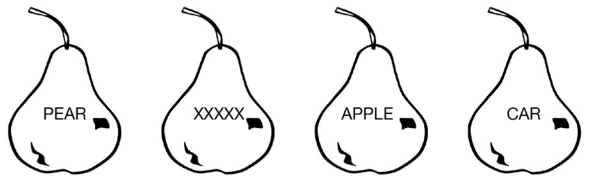


FIGURE 2 | Conditions of a Picture–Word Interference experiment. From left to right: the **congruent**, **neutral**, **semantic**, and **unrelated** conditions (see text).

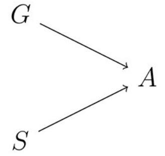


FIGURE 3 | Schematic of an action policy where the behavioral goal *G* and the perceptual state *S* jointly determine the output action *A*.

three decades, and subject to intense controversies about the mechanism that gives rise to the observed interference effect.

The main controversy in the literature is over whether PWI effects are driven by a competitive process during lexical selection, where multiple responses are competing for priority, resulting in slowdown (Roelofs, 1992; Levelt et al., 1999; Damian and Bowers, 2003; Belke et al., 2005; Abdel Rahman and Melinger, 2009) or by the need to exclude the distractor from an articulatory buffer (for example, Mahon et al., 2007). The most extensively documented and tested model of PWI is WEAVER++ (Roelofs, 1992, 2003; Levelt et al., 1999), a model of word production based on production rules and spreading activation where similarity-based interference emerges due to competition in lexical selection.

In contrast to existing computational models, the RDC account of interference in word production is a computational-level model which works by specifying only the problem that is being solved by the cognitive system, without making any commitments to algorithmic-level details (Marr, 1982). The theory and its assumptions are specified completely by (1) the control objective, which is the mathematical statement of the problem that the cognitive system is trying to solve, and (2) the linking function from cognitive costs to observables, such as RT.

As we will see, the control objective that reproduces PWI effects specifies only that there is some computational bottleneck involved in integrating information from bottom-up sensory input and top-down behavioral goals—whether this bottleneck happens in lexical selection, articulation, etc. is unspecified. The computational bottleneck might arise more mechanistically due to dynamics of spreading activation, competing production rules, etc. The question of whether the interference effect arises because of competition or response exclusion does not arise at this level of abstraction.

I am aware of two previous information-theoretic models of the Stroop task. Zénon et al. (2019) present a model of information-processing costs in the Stroop task which predicts that performing an unusual goal (i.e., naming a picture rather than reading a word) results in increased difficulty. Their model does not use bounded-optimal policies and does not account for semantic interference. Also, Christie (2019) models the RT response distribution for congruent, semantic, and neutral trials in a Stroop task using an information-theoretic model in which conflicting control signals are superposed and must be decoded at high cost. This model involves a policy which receives noisy bottom-up and top-down signals and must decide on an action. While this model is based on a noisy channel, rather than rate–distortion theory, it is fundamentally similar to the model presented here because it involves rational action under cognitive constraints modeled using information theory.

4.3. RDC Account

A full model of PWI requires a more complex setup than the simple interference example above. In particular, whereas the interference model given by Equation (4) involved a policy conditional only on an input state, a full model of PWI requires a policy conditional on *two* inputs: a perceptual state and a top-down behavioral goal.

To model PWI, let *G* be a random variable representing a speaker's top-down goals, i.e., whether the goal is to name a picture/color or to read a word. That is, *G* is a random variable taking values in the set $\{\texttt{name}, \texttt{read}\}$. Let *S* be a random variable representing a speaker's perceptual state—that is, the particular word and picture that the speaker is looking at. A speaker then implements a bounded-rational production policy on actions given goals and perceptual states $q(a|g,s)$, subject to information-processing costs. The structure of the model is shown in **Figure 3**.

As the output action is jointly determined by the behavioral goal *G* and the perceptual state *S*, the total mutual information between the inputs to the policy and the output action is given by the formula

$$I[G,S:A] = \left\langle \log \frac{q(a|g,s)}{q(a)} \right\rangle. \tag{7}$$

This quantity gives the total amount of information in the behavioral goal *G* and perceptual state *S* that the policy uses in order to specify the action *A*. The simplest RDC policy would simply take Equation (7) as the computation cost. However, it turns out that in order to model the PWI task, we need to assign different levels of cost to information coming from the two sources, *G* and *S*.

In order to do so, we must first break the quantity in Equation (7) down into two parts, reflecting the contributions of *S* and *G*. Using the chain rule for mutual information (Cover and Thomas, 2006, p. 24, Theorem 2.5.2), we can write:

$$\begin{aligned} \underbrace{I[G,S:A]}_{\text{information transmitted from } G \text{ and } S \text{ to specify } A} &= \underbrace{I[S:A]}_{\text{information from } S} \\ &+ \underbrace{I[G:A|S]}_{\text{information from } G \text{ conditional on } S}, \end{aligned}$$

with the **conditional mutual information** $I[G:A|S]$ defined as

$$I[G:A|S] = \left\langle \log \frac{q(a|g,s)}{q(a|s)} \right\rangle.$$

The conditional mutual information gives the amount of information contributed by G about A in the presence of S, and beyond what is contributed by S alone. Now, following previous work (van Dijk and Polani, 2013; Genewein et al., 2015), we can define a family of computation costs by taking a weighted sum of the information from the two sources:

$$\text{Computation cost} = \alpha I[S:A] + (1-\alpha)\, I[G:A|S], \qquad (8)$$

where $\alpha \in [0,1]$ represents the relative cost of using information from S as opposed to information from G conditional on S. In order to model PWI, it turns out that the minimal information penalty required in the control objective is on the mutual information $I[G:A|S]$—the amount of information that must be "transmitted" from the behavioral goal G to specify the action A in the context of the perceptual state S. So in the computation cost for the PWI simulations, I set $\alpha = 0$ in Equation (8). The substantive hypothesis here is that there is negligible cost for using information from the perceptual state S alone, but high cost for using information from the behavioral goal G in the context of the perceptual state S.

Defining computation cost in this way, the speaker's production policy is a minimum of the control objective:

$$\mathcal{L}(q) = \left\langle d(a_s^g, a) \right\rangle + \frac{1}{\gamma} I[G:A|S], \qquad (9)$$

where a_s^g indicates the correct action to be taken in state s with goal g, and $d: A \times A \to \mathbb{R}^{(+)}$ is a semantic distance measure on production actions A, as defined in section 3.2. The minima of the control objective in Equation (9) have the form:

$$\begin{aligned} q(a|g,s) &= \frac{1}{Z(g,s)} q(a|s) \exp\{-\gamma\, d(a_s^g, a)\} \qquad (10) \\ q(a|s) &= \sum_g p(g|s) q(a|g,s) \\ Z(g,s) &= \sum_a q(a|s) \exp\{-\gamma\, d(a_s^g, a)\}. \end{aligned}$$

Below, I will first analyze the policy in Equation (10) and show that it demonstrates semantic interference under reasonable default parameter settings in a simulation of the PWI task, and then that it can capture some of the major qualitative empirical patterns observed in PWI studies when we vary the parameters of the simulation.

4.4. Simulation Setup

I model the basic PWI task with the following setup. An agent has access to a behavioral goal and a perceptual state, and produces an output action in response to these. The perceptual state consists of a picture and a written word. The behavioral goal specifies whether the agent should read the word or name the picture. Each word and each picture is associated with a single appropriate target action.

TABLE 1 | Default parameters of the simulation of the Stroop task.

Parameter	Value	Meaning
$p_{\texttt{name}}$	0.1	A priori probability of the behavioral goal being to name, rather than read.
N_w	32	Number of different words in possible perceptual states.
N_p	32	Number of different pictures in possible perceptual states.
γ	4	Information processing resources (see Equation 9).

See text for discussion.

More formally, the behavioral goal is a random variable G that can take one of two values, $g \in \{\texttt{name}, \texttt{read}\}$, with the probability of the goal being $\texttt{name}$ equal to a parameter $p_{\texttt{name}} = \frac{1}{10}$, the same value used in Zénon et al. (2019). This low probability is meant to reflect the fact that when one sees some text, the relevant behavioral goal is usually to read the text, not name the object it is displayed or written on, especially when reading a card or a computer screen in a lab environment. As we will see, this low probability will end up driving the asymmetry between reading and naming in the model.

The perceptual state is represented by the random variable S and takes values in *pairs* of discrete objects $\langle w, p \rangle$, representing a state where an agent is seeing word w superimposed on picture p. The number of possible words is N_w and the number of possible pictures is N_p; in all the simulations below, I fix $N_w = N_p = 32$ and assume a uniform distribution on the possible states. The output actions are represented by a random variable A taking one of $N_a = 32$ different values. Each goal g and state s is associated with a target action a_s^g defined as follows: given the goal $g = \texttt{read}$ and the state $s = \langle w, p \rangle$, the target action is w; given the goal $g = \texttt{name}$, the target action is p. The distance metric among output actions $d: A \times A \to \mathbb{R}^{(+)}$ will be defined below, either as an idealized metric or as a metric derived from word embeddings (Mikolov et al., 2013), when we move to modeling experimental data.

The last parameter we need to specify an RDC policy is the scalar γ, which gives the computational resources (inverse cost) available for information processing in the model. With all these parameters in hand, we can compute the RDC policy from the control objective in Equation (9). Simulation parameters are summarized in **Table 1**.

As a more concrete example, suppose the goal $g = \texttt{name}$, and the perceptual state is the pair $\langle \texttt{apple}, \texttt{pear} \rangle$, representing the word "apple" superimposed on a picture of a pear. Because the goal is $g = \texttt{name}$, the target action a_s^g is to say "pear." If the agent takes this action, then the distortion is zero, because $d(\texttt{pear}, \texttt{pear}) = 0$. On the other hand, if the agent takes the action of saying "apple," then the distortion is $d(\texttt{pear}, \texttt{apple})$, which may be small, since these are semantically related words that share many features. Because this distortion is low, an agent may be attracted toward saying "apple," which has higher distortion than "pear," but has lower computation cost because it does not require attending to the costly behavioral goal. Then the probability of producing the correct word "pear" will be low and the decision cost for the correct word "pear" will be high.

Given a state $\langle w, p \rangle$ and a goal g, we can define one part of the state as the "target" and another as the "distractor." When $g = \texttt{name}$, the target is p and the distractor is w. When $g = \texttt{read}$, the target is w and the distractor is p. In each state, there will be a certain semantic distance between the target and distractor, called the **distractor distance**. If a_w represents the action associated with w and a_p is the action associated with p, then when $g = \texttt{name}$, the distractor distance is $d(a_p, a_w)$; when $g = \texttt{read}$, the distractor distance is $d(a_w, a_p)$.

The major conditions of PWI experiments are the congruent, semantic, neutral, and unrelated conditions (defined in **Figure 2**). So far, we have the ability to model three of these: the congruent condition corresponds to the case where the distractor distance is 0 (i.e., the target actions are identical across goals: $a_w = a_p$); the semantic condition corresponds to the case where distractor distance is low; and the unrelated condition means the distractor distance is high. I will return to the neutral condition below.

4.5. Results

4.5.1. Basic Results: Idealized Semantic Distance Metric

First I present simulation results showing the existence of semantic interference effects given an idealized semantic metric among words. This metric is generated randomly by placing $N_w = 32$ words uniformly at random in bounded 2-dimensional space of size 7 × 7. An example such space is shown in **Figure 4**. An RDC policy was computed for picture naming and word reading given this space, considering all possible pairings of words as pictures and as names.

In **Figure 5**, I show the decision cost and the computation cost based on the simulation in this space, as a function of distractor distance. We see a few basic patterns:

- There is no decision cost and low computation cost when the distractor distance $d = 0$, corresponding to the congruent condition in experiments.
- Semantic interference exists in the decision cost. The interference is high for close words (corresponding to the semantic condition), and falls off rapidly at distant words (corresponding to the unrelated condition).
- When the goal is $g = \texttt{read}$, interference of any kind is negligible.

In the simulation, computation cost comes out to be essentially a constant function of the goal, except when the appropriate actions given the two goals coincide (distractor distance 0). In fact, as the distractor distance gets large, the computation cost turns out to approximate the surprisal of the goal given the state $-\log p(g|s)$. In doing so, the computation cost recovers the model of Stroop interference from Zénon et al. (2019)[3].

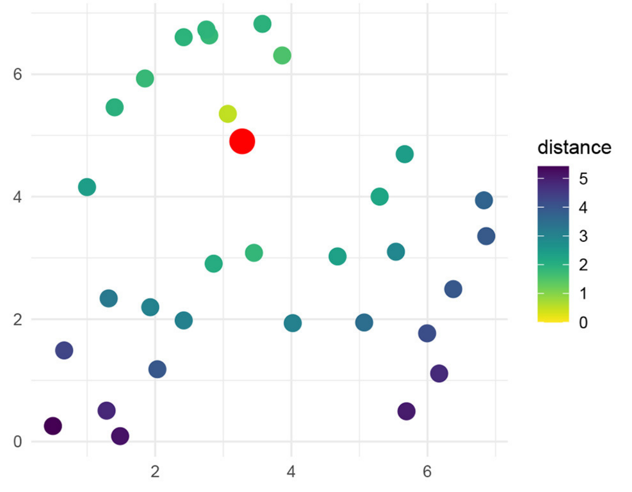


FIGURE 4 | Example of an idealized semantic metric of words as used for basic simulations. Thirty-two words are placed randomly in a two-dimensional bounded Euclidean space of size 7 × 7. A target word is indicated in red. The remaining points are colored according to their distance from the target word.

This most basic simulation already captures several qualitative patterns from the empirical literature (as listed by MacLeod, 1991). First, we recover the fact that naming is generally slower than reading (Cattell, 1886), as indicated by the uniformly higher computation cost for naming. Second, we recover the existence of facilitation in the congruent condition, reflected in lower decision cost and lower computation cost when distractor distance is zero. Third, we recover the existence of interference in the semantic condition relative to the congruent condition and the unrelated condition, as reflected in the decision cost. Fourth, interference exists for the naming task but is negligible in the reading task. Fifth, the interference effect is gradient (Klein, 1964): when the distractor is *more* semantically similar to the target, there is more interference; this is reflected in the decision cost for the naming condition.

The semantic gradient deserves a bit more discussion. There has been controversy in the literature on Picture–Word Interference about whether a semantic gradient really exists, as opposed to a categorical effect for distractors that are in the same category as the target (Hutson and Damian, 2014; Bürki et al., 2020). In the RDC model, there is a semantic gradient observable in the decision cost, but it falls off very rapidly from distance 1 to distance 2, and distance 2 shows only barely more interference than distance 3. Therefore the theory predicts that a semantic

[3] When distractor distance is 0, computation cost comes out to nearly zero. This may seems surprising, but follows from the fact that computation cost here is the pointwise conditional pointwise mutual information $\log \frac{q(a|g,s)}{q(a|s)}$, which is zero when the action a is already fully specified by the perceptual state s, such that the behavioral goal g adds no new information. It should be noted that computation cost zero does not imply a prediction of RT zero—see section 4.5.5.

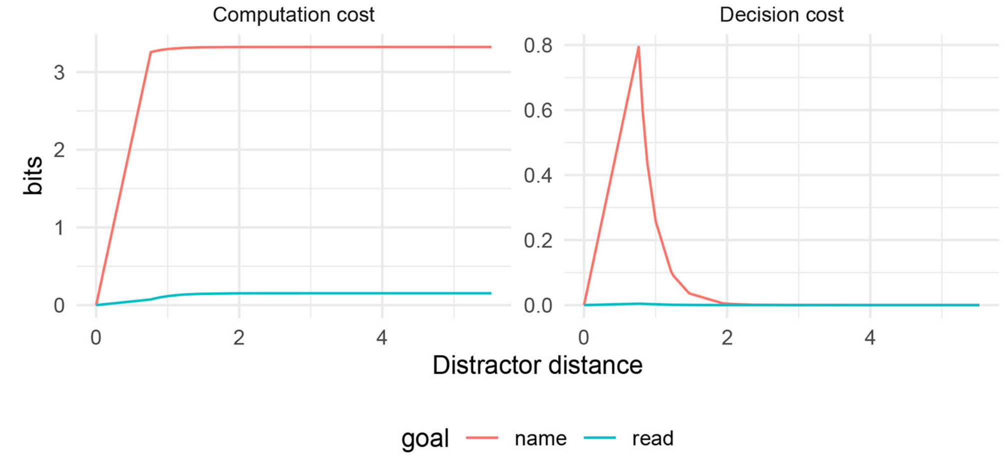


FIGURE 5 | Simulated costs in Picture–Word Interference task, as a function of semantic distance between target and distractor.

gradient does exist, but it is highly concentrated, and might be hard to detect in experiments.

Above, I have shown that RDC can capture the basics of semantic interference in PWI tasks in a simulation with simple and reasonable default parameter settings. Next, I will show how we can recover more of the empirical patterns by varying the parameters of the simulation and the model.

4.5.2. Reverse Stroop

The **reverse Stroop effect** refers to a reversal in the difference between naming and reading in a PWI/Stroop task. Usually, interference happens in the naming task and not in the reading task. However, after a great deal of experience with naming in incongruent trials, two things happen: the interference effect in naming shrinks, and subjects begin to show an interference effect in reading as well as naming (Stroop, 1935; MacLeod, 1991).

While early work hypothesized that the reverse Stroop effect is caused by practice and task familiarity (Stroop, 1935), later work has shown that reverse Stroop effects are more likely related to the difficulty of task switching between naming and reading (Allport and Wylie, 2000; Roelofs, 2021). In terms of simulation parameters, it seems sensible to identify reverse Stroop manipulations with an increase in the parameter $p_{\texttt{name}}$, reflecting increased relevance of the naming goal, perhaps due to recency.

Figure 6 shows computation and decision costs under varying $p_{\texttt{name}}$ in the idealized semantic distance metric. As this value increases, a reverse Stroop effect emerges: the reading task begins to show interference in both costs. Meanwhile, the interference associated with naming is predicted to decrease.

Beyond the Reverse Stroop effect, the simulations here demonstrate the general effects of varying the simulation parameter $p_{\texttt{name}}$. Such results could be used, for example, when modeling picture–picture interference effects, where participants are confronted with two pictures and must name only a certain one (for example, Glaser and Glaser, 1989). In that case, the behavioral goals associated with each of the two pictures would have more similar prior probabilities, and the resulting RDC predictions would look more like the dotted lines in **Figure 6**.

4.5.3. Empirically-Derived Semantic Distance Metric

The results above showed basic qualitative effects in an idealized semantic space. Now I turn to results based on an empirically-derived semantic space, leading to a quantitative comparison to human reaction times. The use of an empirically-derived semantic space brings two advantages over the idealized space above: (1) it allows for a comparison with experimental data on real words, and (2) it shows that the predicted interference effects arise given a realistic geometry for the semantic space and a realistic distribution of words in it.

In the last decade, the field of natural language processing has devoted a great deal of attention to deriving representations of words as points (called **embeddings**) in high-dimensional space, such that the distances among embeddings reflect semantic relationships among words (Mikolov et al., 2013; Pennington et al., 2014). These representations differ in their details, but they are all derived by an optimization process whose goal is to create embeddings such that the *context* of a word can be predicted accurately from its embedding (Goldberg and Levy, 2014), in keeping with the old linguistic intuition that the meaning of a word is related to its distribution with respect to other words (Harris, 1954; Firth, 1957). The result is that the "distance" between two words *A* and *B* reflects the difference between the typical contexts for *A* and *B*. As such, these distributional embeddings provide a distance metric which fits with the RDC framework, which holds that two actions are similar if there is low cost for failing to distinguish them. In particular, the embedding distance between words reflects how badly one would mis-predict the context of one word when it is mistaken for another.

There have been previous attempts to model semantic interference effects in Stroop and PWI using embedding spaces, such as these (de Marchis et al., 2013; Hutson and Damian, 2014). The embedding spaces can broadly distinguish between semantically close words compared against unrelated words,

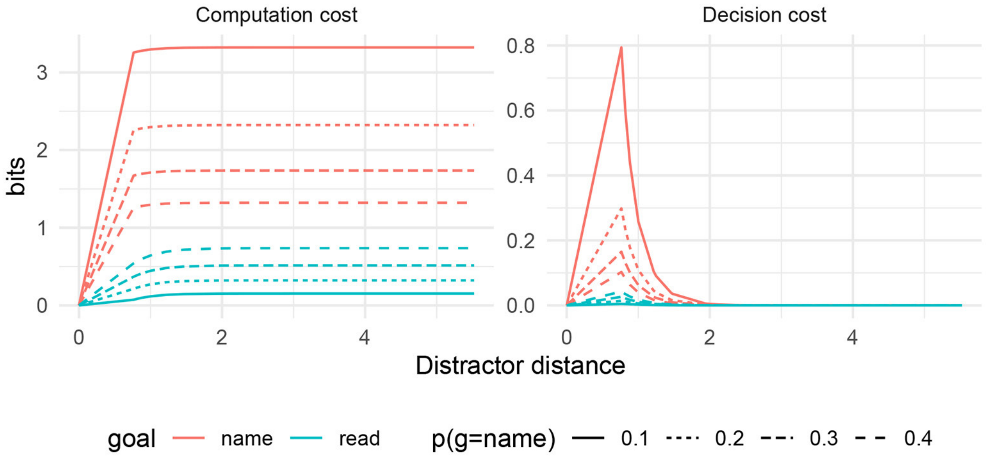


FIGURE 6 | Computation and decision cost for PWI under varying values of $p_{\texttt{name}}$. A reverse Stroop effect emerges in the decision cost under the reading goal.

although they do not seem to be able to make reliable item-level predictions within semantically close words (Hutson and Damian, 2014).

Here, I adopt the English fastText embedding space derived by Facebook[4] as a semantic distance metric among words. In work using these embeddings, the distance between embeddings u and v is usually quantified as **cosine distance**:

$$d_{\cos}(u, v) = 1 - \frac{u \cdot v}{||u||_2 ||v||_2},$$

where $\cdot$ indicates a dot product and $||u||_2$ indicates an L_2 norm. In order to produce distances in the interval $[0, \infty)$, I apply a logit transform to the cosine distance[5].

I use the set of 32 words from the Picture–Word Interference experiment presented in Roelofs and Piai (2017). The items from this experiment consist of picture–word pairings which are either semantically close ("semantic") or semantically unrelated ("unrelated"). Here, I show that RDC with the fastText embedding space predicts higher cognitive cost for the semantic pairings as opposed to the unrelated word pairings, and also lower cost when the word and the picture to be named are identical[6]. Except for the semantic distance metric, all other parameters of the simulation are the same as above.

In **Figure 7**, I show theoretical computation cost and decision cost by distractor distance for the word pairs listed in Roelofs and Piai (2017). Red dots indicate word pairs in the "semantic" condition; green dots indicate word pairs in the "unrelated" condition; and blue dots indicate identical words. Predicted cognitive cost is lowest for identical words. For "unrelated" and "semantic" words, there is high computation cost. For "semantic" words, there is also high decision cost.

The simulation using an empirically-derived semantic distance metric shows the same qualitative patterns as the simulation using an idealized metric. Furthermore, we see that the semantic distances largely correspond (although imperfectly) with the designation of items as "semantic" vs. "unrelated."

4.5.4. Neutral vs. Unrelated Trials

The PWI task has a fourth major condition: the *neutral* condition, where a picture is presented along some kind of neutral orthographic stimulus that would not reasonably be read out loud, such as XXXXX. Here, I will incorporate this condition into the simulation and show that we immediately recover three empirically-attested patterns: (1) there is facilitation in the congruent condition relative to the neutral condition, (2) there is interference in the unrelated condition relative to the neutral condition, and (3) the size of facilitation is small relative to the size of interference (MacLeod, 1991).

Recall that in the basic simulation, the a priori probability that the behavioral goal is $g = \texttt{name}$ rather than $g = \texttt{read}$ is $\frac{1}{10}$. I model the neutral condition by adding into the simulation a set of states s_{neutral} with neutral text distractors, such that $p(g = \texttt{name}|s_{\text{neutral}}) = \frac{9}{10}$ for all neutral states. This models the scenario where a subject sees XXXXX superimposed on an image. The idea is that given such a state, a subject would only expect to actually read the stimulus (saying "eks eks eks eks eks") $\frac{1}{10}$ of

[4] Available for download at https://fasttext.cc/docs/en/pretrained-vectors.html

[5] The logit-transformed distance metric between two word embeddings u and v is

$$d(u, v) = \text{logit}\left(\frac{1}{2} + \frac{1}{2} d_{\cos}(u, v)\right),$$

with the logit function defined as

$$\text{logit}(x) = \log\left(\frac{x}{1 - x}\right).$$

[6] These words were originally in Dutch; I translate them into English in order to get their distances. In preliminary experiments, I also tried using the Dutch fastText vectors, and using the English GloVe vectors (Pennington et al., 2014). I use the English fastText vectors because I found that they most reliably assign lower distances to the "semantic" word pairings compared to the "unrelated" word pairings in the experimental items. Rank-order correlations of semantic distances among the embedding spaces are: English fastText vs. English GloVE $\rho = 0.77$; English fastText vs. Dutch fastText $\rho = 0.59$; English GloVE vs. Dutch fastText $\rho = 0.54$.

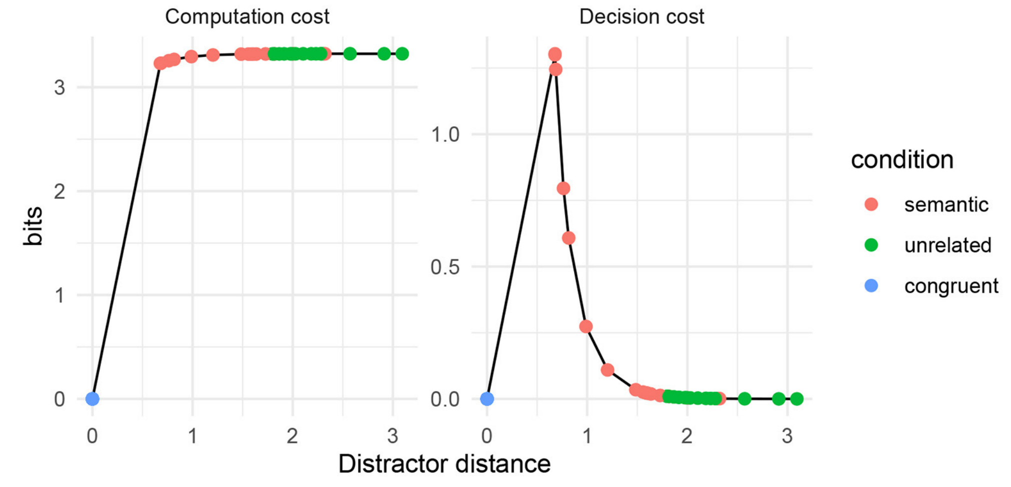


FIGURE 7 | Computation and decision costs for word pairs from the items of Roelofs and Piai (2017), using fastText as the semantic distance metric.

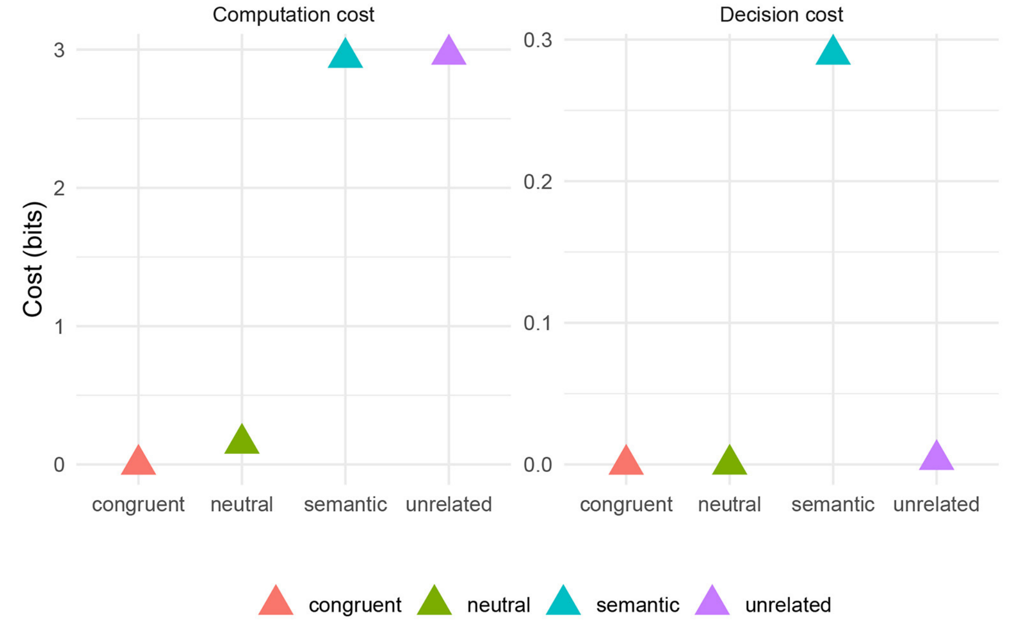


FIGURE 8 | Simulated costs by PWI task condition based on materials from Roelofs and Piai (2017) and fastText word embeddings.

the time. Outside of a state with a neutral distractor s_{neutral}, the probability of naming is still $\frac{1}{10}$.

Figure 8 shows the simulated decision and computation costs for four experimental conditions based on the items from Roelofs and Piai (2017): congruent (the case where the distance $d = 0$), semantic, unrelated, and neutral (simulated as the case where $s = s_{\text{neutral}}$). The three empirical patterns are captured here by the computation cost. The neutral condition has drastically reduced computation cost relative to the semantic and unrelated conditions, indicating facilitation. Also, the computation cost is slightly less in the congruent case relative to the neutral case, indicating facilitation. Also, the size of the facilitation effect (the difference between neutral and congruent conditions) is small relative to the interference effect (the difference between neutral and semantic/unrelated conditions).

The model robustly recovers the existence of facilitation and interference. The relative magnitude of facilitation and interference depends on a model parameter: the probability $p(g = \texttt{name}|s = s_{\text{neutral}})$[7]. Therefore, it is therefore possible to make a prediction: the facilitation effect should get larger under any manipulation that makes the orthographic string in the

[7] The default values for $p(g|s)$ have not been tuned to fit the human data, but were selected a priori and kept constant throughout all simulations.

neutral condition more and more like something that someone would reasonably read. In fact, there is already some evidence in this direction in the literature: pseudowords, which presumably fall somewhere between XXXXX and a real word in terms of $p(g = \texttt{name}|s)$, cause less interference than real words in the Stroop task (Klein, 1964).

4.5.5. Fit to Human RT Data

Here I relate the simulated computation and decision costs to empirical human RT data. To do so, we need a more specific linking function from computation and decision cost to RT.

I propose that RT can be predicted from a linear combination of computation and decision cost. That is, the predicted RT in a condition is given from cognitive costs by a transformation:

$$\mathrm{RT} = a + bX + cY,$$

where X is computation cost, Y is decision cost, and a, b, and c are non-negative scalars. This linking function supposes that computation cost and decision cost are each associated with some fixed rate of information processing, given by b and c, respectively, in terms of milliseconds per bit. The scalar a represents a constant RT delay across conditions (in the model of Zénon et al., 2019, this constant cost corresponds to perceptual information processing).

Figure 9 shows a comparison of empirical mean RTs in a PWI task, drawn from Roelofs and Piai (2017), compared against simulated RTs, with $a = 730$ ms, $b = 30$ ms/bit, and $c = 140$ ms/bit[8]. This mixture gives a good qualitative fit to the human data.

The relationship of information-processing costs to RT may not be so simple, however. In particular, RT distributions appear to follow what is called an Ex-Gaussian distribution (Ratcliff, 1979; Luce, 1986; Balota et al., 2008). An Ex-Gaussian random variable is the sum of a Gaussian random variable with mean μ and an Exponential random variable with rate τ. The resulting distribution is skewed positive when compared with a Gaussian distribution. Interestingly, it has been suggested that the μ and τ parameters of the Ex-Gaussian distribution reflect different aspects of cognitive processing in the PWI task (Heathcote et al., 1991; Mewhort et al., 1992; Spieler et al., 2000; Piai et al., 2011, 2012; Roelofs, 2012; Scaltritti et al., 2015; San José et al., 2021).

Here I present an analysis comparing computation and decision costs to the full Ex-Gaussian analysis of experimental PWI data, including congruent, semantic, neutral, and unrelated conditions, performed by Roelofs and Piai (2017). In **Figure 10**, I show their estimates of the μ parameter compared with a combination of computation cost and decision cost ($a = 615$ ms, $b = 25$ ms/bit, $c = 65$ ms/bit). In **Figure 11**, I compare their τ estimates to decision cost alone ($a = 120$ ms, $b = 0$, $c = 85$ ms/bit)[9]. The reasonable qualitative match suggests that both computation and decision cost are reflected in the μ component of the RT distribution, while only decision cost is reflected in the τ component. It is striking that the τ component seems to reflect only decision cost, suggesting that decision cost is indeed an index of a distinct kind of cognitive cost. This result is in line with the pattern reported by Roelofs and Piai (2017): μ shows a contrast among neutral, unrelated, and semantic conditions, while τ shows a contrast only between the semantic condition and the others (see also Scaltritti et al., 2015; San José et al., 2021).

Summing up, the overall empirical pattern is that computation cost captures basic interference effects in RT, while decision cost captures the additional RT slowdown associated with semantically close distractors. The RT component μ reflects both computation and decision cost, while the additional RT component τ reflects only decision cost.

4.6. Discussion

It is striking that the framework laid out here can successfully model many aspects of PWI, despite being developed nearly entirely for purposes other than cognitive modeling. Rate–distortion theory was developed purely as an abstract theory of lossy communication, and its application to control problems has primarily been confined to the computer science and robotics literature.

Furthermore, RDC captures the major empirical patterns of the Picture–Word Interference task with few free parameters. The degrees of freedom in the specification of the model are (1) the distribution over goals and states, (2) the information-processing resource parameters used to define the control objective (the scalar γ, which was set to a constant value in all simulations reported above), and (3) the similarity metric among actions. All of these degrees of freedom correspond to quantities that can be independently estimated, at least in principle. The distribution over goals and states is set by the frequency of goals and states in a person's everyday experience; the information-processing cost parameters are set by studies of cognitive difficulty; and the similarity metric among actions is determined by the relative cost of the consequences of confusing one action for another. The result is a parsimonious model that captures several patterns naturally.

5. GENERAL DISCUSSION

I have shown that the rate–distortion theory of control can naturally account for similarity-based interference in general, and that it offers a strong model of Picture–Word/Stroop interference effects. Now I turn to the interpretation of the model and how it relates to word production more generally.

[8] All of the scaling factors presented in this section were derived by linear regression on the empirical means, followed by rounding. From the linear regressions, the optimal models before rounding are

$$\text{mean RT} \approx 737 + 28 \times \text{Computation cost} + 139 \times \text{Decision cost}$$
$$\mu \approx 615 + 25 \times \text{Computation cost} + 65 \times \text{Decision cost}$$
$$\tau \approx 123 + 2 \times \text{Computation cost} + 87 \times \text{Decision cost.}$$

[9] The decision to map computation cost to μ and τ, and decision cost to τ alone, was taken *post-hoc* based on regressions on the empirical RTs.

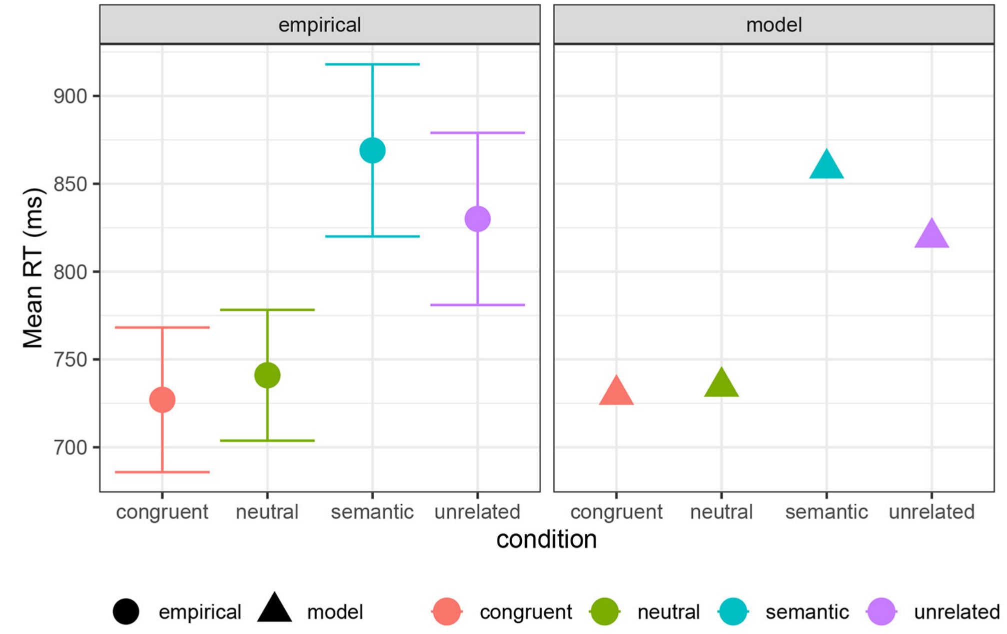


FIGURE 9 | Empirical mean RTs for PWI conditions from Roelofs and Piai (2017), compared with model predictions (see text). Error bars show 95% confidence intervals of the mean in the empirical data.

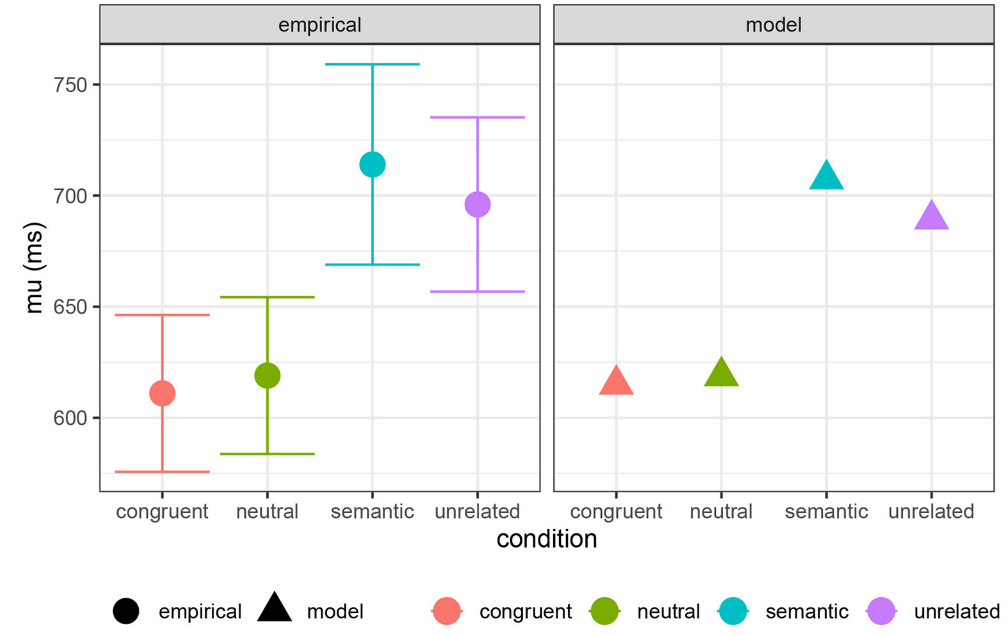


FIGURE 10 | Empirically estimated μ parameter of Ex-Gaussian RT distribution for PWI conditions from Roelofs and Piai (2017), compared with model predictions (see text).

5.1. Interpretation of Computation and Decision Cost

I used two notions of cost: computation cost and decision cost, where computation cost is the cost term that is contained in the control objective, and decision cost is the surprisal of selecting a single action given a probabilistic policy. As a summary, semantic similarity-based interference emerged in the decision cost, while computation cost predicted general interference and difficulty for the less-probable goal in context (naming as opposed to reading).

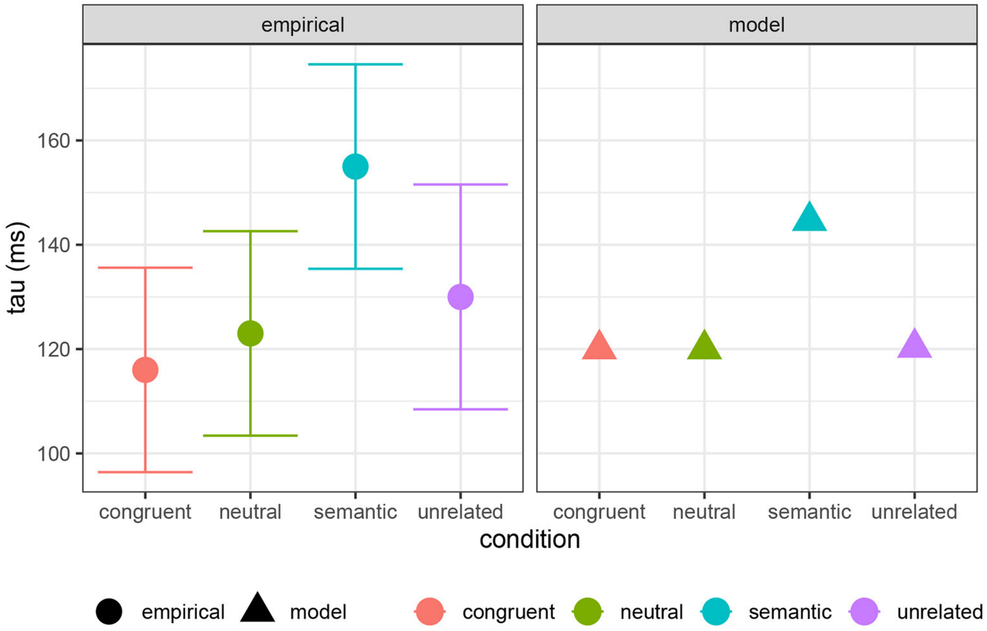


FIGURE 11 | Empirically estimated τ parameter of Ex-Gaussian RT distribution for PWI conditions from Roelofs and Piai (2017), compared with model predictions ($\tau = 120 + 85 \times$ Decision cost).

I proposed that computation cost and decision cost map linearly to RT. The reason for this proposal was simplicity. However, it may be that other linking functions provide a better connection between $q(a|g, s)$ and empirically observable response times, for example by linking RDC components to components of drift–diffusion models (Bogacz et al., 2006; Ortega and Braun, 2013). I leave the exploration of this possibility to future work.

5.2. Relation to Algorithmic-Level Models

As a computational-level theory, RDC specifies only the problem being solved by our cognitive system, and does not make claims about algorithmic or implementational details. It should be hoped, then, that existing successful algorithmic models of PWI can be seen as implementing the core parts of the RDC account.

In this connection, the recent extension of WEAVER++ by San José et al. (2021) is especially interesting, as it adds an element of periodically lapsing attention to the behavioral goal in order to explain the Ex-Gaussian distribution of RTs in PWI experiments. Similarly, the RDC model of picture–word interference crucially works by positing a cost associated with extracting information from the behavioral goal in the presence of the perceptual state. Essentially, the RDC agent can only access the behavioral goal through a channel with limited bandwidth. This limited bandwidth equates to a kind of inattention: because the agent has limited resources with which to attend to the channel, it will often not attend. Indeed, RDC was initially introduced as a model of "rational inattention" in economics with this reasoning (Sims, 2003, 2005, 2010).

Similarly, the production rules and spreading activation dynamics of WEAVER++ can be seen as implementing RDC-like behavior. For example, one production rule used in the WEAVER++ simulation of PWI in San José et al. (2021) states that if the behavioral goal is to name a picture, and a written word is present, then activation relating to the written word is blocked off. Similar logic is instantiated by the RDC policy. Consider the equilibrium probability (following Equation 10) to produce the written word a_w when the behavioral goal is $g = \texttt{name}$:

$$q(a_w|g = \texttt{name}, s) \propto q(a_w|s) \exp\{-\gamma d(a_p, a_w)\},$$

where a_p is the action corresponding to naming the picture. The first factor $q(a_w|s)$ will be relatively large, because the prior is that the behavioral goal is usually to read, not to name. This large value corresponds to activation for the written word. However, this large value will be squashed by the exponentially small value of the second factor $\exp\{-\gamma d(a_p, a_w)\}$ (unless a_p and a_w are close), resulting in an ultimately low probability to name the written word. This corresponds to blocking of activation.

The RDC model presented here shows how similarity-based interference can arise from a very generically-defined computational bottleneck. It achieves this generality without sacrificing quantitative precision. Nevertheless, it is likely that many aspects of PWI and similarity-based interference more generally might only be explainable within more algorithmic and mechanistic frameworks. For example, a great deal of work on PWI has dealt with stimulus-onset asynchrony (SOA) effects, where the distractor word or the picture do not appear at the same time. These effects are naturally captured in spreading-activation models that describe the evolution of activation with time. It is less clear how such time-based effects would be captured within a purely computational-level account, which

$$\begin{tikzcd} G \arrow[r] \arrow[dr] & A_2 \\ S \arrow[r] \arrow[ur] & A_1 \end{tikzcd}$$

FIGURE 12 | Schematic of a policy where the behavioral *G* and the perceptual state *S* determine two actions A_1 and A_2 to be performed by different actuators.

simply models the *function* that is computed by cognitive systems, and not *how* it is computed.

5.3. Further Word Production Phenomena: Facilitation

I intend to advance RDC, or an extension of it, as a model of word production in general. I have presented its application to interference in PWI and Stroop paradigms because these are well-known and challenging phenomena to model. However, there are many other language production phenomena on which an RDC model has yet to be tested, including several that arise within the PWI paradigm. One such set of phenomena is facilitation, both phonological and semantic.

The PWI task exhibits phonological facilitation, meaning that naming time is sped up when the distractor word is *phonologically* similar to the target word (Meyer and Schriefers, 1991). In the simple simulations presented here, the RDC does not predict this kind of facilitation. However, it can when the control objective is specified in more detail, as I sketch below.

Imagine that the goal of the policy is not to output a single atomic output action, but rather to output a large number of actions. For example, one can imagine that the policy must output instructions to a large number of actuators. This kind of policy is illustrated in **Figure 12**. Equivalently, the output of the policy is a vector $\mathbf{a} = [a_1, a_2, \ldots, a_n]$ of actions to be performed by *n* different actuators.

Given this kind of policy, we can define a "phonological" similarity metric among actions $\mathbf{a}_1$ and $\mathbf{a}_2$ in terms of how many elements overlap between $\mathbf{a}_1$ and $\mathbf{a}_2$. For each overlapping element, we will have a facilitation effect, and for each non-overlapping element, we will have an interference effect. The result is overall facilitation when the target action and the distractor have more overlapping elements.

There are other extensions of RDC and other mechanisms that could give rise to facilitation effects, for example multi-stage hierarchical policies where the output of one policy becomes the input to another. Such families of more elaborate RDC policies have been explored in simulations by Genewein et al. (2015).

Facilitation has also been reported in PWI settings for certain semantically similar words, and a great deal of effort has gone into experimentally characterizing when semantically similar words will cause facilitation or interference, often dealing with whether a given target word is in the "response set" for the experiment (e.g., Roelofs, 1992, 2003; Caramazza and Costa, 2000, 2001; Mahon et al., 2007; Piai et al., 2012). While empirical picture remains complex (Bürki et al., 2020), these results have often been taken to reflect dynamics during different stages of word production. While the simple RDC model presented here does not predict these facilitation effects, a more articulated model might: for example, a model with a non-zero penalty on perceptual state information, or a hierarchical policy (Genewein et al., 2015; Zénon et al., 2019). The answer may also lie in the linking function from the RDC policy to observables, such as RT: if computation cost is sometimes the dominant determinant of reaction times, rather than decision cost, then **Figure 5** suggests that we would expect semantic facilitation rather than interference. I leave the investigation of these possibilities to future work.

5.4. Conclusion

This work has extended the reach of information-theoretic models of language processing. Although information-theoretic models have seen broad success in the study of language comprehension (Hale, 2001; Moscoso del Prado Martín et al., 2004; Levy, 2008; Hale et al., 2018; Futrell et al., 2020) and the emergence of linguistic structure (Zaslavsky et al., 2018; Hahn et al., 2020), they have not yet seen much application to language production. This work has taken the first steps toward remedying this gap using the rate–distortion theory of control.

Furthermore, the apparent inability to capture similarity relations among stimuli has been a major barrier for the adoption of information-theoretic models in cognitive science (Luce, 2003, p. 185). This work shows that rate–distortion theory allows us to overcome this difficulty and model some of the most salient similarity-based effects in psychology.

OPEN PRACTICES STATEMENT

All data and code for reproducing the results in this paper can be found online at http://github.com/langprocgroup/wordprodmodel.

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.

AUTHOR CONTRIBUTIONS

RF conceived the research, conducted the research, and wrote the paper.

ACKNOWLEDGMENTS

I would like to thank Ardi Roelofs and Vitória Piai for generously sharing data, the three reviewers for helpful and constructive feedback, and Michael Hahn, Noga Zaslavsky, Seth Frey, Greg Scontras, and the MIT Language Lab and Computational Psycholinguistics Lab for discussion.

REFERENCES

Abdel Rahman, R., and Melinger, A. (2009). Semantic context effects in language production: a swinging lexical network proposal and a review. *Lang. Cogn. Process*. 24, 713–734. doi: 10.1080/01690960802597250

Allport, A., and Wylie, G. (2000). "Task switching, stimulus-response bindings, and negative priming," in *Control of Cognitive Processes: Attention and Performance XVIII*, eds S. Monsell and J. Driver (Cambridge, MA: MIT Press), 35–70.

Anderson, J. R., and Lebiere, C. (1998). *Atomic Components of Thought*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Balota, D. A., Yap, M. J., Cortese, M. I., and Watson, J. M. (2008). Beyond mean response latency: response time distributional analyses of semantic priming. *J. Mem. Lang*. 59, 495–523. doi: 10.1016/j.jml.2007.10.004

Belke, E., Brysbaert, M., Meyer, A. S., and Ghyselinck, M. (2005). Age of acquisition effects in picture naming: evidence for a lexical-semantic competition hypothesis. *Cognition* 96, B45–B54. doi: 10.1016/j.cognition.2004.11.006

Berger, T. (1971). *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Eaglewood Cliffs, NJ: Prentice-Hall.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev*. 113:700. doi: 10.1037/0033-295X.113.4.700

Braun, D. A., and Ortega, P. A. (2014). Information-theoretic bounded rationality and ε-optimality. *Entropy* 16, 4662–4676. doi: 10.3390/e16084662

Bürki, A., Elbuy, S., Madec, S., and Vasishth, S. (2020). What did we learn from forty years of research on semantic interference? A Bayesian meta-analysis. *J. Mem. Lang*. 114:104125. doi: 10.1016/j.jml.2020.104125

Caramazza, A., and Costa, A. (2000). The semantic interference effect in the picture-word interference paradigm: does the response set matter? *Cognition* 75, B51–B64. doi: 10.1016/S0010-0277(99)00082-7

Caramazza, A., and Costa, A. (2001). Set size and repetition in the picture-word interference paradigm: Implications for models of naming. *Cognition* 80, 291–298. doi: 10.1016/S0010-0277(00)00137-2

Cattell, J. M. (1886). The time it takes to see and name objects. *Mind* 11, 63–65. doi: 10.1093/mind/os-XI.41.63

Christie, S. (2019). *Information-theoretic bounded rationality: timing laws and cognitive costs emerge from rational bounds on information coding and transmission* (Ph.D. thesis), University of Minnesota, Minneapolis, MN, United States.

Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory*. Hoboken, NJ: John Wiley & Sons.

Damian, M. F., and Bowers, J. S. (2003). Locus of semantic interference in picture-word interference tasks. *Psychonom. Bull. Rev*. 10, 111–117. doi: 10.3758/BF03196474

Damian, M. F., and Martin, R. C. (1999). Semantic and phonological codes interact in single word production. *J. Exp. Psychol. Learn. Mem. Cogn*. 25:345. doi: 10.1037/0278-7393.25.2.345

de Marchis, G., Expósito, M. d. P. R., and Avilés, J. M. R. (2013). Psychological distance and reaction time in a Stroop task. *Cogn. Process*. 14, 401–410. doi: 10.1007/s10339-013-0569-x

Fan, J. (2014). An information theory account of cognitive control. *Front. Hum. Neurosci*. 8:680. doi: 10.3389/fnhum.2014.00680

Firth, J. R. (Ed.). (1957). "A synopsis of linguistic theory, 1930–1955," in *Studies in Linguistic Analysis* (Oxford: Basil Blackwell), 1–32.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci*. 11:127. doi: 10.1038/nrn2787

Futrell, R., Gibson, E., and Levy, R. P. (2020). Lossy-context surprisal: an information-theoretic model of memory effects in sentence processing. *Cogn. Sci*. 44:e12814. doi: 10.1111/cogs.12814

Genewein, T., Leibfried, F., Grau-Moya, J., and Braun, D. A. (2015). Bounded rationality, abstraction, and hierarchical decision-making: an information-theoretic optimality principle. *Front. Robot. AI* 2:27. doi: 10.3389/frobt.2015.00027

Gershman, S. J. (2020). Origin of perseveration in the trade-off between reward and complexity. *Cognition* 204:104394. doi: 10.1016/j.cognition.2020.104394

Gershman, S. J., and Bhui, R. (2020). Rationally inattentive intertemporal choice. *Nat. Commun*. 11:3365. doi: 10.1038/s41467-020-16852-y

Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. (2015). Computational rationality: a converging paradigm for intelligence in brains, minds, and machines. *Science* 349, 273–278. doi: 10.1126/science.aac6076

Glaser, W. R., and Glaser, M. O. (1989). Context effects in stroop-like word and picture processing. *J. Exp. Psychol. Gen*. 118:13. doi: 10.1037/0096-3445.118.1.13

Goldberg, Y., and Levy, O. (2014). Word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv* 1402.3722.

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. T. (2001). "A probabilistic Earley parser as a psycholinguistic model," in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics and Language Technologies*, (Pittsburgh, PA), 1–8. doi: 10.3115/1073336.1073357

Hale, J. T., Dyer, C., Kuncoro, A., and Brennan, J. (2018). "Finding syntax in human encephalography with beam search," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Melbourne, VIC: Association for Computational Linguistics), 2727–2736. doi: 10.18653/v1/P18-1254

Harris, Z. S. (1954). Distributional structure. *Word* 10, 146–162. doi: 10.1080/00437956.1954.11659520

Heathcote, A., Popiel, J., and Mewhort, D. J. K. (1991). Analysis of reposne time distributions: an example using the Stroop task. *Psychol. Bull*. 109, 340–347. doi: 10.1037/0033-2909.109.2.340

Hick, W. E. (1952). On the rate of gain of information. *Q. J. Exp. Psychol*. 4, 11–26. doi: 10.1080/17470215208416600

Howes, A., Lewis, R. L., and Vera, A. (2009). Rational adaptation under task and processing constraints: implications for testing theories of cognition and action. *Psychol. Rev*. 116:717. doi: 10.1037/a0017187

Hutson, J., and Damian, M. (2014). Semantic gradients in picture-word interference tasks: is the size of interference effects affected by the degree of semantic overlap? *Front. Psychol*. 5:872. doi: 10.3389/fpsyg.2014.00872

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *J. Exp. Psychol*. 45:188. doi: 10.1037/h0056940

Jäger, L., Engelmann, F., and Vasishth, S. (2017). Similarity-based interference in sentence comprehension: literature review and Bayesian meta-analysis. *J. Mem. Lang*. 94, 316–339. doi: 10.1016/j.jml.2017.01.004

Kahneman, D. (2003). Maps of bounded rationality: psychology for behavioral economics. *Am. Econ. Rev*. 93, 1449–1475. doi: 10.1257/000282803322655392

Klein, G. S. (1964). Semantic power measured through the interference of words with color-naming. *Am. J. Psychol*. 77, 576–588. doi: 10.2307/1420768

Koechlin, E., and Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends Cogn. Sci*. 11, 229–235. doi: 10.1016/j.tics.2007.04.005

Laming, D. R. J. (1968). *Information Theory of Choice-Reaction Times*. Academic Press.

Laming, D. R. J. (2003). *Human Judgment: The Eye of the Beholder*. Cengage Learning EMEA.

Levelt, W. J. M., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behav. Brain Sci*. 22, 1–38. doi: 10.1017/S0140525X99001776

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

Lewis, R. L., Howes, A., and Singh, S. (2014). Computational rationality: linking mechanism and behavior through bounded utility maximization. *Top. Cogn. Sci*. 6, 279–311. doi: 10.1111/tops.12086

Lieder, F., and Griffiths, T. L. (2019). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci*. 43, 1–60. doi: 10.1017/S0140525X1900061X

Luce, R. D. (1986). *Response Times*. New York, NY: Oxford University Press.

Luce, R. D. (2003). Whatever happened to information theory in psychology? *Rev. Gen. Psychol*. 7, 183–188. doi: 10.1037/1089-2680.7.2.183

Lupker, S. J. (1979). The semantic nature of response competition in the picture-word interference task. *Mem. Cogn*. 7, 485–495. doi: 10.3758/BF03198265

Lynn, C. W., Kahn, A. E., Nyema, N., and Bassett, D. S. (2020). Abstract representations of events arise from mental errors in learning and memory. *Nat. Commun*. 11:2313. doi: 10.1038/s41467-020-15146-7

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychol. Bull.* 109:163. doi: 10.1037/0033-2909.109.2.163

Mahon, B. Z., Costa, A., Peterson, R., Vargas, K. A., and Caramazza, A. (2007). Lexical selection is not by competition: a reinterpretation of semantic interference and facilitation effects in the picture-word interference paradigm. *J. Exp. Psychol. Learn. Mem. Cogn.* 33:503. doi: 10.1037/0278-7393.33.3.503

Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. W.H. Freeman & Company.

Mewhort, D. J., Braun, J. G., and Heathcote, A. (1992). Response time distributions and the Stroop task: a test of the Cohen, Dunbar, and McClelland (1990) model. *J. Exp. Psychol. Hum. Percept. Perform.* 18:872. doi: 10.1037/0096-1523.18.3.872

Meyer, A. S., and Schriefers, H. (1991). Phonological facilitation in picture-word interference experiments: effects of stimulus onset asynchrony and types of interfering stimuli. *J. Exp. Psychol. Learn. Mem. Cogn.* 17:1146. doi: 10.1037/0278-7393.17.6.1146

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*, Vol. 26, eds C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Inc.), 3111–3119.

Moscoso del Prado Martín, F., Kostić, A., and Baayen, R. H. (2004). Putting the bits together: an information theoretical perspective on morphological processing. *Cognition* 94, 1–18. doi: 10.1016/j.cognition.2003.10.015

Ortega, P. A., and Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proc. R. Soc. A Math. Phys. Eng. Sci.* 469:20120683. doi: 10.1098/rspa.2012.0683

Ortega, P. A., and Stocker, A. A. (2016). "Human decision-making under limited time," in *Proceedings of the 30th Conference on Neural Information Processing Systems* (Barcelona).

Pennington, J., Socher, R., and Manning, C. D. (2014). "GloVe: global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, (Doha: Association for Computational Linguistics), 1532–1543. doi: 10.3115/v1/D14-1162

Piai, V., Roelofs, A., and Schriefers, H. (2011). Semantic interference in immediate and delayed naming and reading: attention and task decisions. *J. Mem. Lang.* 64, 404–423. doi: 10.1016/j.jml.2011.01.004

Piai, V., Roelofs, A., and Schriefers, H. (2012). Distractor strength and selective attention in picture-naming performance. *Mem. Cogn.* 40, 614–627. doi: 10.3758/s13421-011-0171-3

Ratcliff, R. (1978). A theory of memory retrieval. *Psychol. Rev.* 85:59. doi: 10.1037/0033-295X.85.2.59

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychol. Bull.* 86, 446–461. doi: 10.1037/0033-2909.86.3.446

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition* 42, 107–142. doi: 10.1016/0010-0277(92)90041-F

Roelofs, A. (2003). Goal-referenced selection of verbal action: modeling attentional control in the stroop task. *Psychol. Rev.* 110:88. doi: 10.1037/0033-295X.110.1.88

Roelofs, A. (2012). Attention, spatial integration, and the tail of response time distributions in stroop task performance. *Q. J. Exp. Psychol.* 65, 135–150. doi: 10.1080/17470218.2011.605152

Roelofs, A. (2021). How attention controls naming: lessons from Wundt 2.0. *J. Exp. Psychol. Gen.* doi: 10.1037/xge0001030. [Epub ahead of print].

Roelofs, A., and Piai, V. (2017). Distributional analysis of semantic interference in picture naming. *Q. J. Exp. Psychol.* 70, 782–792. doi: 10.1080/17470218.2016.1165264

Rubin, J., Shamir, O., and Tishby, N. (2012). "Trading value and information in MDPs," in *Decision Making With Imperfect Decision Makers*, eds T. V. Guy, M. K1rn, and D. H. Wolpert (Berlin; Heidelberg: Springer), 57–74. doi: 10.1007/978-3-642-24647-0_3

San José, A., Roelofs, A., and Meyer, A. S. (2021). Modeling the distributional dynamics of attention and semantic interference in word production. *Cognition* 211:104636. doi: 10.1016/j.cognition.2021.104636

Scaltritti, M., Navarrete, E., and Peressotti, F. (2015). Distributional analyses in the picture-word interference paradigm: exploring the semantic interference and the distractor frequency effects. *Q. J. Exp. Psychol.* 68, 1348–1369. doi: 10.1080/17470218.2014.981196

Schach, S., Gottwald, S., and Braun, D. A. (2018). Quantifying motor task performance by bounded rational decision theory. *Front. Neurosci.* 12:932. doi: 10.3389/fnins.2018.00932

Schriefers, H., Meyer, A. S., and Levelt, W. J. M. (1990). Exploring the time course of lexical access in language production: picture-word interference studies. *J. Mem. Lang.* 29, 86–102. doi: 10.1016/0749-596X(90)90011-N

Shannon, C. E. (1959). "Coding theorems for a discrete source with a fidelity criterion," in *IRE National Convention Record*, 142–163.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science* 237, 1317–1323. doi: 10.1126/science.3629243

Simon, H. A. (1955). A behavioral model of rational choice. *Q. J. Econ.* 69, 99–118. doi: 10.2307/1884852

Simon, H. A. (1972). "Theories of bounded rationality," in *Decision and Organization*, eds C. B. McGuire and R. Radner (Amsterdam: North-Holland Publishing Company), 161–176.

Sims, C. A. (2003). Implications of rational inattention. *J. Monet. Econ.* 50, 665–690. doi: 10.1016/S0304-3932(03)00029-1

Sims, C. A. (2005). "Rational inattention: a research agenda," in *Deutsche Bundesbank Spring Conference, Number 4* (Berlin).

Sims, C. A. (2010). "Rational inattention and monetary economics," in *Handbook of Monetary Economics*, Vol. 3, Chapter 4, eds B. M. Friedman and M. Woodford (Elsevier), 155–181. doi: 10.1016/B978-0-444-53238-1.00004-1

Sims, C. R. (2016). Rate-distortion theory and human perception. *Cognition* 152, 181–198. doi: 10.1016/j.cognition.2016.03.020

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science* 360, 652–656. doi: 10.1126/science.aaq1118

Spieler, D. H., Balota, D. A., and Faust, M. E. (2000). Levels of selective attention revealed through analyses of response time distributions. *J. Exp. Psychol. Hum. Percept. Perform.* 26:506. doi: 10.1037/0096-1523.26.2.506

Starreveld, P. A., and La Heij, W. (2017). Picture-word interference is a Stroop effect: a theoretical analysis and new empirical findings. *Psychonom. Bull. Rev.* 24, 721–733. doi: 10.3758/s13423-016-1167-6

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *J. Exp. Psychol.* 18:643. doi: 10.1037/h0054651

Tishby, N., and Polani, D. (2011). "Information theory of decisions and actions," in *Perception-Action Cycle*, eds V. Cutsuridis, A. Hussain, and J. Taylor (New York, NY: Springer), 601–636. doi: 10.1007/978-1-4419-1452-1_19

van Dijk, S. G., and Polani, D. (2011). "Grounding subgoals in information transitions," in *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)* (IEEE), 105–111. doi: 10.1109/ADPRL.2011.5967384

van Dijk, S. G., and Polani, D. (2013). Informational constraints-driven organization in goal-directed behavior. *Adv. Complex Syst.* 16:1350016. doi: 10.1142/S0219525913500161

van Dijk, S. G., Polani, D., and Nehaniv, C. L. (2009). "Hierarchical behaviours: getting the most bang for your bit," in *European Conference on Artificial Life* (Springer), 342–349. doi: 10.1007/978-3-642-21314-4_43

van Maanen, L., van Rijn, H., and Borst, J. P. (2009). Stroop and picture-word interference are two sides of the same coin. *Psychonom. Bull. Rev.* 16, 987–999. doi: 10.3758/PBR.16.6.987

Watkins, O. C., and Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *J. Exp. Psychol. Hum. Learn. Mem.* 1:442. doi: 10.1037/0278-7393.1.4.442

Wood, W., and Rünger, D. (2016). Psychology of habit. *Annu. Rev. Psychol.* 67, 289–314. doi: 10.1146/annurev-psych-122414-033417

Zaslavsky, N., Kemp, C., Regier, T., and Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proc. Natl. Acad. Sci. U.S.A.* 115, 7937–7942. doi: 10.1073/pnas.1800521115

Zénon, A., Solopchuk, O., and Pezzulo, G. (2019). An information-theoretic perspective on the costs of cognition. *Neuropsychologia* 123, 5–18. doi: 10.1016/j.neuropsychologia.2018.09.013

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 Psychology

ORIGINAL RESEARCH
published: 18 June 2021
doi: 10.3389/fpsyg.2021.661415

Lexical Diversity, Lexical Sophistication, and Predictability for Speech in Multiple Listening Conditions

*Melissa M. Baese-Berk**, *Shiloh Drake*, *Kurtis Foster*, *Dae-yong Lee*, *Cecelia Staggs and Jonathan M. Wright*

Speech Perception and Production Lab, Department of Linguistics, University of Oregon, Eugene, OR, United States

When talkers anticipate that a listener may have difficulty understanding their speech, they adopt a speaking style typically described as "clear speech." This speaking style includes a variety of acoustic modifications and has perceptual benefits for listeners. In the present study, we examine whether clear speaking styles also include modulation of lexical items selected and produced during naturalistic conversations. Our results demonstrate that talkers do, indeed, modulate their lexical selection, as measured by a variety of lexical diversity and lexical sophistication indices. Further, the results demonstrate that clear speech is not a monolithic construct. Talkers modulate their speech differently depending on the communication situation. We suggest that clear speech should be conceptualized as a set of speaking styles, in which talkers take the listener and communication situation into consideration.

Keywords: lexical sophistication, lexical diversity, non-native speech, speech in noise, adverse listening conditions

Edited by:
Rory Turnbull,
Newcastle University, United Kingdom

Reviewed by:
McNeel Gordon Jantzen,
Western Washington University,
United States
Carlos Romero-Rivas,
Autonomous University of
Madrid, Spain

Correspondence:
Melissa M. Baese-Berk
mbaesebe@uoregon.edu

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

Received: *30 January 2021*
Accepted: *21 May 2021*
Published: *18 June 2021*

Citation:
Baese-Berk MM, Drake S, Foster K, Lee D, Staggs C and Wright JM (2021) Lexical Diversity, Lexical Sophistication, and Predictability for Speech in Multiple Listening Conditions.
Front. Psychol. 12:661415.
doi: 10.3389/fpsyg.2021.661415

INTRODUCTION

When communicating in natural situations, talkers modulate their speech for their audience (e.g., Clark et al., 1983). Modulation can take many forms, including choosing appropriate lexical items for the audience, modulating syntactic structure, and modifying acoustic properties (Clark and Carlson, 1981; Clark and Murphy, 1982; Arnold et al., 2012). This type of modulation typically happens without explicit instruction or feedback and is robust across talker populations and contexts (Beckford Wassink et al., 2007; Androutsopoulos, 2014; Ferreira, 2019).

The most famous example of this type modulation is child- or infant-directed speech (IDS), a speaking style used, as the name suggests, when communicating with infants or children (Snow, 1977; Stern et al., 1982; Fernald and Simon, 1984). While individual talkers may differ in their exact implementation of IDS, common properties of this speaking style include higher average pitch, a broader pitch range, and shorter utterance durations. Infant-directed speech is not universal (e.g., Pye, 1986; Ingram, 1995); however, it is widely used in many cultures, without explicit instruction (e.g., Grieser and Kuhl, 1988; Fernald et al., 1989; Kuhl et al., 1997). Infant-directed speech also changes in both syntactic and lexical complexity as the infant grows older, presumably in response to increases in infants' receptive abilities as well as their ability to communicate with adult interlocutors (Genovese et al., 2020). Even children can produce IDS in situationally appropriate ways (Dunn and Kendrick, 1982; Warren-Leubecker and Bohannon, 1983; Weppelman et al., 2003), suggesting that the ability to modulate our speech for our audience develops rather quickly and is robust.

However, even when speaking to adult listeners, talkers modulate their speech in a variety of ways. For example, a wide range of speaking styles have often been included under the umbrella of "clear speech." Clear speech is typically defined as a listener-oriented speaking style, characterized primarily by a variety of acoustic modifications. However, recent work has suggested that clear speech differs as a function of the intended audience or communication style (Hazan and Baker, 2011; Hazan et al., 2012) and that clear speech produced in naturalistic communication scenarios differs from clear speech elicited in a laboratory in more artificial communication scenarios (Moon and Lindblom, 1994; Scarborough et al., 2007; Hazan and Baker, 2011; Scarborough and Zellou, 2013).

The bulk of work on clear speech has focused on acoustic modifications of the speech signal, which are thought to make the signal easier for the listener to understand. However, other lines of research have demonstrated that there are multiple other factors that impact how easy it is for listeners to understand the speech they are exposed to. For example, semantic predictability impacts how accurately listeners perceive speech in a variety of challenging listening situations, including speech in noise (Signoret et al., 2018), non-native speech (Baese-Berk et al., 2021), hearing-impaired listeners (Holmes et al., 2018), and cochlear implant users (Winn, 2016).

Indeed, predictability is crucially important for speech understanding and communication in general (see Kutas et al., 2011 for a review). For example, many studies have suggested that listeners use prediction to determine when a speaker is likely to complete their turn so that they can begin the next conversational turn as a speaker (Schegloff et al., 1974). On even shorter time scales, listeners make eye-movements toward relevant targets before they are produced if the target is syntactically or semantically predictable (e.g., Altmann and Kamide, 1999, 2007). Predictability, in various forms, has also been shown to impact language processing. Less predictable words are read more slowly than their more predictable counterparts (Ehrlich and Rayner, 1981; Levy, 2008), and predictability of lexical items is evident in event related potentials (ERPs) to unpredictable lexical items (e.g., N400 responses to semantically less predictable nouns, Kutas and Hillyard, 1980).

In the current paper, we do not directly investigate predictability per se. Instead, we examine lexical factors that could affect the predictability of the speech that listeners hear and could impact the ease of understanding speech. Specifically, we examine speech produced in naturalistic communication scenarios across a variety of contexts known to elicit a clear speech style. We ask whether, in addition to acoustic modifications previously reported, speakers modulate the lexical content of their speech—including a variety of measures of lexical diversity and lexical sophistication. We ask whether these measures differ both when (1) comparing scenarios that naturally elicit clear speech to those that do not elicit such a style and (2) comparing within distinct communication situations that may each elicit clear speech, but differ in their specific challenges for the talker and listener (e.g., speech to a non-native talker vs. to someone hearing the speech through noise).

Below, we briefly review the previous literature on modifications found in clear speech, measures of lexical diversity, and measures of lexical sophistication before turning our attention to the current study.

RELATED WORK

Communication in Adverse Listening Situations/Clear Speech

As described above, clear speech is a speaking style adopted by speakers, usually in situations where they anticipate that their listener may have trouble understanding their speech. Substantial previous work has examined the acoustic properties of clear speech. Typical modifications include slower speaking rates, higher average intensity, greater fundamental frequency range, and larger vowel spaces compared to plain or conversational speech (Picheny et al., 1986; Krause and Braida, 2004; Smith, 2007; Maniwa et al., 2009).

Importantly, these modifications result in a benefit for the listener. That is, listeners are able to more accurately transcribe speech (i.e., intelligibility) when the speech is produced in a clear speaking style (Bradlow and Bent, 2002; Krause and Braida, 2002; Maniwa et al., 2008; Hazan and Baker, 2011). These benefits emerge for a variety of listener populations including normal-hearing listeners (Krause and Braida, 2002; Liu and Zeng, 2006; Hazan et al., 2018), hearing-impaired listeners (Picheny et al., 1985; Ferguson and Kewley-Port, 2002), listeners with cochlear implants (Liu et al., 2004), non-native listeners (Bradlow and Bent, 2002), and for speech-in-noise in a variety of populations (Payton et al., 1994; Bradlow and Alexander, 2007; Calandruccio et al., 2020).

Primarily, the studies cited above elicited speech in the laboratory using instructions to produce speech for a hypothetical listener who may have challenges understanding the speech. Some previous work has elicited clear speech with naturalistic methods (Moon and Lindblom, 1994; Scarborough et al., 2007; Hazan and Baker, 2011; Scarborough and Zellou, 2013). In these situations, talkers typically do not receive instructions to modify their speech or to speak clearly. Instead, they are placed in communication situations where their speech will be harder for their listener to understand. There have been some differences reported between these two elicitation types with some showing more hyperarticulation in speech elicited in naturalistic conditions and others showing more hyperarticulation for speech elicited with a hypothetical listener. Importantly, compared to plain speech, both types of elicitation methods result in acoustic modifications and perceptual benefits (see e.g., Hazan and Baker, 2011; Hazan et al., 2015; Lee and Baese-Berk, 2020).

While these previous findings have demonstrated that this listener-oriented speaking style tends to result in both acoustic modifications by the talker and perceptual benefits for listeners, much less attention has been paid to other properties of the language produced by speakers in these situations, especially in clear speech elicited in naturalistic situations. That is, one could imagine that when in a naturalistic environment where communicative success is imperative, talkers may modify their speech in multiple ways, including lexical, syntactic, or pragmatic selection. These modifications could result in even greater ease for listeners. This type of investigation is critically important because some previous work has demonstrated that intelligibility benefits for listeners are not necessarily reflected in acoustic modifications of clear speech (e.g., Lee and Baese-Berk, 2020). That is, in some cases listeners understand speech that was elicited in naturalistic scenarios that often result in "clear speech" better than speech elicited as "plain speech," but investigations for acoustic correlates that may be driving these results have not shown significant differences between the two speaking styles (e.g., no significant differences in speaking rate, F0, intensity, etc.). Therefore, it is possible that other, non-acoustic, properties of the signal are impacting ease of understanding for listeners.

Further, most previous studies of clear speech have examined the speaking style as a monolithic construct, and have not directly investigated cases in which the specific properties of clear speech might shift as a function of the audience and the needs of the audience. As a counterexample, Hazan et al. (2012) demonstrated that acoustic properties of clear speech differ as a function of communicative barrier (i.e., vocoded speech vs. speech presented in multi-talker babble). For example, speaking rate and fundamental frequency differ across the two conditions—though both are distinct from plain speech. Also, preliminary work from our lab (Wright and Baese-Berk, 2020) suggests that lexical and syntactic information may shift as a function of the needs of the audience. Using only lexical and syntactic information from the talker's speech in transcriptions of conversations from the LUCID corpus, which included three clear speech eliciting conditions and one plain speech eliciting condition, we found that natural language processing classifiers perform significantly above chance when predicting the listening condition of the audience based solely on the talker's speech. This suggests that there are some non-acoustic properties of the speech that are differentiated among the various clear speech eliciting conditions. However, the factors differentiating lexical and syntactic properties that allowed the classifiers to perform well were not clear.

There is a broad body of work on how interlocutors refer to objects in the world in conversation (see Arnold, 2008 for a review). When speaking, we have the choice of many different ways to refer to the same referent in the world (e.g., *the cat*, *it*, *the striped one*), and the method of reference we select seems to depend on many factors. Among these factors are whether the information being referred to is new or given (i.e., previously referred to in discourse), what a speaker knows about a listener's familiarity with the topic, other information that the speaker infers about the listener (e.g., proficiency in the language of discourse), and ease of retrieval for the speaker. Thus, it seems that the notion of what constitutes "clear speech" can be even further subdivided.

Therefore, here, we investigate one specific aspect that could be modified by talkers during elicitation of clear speech in naturalistic conversations: lexical selection. Below, we briefly describe the two families of measurements used in our analyses: lexical diversity and lexical sophistication. Both families of measures are used widely in assessment of second language writing, among other fields. We believe that they are appropriate for the present study because they provide us with a series of measures capable of directly assessing lexical complexity, which may impact how listeners perceive speech and/or how speakers modify their speech for listeners.

Lexical Diversity

Broadly speaking, lexical diversity is the range of different words used in a text or conversation. A greater range is equivalent to higher diversity. Lexical diversity is used in a variety of assessment tools including as a measure of proficiency in a second language (Engber, 1995; Cumming et al., 2005), vocabulary knowledge (Zareva et al., 2005; Yu, 2010), and even as a marker of onset of neurodegenerative diseases like Alzheimer's disease (Garrard et al., 2005; van Velzen and Garrard, 2008) or in mild cases of aphasia (Cunningham and Haley, 2020). Measures of lexical diversity are important for many reasons. While more diverse texts or speech samples may be indicative of greater proficiency for the speaker or writer, they may also be more challenging for a reader or listener to understand. That is, samples with greater diversity, may also include less repetition, more switches among topics, and use of multiple lexical items to refer to the same concept. Each of these factors could make it *more* challenging for a listener to understand what is being said. Therefore, we may expect *lower* lexical diversity values in clear speech situations than in plain speech situations.

Historically, lexical diversity has been indexed via the type-token ratio (Johnson, 1944; Templin, 1957), in which the total number of unique words (i.e., types) is divided by the total number of words (i.e., tokens). The closer this ratio is to 1, the greater lexical diversity in the sample. However, indices like type-token ratio are often sensitive to length of language sample: longer texts often have disproportionately lower type-token ratios than shorter texts, and this value may not be indicative of lexical diversity more broadly. Further, some measures of lexical diversity (including type-token ratio), make assumptions about textual homogeneity. That is, some measures of lexical diversity fail to recognize that talkers may vary diversity levels in different points of conversation or a text for some specific purpose. For instance, there are particular circumstances in which language that is less lexically diverse is employed as a rhetorical strategy, therefore, indices have been developed that control for the intentional use and variety of particular structures. This serves to ensure that the measure does not treat a single structure or pattern as representative of the text as a whole (see McCarthy and Jarvis, 2010 for a summary of these issues).

In the present study, we present results from the typical type-token ratio analyses. However, given the considerations above, we also report three additional measures, which may provide a more complete understanding of lexical diversity within our sample. First, we report the moving average type-token ratio (MATTR; Covington and McFall, 2010), which uses a 50-word window to continuously calculate type-token ratio throughout a sample. Second, we report the hypergeometric distribution (HD-D; McCarthy and Jarvis, 2007). This represents the probability of drawing a number of tokens with some specific type from a sample of a specific size. Finally, we report a version of the "measure of textual lexical diversity" (MTLD; McCarthy, 2005; McCarthy and Jarvis, 2010). While we refer the reader to previous work for specific descriptions of this index, the measure roughly corresponds to the average length in words that the sample stays at a specific type token ratio.

Taken together, we believe these indices will allow us to better understand the lexical diversity of the samples in the current study. By comparing how these indices differ across a number of conditions that induce clear speech, we will be able to better understand how clear speech may vary across scenarios.

Lexical Sophistication

Lexical sophistication is often simply described as the number of "unusual" words in a sample. As is the case for lexical diversity, a number of constructs can be used for characterizing lexical sophistication, depending on the goals of the researcher (Eguchi and Kyle, 2020). Lexical sophistication is frequently used as an indicator of language proficiency in second language assessments of speaking and writing (Laufer and Nation, 1995; Kyle and Crossley, 2015; McNamara et al., 2015). However, we believe that it could be a tool to characterize the relative lexical complexity of clear speech, as in the current study.

Here, we specifically assess four measures of lexical sophistication (see Crossley et al., 2012), all of which investigate the relative frequency of a word or sets of words. First, we report the lexical frequency for words within our speech samples. This frequency is calculated using a reference corpus. The reference corpus should, ideally, match the properties of the speech sample, given that relative frequency of a word, for example, may differ across language variety or modality (i.e., spoken vs. written). We discuss this issue in more detail below. Second, we report the range, or the number of speech samples in a particular corpus in which a word occurs. Third, we report two measures of bigram frequency in a sample: the mean frequency for bigrams (i.e., pairs of words) and the proportion of bigrams in the sample that are within the most frequent 25,000 bigrams in the corpus. Finally, we report the same two measures for trigrams (sets of three consecutive words).

We interpret measures of lexical sophistication as being indicative of lexical complexity within our clear speech and plain speech samples. We predict that, if talkers modify their lexical complexity for their audience, they will use higher frequency words and higher frequency collocations (i.e., bigrams and trigrams) when producing clear speech than plain speech.

Current Study

In the current study we examine talker speech modulations across naturalistic scenarios in the London UCL Clear Speech in Interaction Corpus (LUCID; Baker and Hazan, 2011; Hazan and Baker, 2011). The LUCID corpus includes naturalistic conversations in a variety of conditions designed to elicit clear speech, as well as a "no-barrier" condition that elicit naturalistic conversation between native English speakers. The clear speech conditions include speech in noise, a simulation of speech through a cochlear implant (i.e., vocoded speech), and conversations between individuals who do not share a language background (i.e., native English speakers and non-native English speakers). Previous studies have used this dataset to demonstrate that talkers make acoustic modifications of their speech in clear-speech situations (Hazan and Baker, 2011) and that speech in clear-speech situations is more easily understood than speech in plain-speech situations (Hazan and Baker, 2011; Lee and Baese-Berk, 2020). To determine how speakers might modulate other aspects of their speech, we use measures of lexical diversity and lexical sophistication to directly investigate how talkers modulate lexical selection across clear-speech eliciting conditions and plain-speech eliciting conditions.

Specifically, we compare lexical selection in clear-speech eliciting conditions to a condition not designed to elicit clear speech. As previous studies have shown robust acoustic differences between the two broad speaking styles, we ask whether lexical diversity and lexical sophistication also differ between these styles.

We also compare clear-speech eliciting conditions with L1 listeners to speech directed to L2 listeners. We ask whether speech to L2 listeners without an additional barrier to communication differs from speech to L1 listeners in communicatively challenging situations (speech in noise; a simulation of speech through a cochlear implant). Most work on clear speech refers to the clear-speech speaking style as "listener-oriented," and groups clear-speech eliciting conditions together under the same umbrella. However, here, we ask whether clear-speech eliciting conditions are actually the same and whether talkers are orienting their speech toward some generic listener who may have difficulty understanding them or whether this modulation is more dynamic in nature. While clear-speech eliciting conditions may share some properties, they may also differ in ways that are important to understand if we are to fully account for how talkers modulate their speech for their audience.

Finally, we compare clear-speech eliciting conditions with L1 speakers directly to each other, asking whether measures of lexical diversity and lexical sophistication reveal differences in lexical selection in speech to L1 listeners as a function of the challenging listening situation, which expands on previous work that has demonstrated that there are acoustic features that differ as a function of the communication challenge faced (Hazan et al., 2012).

METHODS

In this study, we analyze data previously collected for the LUCID corpus. Below, we briefly describe the participants and task before describing more detail the specific stimuli we analyzed in the present paper, the measures we extracted, and the analyses conducted. For more in depth descriptions of the participants and task, we direct the reader to Baker and Hazan (2011). Further, all sound files and transcripts analyzed in this project are publicly available via SpeechBox (Bradlow[1]).

[1] Bradlow, A. R. *SpeechBox*. https://speechbox.linguistics.northwestern.edu.

Participants

Participants in this task were 40 native, monolingual speakers of southern British English, between 18 and 29 years of age. 20 participants identified as female, and 20 participants identified as male. Participants did not self-identify as having a history of speech or hearing disorder and all participants passed a basic hearing screening.

Task

Each participant in the LUCID Corpus completed a set of Diapix tasks (Van Engen et al., 2010). Participants in this task completed a "spot-the-differences" task. Each participant is presented with a different hand-drawn picture that is very similar to their partner's picture but contains several key differences. These differences can include missing items (e.g., a sign being present in one picture but absent in the other) or differences in objects or actions (e.g., a girl sitting on a beach ball in one picture but playing with the beach ball in the other picture). Differences in missing items are equally distributed between picture pairs. Participants are asked to collaborate with their partner to find 10 differences between their pictures without seeing their partner's picture (see Baker and Hazan, 2011 for pictures used in the Diapix tasks). This task requires both partners to contribute to solving the task, resulting in a different balance of speech across talkers than tasks like the Map Task (Anderson et al., 1991), which has a set giver-receiver structure. The range of items in a Diapix picture allows the experimenter to more closely limit the lexical items that will be discussed in the picture than a free-ranging conversation, and, at the same time, the specific structure of the pictures described in the LUCID corpus (i.e., DiapixUK) requires participants to use a variety of linguistic structures to accurately complete the task.

The LUCID corpus includes talkers describing one of three different types of scenes: beach, farm, or shop. Each participant completed each scene with a different partner or communication situation. During session 1, all talkers completed the task in quiet listening conditions. During session 2, the target talkers spoke to partners who heard vocoded speech (i.e., cochlear implant simulations). During session 3, talkers spoke with a partner who either heard the speech in multi-talker babble (i.e., noise) or a partner who is a native speaker of a non-English language and is a low-proficiency English speaker. Therefore, speech was produced in one of four conditions analyzed below. We adopt the terminology used by Hazan and colleagues in their work to refer to these conditions: no-barrier (i.e., conversational/plain speech), vocoded (i.e., cochlear implant simulation), babble (i.e., speech-in-noise), and L2 (i.e., speech with a communication partner who is a non-native speaker). No talkers produced speech in all conditions; however, all talkers produced speech in three of the four conditions. Further, the order of the pictures was counterbalanced across talkers, thus any effects below cannot be accounted for solely by picture content or picture order. By examining speech from the same set of talkers, we also hope to roughly control for individual differences in how talkers modulate their speech for an audience.

Stimuli

The LUCID corpus contains sound files for each conversation and each conversation is orthographically transcribed in time-aligned TextGrids. For this project, we used the Praat TextGrids (Boersma and Weenink, 2021) associated with each sound file to extract the speech from the target talker for each conversation. Here, we define the target talker as the talker who does not experience the communication barrier (i.e., not hearing speech in babble or through a vocoder). The transcriptions were cleaned to prepare them for tokenization (i.e., dividing the transcript into individual words) and lemmatization (i.e., modifying the words into uninflected lexical items) using the Tool for the Automatic Analysis of Lexical Diversity (TAALED) and Tool for the Automatic Analysis of Lexical Sophistication (TAALES) interfaces (described below). All filled and unfilled pauses, as well as other vocal noises (i.e., laughter) were removed from the transcriptions.

Measurements

Using the transcripts described above, we extracted a series of lexical diversity and lexical sophistication measures. For the lexical diversity measures, we used the TAALED (Kyle et al., 2021). This tool allows for extraction of typical measures of lexical diversity (e.g., type-token ratio), but also a variety of more complex measures of diversity (e.g., MTLD). For the lexical sophistication measures, we used the TAALES (Kyle and Crossley, 2015; Kyle et al., 2018).

Tool for the Automatic Analysis of Lexical Diversity calculates lexical diversity within a single spoken or written text, and thus does not require a reference corpus. Tool for the Automatic Analysis of Lexical Sophistication, on the other hand, calculates frequency information and other measures in reference to larger corpora, and thus requires a reference corpus. Because our speakers in this study were all native speakers of southern British English, we used the British National Corpus (BNC Consortium, 2007) as our reference corpus. Specifically, we used the spoken-language sections of the corpus, since we are examining spoken language, not written language[2].

Analyses

We conducted linear mixed models for each measurement of interest. For each measurement, the measurement (e.g., type-token ratio) was the dependent variable. Condition was the fixed factor. We Helmert coded condition to make the following comparisons: (1) no-barrier condition vs. barrier conditions (L2, vocoded, and babble); (2) L2 vs. other barrier conditions (i.e., babble and vocoded speech); and (3) babble vs. vocoded speech[3].

[2] The sound files for the LUCID corpus and the transcriptions in Praat TextGrids are publicly available in SpeechBox (*https://speechbox.linguistics.northwestern.edu/#!/home.*) Tool for the Automatic Analysis of Lexical Sophistication and Tool for the Automatic Analysis of Lexical Diversity are also both publicly available (*https://www.linguisticanalysistools.org.*) Further, our data (exported from TAALES and TAALED) and code and preregistration for our analyses are available via OSF (*https://osf.io/dfhpu/?view_only=49d95d90424941da82217a239ab7450c*).

[3] Note that this analysis (and any analysis with multiple levels for a single factor in a mixed model) does not allow reporting of a "main effect" of condition, as in a traditional ANOVA (see, e.g., Schad et al., 2020). Therefore, these comparisons are not *post-hoc* comparisons but are the (preregistered) comparisons of interest

Our reasoning for including these comparisons was as follows: First, we need to understand whether participants modify these factors when producing speech in challenging listening situations in general vs. in an "easy" listening condition. The first comparison answers this question. Second, the three barrier conditions all differ from each other, but the L2 condition differs from the other two conditions in that both of those conditions have a similar listener (i.e., L1 listener). The second comparison allows us to ask whether the language background of the interlocutor corresponds to specific modifications of lexical selection by the talker. Finally, we ask whether the two conditions with an L1 listener in a challenging situation differ from one another through the third comparison.

In all models, we include talker as a random intercept. Inclusion of other random effects (e.g., scene) resulted in overfitting of the models and are thus not included (Barr et al., 2013).

Significance of each factor was calculated using model comparisons where a model without the factor in question was compared to a model including that factor. Tables containing full model results are included in **Supplementary Material**. Below, we summarize the model comparison results.

RESULTS

Below, we present analyses for each of the indices we have calculated. First, we present the results for lexical diversity, followed by the results for lexical sophistication. In all cases we investigate all words produced, rather than subsetting to content words or function words. In general, content words show similar patterns to the full set of words. Patterns for function words differ slightly, but we believe that this is largely driven by the fact that function words in general are a smaller set of words which skew these measures. Therefore, below we report the analyses for all words.

Lexical Diversity

Before examining specific indices, it is useful to note how much speech is produced in each condition. Because it is clear that some lexical diversity measures are sensitive to length of sample, we begin by reporting the average number of tokens in each sample for each condition. This is shown in **Table 1** below:

TABLE 1 | Average number of words (i.e., tokens) per conversation per condition.

Condition	Average number of tokens
No-barrier	662.78
L2	1,095.92
Babble	756.75
Vocoded	785.21

It is clear that talkers produce the most speech when communicating with an L2 listener and the least speech when speaking in the "no-barrier" condition. The two other "barrier" conditions (babble and vocoded speech) are intermediate, but are closer to the no-barrier condition than to the L2 condition. This suggests that if we find effects of lexical diversity with indices that are sensitive to sample length (e.g., type-token ratio) these effects may be driven by these rather large differences in text length. We still report these results below because we believe that a picture from all metrics is informative.

Type-Token Ratio

All three main effects were significant for the analysis of type-token ratio. The comparison of the no-barrier condition to the other three conditions significantly improved model fit ($\chi^2 = 139.8$, $p < 0.0001$). The comparison of the L2 condition to the other two barrier conditions (babble and vocoded) also significantly improved model fit ($\chi^2 = 60.916$, $p < 0.0001$). Finally, the comparison between the babble and vocoded conditions also significantly improved model fit ($\chi^2 = 6.15$, $p = 0.013$).

Examining **Figure 1** below, it is clear that the type-token ratio is highest for the no-barrier condition, compared to the other conditions. Further, the L2 condition demonstrates the lowest type-token ratio, and the other two conditions are intermediate, with the vocoded condition showing a higher type-token ratio than the babble condition. This is in line with our prediction that talkers might use more repetitive speech in the "barrier" conditions than the no-barrier condition. However, this is also in line with previous findings suggesting that type-token ratio may be sensitive to sample length. Therefore, we now turn our attention to more sophisticated measures of lexical diversity.

Moving Average Type-Token Ratio

As in the case of type-token ratio, all three main effects were significant for the analysis of the MATTR (calculated over a 50-word window). The comparison of the no-barrier condition to the other three conditions significantly improved model fit ($\chi^2 = 149.1$, $p < 0.0001$). The comparison of the L2 condition to the other two barrier conditions also significantly improved model fit ($\chi^2 = 7.85$, $p = 0.005$). Finally, the comparison between the babble and vocoded conditions also significantly improved model fit ($\chi^2 = 20.037$, $p < 0.001$).

As demonstrated in **Figure 2** below, it is clear these results fall in line with those results for the basic type-token ratio described above.

Hypergeometric Distribution

Here, the results differ from the two type-token ratio analyses described above. Two of the main effects significantly improve model fit. The comparison of the no-barrier condition to the other three conditions significantly improved model fit ($\chi^2 = 80.207$, $p < 0.0001$). Further, the comparison between the babble and vocoded conditions also significantly improved model fit ($\chi^2 = 8.9887$, $p = 0.003$). However, the comparison

for this study. At the request of an anonymous reviewer, we also conducted analyses to examine overall effects. The results of these analyses, presented in **Supplemental Materials**, mirror those reported below. The models for all but one metric have t-values >1.85, suggesting a significant difference among conditions. The exception to this is trigram frequency (in Section Trigram Frequency below), which also does not show differences among conditions in our preregistered analysis (reported below in each of the subsections of section Results).

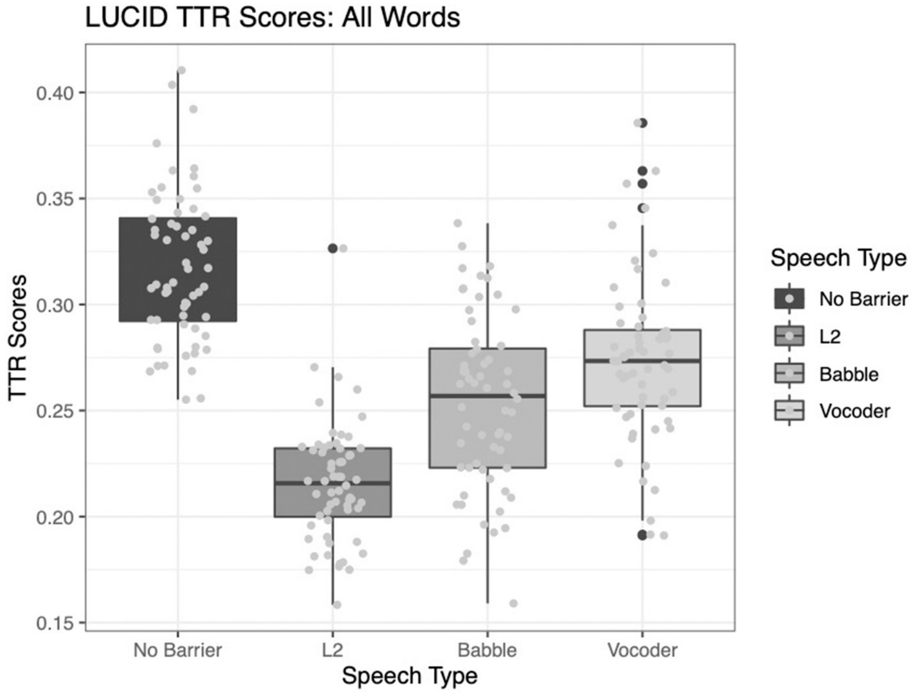


FIGURE 1 | Type-token ratio across four conditions.

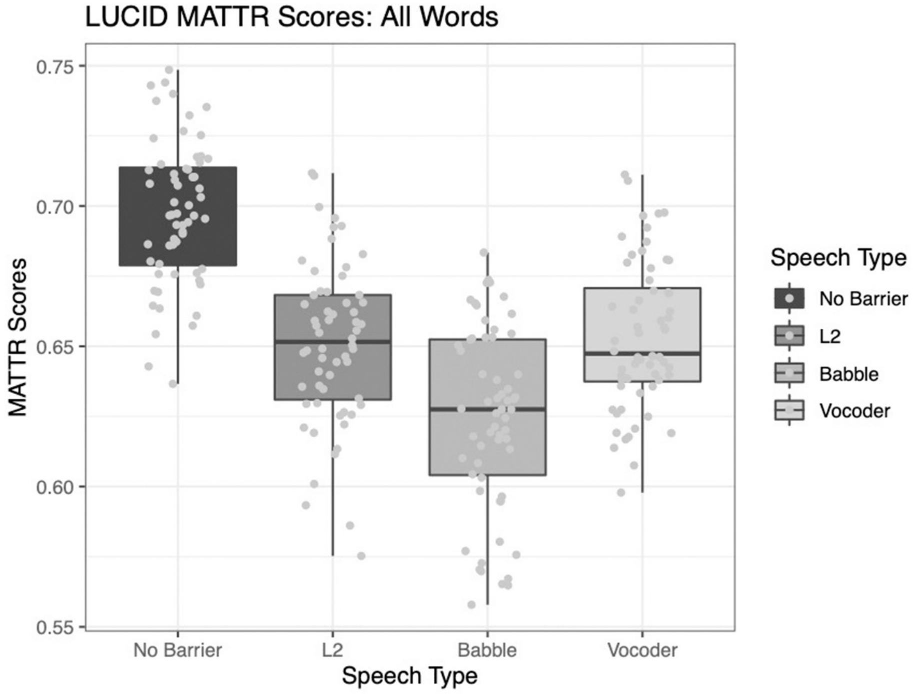


FIGURE 2 | Moving-average type-token ratio calculated over a 50-word window across four conditions.

of the L2 condition to the other two barrier conditions does not significantly improve model fit ($\chi^2 = 0.1698$, $p = 0.6803$).

Figure 3 shows the results for this index. Note that HD-D is designed to control for the assumption of homogeneity in the sample, more than for the imbalance in text size, suggesting that

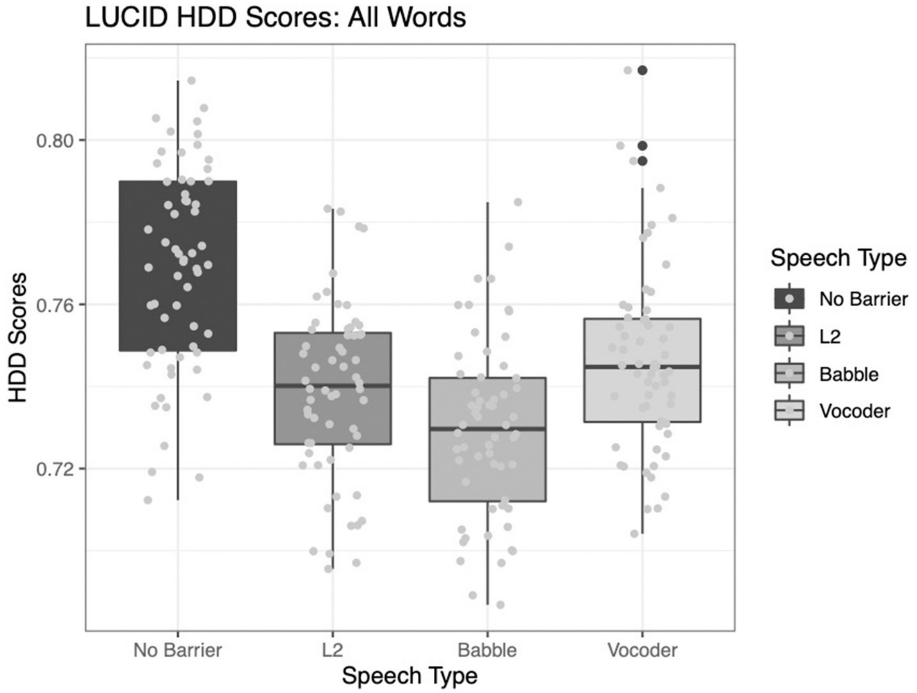


FIGURE 3 | Hypergeometric distribution (from a random sample of 42 tokens); converted to the same scale as type-token ratio across four conditions.

when controlling for homogeneity, speech to L2 listeners may be similar in terms of lexical diversity to speech in the other two barrier conditions.

Measure of Textual Lexical Diversity

As in the case of the type-token ratio indices reported above, all main effects significantly improve model fit. The comparison of the no-barrier condition to the other three conditions significantly improved model fit ($\chi^2 = 119.69$, $p < 0.0001$). The comparison of the L2 condition to the other two barrier conditions (babble and vocoded) also significantly improved model fit ($\chi^2 = 4.3075$, $p = 0.038$). Finally, the comparison between the babble and vocoded conditions also significantly improved model fit ($\chi^2 = 8.2303, p = 0.004$).

Figure 4 depicts the MTLD indices for each condition.

Order Effects

One concern with the results here is that participants perform the task multiple times, and thus the order of conditions may impact the results. The order of conditions was fixed across participants such that all participants first completed the no barrier condition followed by the vocoded condition. Half the participants then completed the babble condition and half of the participants completed the L2 condition. Therefore, condition order is conflated with condition type for this study. However, given the results, we believe that order of condition is not a major concern for our study. That is, one might expect that over time participants would repeat words more often (i.e., have lower lexical diversity measures). If this were the case, we would expect that the L2 and babble conditions should have the least lexical diversity. While it is the case that, in general, these conditions have less lexical diversity than the no barrier condition, they do not differ systematically from the vocoder condition. Therefore, we believe it is unlikely that order of conditions alone explains our results. This interpretation is in line with evidence from Baker and Hazan (2011), who demonstrated that these participants did not appear to improve or "learn" across iterations of completing this task.

A second concern is that the order of pictures within a condition may impact performance. Each participant completed three pictures within each condition. However, order of picture was not a significant predictor of model fit for any of the above metrics, and was therefore not included in the final model fit for any metric. This is consistent with evidence suggesting participants do not complete the task more quickly across iterations of the pictures (Lee and Baese-Berk, 2020).

Interim Summary

Taken together, these results suggest that there are significant differences in lexical diversity between conditions that are and are not designed to elicit clear speech. The no-barrier condition shows the most lexical diversity, whereas the L2 condition, generally, shows the least diversity. There are some differences across metrics in terms of the relative ranking of diversity values for the babble and vocoded conditions, suggesting that these two

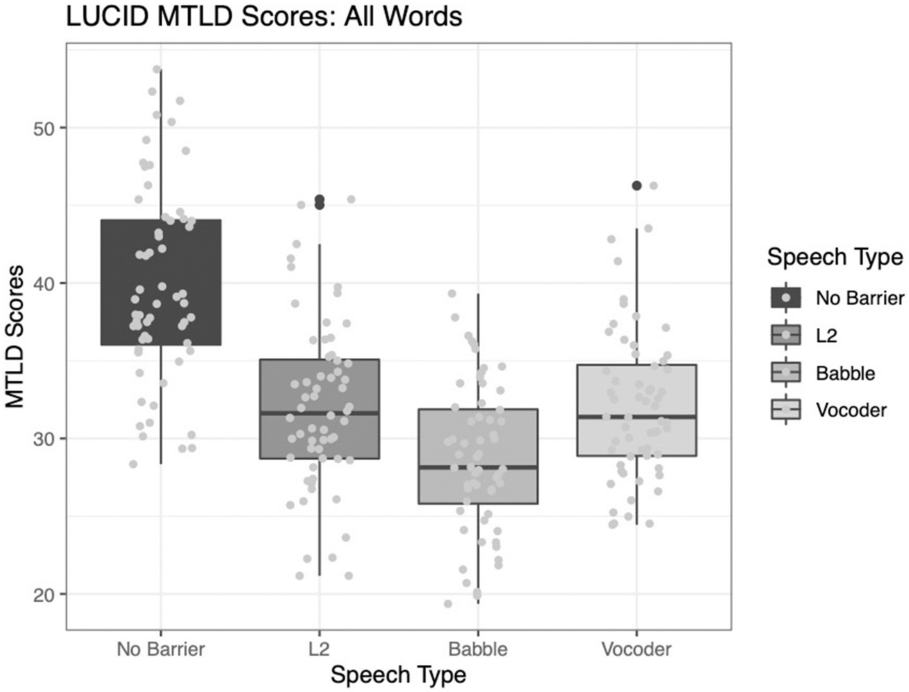


FIGURE 4 | Measure of textual lexical diversity (using a moving average approach, both forward and backward) across four conditions.

conditions may be more similar to one another than to either the no-barrier or L2 conditions.

Lexical Sophistication

As in the case of the lexical diversity results presented above, we describe each index in turn below.

Lexical Frequency

Two of the main effects significantly improved model fit for the analysis of lexical frequency. The comparison of the no-barrier condition to the other three conditions significantly improved model fit ($\chi^2 = 66.666$, $p < 0.0001$). The comparison of the L2 condition to the other two barrier conditions (babble and vocoded) also significantly improved model fit ($\chi^2 = 12.225$, $p = 0.0005$). However, the comparison between the babble and vocoded conditions did not significantly improve model fit ($\chi^2 = 3.3212$, $p = 0.068$).

Examining **Figure 5** below, it is clear that lexical frequency is the lowest for the no-barrier condition and highest for the L2 condition. As in the case of the lexical diversity measures presented above, the other two conditions fall intermediate to these conditions. While numerically the babble condition shows higher frequency than the vocoded condition, this difference was not significant. This result suggests that speakers modify not only the variability in words they produce, but also specifically *which* words they produce. We continue to explore these effects with the indices below.

Range

The pattern of results for range is different from any of the previously reported results. Recall that range here refers to the number of samples in the reference corpus (i.e., BNC) that a word appears in. Another way of describing this metric is how "common" the word is. Here, we see that the comparison of the no-barrier condition to the other three conditions *did not* significantly improve model fit ($\chi^2 = 0.6595$, $p = 0.4168$). This is notable because, thus far, all analyses have suggested significant differences between the conditions designed to elicit clear speech (i.e., barrier conditions) and the condition designed not to elicit clear speech (i.e., no-barrier condition). To further complicate the puzzle, the other two main effects *do* significantly contribute to model fit. The comparison of the L2 condition to the other two barrier conditions (babble and vocoded) significantly improved model fit ($\chi^2 = 59.877$, $p < 0.0001$). Further, the comparison between the babble and vocoded conditions also significantly improved model fit ($\chi^2 = 7.7695$, $p = 0.005$).

Examining **Figure 6** below, it becomes clear that the pattern of results is different from the patterns demonstrated for the other indices. While we continue to observe more common words (i.e., a greater range) for the L2 condition, it is not the case that the no-barrier condition follows the typical patterns observed above. Specifically, instead of the no-barrier condition being the lowest value, the vocoded condition is the lowest. It is not immediately clear why this would be the case; however, it is possible that because the vocoded condition is the least familiar to participants they may demonstrate less consistency across

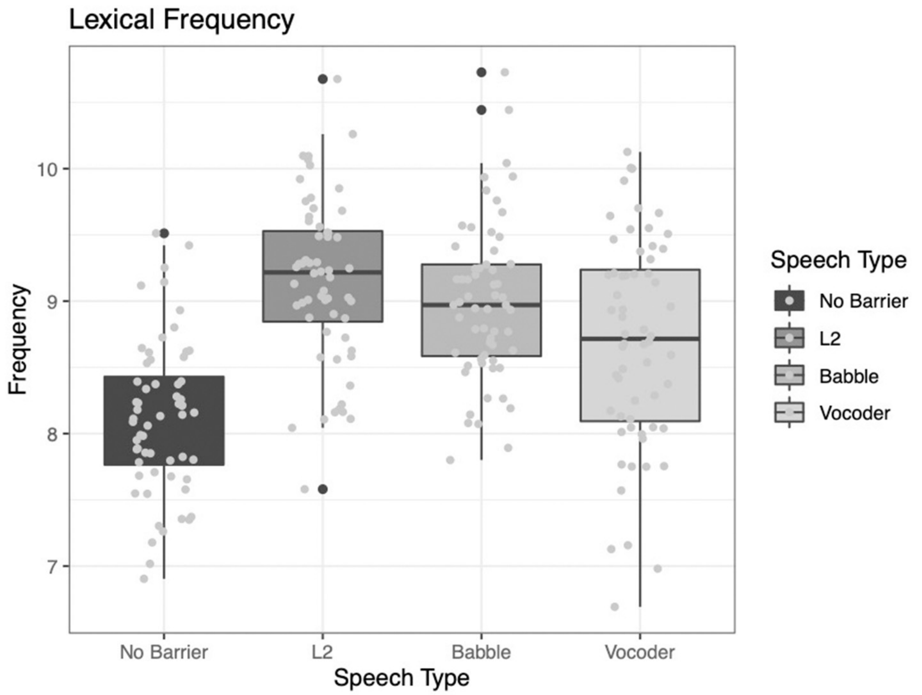


FIGURE 5 | Lexical frequency from the LUCID corpus across four conditions.

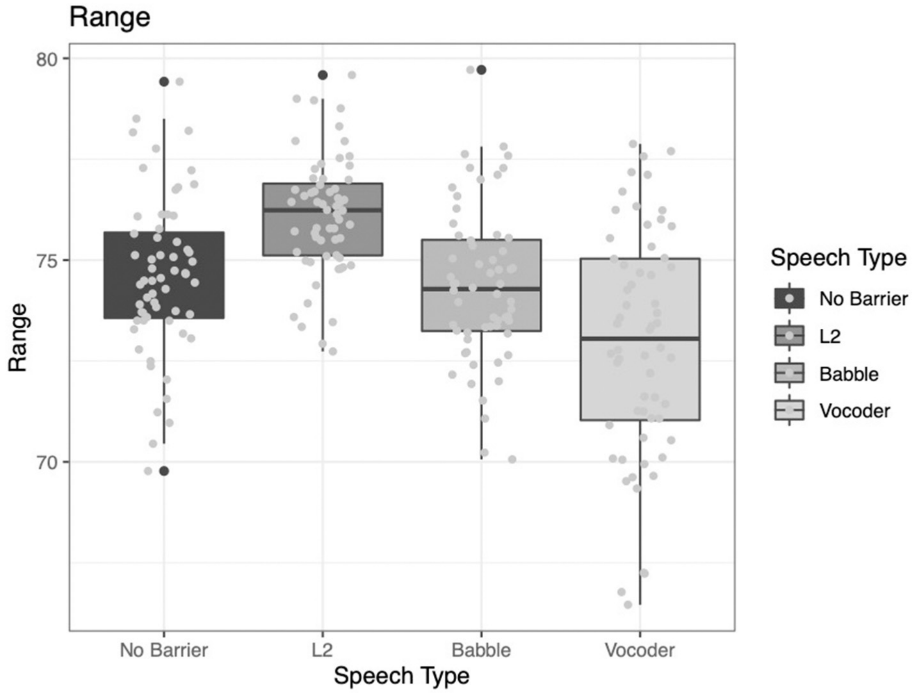


FIGURE 6 | Range of samples from the BNC in which a word from the LUCID corpus was found across four conditions.

indices, compared to the other conditions. That is, the other three conditions are cases that talkers are likely to have at least some familiarity with. Talking to someone in a noisy environment is a common occurrence at a restaurant or party. Speaking with a non-native speaker is also a relatively common occurrence for many talkers in our increasingly globalized society. However, speaking to someone who is perceiving your speech through a vocoder is relatively rare. Even if a person does have experience communicating with someone with a cochlear implant, it is unlikely they would have experience hearing that type of speech as well. Here, all participants are familiarized with how speech sounds when vocoded, which could impact how they modify their speech.

Bigram Frequency

For bigram frequency, we see that the comparison of the no-barrier condition to the other three conditions significantly improved model fit ($\chi^2 = 26.318$, $p < 0.0001$). However, the other two comparisons did not significantly improve model fit (L2 vs. other conditions: $\chi^2 = 0.5585$, $p = 0.4549$; babble vs. vocoded: $\chi^2 = 0.2157, p = 0.6423$).

These results, too, diverge from some of the previously reported results. Examining **Figure 7** below, we see that while the no-barrier condition shows the lowest bigram frequency, the other three conditions do not differ significantly from one another.

Trigram Frequency

The trigram frequency analysis reveals that none of the main effects significantly improve model fit (No-barrier vs. barrier conditions: $\chi^2 = 2.5851$, $p = 0.1079$; L2 vs. other conditions: $\chi^2 = 1.7431$, $p = 0.1867$; babble vs. vocoded: $\chi^2 = 1.9693$, $p = 0.1605$).

While numerically the results fit with our previous observations (see **Figure 8**), because no results are significant, it is difficult to interpret these findings. We are especially cautious not to overinterpret the null results we observe here.

Proportion of Bigrams Within the 25,000 Most Frequent Bigrams

Rather than looking at frequency of the two- and three-word collocations in our samples, here we examine the proportion of these collocations that are among the most frequent bigrams (and then trigrams) in the corpus.

As was the case for lexical frequency, two of the main effects significantly improved model fit for the analysis of lexical frequency. The comparison of the no-barrier condition to the other three conditions significantly improved model fit ($\chi^2 = 13.932$, $p = 0.0002$). The comparison of the L2 condition to the other two barrier conditions (babble and vocoded) also significantly improved model fit ($\chi^2 = 57.597$, $p < 0.0001$). However, the comparison between the babble and vocoded conditions did not significantly improve model fit ($\chi^2 = 2.9483$, $p = 0.086$).

Examining **Figure 9** below, it is clear that this index follows the pattern of many of the indices above. The no-barrier condition reports the lowest proportion of bigrams among the 25,000 most frequent and the L2 condition reports the highest proportion, with the other conditions lying intermediate between the two.

Proportion of Trigrams Within the 25,000 Most Frequent Trigrams

For this index, the only factor that emerged as significantly contributing to model fit was the comparison of the L2 condition to the other barrier conditions ($\chi^2 = 48.845$, $p < 0.0001$). The other two comparisons did not significantly improve model fit (no-barrier vs. barrier conditions: $\chi^2 = 1.1734$, $p = 0.2787$; babble vs. vocoded: $\chi^2 = 1.2454$, $p = 0.2644$).

Examining **Figure 10**, it is clear that the L2 condition results in the highest proportion of trigrams among the most frequent in the corpus; however, the other conditions show less clear patterns.

Order Effects

As described above, it is possible that condition order, which is conflated with condition itself, may impact the lexical sophistication results. However, as in the case of lexical diversity described in Section Order Effects above, we believe that predicted order effects would be the opposite of the condition effects we see in the present data (i.e., the L2 and babble conditions should have higher measures of lexical sophistication if the task is easier as talkers adapt to the task, topics, and their partner).

Further, examining order of pictures, we primarily see no significant impact on picture order for the metrics described above. There is one exception, however. Picture order is a significant predictor of the proportion of bigrams within the 25,000 most frequent bigrams ($\chi^2 = 63.218$, $p < 0.0001$). Picture order does not interact with any other factors in the model (i.e., condition). Examining the data, it appears that talkers use a higher proportion of bigrams among the most frequent bigrams in the first picture of each condition and use a smaller proportion in later pictures. We caution over-interpretation of this particular finding as it is not consistent with the null results for the other metrics. However, it is possible that as listeners adapt to the task they do use slightly less frequent collocations as the task progresses. Some acoustic analyses of this data (Lee and Baese-Berk, 2020) have demonstrated that some acoustic properties of the signal (e.g., vowel duration) also decrease across pictures, suggesting that perhaps some aspects of speech do differ as the speaker adapts to speech within a condition[4].

Interim Summary

Overall, the results of lexical sophistication demonstrate similar results to the lexical diversity results above. On average, the no-barrier condition is different from the barrier conditions across many indices, indicating that conditions designed to elicit clear speech not only elicit different numbers of unique words but also different kinds of words. Further,

[4] Note, however, that many other aspects of the acoustic signal (e.g., $F0$, speaking rate) and intelligibility do NOT differ as a function of the order of the picture in the task (Lee and Baese-Berk, 2020). Therefore, we reiterate our caution in over-interpretation of this specific result.

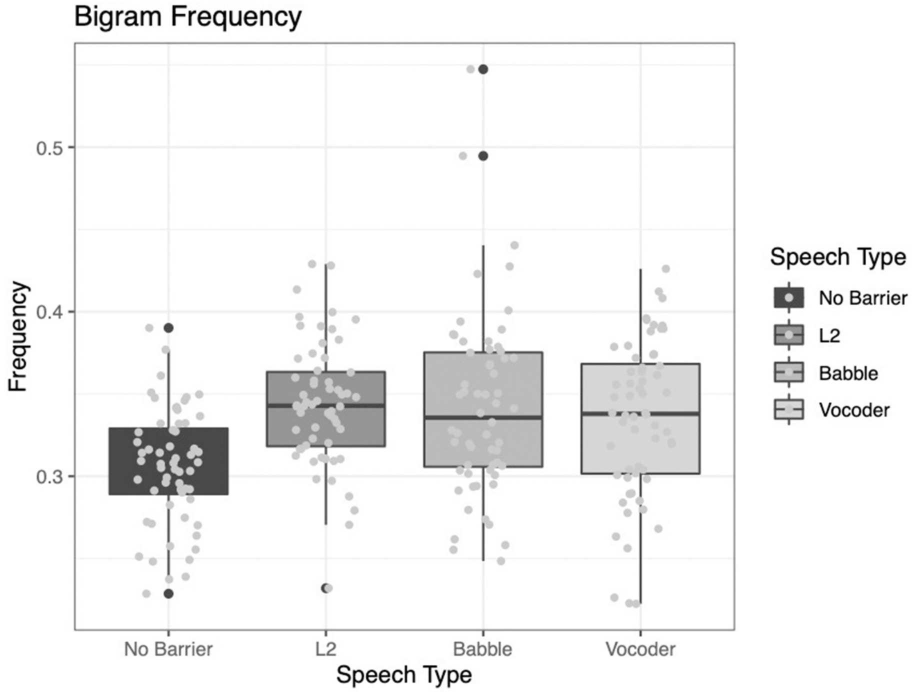


FIGURE 7 | Frequency of bigrams (i.e., pairs of words) from the LUCID corpus across four conditions.

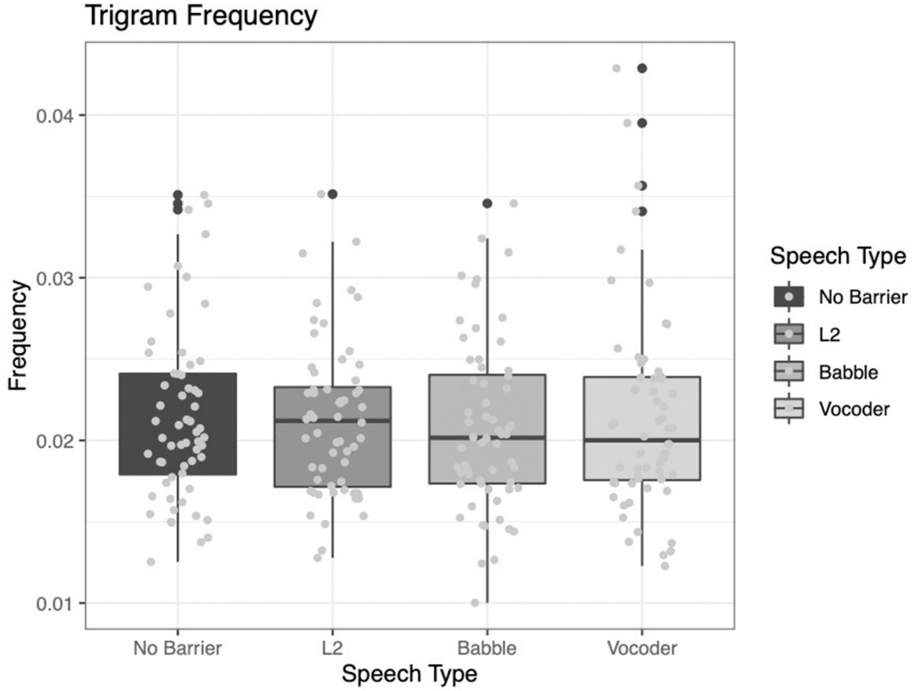


FIGURE 8 | Frequency of trigrams (i.e., sets of three of words) from the LUCID corpus across four conditions.

on many metrics the L2 condition differs from the other conditions designed to elicit clear speech. However, the vocoded and babble conditions demonstrate less clear patterns. Indeed, on some metrics they pattern more closely with the no-barrier condition than with the L2 condition, suggesting that different listeners may elicit different types of clear

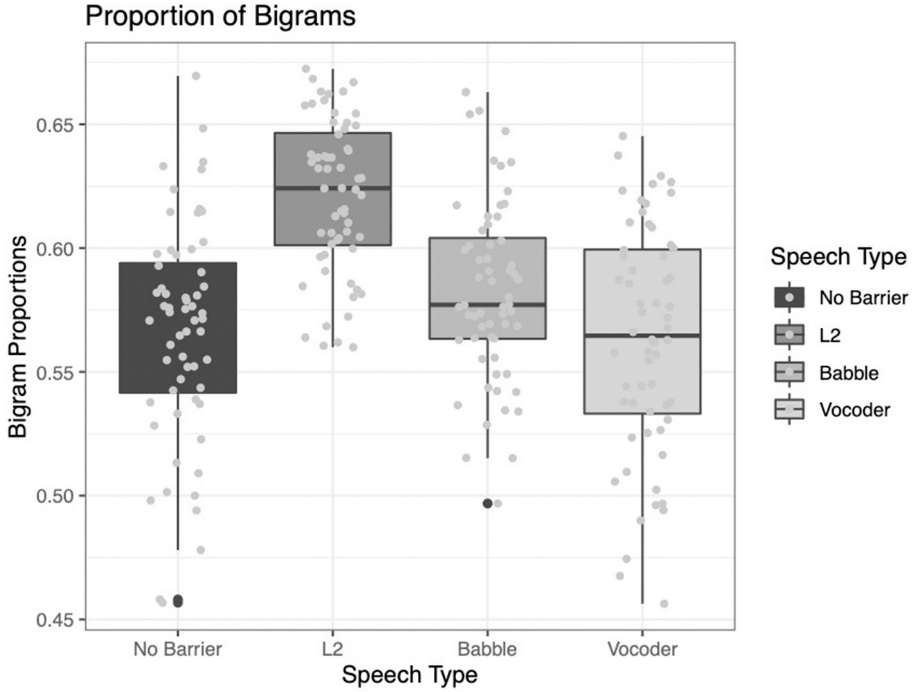


FIGURE 9 | Proportion of bigrams from the LUCID found among the 25,000 most common in the BNC.

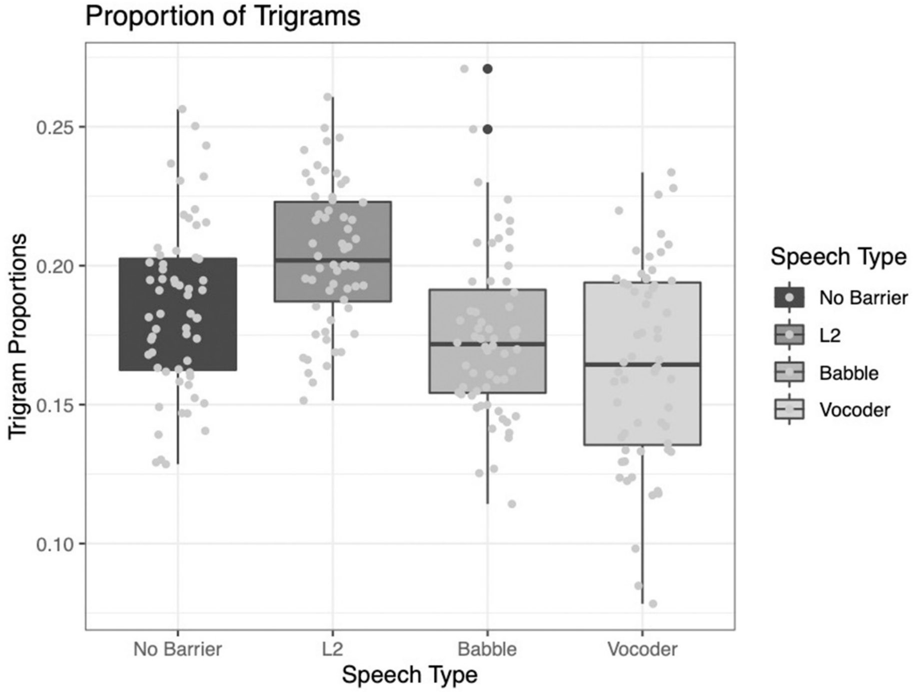


FIGURE 10 | Proportion of trigrams from the LUCID found among the 25,000 most common in the BNC.

speech. These results are particularly remarkable because the semantic content of the speech is relatively constrained by the pictures being described. That is, talkers do not have unlimited access to use any lexical items they would like. Instead, they are at least somewhat constrained by the task.

DISCUSSION

Overall, our findings suggest that talkers do, indeed, modulate the lexical diversity and lexical sophistication of their speech as a function of who they are talking to and in what conditions they are producing their speech. Below, we briefly discuss the implications of these findings for our understanding of clear speech, their implications for our understanding of speech processing and communication more broadly, and propose some future directions for investigation.

Implications for Understanding of Clear Speech

Previous studies of clear speech have largely treated the speaking style as a monolithic construct encompassing all types of scenarios in which a talker might want to produce clearer speech for a listener. Indeed, in studies that have elicited clear speech in the laboratory for a hypothetical listener, the listener is often given a number of options for who they should be envisioning as the recipient of their speech. For example, a common instruction is to "speak as though you are talking to someone who has difficulty hearing or is a non-native speaker of a language" (Picheny et al., 1985; Biersack et al., 2005; Maniwa et al., 2009), which conflates two of the scenarios examined here.

At the same time, clear speech is often described explicitly as a "listener-oriented" speaking style. This is likely largely because the acoustic modifications seen in clear speech correlate with robust improvements in a variety of perceptual measures including objective number of words understood (intelligibility) and subjective difficulty understanding the speech (comprehensibility). However, if this speaking style is truly listener oriented, wouldn't one expect that at least some of the modifications ought to be tailored toward the specific listener one encounters?

Indeed, here we demonstrate that listeners do appear to not only modulate the lexical content of the speech they produce in clear speech conditions, but also modulate this content differently for different types of communication situations. This finding is consistent with previous research suggesting that speakers do alter their speech along different dimensions depending on the identity of the listener. For example, while talkers alter pitch similarly in speech to pets and infants, they only hyperarticulate vowels in IDS (Burnham et al., 2002; Xu et al., 2013). Indeed, other discussions of clear speech research (e.g., Smiljanic and Bradlow, 2011) have suggested the importance of understanding how clear speech might be modulated depending on the audience. While a large body of research has demonstrated that different populations benefit differently from aspects of clear speech (e.g., non-native listeners of differing proficiency levels benefit differently from clear speech), the specific interaction of how talkers specifically modulate their speech (and how listeners may or may not benefit from these modulations) remains understudied.

A skeptical reader may ask whether these results could be due to some factors we are not capturing by comparing across these conditions. However, we believe that the most obvious of these factors are indeed controlled in the current data. One concern, for example, might be that some talkers are more or less likely to modulate their speech for their listener. However, each talker in the corpus used for this analysis appeared in three of the four conditions.

A concern that might be more directly related to the issues of lexical diversity and sophistication investigated here is the influence of topic on these results. That is, if talkers are in truly natural conversations, they can choose the lexical content they produce with relative freedom. Some topics may be more or less likely to elicit more diverse or sophisticated lexical items. One feature that makes this corpus ideal for an analysis like ours is that the semantic content is relatively constrained. For example, one would be relatively surprised to hear a talker discussing nuclear physics when describing the beach scene. This feature, we believe, stacks the cards against us finding the results we did. That is, because the lexical content is relatively constrained, it is even more remarkable to see effects of lexical diversity and sophistication emerge.

We believe that these findings have two important implications for our understanding of clear speech. The first is that typical investigations of clear speech focus on acoustic properties of the speech or on perceptual consequences of clear speech for listeners. Our findings suggest that clear speech encompasses a set of speaking styles that differ from plain speech not only on acoustic dimensions but also on other dimensions, including lexical selection.

The second is that a more nuanced understanding of clear speech is necessary to fully understand the phenomenon (or set of phenomena). That is, while clear speech as an overarching style does, clearly, have some characteristics that are common, it does appear that this speech is listener-oriented in a more specific way. Talkers modify their speech for their listeners (as seen in the differentiation of L2 speech from the other two clear-speech eliciting conditions) and, in some situations, depending on the communication situation with a single listener (i.e., babble vs. vocoded speech). These results open new avenues for exploration, which we describe in more detail in section Future Directions and Open Questions below.

Audience Design, Speech Production, and Predictability

In some ways, these results are unsurprising. As discussed in the introduction of this paper, it has been clear for decades that talkers modulate their speech for their listener. Indeed, this modulation, often described as "audience design" (Clark and Murphy, 1982) can take many forms including modulating speaking style (Bell, 1997) and modulating referents to given or new items (Horton and Gerrig, 2002). However, a speaker's ability to modulate their speech for specific audiences is impacted by many factors, including memory demands (Horton and Gerrig, 2005). Further, it is not fully clear how audience design may impact lexical selection beyond modulating items within the common ground (Horton and Gerrig, 2002, 2005) or entraining on a shared term to refer to an object (Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996; Metzing and Brennan, 2003). That is, it is unclear how much speakers modulate lexical sophistication or lexical diversity as a function of their audience.

Indeed, tracking the frequency of lexical items used is, on its surface, rather complicated. Tracking the frequency and appropriately modulating the frequency of collocations of words appears to be even more complicated. While we do not suggest that speakers are consciously modulating the frequency of words or collocations that they use, it is important to note that speakers do have some metalinguistic awareness of lexical frequency (Carroll, 1971; Verhagen and Mos, 2016). That is, they are aware of what words are relatively higher and lower frequency, suggesting that modulating such factors in their speech may not be as complicated as it initially sounds.

One fundamental question is *why* speakers might modulate their speech in the ways we observe here. We have suggested throughout the paper that this modulation may result in speech that is easier to understand. But easier how, exactly? One way in which the speech may be easier to understand is that it may be more predictable for the listener. It is clear that semantic predictability within a sentence impacts perception. Low predictability sentences (e.g., *mom thinks that it is yellow*) are less well-understood than high-predictability sentences (e.g., *the color of a lemon is yellow*; Kalikow et al., 1977). Similarly, semantically anomalous sentences (e.g., *the black top ran the spring*) are harder to understand than semantically meaningful ones (Miller and Isard, 1963). On a lexical level, high frequency words are perceived more accurately than low frequency words (Carroll, 1971; Verhagen and Mos, 2016). Caregivers use more repetition and a more restricted vocabulary when talking to 6-month-old infants than to 3-month-old infants (Genovese et al., 2020), but a larger and more diverse vocabulary again as infants age and develop more adult-like linguistic abilities (Genovese et al., 2020; Tal et al., 2021). In addition, native talkers, when communicating with non-native talkers, have been found to avoid idiomatic expressions and use more high-frequency words (e.g., Rodriguez-Cuadrado et al., 2018). This suggests that, in both IDS and foreigner-directed speech, talkers make efforts to modulate their lexical choices to avoid confusion, and aid non-native or young listeners through a preference for common words, and phrases that are less semantically ambiguous. Therefore, it could be the case that decreased lexical sophistication results in speech that is slightly more predictable, and thus easier to understand. Another potential argument is supported by claims that talkers may, to an extent, imitate or match certain characteristics or features of infant-speech or foreigner speech when modifying their own speech to aid in communication (Ferguson, 1975). The decrease in lexical diversity and lexical sophistication could be an effort to match the diversity and sophistication of their communicative partner when considering the L2 condition.

Language users modulate their speech in discourse to disambiguate referents as much as possible, which also aids comprehension. Arnold (2008) suggests that modulations in how referents are expressed in discourse are functions of speakers making larger-scale decisions about the level of an addressee's knowledge based on shared social groups or other information that is available about the addressee, and smaller-scale adjustments throughout a conversation depending on the conversation's focus, topic, and whether the information being discussed is given or new. In environments where it is particularly difficult for interlocutors to understand each other, they may resort to different methods of referring to objects in the world than they would in environments where conversation is easier to understand. This would predict increased lexical diversity in the no-barrier condition compared to the other conditions, which is what we observed. These findings potentially support previous literature highlighting the adaptive and instructive nature of foreigner-directed speech, in that talkers seem to modulate their speech in a way that will help with comprehension, and also potentially with acquisition, despite their attitudes toward the speakers themselves (Uther et al., 2007). Thus, given its inherently didactic nature, the trend for lexical diversity to decrease when communicating with non-native talkers may be relatively salient across multiple L2 backgrounds. This trend occurs even though talkers incorporate social information, whether positive or negative, when making judgments about the addressee's prior knowledge.

It is quite clear that the decreased lexical diversity measures also result in more predictable speech. While we have not examined the productions directly, one interpretation of the decreased lexical diversity in the conditions designed to elicit clear speech is that there is an increase in repetition. Previous research on foreigner-directed speech supports this hypothesis by showing that native talkers do tend to employ more repetitions or reduplications in an attempt to help clarify their message (Ferguson, 1975; Rodriguez-Cuadrado et al., 2018). Thus, it is possible that this is what we are seeing through the low lexical diversity scores in the L2 condition. One interesting avenue for future exploration would be whether listeners signal a need for repetition, or whether the speakers choose to provide the repetition without an explicit prompt. It is also possible that clarifications take different forms across conditions. For example, repeating vs. rephrasing may be differently distributed across the conditions. Intuitively, one might expect the L2 condition would result in the most rephrasing, as listeners might be unfamiliar with particular lexical items. However, if our results are due to increased repetition, it appears that we may, in fact, predict the most repetition in those conditions, if we were to directly investigate the conversations in more detail.

Taken together, our results suggest that talkers have extraordinary ability to modify multiple aspects of their speech for their listener. This modulation may impact predictability of speech, making it easier to understand. However, the specific interactions between lexical diversity, sophistication, and predictability in the signal should be investigated in future studies.

Future Directions and Open Questions

Of course, this project leaves many open questions and avenues for future direction. For example, while we investigate lexical selection in the present study, we do not investigate syntactic or other high-level properties of the language produced by talkers in each condition. One might expect that speakers would demonstrate the most syntactic complexity in the no-barrier condition and the least syntactic complexity in speech to non-native listeners. Similarly, one could investigate "burstiness" (Altmann et al., 2009), or how locally frequent words are. That is, one might expect that in the clear speech conditions talkers may produce more bursty speech, which has more productions of similar words in a short period of time before shifting to a new topic with new lexemes presenting as bursty. In the present study, we only investigate a handful of metrics of the lexical selection by talkers. A number of other lexical properties (e.g., neighborhood density) could provide additional information about the lexical content produced in clear speech and how it might vary across listeners and communication scenarios.

Further, it is important to note that the results of the study are somewhat limited because condition and order of condition are conflated. We do not believe that condition order is the driving factor for our results. If condition order (rather than condition per se) were the source of differences, we would expect to see identical patterns for all metrics in the babble and L2 conditions, which is not what we observe[5]. Additional evidence that condition order alone is driving our results can be found in other work using these same stimuli (e.g., Baker and Hazan, 2011; Lee and Baese-Berk, 2020), which failed to find effects of reduction over the course of a task. That is, Baker and Hazan (2011) fail to find evidence of "learning" across conditions or pictures. Lee and Baese-Berk (2020) find that talkers "re-set" at the start of a new picture in terms of intelligibility of their speech. These findings are consistent with work in the area of second mention reduction which demonstrates that a variety of factors (e.g., topic changes, listener changes, and even narrative devices) can "block" such reductions (acoustic or lexical) from occurring (see, e.g., Fowler et al., 1997). Given these converging results, we do believe that condition, not order, is driving these results. However, future work should counterbalance conditions across orders to ensure that differences we observe are, indeed, driven by condition.

An additional area of inquiry is whether the findings demonstrated here hold throughout a conversation. In some previous work from our lab (Lee and Baese-Berk, 2020), we investigated these same conversations in terms of their acoustic properties and the perceptual consequences. We demonstrated that, in general, speakers produce more intelligible speech when communicating with non-native talkers than native talkers; however, they become less clear over the course of a single conversation. When the topic of conversation switches (i.e., talkers switch to a new picture with the same listener), they "reset" starting over with clearer speech. We interpreted these findings as evidence that what has been previously described as clear speech may have both listener- and speaker-oriented motivations. It is possible that similar patterns of becoming less clear occur with lexical items, though it is less clear whether the "reset" would occur for lexical items shown here.

CONCLUSION

In the present study, we investigate speech from naturalistic conversations designed to elicit a clear speaking style. Specifically, we investigate a series of indices of lexical diversity and lexical sophistication in this speech. We find that talkers modulate their speech in terms of both the lexical diversity (i.e., variability of lexical items) and lexical sophistication (i.e., typicality of lexical items). Specifically, talkers show the most lexical diversity and the most lexical sophistication in conversational situations that are designed to elicit plain speech. They demonstrate the least lexical diversity and least lexical sophistication in speech produced for a non-native listener. The results suggest that, in addition to the acoustic modifications previously demonstrated in clear speech work, talkers modulate their lexical selection as well. Further, the results demonstrate that clear speech is not a monolithic construct. Rather, it is a set of speaking styles in which talkers take the listener and communication situation into consideration.

DATA AVAILABILITY STATEMENT

The datasets generated for 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/dfhpu/?view_only=49d95d90424941da82217a239ab7450c.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Oregon, Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

We would like to thank members of the Speech Perception and Production Lab at the University of Oregon for their help in conducting this research. We would like to thank Valerie Hazan and colleagues at University College London for providing the speech files and transcriptions from the LUCID corpus that were used in this project, and Ann Bradlow and colleagues at Northwestern University for providing these recordings and transcriptions via SpeechBox.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.661415/full#supplementary-material

[5] It is important to note, however, that given our preregistered analyses, we do not report direct comparisons between the babble and L2 conditions. We believe this is appropriate given that a null result (the predicted result if order effects were significant) would, itself, be difficult to interpret, as it would not provide conclusive evidence *for* the null hypothesis, it would just fail to provide evidence to reject it.

REFERENCES

Altmann, E. G., Pierrehumbert, J. B., and Motter, A. E. (2009). Beyond word frequency: bursts, lulls, and scaling in the temporal distributions of words. *PLoS ONE* 4:e7678. doi: 10.1371/journal.pone.0007678

Altmann, G. T. M., and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* 73, 247–264. doi: 10.1016/S0010-0277(99)00059-1

Altmann, G. T. M., and Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: linking anticipatory (and other) eye movements to linguistic processing. *J. Mem. Lang.* 57, 502–518. doi: 10.1016/j.jml.2006.12.004

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., et al. (1991). The HCRC map task corpus. *Lang. Speech* 34, 351–366. doi: 10.1177/002383099103400404

Androutsopoulos, J. (2014). Languaging when contexts collapse: audience design in social networking. *Discourse Context Media* 4–5, 62–73. doi: 10.1016/j.dcm.2014.08.006

Arnold, J. E. (2008). Reference production: production-internal and addressee-oriented processes. *Lang. Cognit. Process.* 23, 495–527. doi: 10.1080/01690960801920099

Arnold, J. E., Kahn, J. M., and Pancani, G. C. (2012). Audience design affects acoustic reduction via production facilitation. *Psychon. Bull. Rev.* 19, 505–512. doi: 10.3758/s13423-012-0233-y

Baese-Berk, M. M., Bent, T., and Walker, K. (2021). Semantic predictability and adaptation to nonnative speech. *JASA Express Lett.* 1:015207. doi: 10.1121/10.0003326

Baker, R., and Hazan, V. (2011). DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behav. Res. Methods* 43, 761–770. doi: 10.3758/s13428-011-0075-y

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. (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

Beckford Wassink, A., Wright, R. A., and Franklin, A. D. (2007). Intraspeaker variability in vowel production: an investigation of motherese, hyperspeech, and Lombard speech in Jamaican speakers. *J. Phonet.* 35, 363–379. doi: 10.1016/j.wocn.2006.07.002

Bell, A. (1997). "Language style as audience design," in *Sociolinguistics,* eds N. Coupland and A. Jaworski (London: Palgrave), 240–250. doi: 10.1007/978-1-349-25582-5_20

Biersack, S., Kempe, V., and Knapton, L. (2005). *Fine-Tuning Speech Registers: A Comparison of the Prosodic Features of Child-Directed and Foreigner-Directed Speech.* Isca-Speech.Org. Available online at: http://www.isca-speech.org/archive/archive_papers/interspeech_2005/i05_2401.pdf (accessed January 15, 2021).

BNC Consortium (2007). *British National Corpus [Corpus]*. Available online at: www.natcorp.ox.ac.uk (accessed January 15, 2021).

Boersma, P., and Weenink, D. (2021). *Praat: Doing Phonetics by Computer*. Available online at: https://www.fon.hum.uva.nl/praat/ (accessed January 15, 2021).

Bradlow, A. R., and Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *J. Acoust. Soc. Am.* 121:2339–2349. doi: 10.1121/1.2642103

Bradlow, A. R., and Bent, T. (2002). The clear speech effect for non-native listeners. *J. Acoust. Soc. Am.* 112, 272–284. doi: 10.1121/1.1487837

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

Burnham, D., Kitamura, C., and Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science* 296:1435. doi: 10.1126/science.1069587

Calandruccio, L., Porter, H. L., Leibold, L. J., and Buss, E. (2020). The clear-speech benefit for school-age children: speech-in-noise and speech-in-speech recognition. *J. Speech Lang. Hear. Res.* 63, 4265–4276. doi: 10.1044/2020_JSLHR-20-00353

Carroll, J. B. (1971). Measurement properties of subjective magnitude estimates of word frequency. *J. Verb. Learn. Verb. Behav.* 10, 722–729. doi: 10.1016/S0022-5371(71)80081-6

Clark, H. H., and Carlson, T. B. (1981). "Context for comprehension," in *Attention and Performance IX*, eds A. D. Baddeley and https://www.worldcat.org/search?q=au%3ALong%2C$+$John.&qt=hotauthorJ.Long (Hillsdale, NJ: Lawrence Erlbaum Associate) 313–330.

Clark, H. H., and Murphy, G. L. (1982). "Audience design in meaning and reference," in *Advances in Psychology,* Vol. 9, eds J.-F. Le Ny and W. Kintsch (Amsterdam: North-Holland Publishing Company), 287–299. doi: 10.1016/S0166-4115(09)60059-5

Clark, H. H., Schreuder, R., and Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. *J. Verb. Learn. Verb. Behav.* 22, 245–258. doi: 10.1016/S0022-5371(83)90189-5

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

Covington, M. A., and McFall, J. D. (2010). Cutting the Gordian knot: the moving-average type–token ratio (MATTR). *J. Quan. Linguist.* 17, 94–100. doi: 10.1080/09296171003643098

Crossley, S. A., Salsbury, T., and McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Lang. Test.* 29, 243–263. doi: 10.1177/0265532211419331

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., and Jamse, M. (2005). Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL®. *ETS Res. Rep. Ser.* 2005, 1–77. doi: 10.1002/j.2333-8504.2005.tb01990.x

Cunningham, K. T., and Haley, K. L. (2020). Measuring lexical diversity for discourse analysis in aphasia: moving-average type–token ratio and word information measure. *J. Speech Lang. Hear. Res.* 63, 710–721. doi: 10.1044/2019_JSLHR-19-00226

Dunn, J., and Kendrick, C. (1982). The speech of two- and three-year-olds to infant siblings: 'Baby talk' and the context of communication. *J. Child Lang.* 9, 579–595. doi: 10.1017/S030500090000492X

Eguchi, M., and Kyle, K. (2020). Continuing to explore the multidimensional nature of lexical sophistication: the case of oral proficiency interviews. *Mod. Lang. J.* 104, 381–400. doi: 10.1111/modl.12637

Ehrlich, S. F., and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *J. Verb. Learn. Verb. Behav.* 20, 641–655. doi: 10.1016/S0022-5371(81)90220-6

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *J. Second Lang. Writ.* 4, 139–155. doi: 10.1016/1060-3743(95)90004-7

Ferguson, C. A. (1975). Toward a characterization of English foreigner talk. *Anthropol. Linguist.* 17, 1–14.

Ferguson, S. H., and Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.* 112, 259–271. doi: 10.1121/1.1482078

Fernald, A., and Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Dev. Psychol.* 20, 104–113. doi: 10.1037/0012-1649.20.1.104

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., and Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *J. Child Lang.* 16, 477–501. doi: 10.1017/S0305000900010679

Ferreira, V. S. (2019). A mechanistic framework for explaining audience design in language production. *Annu. Rev. Psychol.* 70, 29–51. doi: 10.1146/annurev-psych-122216-011653

Fowler, C. A., Levy, E. T., and Brown, J. M. (1997). Reductions of spoken words in certain discourse contexts. *J. Mem. Lang.* 37, 24–40. doi: 10.1006/jmla.1996.2504

Garrard, P., Maloney, L. M., Hodges, J. R., and Patterson, K. (2005). The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain* 128, 250–260. doi: 10.1093/brain/awh341

Genovese, G., Spinelli, M., Romero Lauro, L. J., Aureli, T., Castelletti, G., and Fasolo, M. (2020). Infant-directed speech as a simplified but not simple register: a longitudinal study of lexical and syntactic features. *J. Child Lang.* 47, 22–44. doi: 10.1017/S0305000919000643

Grieser, D. L., and Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: support for universal prosodic features in motherese. *Dev. Psychol.* 24, 14–20. doi: 10.1037/0012-1649.24.1.14

Hazan, V., and Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *J. Acoust. Soc. Am.* 130, 2139–2152. doi: 10.1121/1.3623753

Hazan, V., Grynpas, J., and Baker, R. (2012). Is clear speech tailored to counter the effect of specific adverse listening conditions? *J. Acoust. Soc. Am.* 132, EL371–EL377. doi: 10.1121/1.4757698

Hazan, V., Tuomainen, O., Kim, J., Davis, C., Sheffield, B., and Brungart, D. (2018). Clear speech adaptations in spontaneous speech produced by young and older adults. *J. Acoust. Soc. Am.* 144, 1331–1346. doi: 10.1121/1.5053218

Hazan, V., Uther, M., and Granlund, S. (2015). "How does foreigner-directed speech differ from other forms of listener-directed clear speaking styles?," *Proceedings of the 18th International Congress of Phonetic Sciences* (Glasgow).

Holmes, E., Folkeard, P., Johnsrude, I. S., and Scollie, S. (2018). Semantic context improves speech intelligibility and reduces listening effort for listeners with hearing impairment. *Int. J. Audiol.* 57, 483–492. doi: 10.1080/14992027.2018.1432901

Horton, W. S., and Gerrig, R. J. (2002). Speakers' experiences and audience design: knowing when and knowing how to adjust utterances to addressees. *J. Mem. Lang.* 47, 589–606. doi: 10.1016/S0749-596X(02)00019-0

Horton, W. S., and Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition* 96, 127–142. doi: 10.1016/j.cognition.2004.07.001

Ingram, D. (1995). The cultural basis of prosodic modifications to infants and children: a response to Fernald's universalist theory. *J. Child Lang.* 22, 223–233. doi: 10.1017/S0305000900009715

Johnson, W. (1944). Studies in language behavior: a program of research. *Psychol. Monogr.* 56, 1–15. doi: 10.1037/h0093508

Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J. Acoust. Soc. Am.* 61, 1337–1351. doi: 10.1121/1.381436

Krause, J. C., and Braida, L. D. (2002). Investigating alternative forms of clear speech: the effects of speaking rate and speaking mode on intelligibility. *J. Acoust. Soc. Am.* 112, 2165–2172. doi: 10.1121/1.1509432

Krause, J. C., and Braida, L. D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. *J. Acoust. Soc. Am.* 115, 362–378. doi: 10.1121/1.1635842

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., et al. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science* 277, 684–686. doi: 10.1126/science.277.5326.684

Kutas, M., DeLong, K. A., and Smith, N. J. (2011). "A look around at what lies ahead: prediction and predictability in language processing," in *Predictions in the brain: Using our past to generate a future,* ed M. Bar (Oxford: Oxford University Press), 190–207. doi: 10.1093/acprof:oso/9780195395518.003.0065

Kutas, M., and Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* 207, 203–205. doi: 10.1126/science.7350657

Kyle, K., Crossley, S., and Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behav. Res. Methods* 50, 1030–1046. doi: 10.3758/s13428-017-0924-4

Kyle, K., and Crossley, S. A. (2015). Automatically assessing lexical sophistication: indices, tools, findings, and application. *TESOL Q.* 49, 757–786. doi: 10.1002/tesq.194

Kyle, K., Crossley, S. A., and Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Lang. Assess. Q.* 18, 154–170. doi: 10.1080/15434303.2020.1844205

Laufer, B., and Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written production. *Appl. Linguist.* 16, 307–322. doi: 10.1093/applin/16.3.307

Lee, D.-Y., and Baese-Berk, M. M. (2020). The maintenance of clear speech in naturalistic conversations. *J. Acoust. Soc. Am.* 147, 3702–3711. doi: 10.1121/10.0001315

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

Liu, S., Del Rio, E., Bradlow, A. R., and Zeng, F.-G. (2004). Clear speech perception in acoustic and electric hearing. *J. Acoust. Soc. Am.* 116, 2374–2383. doi: 10.1121/1.1787528

Liu, S., and Zeng, F.-G. (2006). Temporal properties in clear speech perception. *J. Acoust. Soc. Am.* 120, 424–432. doi: 10.1121/1.2208427

Maniwa, K., Jongman, A., and Wade, T. (2008). Perception of clear fricatives by normal-hearing and simulated hearing-impaired listeners. *J. Acoust. Soc. Am.* 123, 1114–1125. doi: 10.1121/1.2821966

Maniwa, K., Jongman, A., and Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *J. Acoust. Soc. Am.* 125, 3962–3973. doi: 10.1121/1.2990715

McCarthy, P. M. (2005). An *Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD)*. Ph.D. Thesis, The University of Memphis.

McCarthy, P. M., and Jarvis, S. (2007). vocd: a theoretical and empirical evaluation. *Lang. Test.* 24, 459–488. doi: 10.1177/0265532207080767

McCarthy, P. M., and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. *Behav. Res. Methods* 42, 381–392. doi: 10.3758/BRM.42.2.381

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., and Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Asses. Writ.* 23, 35–59. doi: 10.1016/j.asw.2014.09.002

Metzing, C., and Brennan, S. E. (2003). When conceptual pacts are broken: partner-specific effects on the comprehension of referring expressions. *J. Mem. Lang.* 49, 201–213. doi: 10.1016/S0749-596X(03)00028-7

Miller, G. A., and Isard, S. (1963). Some perceptual consequences of linguistic rules. *J. Verb. Learn. Verb. Behav.* 2, 217–228. doi: 10.1016/S0022-5371(63)80087-0

Moon, S., and Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *J. Acoust. Soc. Am.* 96, 40–55. doi: 10.1121/1.410492

Payton, K. L., Uchanski, R. M., and Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *J. Acoust. Soc. Am.* 95, 1581–1592. doi: 10.1121/1.408545

Picheny, M. A., Durlach, N. I., and Braida, L. D. (1985). Speaking clearly for the hard of hearing I: intelligibility differences between clear and conversational speech. *J. Speech Hear. Res.* 28, 96–103. doi: 10.1044/jshr.2801.96

Picheny, M. A., Durlach, N. I., and Braida, L. D. (1986). Speaking clearly for the hard of hearing II. *J. Speech Lang. Hear. Res.* 29, 434–446. doi: 10.1044/jshr.2904.434

Pye, C. (1986). Quiché Mayan speech to children. *J. Child Lang.* 13, 85–100. doi: 10.1017/S0305000900000313

Rodriguez-Cuadrado, S., Baus, C., and Costa, A. (2018). Foreigner talk through word reduction in native/non-native spoken interactions. *Bilingual. Lang. Cogn.* 21, 419–426. doi: 10.1017/S1366728917000402

Scarborough, R., Dmitrieva, O., Hall-Lew, L., Zhao, Y., and Brenier, J. (2007). An acoustic study of real and imagined foreigner-directed speech. *J. Acoust. Soc. Am.* 121, 3044–3044. doi: 10.1121/1.4781735

Scarborough, R., and Zellou, G. (2013). Clarity in communication: "Clear" speech authenticity and lexical neighborhood density effects in speech production and perception. *J. Acoust. Soc. Am.* 134, 3793–3807. doi: 10.1121/1.4824120

Schad, D. J., Vasishth, S., Hohenstein, S., and Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: a tutorial. *J. Mem. Lang.* 110:104038. doi: 10.1016/j.jml.2019.104038

Schegloff, E., Jefferson, G., and Sacks, H. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696–735. doi: 10.1353/lan.1974.0010

Signoret, C., Johnsrude, I., Classon, E., and Rudner, M. (2018). Combined effects of form- and meaning-based predictability on perceived clarity of speech. *J. Exp. Psychol. Hum. Percept. Perform.* 44, 277–285. doi: 10.1037/xhp0000442

Smiljanic, R., and Bradlow, A. R. (2011). Bidirectional clear speech perception benefit for native and high-proficiency non-native talkers and listeners: intelligibility and accentedness. *J. Acoust. Soc. Am,* 130, 4020–4031. doi: 10.1121/1.3652882

Smith, C. (2007). "Prosodic accommodation by French speakers to a non-native interlocutor," in *Proceedings of the XVIth International Congress of Phonetic Sciences* (Saarbücken), 313–348.

Snow, C. E. (1977). The development of conversation between mothers and babies. *J. Child Lang.* 4, 1–22. doi: 10.1017/S0305000900000453

Stern, D. N., Spieker, S., and MacKain, K. (1982). Intonation contours as signals in maternal speech to prelinguistic infants. *Dev. Psychol.* 18, 727–735. doi: 10.1037/0012-1649.18.5.727

Tal, S., Grossman, E., and Arnon, I. (2021). Infant-directed speech becomes less redundant as infants grow: implications for language learning. *PsyArXiv*. doi: 10.31234/osf.io/bgtzd

Templin, M. C. (1957). *Certain Language Skills in Children; Their Development and Interrelationships*. Minneapolis, MN: University of Minnesota Press. doi: 10.5749/j.ctttv2st

Uther, M., Knoll, M. A., and Burnham, D. (2007). Do you speak ENGLISH? Similarities and differences in speech to foreigners and infants. *Speech Commun.* 49, 1–7. doi: 10.1016/j.specom.2006.10.003

Van Engen, K. J., Baese-Berk, M. M., Baker, R. E., Choi, A., Kim, M., and Bradlow, A. R. (2010). The wildcat corpus of native-and foreign-accented english: communicative efficiency across conversational dyads with varying language alignment profiles. *Lang. Speech* 53, 510–540. doi: 10.1177/0023830910372495

van Velzen, M., and Garrard, P. (2008). From hindsight to insight – retrospective analysis of language written by a renowned Alzheimer's patient. *Interdiscipl. Sci. Rev.* 33, 278–286. doi: 10.1179/174327908X392852

Verhagen, V., and Mos, M. (2016). Stability of familiarity judgments: individual variation and the invariant bigger picture. *Cognit. Linguist.* 27, 307–344. doi: 10.1515/cog-2015-0063

Warren-Leubecker, A., and Bohannon, J. N. (1983). The effects of verbal feedback and listener type on the speech of preschool children. *J. Exp. Child Psychol.* 35, 540–548. doi: 10.1016/0022-0965(83)90026-7

Weppelman, T. L., Bostow, A., Schiffer, R., Elbert-Perez, E., and Newman, R. S. (2003). Children's use of the prosodic characteristics of infant-directed speech. *Lang. Commun.* 23, 63–80. doi: 10.1016/S0271-5309(01)00023-4

Winn, M. B. (2016). Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. *Trends Hear.* 20:2331216516669723. doi: 10.1177/2331216516669723

Wright, J. M., and Baese-Berk, M. M. (2020). "The impact of pause types on adverse listening condition classification with convolutional neural networks and naive Bayes," in *Annual Meeting of the Psychonomic Society, Virtual Meeting* (Chicago, IL).

Xu, N., Burnham, D., Kitamura, C., and Vollmer-Conna, U. (2013). Vowel hyperarticulation in parrot-, dog- and infant-directed speech. *Anthrozoos* 26, 373–380. doi: 10.2752/175303713X13697429463592

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Appl. Linguist.* 31, 236–259. doi: 10.1093/applin/amp024

Zareva, A., Schwanenflugel, P., and Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: variable sensitivity. *Stud. Second Lang. Acquisit.* 27, 567–595. doi: 10.1017/S0272263105050254

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 Communication

ORIGINAL RESEARCH
published: 21 June 2021
doi: 10.3389/fcomm.2021.643528

Dynamic Formant Trajectories in German Read Speech: Impact of Predictability and Prominence

Erika Brandt[1]*, *Bernd Möbius*[2] *and Bistra Andreeva*[2]

[1]*Leibniz-Centre General Linguistics, Berlin, Germany,* [2]*Language Science and Technology, Saarland University, Saarbrücken, Germany*

Phonetic structures expand temporally and spectrally when they are difficult to predict from their context. To some extent, effects of predictability are modulated by prosodic structure. So far, studies on the impact of contextual predictability and prosody on phonetic structures have neglected the dynamic nature of the speech signal. This study investigates the impact of predictability and prominence on the dynamic structure of the first and second formants of German vowels. We expect to find differences in the formant movements between vowels standing in different predictability contexts and a modulation of this effect by prominence. First and second formant values are extracted from a large German corpus. Formant trajectories of peripheral vowels are modeled using generalized additive mixed models, which estimate nonlinear regressions between a dependent variable and predictors. Contextual predictability is measured as biphone and triphone surprisal based on a statistical German language model. We test for the effects of the information-theoretic measures surprisal and word frequency, as well as prominence, on formant movement, while controlling for vowel phonemes and duration. Primary lexical stress and vowel phonemes are significant predictors of first and second formant trajectory shape. We replicate previous findings that vowels are more dispersed in stressed syllables than in unstressed syllables. The interaction of stress and surprisal explains formant movement: unstressed vowels show more variability in their formant trajectory shape at different surprisal levels than stressed vowels. This work shows that effects of contextual predictability on fine phonetic detail can be observed not only in pointwise measures but also in dynamic features of phonetic segments.

Keywords: information theory, surprisal, predictability, formant trajectories, German, read speech, prominence

OPEN ACCESS

Edited by:
Rory Turnbull,
Newcastle University, United Kingdom

Reviewed by:
Fabian Tomaschek,
University of Tübingen, Germany
Joey Stanley,
Brigham Young University,
United States

***Correspondence:**
Erika Brandt
brandt@leibniz-zas.de

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

Received: *18 December 2020*
Accepted: *31 May 2021*
Published: *21 June 2021*

Citation:
Brandt E, Möbius B and Andreeva B (2021) Dynamic Formant Trajectories in German Read Speech: Impact of Predictability and Prominence. Front. Commun. 6:643528. doi: 10.3389/fcomm.2021.643528

1 INTRODUCTION

Probabilistic reduction of predictable words and subword units has been observed in many languages (e.g., Gahl, 2008; Bell et al., 2009; Bürki et al., 2011; Kuperman et al., 2007; Pellegrino et al., 2011; Pluymaekers et al., 2005a, b). Specifically, vowels are more reduced in their spectral distinctiveness when they are difficult to predict from their context compared to easily predictable vowels (Jurafsky et al., 2001; Wright, 2004; Aylett and Turk, 2006; Clopper and Pierrehumbert, 2008; Scarborough, 2010). This effect of contextual predictability (henceforth, for brevity—predictability) on segmental properties prevails even after controlling for known prosodic effects on phonetic structures, such as lexical stress (Brandt, 2019). For instance, stressed vowels that are difficult to predict tend to be more dispersed, that is, distant from the center of the vowel space, than unstressed vowels that are easily predictable, beyond the extent to which the dispersion would be predicted by stress alone (Brandt et al., 2019). Conversely, the degree of dispersion will be attenuated for stressed vowels in high-predictability contexts and enlarged for unstressed vowels that are hard to predict. Predictability thus affects form encoding. The smooth signal redundancy (SSR) hypothesis (Aylett and Turk, 2004, 2006) proposes that the impact of the predictability of linguistic events on the phonetic encoding of these events is mediated by the prosodic structure, in particular by lexical stress. An alternative interpretation is that the assignment of the prosodic structure is conditioned by predictability (Tang and Shaw, 2020). Both perspectives entail that predictability is tightly interwoven with the prosodic structure.

Aylett and Turk (2006) investigated the effects of predictability and stress on the first and second formants of American English vowels and observed a large amount of shared contribution of predictability and stress to explaining the formant patterns, generally supporting the SSR hypothesis. Crucially, they also found an unexpected unique contribution of predictability in their statistical models. On average, however, prominence is found to be more effective in explaining variability in F1/F2 patterns than predictability. Malisz et al. (2018) analyzed the sensitivity of different prosodic characteristics to predictability and prominence in six languages: American English, Czech, Finnish, French, German, and Polish. They observed a positive interaction effect of these two factors on the segmental duration and the consonantal center of gravity (COG): stressed segments in low-predictability contexts are longer and show higher mean COG than unstressed segments in high-predictability contexts. There was no significant interaction effect between predictability and prominence on vowel dispersion.

Taken together, there is evidence that the mediation of the effects of predictability on the segmental structure by the prosodic structure is not comprehensive and that predictability effects are not entirely consumed by prosodic prominence (Malisz et al., 2018).

However, research so far has neglected the impact of information-theoretic factors on the dynamic characteristics of vowels. The present study therefore focuses on the effect of predictability on formant dynamics using generalized additive mixed models (GAMMs) while controlling for known effects of prosodic prominence on vocalic characteristics. Most literature on predictability effects on segmental properties of speech has focused on (American) English. It is important to replicate results for other languages because of the implications they may have for explaining the production and perception of the phonetic structure. This work investigates dynamic formant trajectory patterns in German vowels in different predictability contexts.

1.1 Dynamic Structure of German Vowels

The German vowel inventory consists of a rather large number of monophthongs with seven tense/lax vowel phoneme pairs [/i–I, y–Y, e(ε)–ε, ø–œ, a–a, o–ɔ, and u–Ʊ/] (Pätzold and Simpson, 1997). In contrast to American or Canadian English, German does not use diphthongization, that is, significant formant change over time within vowels considered as monophthongs (Nearey and Assmann, 1986), to distinguish between tense and lax monophthongs (Strange et al., 2004).

There is, however, still considerable formant movement in German monophthongs, with distinct patterns for tense and lax pairings (Strange and Bohn, 1998). Most of the variance in dynamic formant changes in German monophthongs reflects formant movement toward the place of articulation of neighboring consonants. This coarticulatory effect is observed throughout the entire duration of the vowel and therefore is not restricted to the beginning or end of the vowel (Möbius, 1999). Lax vowels are more strongly influenced by context than tense vowels. Alveolar contexts induce stronger coarticulatory behavior in German vowels than labial contexts. Also, low and back vowels show more contextual variation than front vowels (Strange et al., 2007).

Although formant movement in German monophthongs, and especially in tense vowels, may be more subtle than that in English varieties, native German listeners show the same performance in vowel identification when listening to vocalic nuclei from CVC sequences as they do when hearing silent center syllables with only the onset and offset of the vowel being presented. Additional information about intrinsic vowel length reduces the error rate in identification and discrimination tasks (Strange and Bohn, 1998; Bohn and Polka, 2001). This indicates that German listeners rely on information about formant movement similarly to English natives, who use diphthongization as a cue to differentiate tense and lax monophthongs.

Vowel phonemes may show more or less variability and movement in their formants depending on the denseness of the vowel space in their direct vicinity (Wedel et al., 2018). This idea of competition between neighboring vowel phonemes has the following implications for German. Here, the front, close to mid-close vowel space is rather dense with a high number of vowel phonemes, while the open, mid vowels and the close, back vowels have considerably less competition from neighboring vowel phonemes (Möbius, 2001).

1.2 Information-Theoretic Measures

Information-theoretic measures (Shannon, 1948), such as frequency or predictability, have been linked to the realization of linguistic structures (for review, see Hale, 2016; Jaeger and Buz, 2017). In this context, surprisal S (unit_i), which estimates the predictability of local structures, has been shown to correlate with human processing difficulty pertaining to linguistic units at different levels (Demberg et al., 2012; Levy, 2008; Hale, 2001; Levy, 2011). Surprisal is measured in bits of information and calculated as the negative log to the base two of the probability (P) of a linguistic unit (unit_i) appearing in a specific context (*context*), which can be the preceding or following context of that unit or both (**Eq. 1**).

$$S(unit_i) = -log_2 P(unit_i|context). \tag{1}$$

The surprisal measure reflects the intuition that linguistic units that are difficult to predict from context are more surprising when they occur, and conversely, the occurrence of easily predictable units is less surprising. Surprisal quantifies the predictability of local structures and is usually estimated from language models (LMs) based on large text corpora. In this study, we measure predictability as surprisal based on phoneme-level LMs because we investigate phonetic structures whose variability is thought to be best reflected by predictability estimated at the phoneme level (Oh et al., 2015). Hierarchical structural information, such as syllable or word boundaries, which also affect segmental properties, is implicitly reflected in sequences of phones (Raymond et al., 2006).

When investigating the impact of information-theoretic measures on linguistic structures, it is important to distinguish predictability from pure frequency effects, although frequency and predictability are not independent measures (Cohen Priva and Jaeger, 2018). Frequently used linguistic elements are under greater pressure to be efficient than less frequent ones (Zipf, 1949). More recent crosslinguistic studies have found that it is not frequency of occurrence but contextual predictability that is more efficient in explaining variability in word length, especially for lower-frequency words (Dautriche et al., 2017; Piantadosi et al., 2011). This line of research suggests that the effect of frequency is subordinate to that of predictability.

In studies on predictability effects on phonetic structures, word frequency is usually included as a control variable to tease apart effects of the two information-theoretic measures, viz. predictability and word frequency, on linguistic variability (e.g., Bell et al., 2009; Gahl et al., 2012; Jurafsky et al., 2001). On average, low-frequency words include vowels with increased dispersion, or distance from the center of the vowel space, compared to high-frequency words (Jurafsky et al., 2001; Zhao and Jurafsky, 2009). Vowels in frequent syllables have been shown to have faster formant transitions, that is, to show stronger coarticulatory influences, than vowels in infrequent syllables (Benner et al., 2007). This frequency effect has been found to be consistent in different lexical stress conditions. In accordance with the current literature, we therefore include word frequency as an additional information-theoretic measure in our models.

1.3 Research Questions and Hypotheses

The main aim of this study is to investigate whether German formant trajectories differ in their curvature when vowels stand in different surprisal contexts or appear in words with different frequencies of occurrence. We test for the effect of surprisal on formant movement by including the factor in interaction with the measurement point in the nonparametric part of our statistical model (**Section 2.2.4**). Given our previous findings that vowel dispersion in German is significantly affected by surprisal and word frequency (Brandt et al., 2019), we expect to find differences in formant trajectories between vowels in these different contexts, too.

Following the SSR hypothesis (Aylett and Turk, 2004, 2006), we investigate whether the effect of predictability on formant movement is modulated by a word-level effect of prominence, that is, primary lexical stress. We also control for the known effect of the place of articulation of directly preceding and following speech sounds on formant movements in the statistical models. Moreover, our models take into account that vowels located in less densely populated regions of the German vowel space are more variable in their formants, especially in F1 (Möbius, 2001), by including vowel phonemes as a predictor. We predict that the information-theoretic measure of surprisal affects formant trajectories above and beyond the effects of stress and coarticulation captured by the control factors.

2 MATERIALS AND METHODS

2.1 Materials

2.1.1 Speech Corpus

The Siemens Synthesis corpus (SI1000P) (Schiel, 1997) is used as speech material. These recordings were done to provide high-quality material for concatenative speech synthesis. The corpus contains audio recordings from two professional, middle-aged, male speakers of Standard German. Both speakers are trained and experienced broadcast announcers who worked at a German local state broadcasting station (BR) at the time of the recording. They were asked to read as if in a broadcasting setting. Both speakers read the same speech material. Each speaker recorded 992 sentences selected from the Frankfurter Allgemeine newspaper corpus (SI1000) in an echo-canceling studio using a Sennheiser MKH20 omnidirectional microphone with a controlled distance of 30 cm to the mouth, at a sampling rate of 48 kHz and 16 bits, filtered and down-sampled to 16 kHz. Canonical transcriptions and automatic word and phoneme segmentations are available.

2.1.2 Language Modeling Corpus

For the purpose of language modeling and extraction of word frequency values, we used a large text corpus with a sufficient amount of data. A German language model was trained using the web-crawled DeWaC corpus (Baroni et al., 2009), which comprises 1.2 billion running words and 9.3 million lexical types from a diverse range of genres.

2.2 Data Analysis

2.2.1 Speech Data Analysis

The automatic annotations provided in the speech corpus were manually verified by two phonetically trained annotators in the Phonetics laboratory at Saarland University who showed a very strong inter-rater agreement in the choice of their segment boundaries based on a Spearman's rho correlation test ($\rho = 0.93, S = 1427500000, p < 0.001$). The beginning of vowels was marked when F1 is clearly visible in the broadband spectrogram, and ends of vowels were marked at the end of a visible F2 structure.

The first and second formants were extracted using the Burg algorithm in Praat using a time step of 0.01 s, a maximum number of five formants, a ceiling of 5,000 Hz for the formant search range which is the default for adult male speakers, a window length of 25 ms, and preemphasis from 50 Hz at every 10% of the time-normalized vowel duration, yielding a formant trajectory defined by 11 samples for each vowel. The number of measurement points is sufficient for formant trajectory estimation since male speakers produce speech at an average

TABLE 1 | Number of tokens per vowel phoneme and primary lexical stress position in the dataset.

Vowel	Tokens	Stressed	Unstressed
i:	4,470	1,905	2,565
ɪ	5,650	1,965	3,685
e:	3,753	1,941	1,812
a:	3,040	1,808	1,232
a	5,964	2,859	3,105
o:	3,160	1,387	1,773
u:	1,176	480	696
ʊ	3,288	930	2,358

fundamental frequency of about 100–120 Hz, which means that formant values change at about every 8–10 ms. The average vowel duration in our data is 77 ms (SD = 33 ms), yielding a sufficiently dense sample of formant measurements per vowel.

Vowels in function words were excluded from the analysis following Bell et al. (2009). We also excluded diphthongs from the dataset because they inherently show more movement in their formants than monophthongs. The starting point at 0% and the end point at 100% of the vowel duration were discarded in the analysis because here formant extraction is potentially heavily influenced by the preceding or following speech sound. Formant values were cleaned using the interquartile ranges for F1 and F2 for German male speakers in the study by Pätzold and Simpson (1997) as a guideline. Since we model formant trajectories and are not limited to formant values at the temporal midpoint, we used more generous ranges for F1 (200–700 Hz) and F2 (450–2,400 Hz). Vowel tokens with formant values outside of these ranges were excluded from the analysis (n = 195, 0.34%). Formant values were not normalized because the statistical analysis applied here incorporates smoothing (see **Section 2.2.4**).

Only a subset of the German vowels was used in the modeling of German formant movement: front, close vowels: /i, I, e/; open, mid vowels: /a, a/; and back, close vowels: /u, ʊ, o/. This strategy allowed us to make inferences about vowel-specific formant movement depending on the placement in the vowel space. We decided to focus on peripheral vowels because they span the entirety of the German vowel space and are possibly very different in the extent of their formant movement and variability of their formant values in general. We analyzed a total of 30,501 vowel tokens, with 13,275 in stressed and 17,226 in unstressed positions (**Table 1**).

2.2.2 Language Modeling Procedure

Data preprocessing of the DeWaC corpus included lowercasing, punctuation removal, and grapheme-to-phoneme (g2p) conversion (Möhler et al., 2000). The transcriptions of the most frequent 1,000 words in the corpus were manually verified by the first author. Systematic errors in the g2p conversion were identified and corrected.

The training corpus (80% of the data) was used to train n-phone LMs using the SRILM toolkit (Stolcke, 2002). All LMs include sentence and word boundary markers and are based on both function and content words. By default, SRILM calculates the conditional probability of a linguistic unit based on its preceding context. In order to calculate conditional probabilities based on the following context, we used the built-in SRILM function *reverse-text*, which reverses the order of the linguistic units in each sentence. Models were smoothed using Witten–Bell smoothing. Because of the limited lexicon of the LM, count-of-counts statistics, such as Kneser–Ney, produced erroneous output.

The output for contextual predictability of the n-phone LMs was then transferred into surprisal (**Eq. 1**). We also extracted word frequency. Surprisal and word frequency were log-transformed because of their pronounced positive skewness. Surprisal values based on small n-phone sizes, as used in this study, express the probability of the phonotactic structure of a language, rather than simply giving information about preceding or following speech sounds. When segments are in word-initial or -final positions, the surprisal values reflect the word boundary marker. Other linguistic levels that potentially affect acoustic variability, even on the subword level, are only implicitly expressed in the surprisal values used here. We aimed to control for these effects by including word identity in the random structure of the statistical models.

We limit our investigation of formant movement to bi- and triphone surprisal for several reasons. First, the statistical models calculated in **Section 3** explain a large quantity of deviance in the formant trajectory data (about 85%). Second, the increasing n-phone size leads to higher sparsity in the data, that is, vowels that are close to the beginning or end of a sentence are not matched with a respective surprisal value (sentences were read as separate prompts), and certain unusual combinations of longer n-phone strings are not represented in the language model. Third, in a different investigation of the effect of surprisal on vowel dispersion, we have tested different n-phone sizes up to six and shown that the correlation between these two measures drops distinctively from the triphone level to the six-phone level (Brandt, 2018).

The bi- and triphones that are used for surprisal extraction are based on a transcription of the actual produced utterance, in contrast to using the normative, dictionary forms. We follow Tucker et al. (2019) in this approach, who found that the prediction accuracy of vowel duration decreases when using diphones based on dictionary transcriptions compared to using diphones based on transcriptions of actual productions.

In addition, it should be noted that higher order n-phones always contain the string of their respective lower order n-phones, that is, the information of the biphone is contained within the triphone. For that reason, we expect biphone and triphone surprisal values that share the same context direction to be correlated to some extent.

2.2.3 Primary Lexical Stress

Prominence was coded as a binary factor based on primary lexical stress (levels: stressed vs. unstressed) in the corpus text. Monosyllabic content words were classified as stressed.

2.2.4 Generalized Additive Mixed Modeling

We used generalized additive mixed models (GAMMs) to investigate dynamic changes in the formant trajectories of F1 and F2 as provided by the R package mgcv (R Development Core Team, 2008; Wood, 2011, 2017), visualized with itsadug (van Rij et al., 2017). GAMMs combine parametric terms and smooth terms in their structure, that is, they allow investigation of the relations between a response and one or more covariates in average values and also in nonlinear terms. In addition, they incorporate random effects, that is, random intercepts, slopes, and smooths. Random smooths allow us to model nonlinear by-group variation in the response variable (Sóskuthy, 2017). Recently, GAMMs have gained popularity in phonetic studies with a focus on speech articulation (Tomaschek et al., 2018b; Carignan et al., 2020) and acoustic–phonetic measures (Kirkham et al., 2019). In addition to their advantage of modeling nonlinear data, GAMMs are also able to capture interaction effects of two continuous variables by means of tensor product interaction [ti ()]. In the field of phonetics, this is particularly useful for modeling articulatory or acoustic data because they are conditioned by the interaction of time (temporal dimension or duration) and other continuous dimensions, such as space or measurement points.

Prior to model fitting, we checked for collinearity between the variables by using the pairs. panels () function of the *psych* package (Revelle, 2021). As expected, surprisal values that share the same context direction were moderately correlated (preceding context: r = 0.47, following context: r = 0.62), which was why we decided to calculate separate models for each surprisal variable. Word frequency and surprisal values, however, only showed a very weak (surprisal (X|X-1): r = −0.08, surprisal (X|X+1): r = −0.09, surprisal (X|X+2): r = −0.1) or weak (surprisal (X|X-2): r = −0.2) negative relationship, that is, vowels in high surprisal contexts show a slight tendency to appear in low-frequency words.

Surprisal values and word frequency were log-transformed. Vowel phonemes with three factor levels, front (/i, I, e/), mid (/a, a/), and back (/u, Ʊ, o/), were deviation-coded, comparing each level to the grand mean. The two-level factor stress (levels: unstressed and stressed) was treatment-coded.

We followed the modeling approach presented in the GAMM tutorial article by Wieling (2018). The model structure is given in listing 1. GAMMs were fitted using the bam () function of the *mgcv* package (Wood, 2019) because our dataset has more than 10,000 data points. Autocorrelation in the formant values can be expected for the temporal dimension vowel duration and also for the measurement point. Therefore, we included the autoregression function provided in the *mgcv* package. An AR (1) autoregressive error model for the residuals in a Gaussian model was included by using the rho parameter and setting the start event as 10% of the normalized vowel duration on an ordered dataset.

The smooth terms were fit with 'thin plate regression splines' (bs = "tp") (Wood, 2003). The interaction of the measurement point and duration and the interaction of surprisal and stress were fitted with "tensor product smooths" [ti ()], and we used "factor smooth interactions" (bs = "fs") to fit random effects. The smoothing parameter (k) for each smooth was set *via* model diagnosis [gam.check ()]. Since there are less than 10 unique values for the response variable, smooths for the measurement point were set at k = 9 to avoid overfitting of the data. This approach allowed for the right amount of wiggliness in the data.

Model comparison was performed using the *itsadug* function compareML (), which compares two models that vary in one term using the Akaike information criterion (AIC). Models with significantly lower AIC value were preferred. Concurvity of smooth terms was checked [concurvity ()] looking at pairwise concurvity between the terms.

We included fixed effects for deviation-coded vowel phonemes (levels: front, mid, and back), treatment-coded stress (levels: yes and no), and an interaction between both terms. Smooth terms [s ()] for the measurement point were included in the model using ordered by-terms (by =) for stress and vowel phonemes as ordered factors (oVowel, oStress). We were also interested in differences in formant trajectory shape due to different surprisal values by stress and vowel phonemes. Additionally, we included a smooth for word frequency by ordered vowel phonemes. The smooth for word frequency by stress did not increase model performance significantly.

In addition, the smooth term for the measurement point, the smooth of duration, and a tensor product interaction (ti) for the measurement point and duration were added to account for the influence of the temporal structure on the trajectories (Sóskuthy, 2017). Another tensor product interaction for the measurement point and surprisal and a smooth term for surprisal were added to capture how the measurement point and surprisal interact in their effect on first and second formant trajectory. We also tested the tensor product interaction of the measurement point and word frequency, but it did not increase model performance. Including the smooth term for word frequency increased model performance.

To capture the speaker and vowel phoneme variation as well as the effect of following and preceding context on formant trajectory shape, random smooths were included in the model (bs = "fs"). The order of the nonlinearity penalty (m) for the random smooths was set to 1.

3 RESULTS

The results of the GAMMs for F1 and F2 trajectories are presented by the terms in the models, providing a cohesive summary of the effects of surprisal and primary lexical stress and their interaction, word frequency, and the smooth terms on average formant values and the formant trajectory shapes. The GAMM output for each model is given in the supplementary material (**Supplementary Tables S1–S8**). Significant effects are reported when the significance level reaches p < 0.001. Since formant movement is heavily influenced by vowel duration, the average formant trajectory shapes are plotted for the mean vowel duration.

Differences in formant movement are visualized using difference smooth plots using the R package *itsadug* (van Rij et al., 2017). These plots convey the difference in formant trajectory shape between two factor levels (e.g., estimated difference of formant movement between unstressed and stressed vowels). Time windows with significant difference in trajectory shape are marked red and with dashed vertical lines,

LISTING 1 | Structure of generalized additive mixed models used to model response variable (F1/F2) trajectory.

```
#Main effect of deviation coded vowel (levels: front, mid, back),
#stress (levels: unstressed, stressed), and their interaction on F1/F2
F1/F2 ~ Vowel * Stress

#Separate smooth terms for measurement point, duration,
#word frequency, and surprisal
+ s(Percentage, k=9) + s(Duration, k=4)
+ s(WordFrequency, k=4) + s(Surprisal, k=4)

#Smooth terms for measurement point and word frequency
#by ordered vowel
+ s(Percentage, by=oVowel, k=9) + s(Percentage, by=oStress, k=9)
#Smooth terms for surprisal by ordered stress and by ordered vowel
+ s(Surprisal, by=oStress, k=4) + s(Surprisal, by=oVowel, k=4)
#Smooth term for word frequency by ordered vowel
+ s(WordFrequency, by=oVowel, k=4)

#Tensor product smooths for the interaction measurement point and
#duration, and measurement point and surprisal
+ ti(Percentage, Duration) + ti(Percentage, Surprisal)

#Random smooths to account for variability in formant movement per
#measurement point and speaker/preceding context/following context
+ s(Percentage, Speaker, bs="fs", m=1)
+ s(Percentage, Following, bs="fs", m=1)
+ s(Percentage, Preceding, bs="fs", m=1)

#Restricted maximum likelihood approach for model fitting
method = 'REML',

#Rho value is set as to the autocorrelation at lag 1, AR.start to set
#starting point for formant trajectory
rho = rhoval, AR.start=df$start.event, data = df)
```

while those parts of the trajectory with no significant difference in shape are left unmarked. If the estimated difference with a 95% confidence interval of the dependent variable, that is, the first or second formant, is below zero in the difference smooth plot, the dependent variable in the reference level has higher values than the factor level that the reference level is compared to, and *vice versa*. The difference smooth plot only shows the difference between two levels of a factor, that is, multiple plots are needed if the factor has more than two levels.

3.1 Vowel Phonemes

In our analysis of formant movement in German vowels, we focus on vowel phonemes in the periphery of the vowel space. We define three levels for the factor vowel: front: /i:, I, e:/; mid: /a:, a/:, and back vowels: /u:, ʊ, o:/. This factor is deviation-coded, which allows us to compare each level to the grand mean (see **Section 2.2.4**).

As can be seen in **Figure 1**, F1 is lower in back and front vowels compared to the grand mean, and F2 is lower in back vowels and higher in front vowels compared to the grand mean. Including an

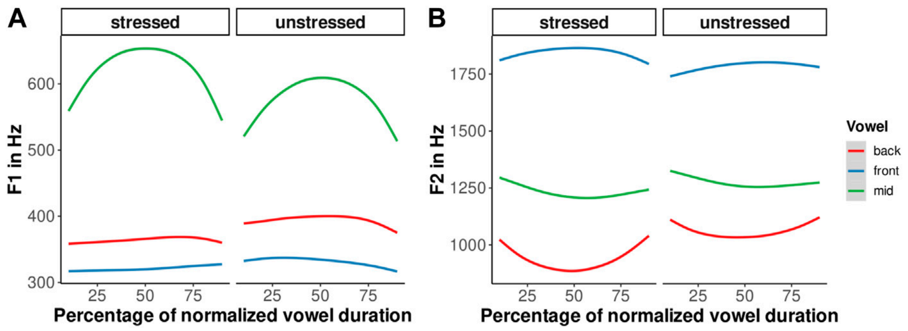


FIGURE 1 | Mean first **(A)** and second **(B)** formant trajectories per vowel phoneme category (front, mid, and back) and primary lexical stress (unstressed and stressed).

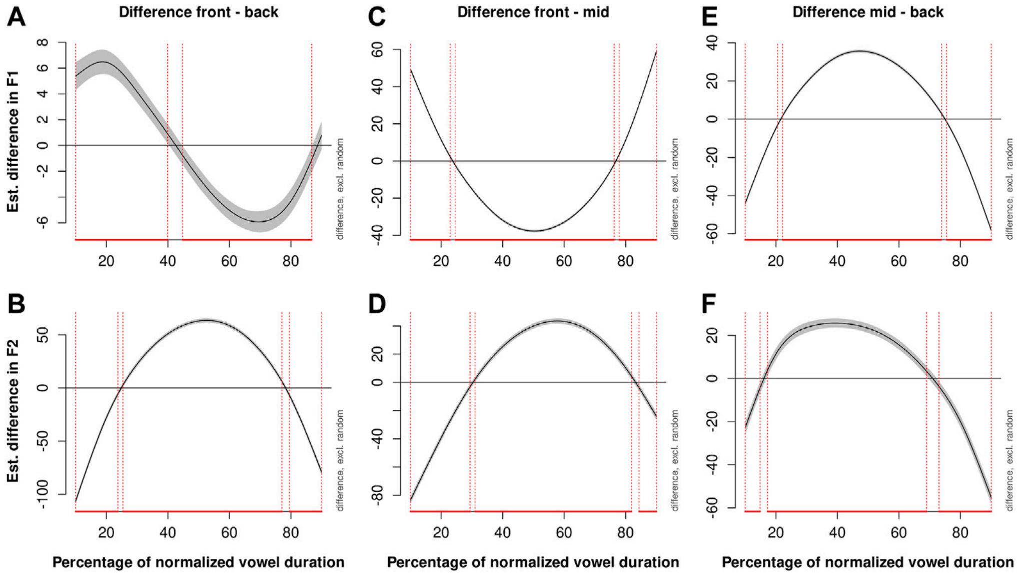


FIGURE 2 | Difference smooth for first **(A)** and second **(B)** formants between front and back vowels **(C)**, front and mid vowels **(D)**, and mid and back vowels **(E,F)** with a 95% confidence interval.

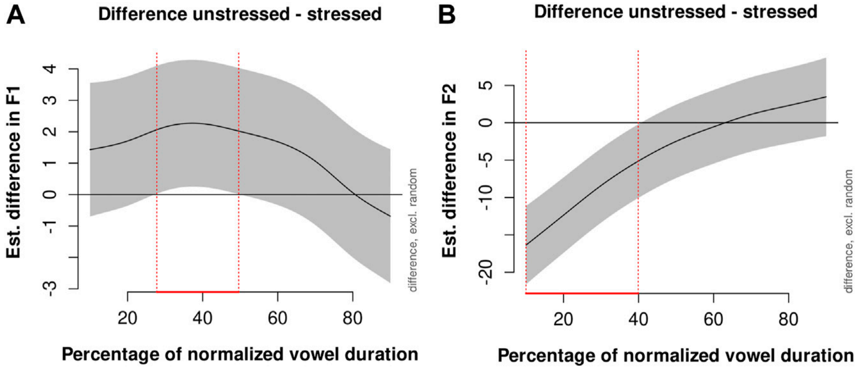


FIGURE 3 | Difference smooth in first **(A)** and second **(B)** formants between vowels in unstressed and stressed positions with a 95% confidence interval.

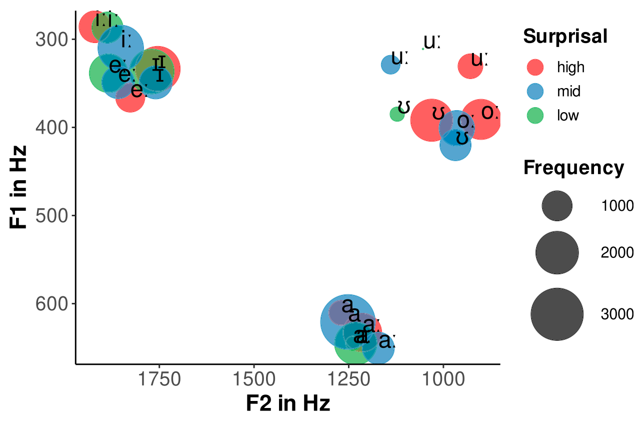


FIGURE 4 | Vowel chart of the subset of peripheral vowels with frequency of vowel tokens in the three bins of high, mid, and low biphone surprisal of the preceding context.

interaction between vowel phonemes and stress improves the model performance of all GAMMs tested in the current study. This interaction effect is also visualized in the average first and second formant trajectories given per stress condition and vowel phoneme in **Figure 1**. According to the GAMM output, F2 in stressed back vowels is significantly lower than in unstressed back vowels. For front vowels, F2 is higher in stressed than in unstressed vowels. F1 is lower in back and front vowels that stand in the stressed position than in those in the unstressed position, that is, these stressed vowels are more close and dispersed in the vowel space than their unstressed counterparts.

The vowel is then also included as an ordered factor in a smooth term with the measurement point to compare first and second formant movement between the three factor levels. Here, "back" is set as the reference level. According to the GAMM output, both F1 and F2 formant movements differ significantly between mid and back vowels and between front and back vowels. The first formant trajectory in German open, mid /a, a/ is significantly more concave with a steeper increase and fall than in back or front vowels throughout almost the entire normalized duration of the vowel (**Figures 2A,D,E,F**). The estimated difference in F1 movement between front and back vowels is smaller than the difference in F1 between mid and back vowels or mid and front vowels but still statistically significant. Front vowels have higher F1 values in the first half of the normalized vowel duration than back vowels and higher F1 values in the second half of the vowel (**Figure 2C**). The F2 trajectory is shaped convex in mid and back vowels, while front vowels, on average, are produced with a concave F2 trajectory. These significant differences in F2 formant movement per vowel category are visualized in the difference smooth plots in **Figure 2B**.

3.2 Stress

The main effect of stress on average F1 and F2 reaches the significance level in almost all models calculated in the current study. Mean F1 and F2 are slightly lower in stressed vowels than in unstressed vowels. Stress is also included in the smooth term for the measurement point, accounting for variability in F1 or F2 movement in different stress conditions. This smooth term reaches the significance level in all models. However, it can be seen in **Figure 3** that only a section of the formant trajectories (marked with vertical, dashed lines) in unstressed vowels is significantly different from that in stressed vowels: F1 movement differs significantly as a function of stress from around 25–50% of the normalized vowel duration (**Figure 3A**); F2 movement in unstressed vowels is different from that in stressed vowels only in the first part of the vowel up to 40% of its normalized duration (**Figure 3B**).

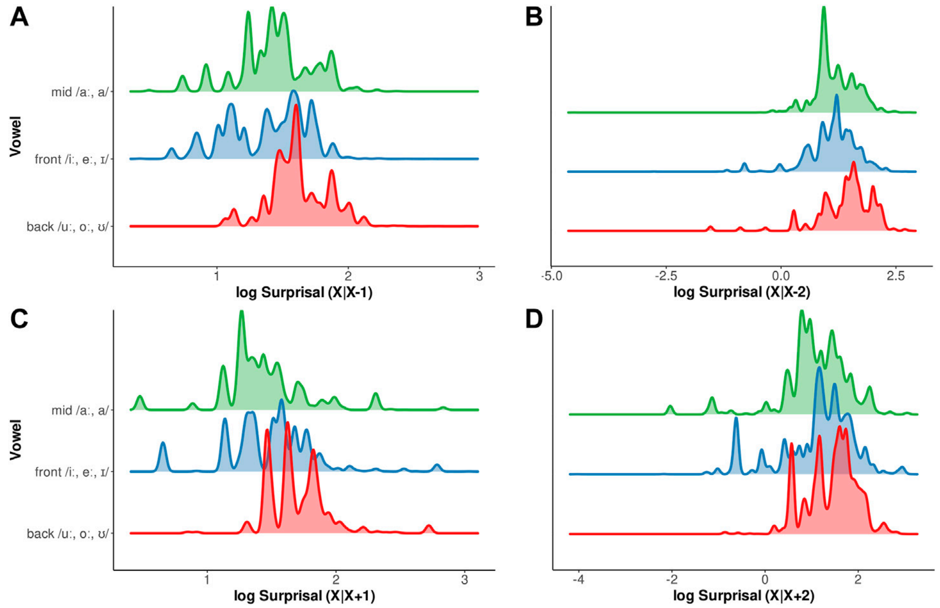


FIGURE 5 | Density plot of front, mid, and back vowels in different surprisal conditions of the preceding **(A,B)** or following **(C,D)** contexts.

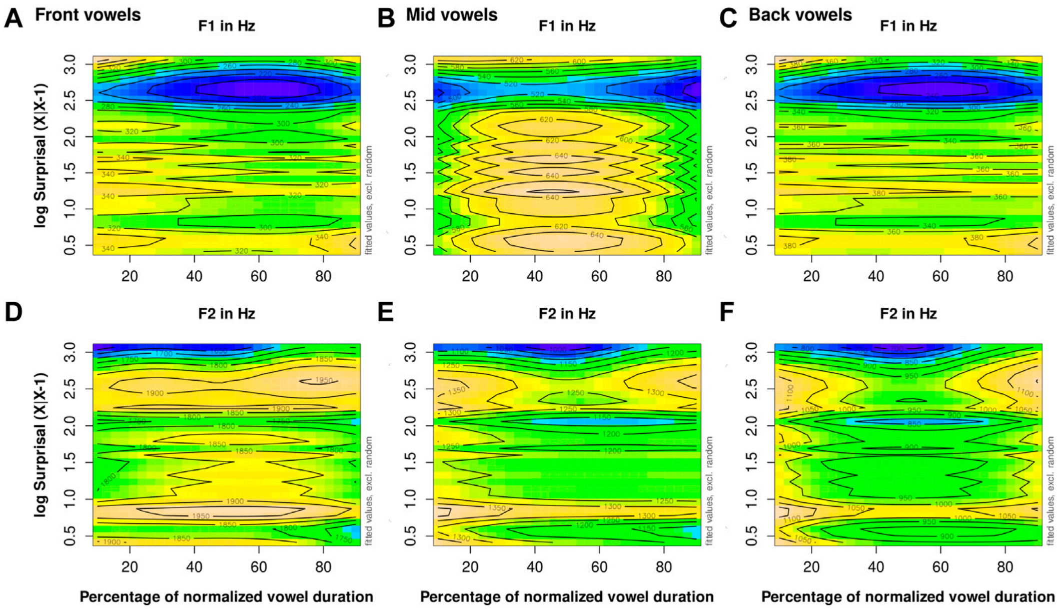


FIGURE 6 | Stressed vowels: heatmap of the interaction of biphone surprisal of the preceding context (log Surprisal (X|X-1)) and the measurement point on the trajectory of the first formant **(A)** and the second formant **(B)** per the vowel categories front **(C)**, mid **(D,E)**, and back **(F)**.

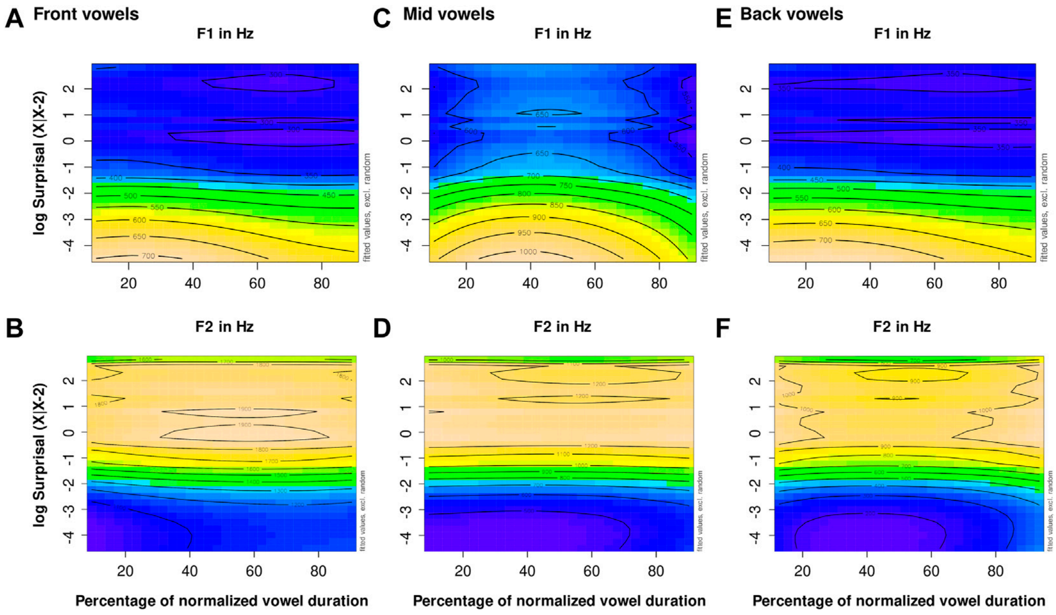


FIGURE 7 | Stressed vowels: heatmap of the interaction of triphone surprisal of the preceding context (log Surprisal (X|X-2)) and the measurement point on the trajectory of the first formant **(A)** and the second formant **(B)** per the vowel categories front **(C)**, mid **(D,E)**, and back **(F)**.

3.3 Surprisal

We present the results for the effect of surprisal on the first and second formant trajectories of peripheral German vowels. Surprisal values are based on bi- and triphones of the preceding and following contexts of the vowel.

For the purpose of visual inspection of our data, we bin biphone surprisal of the preceding context into three equally sized categories of "low," "mid," and "high" and plot the frequency of the peripheral German vowels used in our subset per surprisal category (**Figure 4**).

Although **Figure 4** only shows the frequency of vowel tokens in different categories of biphone surprisal of the preceding context and does not allow for general statements about the distribution of vowel phonemes in different surprisal contexts,

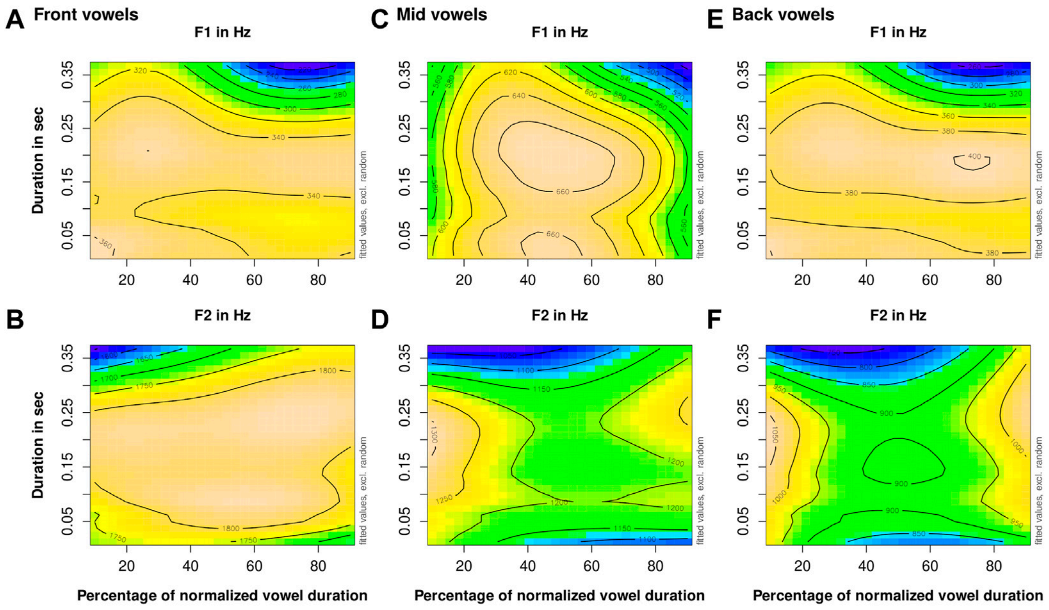


FIGURE 8 | Stressed vowels: heatmap of the interaction of the vowel duration and measurement point on the trajectory of the first formant **(A)** and the second formant **(B)** per the vowel categories front **(C)**, mid **(D,E)**, and back **(F)**.

there are two general observations that can be made. First, the average position of the vowel phoneme in the vowel space changes with regard to surprisal. Second, vowel phonemes are not equally distributed in the range of surprisal values observed in our data.

The second observation becomes even more apparent when investigating the distribution of vowel phonemes per bi- and triphone surprisal of the preceding or following context (**Figure 5**). On average, German back vowels have higher surprisal values, irrespective of n-phone order or direction, than mid or front vowels. Front vowels show slightly higher surprisal values than mid vowels. The difference between the distributions is more pronounced for biphone surprisal values than for triphone surprisal values.

All GAMMs calculated here include a tensor product interaction [ti ()] of the measurement point and surprisal, as well as two separate, simple smooth terms of the measurement point and surprisal, in order to tease apart the interaction effect from the main effect of the two smooth terms. The interaction of the measurement point and surprisal on the first formant trajectory reaches the significance level in all GAMMs. This means that first and second formant movement in German vowels is significantly impacted by the interaction of the bi- and triphone surprisal of both context directions and the measurement point for formant extraction in the normalized vowel duration.

Figure 6 shows how F1 and F2 trajectory shapes for stressed vowels in the front, mid, and back positions vary. We observe that F1 shows the lowest values for all vowels in the data with high surprisal values (≥ 2.5), that is, stressed front and back vowels are more close in high surprisal contexts than in low surprisal contexts, and stressed mid vowels show more pronounced F1 movement in their previously observed concave trajectory shape due to distinctly low F1 values (around 500 Hz) in the first and last third of the normalized vowel duration.

When we plot the GAMM heatmaps (**Figure 7**) for the interaction of the measurement point and triphone surprisal of the preceding context for all stressed vowels in the corpus, we find quite different patterns in the formant trajectories from those observed for biphone surprisal of the preceding context (**Figure 6**). Stressed high surprisal (≥ −1.5) front, mid, and back vowels have lower F1 values than stressed low surprisal vowels. For F2, however, high surprisal vowels (≥ −2) overall have higher formant values than low surprisal vowels, again irrespective of their position in the vowel space. This means that high surprisal vowels are produced with more frontness than low surprisal vowels. It should be noted that triphone surprisal values of the preceding context (R = −4.5–2.8) have a larger range than biphone surprisal values of the same context direction (R = 0.4–3.1). Judging from visual inspection alone, the average first and second formant trajectories per vowel category (**Figure 1**) seem to be better presented by the interaction plots for the measurement point and biphone surprisal (**Figure 6**) than by the heatmaps displaying the interaction effect of triphone surprisal and the measurement point (**Figure 7**).

Since we control for stress in the GAMMs, we can also investigate the impact of stress on the interaction between surprisal and the measurement point for different vowel phonemes, n-phone sizes, and forward and backward contextual predictability. For instance, **Supplementary Figure S2** shows the GAMM heatmaps of the interaction of biphone surprisal of the preceding context and the measurement point for unstressed vowels. When comparing these heatmaps to their counterparts for stressed vowels (**Figure 6**), we see that F1 in front and back vowels has overall lower values for higher surprisal contexts in unstressed vowels compared to stressed vowels. F1 in unstressed mid vowels shows more pronounced movement than in stressed conditions, that is, lower F1 values, at the edges of the vowel. F2 values in unstressed high surprisal (≥ 2.5) vowels, especially in the beginning of the vowel, are much lower than the F2 trajectories for stressed vowels.

We proceed to make the same comparative analysis between the GAMM heatmaps of unstressed and stressed vowels for the interaction effect of triphone surprisal of the preceding context and measurement on formant movement in order to investigate potential influences of the n-phone size. Overall, we find that the relationship between the two factors surprisal and measurement point shows a higher degree of variability in formant movement in the GAMM heatmaps for unstressed vowels than that for stressed vowels. Interestingly, for stressed vowels, we observe that average first and second formant values are closely related to the triphone surprisal level (X|X-2). In the unstressed condition, however, vowel frontness, expressed by F2, shows less of a clear-cut relationship to the surprisal level. Close unstressed vowels are produced with a more pronounced close articulatory setting at lower levels of triphone surprisal of the preceding context than stressed vowels.

We test for surprisal with preceding and following context direction. The GAMM heatmaps for the interaction between surprisal and the measurement point look quite different when comparing different context directions (**Supplementary Figures S1–S7**). For instance, average formant values in stressed vowels are strongly influenced by biphone surprisal of the preceding context but less by the temporal domain expressed by the measurement point, while unstressed vowels in the same surprisal condition show more variability in their formant movement depending on surprisal and the measurement point. We saw a similar pattern for formant trajectories in unstressed vs. stressed vowels in models with triphone surprisal of the preceding context.

3.4 Word Frequency

During the modeling procedure, we excluded a tensor product interaction of the measurement point and word frequency and a smooth of word frequency by ordered stress from the model because they did not add to model performance. However, the simple smooth term for word frequency and the smooth for word frequency by ordered vowel added to the model. This means that F1 and F2 movements do not vary significantly per measurement point in vowels occurring in words with different frequencies, nor do they vary as a function of differences in word frequencies in stressed and unstressed vowels. The model output does, however, show that formant movement is explained by differences in word frequencies and differences in word frequencies by vowel phoneme.

3.5 Interaction Between Duration and Percentage

The interaction term between the vowel duration and measurement point adds to the explained variance in the F1 and F2 data modeled here. Formant movement is heavily influenced by the duration of the vowel and the measurement point during vowel duration.

Figure 8 shows GAMM heatmaps for the first and second formant trajectories in stressed German vowels as an interaction between the vowel duration and measurement point which is modeled by the tensor product interaction of the measurement point and duration.[1]

In the GAMM heatmaps (**Figure 8**), we can observe the same overall formant trajectory shape for each vowel category that is given in **Figure 1**. The heat maps allow us to make more detailed observations about this overall shape, depending on vowel duration. Longer vowels above 0.25 s appear to show more pronounced first and second formant movement with lower minima than vowels with average or short duration. The peak of the F1 trajectory appears earlier in the vowel as a function of vowel duration when the vowel is longer than 0.25 s. We can also see that the average concave F2 trajectory shape for front vowels is mainly due to movement in long vowels, again above 0.25 s, while shorter front vowels show very little F2 movement. Very short vowels show surprisingly low F2 values for mid (around 1,100 Hz) and back (around 850 Hz) vowels.

3.6 Random Effects

The random smooths for the measurement point per speaker and the preceding and following contexts significantly add to the explained variability in F1 and F2 movement in all models.

4 DISCUSSION

This study investigated whether variability in German formant trajectories can be explained by contextual predictability, measured as surprisal, and prominence, that is, primary lexical stress, as well as an interaction of both factors. We also include word frequency as an additional information-theoretic measure in our models. We use generalized additive mixed models (GAMMs) to compare the shape of formant trajectories in different surprisal contexts. Surprisal values are based on the biphone or triphone of the preceding or following context of the vowel. Only monophthongs in content words were considered in the study.

For average F1 and F2, we find expected results for different vowel phonemes that determine the position of the vowel within the acoustic vowel space. The significant interaction effect between the factors vowel and stress in the F1 and F2 models confirms that vowels in the stressed position are more dispersed in the vowel space than vowels in the unstressed position.

For the purpose of the study, we are particularly interested in the results of the smooth terms including surprisal. The GAMM output shows that the first and second formant trajectories in German are

[1] The equivalent GAMM heatmaps for the first and second formant trajectories in unstressed German vowels as an interaction between vowel duration and the measurement point can be found in **Supplementary Figure S1**. This allows us to investigate differences in formant trajectory due to the temporal domain. We include separate heatmaps of this interaction per vowel category since this factor significantly impacts formant movements (**Section 3.1**).

affected by surprisal in both context directions, that is, forward and backward, and by the interaction of surprisal and stress. We analyze these results in more detail *via* visual inspection of GAMM heatmaps that show the interaction effect of surprisal per measurement point (temporal domain) on the formant trajectory. We plot these heatmaps per vowel phoneme and stress condition because we find that these additional factors impact formant movement significantly. This procedure shows that the interaction effects of these factors on formant movement are highly complex. However, there are some general observations that we can make: unstressed vowels seem to show higher variability in their formant trajectory at different surprisal levels than stressed vowels. Differences in average formant values are also more readily expressed as a function of surprisal in stressed vowels than in unstressed vowels.

Our results show that effects of contextual predictability on formant variability are not limited to pointwise measurements of the vowel, as seen in studies on the effect of predictability on vowel dispersion (Malisz et al., 2018), but affect the dynamics throughout the entire vowel duration. When interpreted against the background of the uniform information density (UID) hypothesis (Levy and Jaeger, 2007), our findings add to the concept that the rational speaker uses optimization strategies in speech production throughout the entire utterance to ensure successful communication. This strategic behavior of the speaker also has an effect on the characteristics of formant movement and is observed while controlling for linguistic factors that are known to affect formant movement, such as vowel duration or phonetic context.

We proceed by further discussing our results with respect to the relation of prosodic prominence and predictability, especially in light of the smooth signal redundancy (SSR) hypothesis (Aylett and Turk, 2004, 2006). In addition, possible accounts of the effect of predictability on the phonetic structure are discussed.

4.1 Prosodic Prominence and Predictability Based Formant Movement

We test interaction effects between prosodic prominence and predictability on average first and second formant values and on formant movement in German vowels to investigate the effect of predictability and the prosodic structure, here primary lexical stress, on phonetic variability. This research goal is motivated by the smooth signal redundancy hypothesis (Aylett and Turk, 2004, 2006), which postulates that the effects of language redundancy or predictability on phonetic structures are moderated by the prosodic structure (prosodic prominence), that is, there are no independent or additive effects of predictability on phonetic variability. We can confirm this expected interaction effect between stress and surprisal on first and second average formant values and on formant trajectories.

Since German vowels in the stressed position and under high surprisal are known to be more dispersed in the vowel space (Malisz et al., 2018; Schulz et al., 2016), we would expect higher average F2 and lower average F1 values for front vowels in the stressed position and under high surprisal than for those in the unstressed position. Judging from the GAMM heatmaps for biphone surprisal of the preceding context, that is, the same surprisal measure as that used in our previous studies, we find the predicted pattern for front vowels. For back vowels, on the other hand, we expect lower average F1 and F2 values for stressed vowels in high surprisal contexts than for unstressed vowels. From visual inspection of the GAMM heatmaps in **Figure 6** and **Supplementary Figure S2**, we cannot confirm this expectation for back vowels. For mid vowels /a, aː/, we find that they are produced with more frontness in the unstressed condition under high surprisal than in stressed and high surprisal contexts.

We include an analysis of the impact of the temporal domain (interaction of the vowel duration and measurement point) on first and second formant trajectories, again distinguishing stress condition and vowel phoneme. While there are vast differences in formant movement depending on vowel duration, with longer vowels showing more formant movement than shorter vowels, the effect of stress on this relation appears to be small. This observation is partially in line with work that highlights the importance of time as a crucial factor for articulatory effort (Xu and Prom-on, 2010). The authors found that time constraints determine how much information speakers can convey in a conversational turn and hypothesized that speakers maximize their articulatory effort in unstressed vs. stressed vowels, which can also lead to increased dynamics for unstressed vowels compared to stressed vowels. Tang and Shaw (2020) noted that this principle applies to their findings on word duration as a function of predictability in Mandarin Chinese. The amount of time speakers allocate to a linguistic unit is a function of its importance, that is, less predictable words are produced with longer durations. In our study, we find more pronounced formant movement in unstressed vowels when investigating formant movement as a function of the surprisal and measurement point. Vowel duration and surprisal are, however, known to be correlated (Malisz et al., 2018).

Prosodic prominence, here estimated as primary lexical stress, was found to have a significant impact on the mean values of the first and second formants in German vowels in almost all GAMMs. In our models, the average F1 and F2 in stressed vowels are lower than those in unstressed vowels.

Lexically stressed American English vowels that are perceived as prominent are produced with a more open vocal tract than those vowels that are not perceived as prominent, resulting in higher F1 values for these vowels (Mo et al., 2009). Speakers are assumed to use this strategy to increase the sonority of prominent syllables (Beckman et al., 1992). For F2, or vowel frontness, vowels are hyperarticulated when they stand in a prominent position (Mo et al., 2009), supporting the hypo- and hyperarticulation hypothesis (Lindblom, 1996). This means that prominent back vowels are produced with lower F2 values and prominent front vowels are produced with higher F2 values than their non-prominent counterparts. This effect is captured by expanded vowel dispersion for stressed vowels in German (Schulz et al., 2016) and could also be replicated in our study.

The German vowel system, however, differentiates between tense and lax vowels, which can both stand in stressed or unstressed positions. German formant movement is largely influenced by vowel tenseness and frontness, that is, vowel identity. There are also known effects of stress on German tense vs. lax vowels: stressed tense vowels are longer and more peripheral in their position in the vowel space than unstressed tense vowels. Lax vowels, however, are not significantly affected by stress in their length or average formant values (Jessen, 1995; Mooshammer et al., 1999). Therefore, stress alone is possibly not an ideal factor to predict formant movement in German.

4.2 Accounts of the Effect of Predictability on Speech Variability

This study adds to previous accounts of predictability-based variability in the speech signal at the subword level. There are different accounts of these observed effects: the production ease account and the listener-oriented communicative account. Seminal work advocating the production ease account (e.g., Gahl, 2008; Bell et al., 2009) demonstrated the effect of frequency and predictability on word duration. The production planning hypothesis (Kilbourn-Ceron et al., 2020) views predictability as one of the factors that impact speech planning. Easily predictable phonological information in an upcoming word can facilitate the speech production process of pronunciation variants. The production ease account therefore relies on the contextual predictability of a linguistic structure based on both context directions, as it is known that coarticulatory processes have an effect on preceding and following neighboring phonemes. An alternative, but compatible, explanation has been offered by Tomaschek and others (Tomaschek et al., 2018a, b), who proposed that it is linguistic experience and articulation practice, rather than predictability as such, that shape articulatory trajectories.

The listener-oriented or communicative account, on the other hand, proposes that communication is a balancing act for the speaker between making the least possible amount of effort and attending to the listener's need. As a result, predictable linguistic structures can be reduced because they are easily retrievable from their context, while structures that are difficult to predict from their context must be preserved. Therefore, both context directions (backward and forward) of contextual predictability play a role in this account. A strong interpretation of listener orientation in speech production is challenged by the finding that the speaker's capacity to attribute mental states to others, also known as theory of mind (ToM) (Premack and Woodruff, 1978), does not necessarily lead to the phonetic reduction of predictable linguistic structures (Turnbull, 2019). It should be kept in mind however that high scores in ToM ability, as tested in the study by Turnbull (2019), estimate the speaker's capacity of ToM but not their willingness to apply their ability to attribute mental states to others in a specific communicative setting. In our interpretation of these two accounts of the effect of predictability on speech variability, we note that both the listener-oriented and the production ease accounts rely on contextual predictability of linguistic structures that is based on the preceding and following contexts. There is also evidence from perception studies that listeners do not only utilize preceding information for word recognition in running speech but also following contextual information (Szostak and Pitt, 2013). This process seems to be modulated by contextual predictability in both directions. Listeners pay less attention to the phonetic details of easily predictable words (Manker, 2017).

With regard to our findings, surprisal based on the following context significantly explains the formant trajectory shape in German. This result is not necessarily expected since we also know from previous work that the effect of surprisal in different context directions depends on which acoustic measure is investigated. Segment duration can be explained by surprisal of the preceding and following contexts, whereas vowel dispersion is only predicted by surprisal of the preceding context (Malisz et al., 2018).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The monolingual German speech dataset analyzed for this study, the Siemens Synthesis Corpus (SI1000P), can be found in http://catalog.elra.info/en-us/repository/browse/ELRA-S0082/. The German text corpus used for language modeling is available at https://wacky.sslmit.unibo.it/doku.php?id = corpora.

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

BM, BA, and EB contributed to the conception and design of the study. EB organized the database, performed the statistical analysis, and wrote the first draft of the manuscript. BM and BA wrote sections of the manuscript. All authors contributed to manuscript revision and read and approved the submitted version. Funding awarded to BM and BA.

FUNDING

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 232722074—SFB 1102 Information Density and Linguistic Encoding at Saarland University.

ACKNOWLEDGMENTS

We wish to thank Melanie Weirich for her helpful practical advice regarding GAMMs.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2021.643528/full#supplementary-material

REFERENCES

Aylett, M., and Turk, A. (2006). Language Redundancy Predicts Syllabic Duration and the Spectral Characteristics of Vocalic Syllable Nuclei. *The J. Acoust. Soc. America* 119, 3048–3058. doi:10.1121/1.2188331

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

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

Beckman, M., Edwards, J., and Fletcher, J. (1992). "Prosodic Structure and Tempo in a Sonority Model of Articulatory Dynamics," in *Laboratory Phonology II: Gesture, Segment, Prosody*. Editors G. J. Docherty and D. R. Ladd (Cambridge, United Kingdom: Cambridge University Press), 68–89. doi:10.1017/cbo9780511519918.004

Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability Effects on Durations of Content and Function Words in Conversational English. *J. Mem. Lang.* 60, 92–111. doi:10.1016/j.jml.2008.06.003

Benner, U., Flechsig, I., Dogil, G., and Möbius, B. (2007). "Coarticulatory Resistance in a Mental Syllabary," in *Proceedings of the International Congress of Phonetic Sciences* (Saarbrücken, 485–488.

Bohn, O.-S., and Polka, L. (2001). Target Spectral, Dynamic Spectral, and Duration Cues in Infant Perception of German Vowels. *J. Acoust. Soc. America* 110, 504–515. doi:10.1121/1.1380415

Brandt, E., Andreeva, B., and Möbius, B. (2019). "Information Density and Vowel Dispersion in the Productions of Bulgarian L2 Speakers of German," in *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS 2019)* Melbourne, Australia, 3165–3169.

Brandt, E. (2018). Information Density and Phonetic Structure: Explaining Segmental Variability. Ph.D. thesis. Saarbrücken: University of Saarland.

Brandt, E. (2019). Information Density and Phonetic Structure: Explaining Segmental Variability. Ph.D. thesis. Saarbrücken: Saarland University.

Bürki, A., Ernestus, M., Gendrot, C., Fougeron, C., and Frauenfelder, U. H. (2011). What Affects the Presence versus Absence of Schwa and its Duration: a Corpus Analysis of French Connected Speech. *J. Acoust. Soc. America* 130, 3980–3991. doi:10.1121/1.3658386

Carignan, C., Hoole, P., Kunay, E., Pouplier, M., Joseph, A., Voit, D., et al. (2020). Analyzing Speech in Both Time and Space: Generalized Additive Mixed Models Can Uncover Systematic Patterns of Variation in Vocal Tract Shape in Real-Time MRI. *Lab. Phonology: J. Assoc. Lab. Phonology* 11. doi:10.5334/labphon.214

Clopper, C. G., and Pierrehumbert, J. B. (2008). Effects of Semantic Predictability and Regional Dialect on Vowel Space Reduction. *J. Acoust. Soc. America* 124, 1682–1688. doi:10.1121/1.2953322

Cohen Priva, U., and Jaeger, T. F. (2018). The Interdependence of Frequency, Predictability, and Informativity. *Linguistics Vanguard* 4, 1–17. doi:10.1515/lingvan-2017-0028

Dautriche, I., Mahowald, K., Gibson, E., and Piantadosi, S. (2017). Wordform Similarity Increases with Semantic Similarity: an Analysis of 100 Languages. *Cogn. Sci.* 41, 2149–2169. doi:10.1111/cogs.12453

Demberg, V., Sayeed, A. B., Gorinski, P. J., and Engonopoulos, N. (2012). "Syntactic Surprisal Affects Spoken Word Duration in Conversational Contexts," in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (Jeju Island, Korea: Association for Computational Linguistics), 356–367.

Gahl, S. (2008). *Thyme* and *Time* Are Not Homophones: The Effect of Lemma Frequency on Word Durations in Spontaneous Speech. *Language* 84, 474–496. doi:10.1353/lan.0.0035

Gahl, S., Yao, Y., and Johnson, K. (2012). Why Reduce? Phonological Neighborhood Density and Phonetic Reduction in Spontaneous Speech. *J. Mem. Lang.* 66, 789–806. doi:10.1016/j.jml.2011.11.006

Hale, J. (2001). "A Probabilistic Early Parser as a Psycholinguistic Model," in *Proceedings of NAACL* Stroudsburg, PA, 1–8.

Hale, J. (2016). Information-theoretical Complexity Metrics. *Lang. Linguistics Compass* 10, 397–412. doi:10.1111/lnc3.12196

Jaeger, T. F., and Buz, E. (2017). "Signal Reduction and Linguistic Encoding," in *Handbook of Psycholinguistic*. Editors E. M. Fernandez and H. M. I. Cairns (Oxford, United Kingdom: Wiley-Blackwell), 38–81. doi:10.1002/9781118829516.ch3

Jessen, M. (1995). Acoustic Correlates of Word Stress and the Tense/lax Opposition in the Vowel System of German. Int. Congress Phonetic Sci. *(Stockholm)* 4, 428–431.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. (2001). "Probabilistic Relations between Words: Evidence from Reduction in Lexical Production," in *Frequency and the Emergence of Linguistic Structure*. Editors J. Bybee and P. Hopper (Amsterdam: Benjamins), 229–254. doi:10.1075/tsl.45.13jur

Kilbourn-Ceron, O., Clayards, M., and Wagner, M. (2020). Predictability Modulates Pronunciation Variants through Speech Planning Effects: A Case Study on Coronal Stop Realizations, *Lab. Phonology: J. Assoc. Lab. Phonology*, 11. doi:10.5334/labphon.168

Kirkham, S., Nance, C., Littlewood, B., Lightfoot, K., and Groarke, E. (2019). Dialect Variation in Formant Dynamics: The Acoustics of Lateraland Vowel Sequences in manchester and liverpool English. *J. Acoust. Soc. America* 145, 784–794. doi:10.1121/1.5089886

Kuperman, V., Pluymaekers, M., Ernestus, M., and Baayen, H. (2007). Morphological Predictability and Acoustic Duration of Interfixes in Dutch Compounds. *J. Acoust. Soc. America* 121, 2261–2271. doi:10.1121/1.2537393

Levy, R. (2008). "A Noisy-Channel Model of Rational Human Sentence Comprehension under Uncertain Input," in *Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing*. Honolulu: Waikiki, 234–243.

Levy, R. (2011). "Integrating Surprisal and Uncertain-Input Models in Online Sentence Comprehension: Formal Techniques and Empirical Results," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (Portland, Oregon: Human Language Technologies), 1055–1065.

Levy, R., and Jaeger, T. F. (2007). Speakers Optimize Information Density through Syntactic Reduction. *Adv. Neural Inf. Process. Syst.* 19, 849–856.

Lindblom, B. (1996). Role of Articulation in Speech Perception: Clues from Production. *J. Acoust. Soc. America* 99, 1683–1692. doi:10.1121/1.414691

Möhler, G., Schweitzer, A., Breitenbücher, M., and Barbisch, M. (2000). IMS German Festival (Version: 1.2-os)

Malisz, Z., Brandt, 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

Manker, J. T. (2017). Phonetic Attention and Predictability: How Context Shapes Exemplars and Guides Sound Change. Ph.D. thesis. Berkeley: University of California.

Mo, Y., Cole, J., and Hasegawa-Johnson, M. (2009). "Prosodic Effects on Vowel Production: Evidence from Formant Structure," in *Proceedings of Interspeech*. Brighton, UK), 2535–2538.

Möbius, B. (2001). German and Multilingual Speech Synthesis, *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung, AIMS*, 7.

Möbius, B. (1999). The Bell Labs German Text-To-Speech System. *Computer Speech Lang.* 13, 319–358. doi:10.1006/csla.1999.0127

Mooshammer, C., Fuchs, S., and Fischer, D. (1999). "Effects of Stress and Tenseness on the Production of CVC Syllables in German," in *International Congress of Phonetic Sciences* (San Francisco), 409–412.

Nearey, T. M., and Assmann, P. F. (1986). Modeling the Role of Inherent Spectral Change in Vowel Identification. *J. Acoust. Soc. America* 80, 1297–1308. doi:10.1121/1.394433

Oh, Y. M., Christophe, Coupé., Marsico, E., and Pellegrino, F. (2015). Bridging Phonological System and Lexicon: Insights from a Corpus Study of Functional Load. *J. Phonetics* 53, 153–176. doi:10.1016/j.wocn.2015.08.003

Pätzold, M., and Simpson, A. P. (1997). Acoustic Analysis of German Vowels in the Kiel Corpus of Read Speech. *The Kiel Corpus Of Read/Spontaneous Speech Acoustic Data Base, Processing Tools and Analysis Results. Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK)* 32, 215–247.

Pellegrino, F., Coupé, C., and Marisco, E. (2011). A Cross-Language Perspective on Speech Information Rate. *Language* 87, 539–558. doi:10.1353/lan.2011.0057

Piantadosi, S., Tily, H., and Gibson, E. (2011). Word Lengths Are Optimized for Efficient Communication. *Proc. Natl. Acad. Sci.* 108, 3526–3529. doi:10.1073/pnas.1012551108

Pluymaekers, M., Ernestus, M., and Baayen, R. H. (2005a). Articulatory Planning Is Continuous and Sensitive to Informational Redundancy. *Phonetica* 62, 146–159. doi:10.1159/000090095

Pluymaekers, M., Ernestus, M., and Baayen, R. H. (2005b). Lexical Frequency and Acoustic Reduction in Spoken Dutch. *J. Acoust. Soc. America* 118, 2561–2569. doi:10.1121/1.2011150

Premack, D., and Woodruff, G. (1978). Does the Chimpanzeehave a Theory of Mind?. *Behav. Brain Sci.* 1, 515–526. doi:10.1017/s0140525x00076512

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raymond, W., Dautricourt, R., and Hume, E. (2006). Word-internal/t,d/Deletion in Spontaneous Speech: Modeling the Effects of Extra-linguistic, Lexical, and Phonological Factors. *Lang. Variation Change* 18, 55–97. doi:10.1017/s0954394506060042

Revelle, W. (2021). *Psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois: Northwestern University.R package version 2.1.3.

Scarborough, R. (2010). "Lexical and Contextual Predictability: Confluent Effects on the Production of Vowels," in *Laboratory Phonology 10*. Editors C. Fougeron, B. Kühnert, M. D'Imperio, and N. Vallee (Scarborough: Berlin: De Gruyther Mouton), 557–586.

Schiel, F. (1997). *Siemens Synthesis Corpus - SI1000P*. University of Munich.

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, 350–354.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell Syst. Tech. J.* 27 (379–423), 623–656. doi:10.1002/j.1538-7305.1948.tb00917.x

Sóskuthy, M. (2017). Generalised Additive Mixed Models for Dynamic Analysis in Linguistics: a Practical Introduction. *Working Paper* 1–47.

Stolcke, A. (2002). Srilm - an Extensible Language Modeling Toolkit. *Proc. Interspeech* 2, 901–904.

Strange, W., and Bohn, O.-S. (1998). Dynamic Specification of Coarticulated German Vowels: Perceptual and Acoustical Studies. *J. Acoust. Soc. America* 104, 488–504. doi:10.1121/1.423299

Strange, W., Bohn, O.-S., Trent, S. A., and Nishi, K. (2004). Acoustic and Perceptual Similarity of North German and American English Vowels. *J. Acoust. Soc. America* 115, 1791–1807. doi:10.1121/1.1687832

Strange, W., Weber, A., Levy, E. S., Shafiro, V., Hisagi, M., and Nishi, K. (2007). Acoustic Variability within and across German, French, and American English Vowels: Phonetic Context Effects. *J. Acoust. Soc. America* 122, 1111–1129. doi:10.1121/1.2749716

Szostak, C. M., and Pitt, M. A. (2013). The Prolonged Influence of Subsequent Context on Spoken Word Recognition. *Attention, Perception, Psychophysics* 75, 1533–1546. doi:10.3758/s13414-013-0492-3

Tang, K., and Shaw, J. A. (2020). Prosody Leaks into the Memory of Words. *Cognition* 210, 104601. doi:10.1016/j.cognition.2021.104601

Tomaschek, F., Arnold, D., Bröker, F., and Baayen, R. H. (2018a). Lexical Frequency Co-determines the Speed-Curvature Relation in Articulation. *J. Phonetics* 68, 103–116. doi:10.1016/j.wocn.2018.02.003

Tomaschek, F., Tucker, B. V., Fasiolo, M., and Baayen, H. (2018b). Practice Makes Perfect: the Consequences of Lexical Proficiency for Articulation. *Linguistic Vanguard* 4. doi:10.1515/lingvan-2017-0018

Tucker, B. V., Sims, M., and Baayen, H. (2019). *Opposing Forces on Acoustic Duration*. doi:10.31234/osf.io/jc97w

Turnbull, R. (2019). Listener-oriented Phonetic Reduction and Theory of Mind. *Lang. Cogn. Neurosci.* 34, 747–768. doi:10.1080/23273798.2019.1579349

van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, H. (2017). Itsadug: Interpreting Time Series and Autocorrelated Data Using Gamms. R package version 2.3.

Wedel, A., Nelson, N., and Sharp, R. (2018). The Phonetic Specificity of Contrastive Hyperarticulation in Natural Speech. *J. Mem. Lang.* 100, 61–88. doi:10.1016/j.jml.2018.01.001

Wieling, M. (2018). Analyzing Dynamic Phonetic Data Using Generalized Additive Mixed Modeling: A Tutorial Focusing on Articulatory Differences between L1 and L2 Speakers of English. *J. Phonetics* 70, 86–116. doi:10.1016/j.wocn.2018.03.002

Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. 2 edn. Chapman and Hall/CRC.

Wood, S. (2019). *Mgcv: Mixed GAM Computation Vehicle With Automatic Smoothness Estimation*.

Wood, S. N. (2011). Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models. *J. R. Stat. Soc.* 73, 3–36. doi:10.1111/j.1467-9868.2010.00749.x

Wood, S. N. (2003). Thin-plate Regression Splines. *J. R. Stat. Soc. (B)* 65, 95–114. doi:10.1111/1467-9868.00374

Wright, R. (2004). "Factors of Lexical Competition in Vowel Articulation," in *Papers in Laboratory Phonology VI*. Editors J. Local, R. Ogden, and R. Temple (Cambridge: Cambridge University Press), 26–50.

Xu, Y., and Prom-on, S. (2010). Economy of Effort or Maximum Rate of Information? Exploring Basic Principles of Articulatory Dynamics. *Front. Psychol.* doi:10.3389/fpsyg.2019.02469

Zhao, Y., and Jurafsky, D. (2009). The Effect of Lexical Frequency and Lombard Reflex on Tone Hyperarticulation. *J. Phonetics* 37, 231–247. doi:10.1016/j.wocn.2009.03.002

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. New York: Addison-Wesley.

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: 23 June 2021
doi: 10.3389/fpsyg.2021.657705

Parsing Model and a Rational Theory of Memory

Jakub Dotlačil[1]* *and Puck de Haan*[2]

[1] Utrecht Institute of Linguistics, Utrecht University, Utrecht, Netherlands, [2] Artificial Intelligence, University of Amsterdam, Amsterdam, Netherlands

This paper explores how the rational theory of memory summarized in Anderson (1991) can inform the computational psycholinguistic models of human parsing. It is shown that transition-based parsing is particularly suitable to be combined with Anderson's theory of memory systems. The combination of the rational theory of memory with the transition-based parsers results in a model of sentence processing that is data-driven and can be embedded in the cognitive architecture Adaptive Control of Thought-Rational (ACT-R). The predictions of the parser are tested against qualitative data (garden-path sentences) and a self-paced reading corpus (the Natural Stories corpus).

Keywords: computational psycholinguistics, cognitively constrained parsers, memory retrieval, rational theory of memory, modeling reading data

OPEN ACCESS

Edited by:
Matthew W. Crocker,
Saarland University, Germany

Reviewed by:
John Hale,
University of Georgia, United States
Christian Huyck,
Middlesex University, United Kingdom
Richard Futrell,
University of California, Irvine,
United States

****Correspondence:***
Jakub Dotlačil
j.dotlacil@gmail.com

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

Received: *23 January 2021*
Accepted: *03 May 2021*
Published: *23 June 2021*

Citation:
Dotlačil J and de Haan P (2021)
Parsing Model and a Rational Theory
of Memory.
Front. Psychol. 12:657705.
doi: 10.3389/fpsyg.2021.657705

1. INTRODUCTION

In the rational theory of cognition, it is argued that cognitive functions are largely shaped by our adaptation to the environment. In this view, it is assumed that various aspects of our behavior can be explained as the result of the optimization to the structure of the environment. The rational theory of cognition has been fruitful in explaining regularities in categorization, learning, communication and reasoning, among others (Anderson, 1990, 1991; Oaksford and Chater, 1994, 2007; Tenenbaum et al., 2011; Franke and Jäger, 2016; Piantadosi et al., 2016).

One particularly successful case of the rational theory was its application to the study of human memory, as summarized in Anderson (1991). Assuming that the human memory should strive to provide information that is needed at a particular situation and that it is costly and takes time to retrieve elements from memory, we would expect that the retrieval of an element be related to the probability that it is needed. That is, elements that are most likely to be needed at a particular situation will be prioritized in retrieval. Since retrieval is ordered by need probabilities, it is expected that less needed items require more time to be recalled. Furthermore, if retrieval is abandoned when the cost for retrieval exceeds some threshold, we expect the less needed an item is, the more likely it is that its recall fails. These predictions have been largely confirmed, see Anderson (1991).

The rational theory of memory played an important role in the development of the cognitive architecture Adaptive Control of Thought-Rational, ACT-R (Anderson and Lebiere, 1998; Anderson et al., 2004), which in turn played an important role in psycholinguistic models of parsing (Lewis and Vasishth, 2005; Lewis et al., 2006; Reitter et al., 2011; Engelmann et al., 2013; Vogelzang et al., 2017; Brasoveanu and Dotlačil, 2020). Lewis and Vasishth (2005) and subsequent works showed, in particular, that the rational theory of memory implemented in ACT-R is insightful in analyzing the pattern of recall in forming dependencies during parsing, for example, subject-verb dependency as in (1-a) and antecedent-reflexive dependency as in (1-b) (see also Lewis et al., 2006; Van Dyke, 2007; Wagers et al., 2009; Dillon et al., 2013; Kush et al., 2015; Lago et al., 2015; Jäger et al., 2017; Jäger et al., 2020; Nicenboim et al., 2018; Villata et al., 2018; Engelmann et al., 2019; Vasishth et al., 2019; Smith and Vasishth, 2020, among others).

(1) Example of dependencies, signaled by arrows.

a. Students rarely know the answer. (subject-verb)

b. The man told the woman about himself.

(antecedent-reflexive)

This brings us to the research topic of this paper, namely, studying whether other aspects in which parsing has to rely on memory can also be seen as fitting the research programme of the rational theory of cognition. In particular, during comprehension and production, native speakers have to continuously rely on their past knowledge of parsing rules. For instance, in (1), readers would not be able to comprehend the sentences correctly unless they recall that subjects normally precede verbs in English, verbs are followed by objects, English has prepositions (not postpositions) etc. From the perspective of the rational theory of memory, it is expected that the retrieval of parsing rules, such as these should follow the general considerations highlighted above, i.e., parsing rules should be retrieved in the order of their need probability and the order should monotonically correlate with latencies and accuracies. We will show that it is indeed possible to construct parsing on the basis of the rational theory of memory. The resulting model can furthermore correctly predict qualitative data in psycholinguistics (garden-path phenomena) and its predictions match behavioral measures in a psycholinguistic corpus (Natural Stories Corpus, Futrell et al., 2018).

The structure of the paper is as follows: in the following section, we briefly introduce the rational theory of memory as part of the cognitive architecture ACT-R. Next, we present transition-based parsers developed in computational linguistics and show how transition-based parsing and cognitive architectures can be combined. The cognitively informed parser is then evaluated on garden-path examples and data from Natural Stories Corpus. Finally, our research is briefly compared to related works in computational psycholinguistics.

2. MODELING MEMORY RETRIEVAL IN RATIONAL THEORY

Adaptive Control of Though-Rational assumes, true to its name, that various cognitive functions should be modeled as a case of rational theory of cognition. Here, we will focus on how memory and memory retrieval are formalized in ACT-R.

ACT-R assumes two types of memory: procedural memory and declarative memory. We focus here on the latter, the declarative memory, which is used for the storage of factual knowledge.[1]

The goal of the declarative memory system should be to recall a piece of information i that is needed to achieve the current goal. As is common in ACT-R, we will formalize pieces of information as chunks. These are attribute-value matrices, or, in the terminology of ACT-R, slot-value matrices. An example of a chunk, representing a simplified piece of information retrieved in the dependency in (1-a), is shown in (2). In this notation, slot names appear on the left side and their values on the right side. The chunk represents the knowledge that a plural subject of the form *students* was encountered and stored in memory.

(2) Example of a chunk stored and retrieved in (1-a):

$$\begin{bmatrix} \text{Form} & \textit{students} \\ \text{Function} & \text{SUBJECT} \\ \text{Number} & \text{PLURAL} \end{bmatrix}$$

Assuming that retrieving a chunk is costly and takes time, retrieval from memory must be constrained. An optimal retrieval system would prioritize those chunks that are more likely needed for the current goal. In general, it should hold that the recall of a piece of information, chunk i, adjusted by the value of the current goal G should not exceed the cost of the retrieval C.

(3) $P(i) \cdot G < C$

The task of the rational theory of memory is to find a reasonable estimation of $P(i)$. In ACT-R, it is assumed that $P(i)$, the probability that i is needed, is conditionalized on two sources of information: (i) the history H_i, that is, the past use of i;and (ii) the current context Q. We thus need to estimate $P(i|H_i, Q)$, which can be easily done using Bayes' rule. However, rather than expressing the conditional probability directly, it is standard in ACT-R to estimate log-odds. The estimation is expressed in (4) (i^c is the complement of i, i.e., $P(i^c)$ is the probability that i is not needed; Q, the current context, consists of indices j, which we call cues).

(4) $\log(\frac{P(i|H_i,Q)}{P(i^c|H_i,Q)}) = \log(\frac{P(i|H_i)}{P(i^c|H_i)} \cdot \frac{P(Q|i)}{P(Q|i^c)}) = \log(\frac{P(i|H_i)}{P(i^c|H_i)}) + \log(\prod_{j \in Q} \frac{P(j|i)}{P(j|i^c)})$

The inference in (4) makes the common assumption that while the probability that i is being needed is dependent on H_i and Q, the probabilities of the cues j in the current context Q are mutually independent and not dependent on the history H_i, conditional on i (see Anderson, 1991). The log-odds in (4) have a special status in ACT-R. They are called the activation of i, written as A_i. The activation consists of two parts: the history component [the first addend in(4)] and the context component [the second addend in(4)]. In ACT-R, the history component is called base-level activation, abbreviated as B_i, and the context component is called spreading activation, which we will abbreviate as S_i. We can rewrite the formula as follows[2]:

(5) ACT-R activation: $\log(\frac{P(i|H_i,Q)}{P(i^c|H_i,Q)}) = A_i = B_i + S_i$

Let us see how ACT-R estimates the history and the context components. Before doing so, we want to stress two things. First, the theory we are to discuss is generally and widely accepted by the ACT-R research community. Second, it is important to realize that the estimations of both the history component and the context component are not just arbitrary equations that happen to fit memory data. They should reflect the estimations that the

[1] The procedural memory system stores the knowledge exhibited in automatized, sequential behavior. This type of memory plays less important role in our models of parsing. We will briefly come back to it in section 4.

[2] ACT-R activation also standardly includes noise parameter. In (5), we ignore the noise parameter ϵ, so that activation is deterministic.

mind draws from the structure of the environment in order to arrive at the best estimation of $P(i)$ used in (3), just as a rational theory of cognition would make us expect. However, we will not present evidence that the following estimations are generalized from the structure of the environment since this has been done elsewhere (see Anderson, 1991).

The base-level activation B_i of a chunk is given in (6) and captures the fact that the probability that a chunk will be used next time decreases as a power function of the time since the last use, but it is also affected by the number of times that the chunk has been used. The base-level activation is expressed as the log of the sum of t_k^{-d}, where t_k is the time elapsed between the time of presentation k and the time of retrieval. d is a negative exponent (decay), a free parameter of ACT-R, often set at its default value of 0.5. "Presentation" in ACT-R means two things. Either it refers to the moment that the chunk was created for the first time (i.e., someone learns a particular fact), or the moment when the chunk was successfully recalled from declarative memory to be used in some context, after which it is stored in declarative memory again.

(6) Base-level activation: $B_i = \log\left(\sum_{k\in H} t_k^{-d}\right)$ (d – decay, free parameter)

The second element in the calculation of activation is given in (7). To keep the calculation manageable, some simplifying assumptions are introduced (see Anderson, 1991; Anderson and Lebiere, 1998). First, it is assumed that the cues j in the current context are independent of each other (and of the history H_i), conditional on i. Second, the denominator, which should be $P(j|i^c)$, is simplified into $P(j)$ since conditionalizing j on the irrelevant piece of information i^c should not affect probabilities significantly and can be ignored. The resulting log of probability ratios, $\log \frac{P(j|i)}{P(j)}$ is called the associative strength and is standardly abbreviated as S_{ji}. The equation also includes the weight W, which is a free parameter weighing the context component of the activation.

(7) Spreading activation: $S_i = W \cdot \log \prod_{j\in Q} \frac{P(j|i)}{P(j)} = \sum_{j\in Q} W \cdot \log \frac{P(j|i)}{P(j)} = \sum_{j\in Q} W \cdot S_{ji}$ (W – weight, free parameter)

Finally, the equation in (8) shows how ACT-R estimates the associative strength S_{ji}. This equation is only used if the cue j is predictive of the chunk i. If it is not, S_{ji} is set at 0. Simplifying somewhat, ACT-R assumes that a cue is predictive of a chunk if the cue appears as a value in the chunk.

(8) $S_{ji} = S - \log(fan_j)$ (S – maximum associative strength, free parameter)

S is the log of the size of the declarative memory, but commonly, it is hand-selected as a large enough value to ensure that S_{ji} is always positive (see Bothell, 2017). fan_j is the number of chunks in memory that have the cue j as its value. For discussion as to why (8) approximates $\log \frac{P(j|i)}{P(j)}$, see Brasoveanu and Dotlačil (2020). It might also help to notice that the formula S_{ji} also expresses the following intuition: the associative strength (and consequently, activation) will be large when j appears only in a few chunks since in that case j is highly predictive for each of those chunks; the associative strength will decrease if there are more chunks in declarative memory that carry j as its value (see Anderson, 1974; Anderson and Lebiere, 1998; Anderson and Reder, 1999 for empirical evidence).

Finally, the formula in (9) shows how A_i is related to the time it takes to retrieve a chunk from declarative memory, T_i. The relation between A_i and T_i is modulated by two free parameters, F, latency factor, and f, latency exponent.

(9) Retrieval time: $T_i = Fe^{-fA_i}$ (F, f – free parameters)

When both parameters are set at 1 (their default value), the retrieval time of a chunk i is simply the exponential of its negative activation, which is the reverse odds that the chunk i is needed in the current context [see (4)]:

(10) Retrieval time if $F, f = 1$: $T_i = e^{-A_i} = \frac{P(i^c|H_i,Q)}{P(i|H_i,Q)}$

It follows from (10) that the more a chunk is needed to achieve the current goal, the faster it will be retrieved.

Let us illustrate how all the equations are put together on an example from the introduction, the subject-verb dependency.

Assume we comprehend or produce the verb in (11-a) and want the retrieve the chunk *students* to resolve the subject-verb dependency. For the purposes of this illustration, we assume that the chunk is represented in memory as shown in (11-b), repeated from (2). The dependency needs to be resolved for interpretational purposes since listeners need to know who the agent of *know* is. It is also necessary for production purposes since speakers need to know what inflectional form the verb should have.

(11) a. Dependency:
Students rarely know the answer. (subject-verb)
b. The chunk to be retrieved:

$$\begin{bmatrix} \text{Form} & \textit{students} \\ \text{Function} & \textsc{Subject} \\ \text{Number} & \textsc{Plural} \end{bmatrix}$$

The activation of the subject *students*, its log-odds that the chunk is needed, consists of the base-level activation and the spreading activation. Suppose that 1 s elapsed since storing the chunk in memory and the chunk was not re-used. Then the base-level activation, calculated using the equation in (6), is:

(12) $B_{students} = \log(1^{-d}) = 0$

The spreading activation, calculated using the Equations (7) and (8), is given in (13). Note that the cues [subject], [plural] are the cues in the current context, i.e., we assume for this example that these two cues are present in the cognitive context when resolving the subject-verb dependency.

(13) $S_{students} = W \cdot S_{[subject],students} + W \cdot S_{[plural],students}$

Let us assume that the free parameter S is set at 1 and so is the weight W. Since both cues appear in the chunk *students*, we have to calculate both addends as:

(14) $$S_{students} = 1 \cdot (1 - \log(fan_{[subject]})) + 1 \cdot (1 - \log(fan_{[plural]}))$$

The only part that needs to be decided is the value of the fan for two cues. Let us assume that in the memory, there is no other subject and one other plural element. Then the calculation proceeds as follows:

(15) $$S_{students} = 1 \cdot (1 - \log(1)) + 1 \cdot (1 - \log(2)) \approx 1.31$$

Finally, we can calculate retrieval times as follows:

(16) $$T_{students} = F \cdot e^{-f \cdot (0+1.31)} \approx 0.27 \text{ s (if } F \text{ and } f \text{ set at 1)}$$

Based on the discussion of this example, one might note that the ACT-R model of declarative memory makes several predictions regarding retrieval times. Some of those are summarized in the bullet points below:

- The longer the time elapsed since a chunk was used last time, the lower base-level activation the chunk has. Consequently, chunks that were used a long time ago will be retrieved slower than chunks used recently.
- The less often a chunk was used, the lower base-level activation the chunk has. Consequently, chunks that are rarely used will be retrieved slower than chunks used often.
- The more a chunk matches cues of the current context, the higher the boost from spreading activation. Consequently, chunks with higher matches with cues should be retrieved faster.
- Increasing the fan of a cue will increase the time to retrieve an element. For example, imagine that more chunks with the value *plural* were stored in declarative memory. Then, the associative strength of any chunk with *plural* would be lower and consequently, it would take more time to retrieve such chunks.

To the extent that these qualitative predictions are confirmed, we have supporting evidence for the rational theory of memory as implemented in ACT-R. To the extent that quantitative predictions of the model can be well fit to retrieval data, we also have evidence that the estimates of the history and the context component of (4) in ACT-R are on the right track.

Various evidence has been collected showing that qualitative as well as quantitative predictions of the retrieval model in ACT-R are justified. Anderson (1991) and Anderson and Lebiere (1998) present supporting evidence from general cognitive tasks (independent of language). In psycholinguistics, Lewis et al. (2006), Jäger et al. (2017); Jäger et al. (2020), among others, summarize evidence that at least some cases of the retrieval of dependencies can be modeled as a case of ACT-R retrieval.

The goal of this paper is to apply the retrieval and memory model of ACT-R to a new domain. We will investigate how the rational theory of memory can model parsing knowledge and how the model of parsing can be embedded in ACT-R. We will show that once one thinks of parsing steps as chunks in declarative memory whose retrieval is driven by the same rules as other memory elements, the ACT-R model of memory becomes directly applicable to syntactic parsing. The activation that is associated with retrieved parsing steps can then be used to model the effect of context on processing, e.g., investigations that are mainly the domain of psycholinguistic parsing theories, such as the Surprisal Theory (Hale, 2001). To the extent that the resulting model of parsing makes correct quantitative and qualitative predictions, we construct evidence that processing difficulties observed during parsing can be approached from the vantage point of the rational theory of memory. The hypothesis explored in this paper is further investigated in Dotlačil (accepted)[3], which also studies how individual components of ACT-R retrieval system affect the retrieval of parsing steps and how the retrieval of parsing knowledge interacts with the retrieval of dependencies in processing.

In section 3, we introduce transition-based parsing and show how such parsers can be built as a case of declarative memory in ACT-R. In section 4, we show how the model can be linked to reaction time data and evaluate its qualitative and quantitative predictions.

3. TRANSITION-BASED PARSING

We introduce transition-based parsers and show that they can be, to a large extent, embedded in ACT-R and combined with the memory structures discussed in section 2. Such a combination directly delivers behavioral predictions to be tested in the following sections.

Transition-based parsers are parsing systems that predict transitions from one state to another, following decisions made by a classifier. Since the classifier plays a crucial role in this type of parsers, these parsers are also called classifier-based parsers.

Transition-based parsers are most commonly implemented for dependency grammars and arguably, they are most successful and widespread when constructing dependency graphs (Nivre et al., 2007). However, they have also been applied to phrase-structure parsing (Kalt, 2004; Sagae and Lavie, 2005; Liu and Zhang, 2017; Kitaev and Klein, 2018, a.o.). This paper also develops a phrase-structure transition-based parser. We introduce a shift reduce variant of the transition-based parsing algorithm, which is arguably the most common type of transition-based parser for phrase structures, and show how it can be understood in terms of memory systems discussed in the previous section.

3.1. Algorithm of Transition-Based Phrase-Structure Parsing

The parsing algorithm works with two databases, a stack of constructed trees $\mathcal{S}$ and a stack of upcoming words with their POS (part-of-speech tags) $\mathcal{W}$. When parsing begins, $\mathcal{S}$ is empty and $\mathcal{W}$ carries the upcoming words as they appear in the sentence, so that the first word appears at the beginning of the stack, followed by the second word, etc.

Parsing proceeds by selecting actions based on the content of $\mathcal{S}$ and $\mathcal{W}$. Every parsing step $\mathcal{P}$ is a function from $\mathcal{S}, \mathcal{W}$ to actions $\mathcal{A}$, that is, $\mathcal{P} : \mathcal{S} \times \mathcal{W} \rightsquigarrow \mathcal{A}$. In the variant of the parser that we consider, there are three actions that the parser can select:

- shift

[3] Dotlačil, J. (accepted). Parsing as a cue-based retrieval model. *Cogn. Sci.*

- reduce
- postulate gap

The first action, *shift*, pops the top element from the stack $\mathcal{W}$ and pushes it as a trivial tree onto stack $\mathcal{S}$. An element in $\mathcal{W}$ is a pair $\langle \text{word, POS} \rangle$, the tree moved onto the stack is just the POS tag with the terminal the actual word.

The second action, *reduce*, pops the top element (if the reduction is unary) or it pops the top two elements (if the reduction is binary) in the stack of constructed trees $\mathcal{S}$ and creates a new tree. If the reduction is unary, the new tree has just one daughter under the root, the tree that was just popped from the stack. If the reduction is binary, the newly created tree has two daughters, the two trees that were just popped from the stack. In either case, the newly constructed tree is pushed on top of the stack $\mathcal{S}$. It is assumed that all trees are at most binary, so no further reductions beyond binary reductions are necessary.

Finally, the third action, *postulate gap*, postulates a gap and resolves it to its antecedent. Not every parser in computational linguistics assumes this action, i.e., implemented parsers can proceed just by shifting and reducing (but see Crabbé, 2015; Coavoux and Crabbé, 2017a,b as examples of transition-based parsers that do consider gap resolution). We add gap resolution to our parser since ignoring gaps would make the parser less useful for psycholinguistics, which often studies the effect of gap resolution on processing.

There are several restrictions on the three actions. First, no shift can be applied when $\mathcal{W}$ is empty. When $\mathcal{S}$ is empty, no reduce can be applied and when it has only one tree, reduce binary cannot be applied. Finally, no more than two postulate gaps actions can be applied between two shifts. This last restriction ensures that the system does not fall into the infinite regress of gap postulation.

We illustrate the steps of the shift-reduce parser on a simple example: parsing of *a boy dances*. The phrase structure is shown in **Figure 1** and the parsing steps are:

1. Starting position: $\mathcal{S} = []$, $\mathcal{W} = [\langle \text{a, DT} \rangle, \langle \text{boy, N} \rangle, \langle \text{dances, V} \rangle]$
2. shift $\mathcal{S} = [\langle \begin{smallmatrix}\mathrm{DT}\\ |\\ \mathrm{a}\end{smallmatrix} \rangle], \mathcal{W} = [\langle \text{boy, N} \rangle, \langle \text{dances, V} \rangle]$
3. shift $\mathcal{S} = [\langle \begin{smallmatrix}\mathrm{DT}\\ |\\ \mathrm{a}\end{smallmatrix} \rangle, \langle \begin{smallmatrix}\mathrm{N}\\ |\\ \mathrm{boy}\end{smallmatrix} \rangle], \mathcal{W} = [\langle \text{dances, V} \rangle\rangle]$
4. reduce (binary) with label NP $\mathcal{S} = [\langle \begin{smallmatrix} & \mathrm{NP} & \\ / & & \backslash \\ \mathrm{DT} & & \mathrm{N} \\ | & & | \\ \mathrm{a} & & \mathrm{boy}\end{smallmatrix} \rangle]$, $\mathcal{W} = [\langle \text{dances, V} \rangle]$
5. shift $\mathcal{S} = [\langle \begin{smallmatrix} & \mathrm{NP} & \\ / & & \backslash \\ \mathrm{DT} & & \mathrm{N} \\ | & & | \\ \mathrm{a} & & \mathrm{boy}\end{smallmatrix} \rangle, \langle \begin{smallmatrix}\mathrm{V}\\ |\\ \mathrm{dances}\end{smallmatrix} \rangle]$
6. reduce (unary) with label VP $\mathcal{S} = [\langle \begin{smallmatrix} & \mathrm{NP} & \\ / & & \backslash \\ \mathrm{DT} & & \mathrm{N} \\ | & & | \\ \mathrm{a} & & \mathrm{boy}\end{smallmatrix} \rangle, \langle \begin{smallmatrix}\mathrm{VP}\\ |\\ \mathrm{V}\\ |\\ \mathrm{dances}\end{smallmatrix} \rangle]$
7. reduce (binary) with label S $\mathcal{S} = [\langle \begin{smallmatrix} & & & \mathrm{S} & \\ & & / & & \backslash \\ & \mathrm{NP} & & & \mathrm{VP} \\ / & & \backslash & & | \\ \mathrm{DT} & & \mathrm{N} & & \mathrm{V} \\ | & & | & & | \\ \mathrm{a} & & \mathrm{boy} & & \mathrm{dances}\end{smallmatrix} \rangle]$

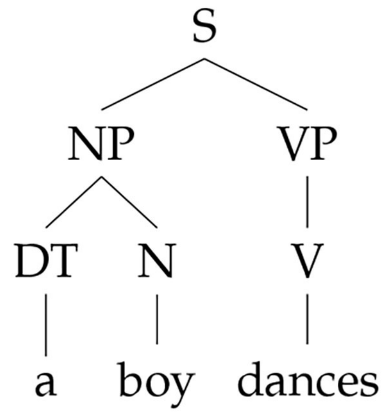


FIGURE 1 | Phrase structure of *a boy dances*.

In this illustrative example, we assume that the parser knows what the right phrase structure is and parses toward that structure. Of course, the crucial question is what happens when the phrase structure is unknown and the parser needs to predict what action to take. This is discussed in the next section.

3.2. Parsing Steps as Memory Retrievals

The parsing step has to decide which action (among *shift*, *reduce*, and *postulate gap*) should be taken, and, if *reduce* is selected, how should the reduction be done: should it be unary or binary and what should the root label of the newly constructed tree be?

We investigate the hypothesis that the parsing step can be treated as a case of memory retrieval. The past parsing steps form the declarative memory of the parser. The parser retrieves a parsing step (or parsing steps) from memory that has the highest probability of being needed given the current goal. The current goal, in turn, is to parse the sentence. From this perspective, parsing is just a particular instantiation of rational theory of memory and can be embedded in ACT-R. The activation of a parsing step, i.e., the log-odds that a step is needed, is calculated from the history component and the context component. The former is derived from the time elapsed since the step has been used and re-used, the latter is calculated based on the cues in the current context and the spreading activation from these cues to chunks in declarative memory.

While it might be possible to think of the context as complete trees in $\mathcal{S}$ and all information in $\mathcal{W}$, we will limit the amount of information in the two databases significantly. It will be assumed that $\mathcal{S}$ and $\mathcal{W}$ carry only some features about the trees and upcoming words, listed in (17). Thus, the parser itself never has a full snapshot of the phrase structure that it is deriving. It only carries some minimal, local information. The phrase structure can always be reconstructed through parsing steps the ACT-R agent (and, arguably, humans) took but there is no single snapshot in which all the information is available to the agent. This position is common in ACT-R parsing, see for example, Lewis and Vasishth (2005).

(17) Features representing context:

a. 1 upcoming word with its POS.
b. root labels of top 4 elements in $\mathcal{S}$
c. lexical head and the POS of the lexical head for top 4 elements in $\mathcal{S}$
d. left and right children in top 2 elements in $\mathcal{S}$
e. antecedent carried (yes or no), i.e., is there an antecedent (like a wh-phrase) that needs to be resolved through a gap and was not resolved yet?

The features should be familiar, maybe with the exception of the lexical head. The head is a terminal that projects its phrase (a verb is the head of a verb phrase, a noun is the head of a noun phrase etc.; see Collins, 1997 on head projection in computational parsers, which this work follows).

All the features in (17) spread activation to chunks stored in declarative memory, which in turn represent all parsing steps completed in the past. Recalling the right parsing step is a case of memory retrieval that follows the rules in section 2. Consequently, it is predicted that different parsing steps might require different amounts of time depending on the time it takes to retrieve them. Parsing steps with higher activations will be recalled faster than parsing steps with lower activations. Activations, in turn, are based on the base-level activation and spreading activation, i.e., the ACT-R estimates of the history and the context component in calculating the need log-odds of a chunk.

4. MODELING READING DATA

We present an implementation of the model of sentence parsing built on the rational approach to memory and discuss two case studies testing the implementation.[4] Section 4.1 introduces the model. Section 4.2 investigates whether the parser can predict processing difficulties for selected garden-path phenomena. Section 4.3 investigates whether the parser can be used to model self-paced reading time data from the Natural Stories Corpus (Futrell et al., 2018).

4.1. Parsing Model

We assume that a declarative memory consists of chunks that represent correct past parsing steps. These chunks are collected from the data in the Penn Treebank (PTB) (Marcus et al., 1993). As is standard, we split the section of the PTB data as follows: all the sections up to and including section 21 are used to train the parser, i.e., to collect the correct parsing steps; section 22 is used for development; section 23 is used to test the accuracy of the parser. Before training we pre-process and prepare the phrase structure by (i) transforming phrases into binary structures in the way described in Roark (2001) (see Roark, 2001; Sagae and Lavie, 2005 on the reasons to do), (ii) annotating phrases with head information, (iii) removing irrelevant information (coreference indices on phrases), (iv) lemmatizing tokens so that lexical heads are stored as lemmas, not as inflected tokens.

Parsing novel sentences consist of recalling the needed chunks, i.e., parsing steps collected from the PTB, from declarative memory. The recall is driven by the activation of the chunks. To calculate the activation of each chunk, formulas in section 2 are applied. We assume that the parser will recall the three chunks with the highest activations and choose the action that is the most common one among those three chunks.[5] The parser repeats this procedure until it encounters *shift*. At that moment, the parser is done with integrating word n and can move its attention to word $n+1$. The activations collected during the parsing are averaged. They can be used to directly predict processing difficulties, as in section 4.2, or used to calculate reaction times, as in section 4.3.

The activation of a chunk is the sum of base-level activation and spreading activation. For base-level activation, we need to estimate how often a parsing step has been used in the past and how much time elapsed. The estimation comes from the frequency of parsing steps, collected from the PTB. The frequencies can be transformed into base-level activation according to the procedure described in Reitter et al. (2011), see also Dotlačil (2018) and Brasoveanu and Dotlačil (2020). The procedure is summarized in **Appendix A**.

The spreading activation is calculated based on the match between values in chunks and features in the current cognitive context at the moment when the parsing step is recalled. The features are summarized in (17).

4.2. Case 1: Garden-Path Sentences

We start the investigations of the predictions of the parser by considering selected garden-path phenomena, taken from previous literature (Bever, 1970; Frazier, 1978; Marcus, 1978; Gibson, 1991; Pritchett, 1992).

We model the predictions for the pairs in (18)–(21). In each pair, the (a) sentence is a classical example of a garden path. The (b) sentence carries the same or almost identical interpretation as the garden path. However, since the disambiguation takes place early in (b) sentences, no garden-path effect is observed.

(18) a. The horse raced past the barn fell.
b. The horse which raced past the barn fell.

(19) a. While she mended the sock fell on the floor.
b. While she mended, the sock fell on the floor.

(20) a. He convinced her tired children are noisy.
b. He convinced her that tired children are noisy.

(21) a. She gave the boy the dog bit a bandage.
b. She gave the boy that the dog bit a bandage.

We want to see how the parser parses (18)–(21) and what activation values are predicted for the words in the sentences. We expect that the activation of the retrieved parsing steps should be lower for garden-path cases [(a) examples] compared to the (b) cases. This should happen at the target words, the words at

[4] The code for the model is available here: https://github.com/jakdot/parsing-model-and-a-rational-theory-of-memory.

[5] Using three chunks, rather than a single chunk, to inform about the action, makes the parser less error-prone and sensitive to outliers. Adding more than three chunks does not improve the accuracy of the parser. We briefly discuss the accuracy of the parser in section 5.

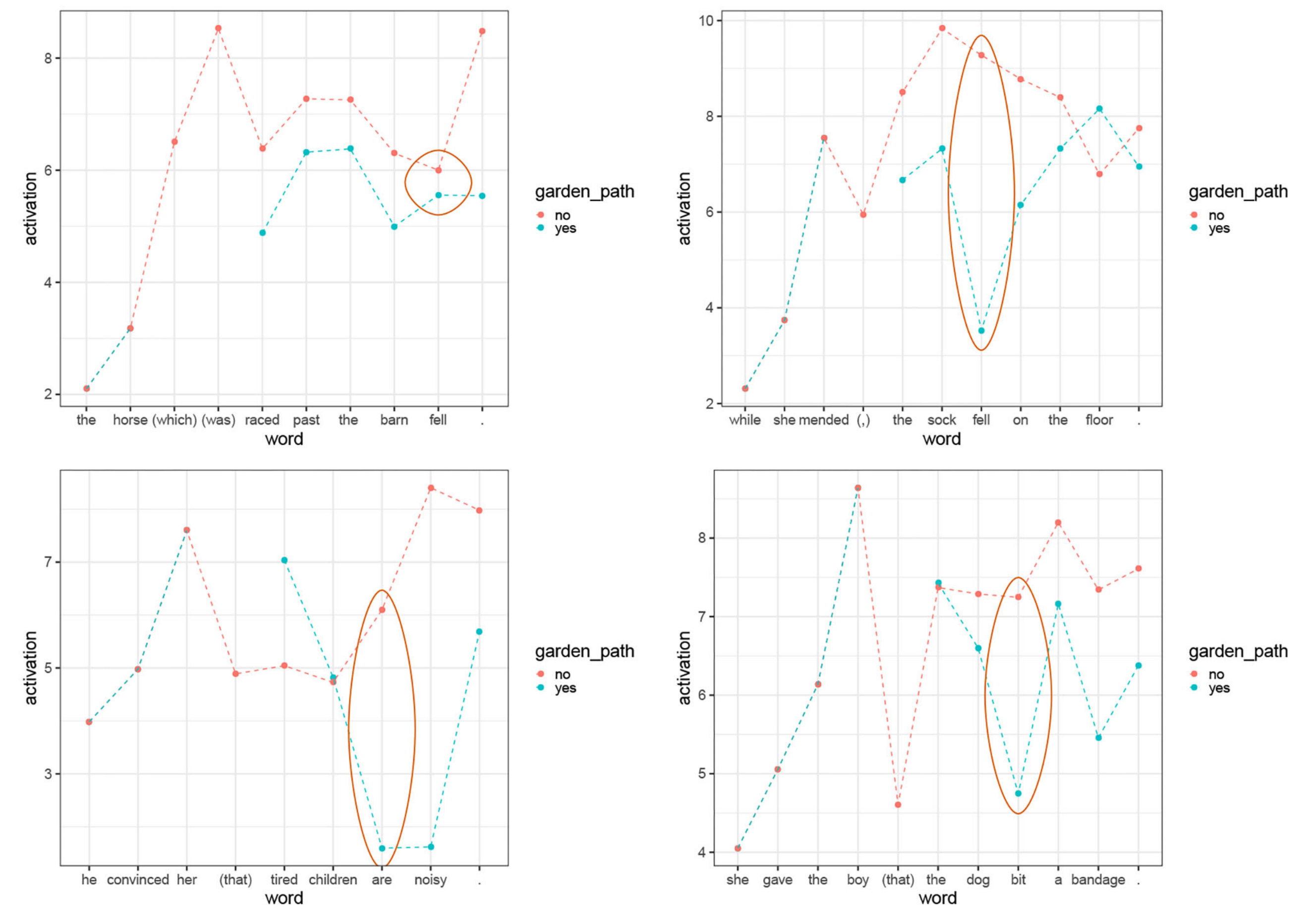


FIGURE 2 | Activations per word for sentence pairs (18)–(21). The yellow bars represent the activations in the sentences that disambiguate early. The blue bars are the activations of the garden-path sentences. The ellipses highlight the activations on the words that trigger the garden-path effect.

which processing difficulties should be located in garden-path sentences. The target words are *fell* for (18), *fell* for (19), *are* for (20), and *bit* for (21). We expect the activation to decrease for garden-path sentences at the disambiguation point because the base-level activation of parsing steps should be low (garden-path sentences should not be very frequent in natural data) and because the spreading activation should be low (garden-path sentences move us to the syntactic context that cannot find a good match in the past parsing steps hence not many cues will spread activation).

The activations per word are graphically summarized in **Figure 2**. For this calculation, we assumed default values of free parameters and we set the maximum associative strength, *S*, from the Equation (8) at 20. As we can see, the (a) examples show lower activations than (b) examples at the target word. Furthermore, with one exception, the classical pair in (18), the difference not only goes in the predicted direction, but it is large at the critical word (2 points of activations or more). Note also that the contrast in activations usually spills over to the following words. Since lower activations translate into higher retrieval times we see that the model is able to predict increased reading times in garden-path sentences. Furthermore, chunks with lower activations have higher probability of retrieval failures (Anderson, 1991; Anderson and Lebiere, 1998). Consequently, the decrease in activation can explain processing difficulties in general, in particular, the failure to provide a correct parse for garden-path sentences (Pritchett, 1992).[6]

The phrase structures built by the parser are correct for all the (b) examples with the exception of (21-b) in which the parser wrongly attaches the noun phrase *a bandage* inside the relative clause. For the (a) sentences, the parser struggles at the disambiguation point and the parsing steps that it retrieves are not adequate phrase structures. It provides phrase structures that are incorrect but in which locally built phrases are combined in a plausible way. The incorrect parses for the (a) sentences were selected by the parser because they had the highest activations

[6] The activations are also very low at the beginning of each sentence, irrespective of whether we deal with a garden-path sentence or not. This is an artifact of the selected model. Most cues for spreading activation come from the tree structures already built. Of course, nothing or almost nothing has been built at the beginning of a sentence, hence there are few cues at the start and consequently, spreading activation is low. It is possible to avoid this property of the model, for example, by not counting just matches in built trees, but also matches by the position in a sentence as cues that can boost activations.

in the context. This means that if we restricted our attention to *correct* parses, the contrast between garden-path sentences and their (b) counterparts would be even larger at the critical words.

One pair in which the contrast between the (a) and the (b) examples goes in the right direction but is so small that the activation contrast is almost irrelevant is the case (18). The fact that the garden-path sentence almost does not differ from the baseline might be caused by the fact that we do not model discourse and semantic phenomena, while Crain and Steedman (1985) showed convincingly that this garden path is sensitive to its context. Since the model does not take context into account, it misses out on discourse effects affecting activations.

To conclude, we see that the contrasts in the activation of retrieved parsing steps can be tied to processing difficulties and predict cognitive difficulties observed in garden-path sentences.

4.3. Modeling Corpus Reading Data

4.3.1. Introduction

We study the predictions of the parsing model for the Natural Stories Corpus (NSC, Futrell et al., 2018). The NSC is a corpus containing 10 English narrative texts with 10,245 lexical tokens in total. The texts were edited to contain various syntactic constructions, including constructions that are very rare. The corpus was read by 181 English speakers using a self-paced reading moving-window paradigm and the self-paced reading data were released along with the texts. Furthermore, all the sentences were annotated according to PTB notational conventions by the Stanford Parser (Klein and Manning, 2003) and checked and hand-corrected. The fact that the NSC has a plethora of syntactic constructions and includes manually controlled PTB-compatible syntactic parses makes the corpus particularly usable for the computational modeling of parsing.

4.3.2. Reading Model

The parser as specified in sections 2 and 3 and implemented in section 4.1 will be used to model the self-paced reading of sentences in the corpus. However, to make sure that the parser does not go astray, at every word, we collect the correct parse provided by the NSC. This correct parse is used as the context for retrieval: based on this parse, the parser attempts to retrieve a parsing step from declarative memory. The declarative memory consists of parsing steps collected from the PTB, see section 4.1 for details. Then, the average activation of the retrieved chunks is recorded. After the parse for the word is finished, the correct parse is considered again for the next word. That means that the parser will have the correct syntactic structure at every word and will use the correct context for retrieval.

Importantly, in a self-paced reading task, readers do much more than just retrieving and applying parsing steps. It seems uncontroversial that a model simulating self-paced reading should, at least, attend visually to word n, retrieve lexical information on that word, parse, press a key (to reveal the next word) and move visual attention to the next word, word $n+1$. We will add these parts and combine them with the parsing model to construct a more realistic model of reading. The added parts are not created *ad hoc*, they are based on the (simplified) models of visual attention and self-paced reading (Anderson and Lebiere, 1998; Brasoveanu and Dotlačil, 2020).

The sequential behavior like reading is modeled in ACT-R as a case of procedural knowledge, which sequences processes, such as the ones mentioned above and calls various sub-modules (visual, declarative memory, motor module) to carry out task specifics. The processes are linked together and controlled by the procedural system. In **Figure 3**, we represent the processes as boxes, which the procedural system lets fire in the order as signaled by the arrows. It is assumed that these processes are repeated on every word. Firing each of these processes takes the same amount of time in the procedural system, specified in (22).

(22) Time to start process: r (r – free parameter)

In addition to that, submodules involved in a process incur extra processing time based on their own properties.

The process *attend word* visually attends to a word. To keep the model simple, we will assume that visual attention takes a fixed amount of time, in line with basic models of ACT-R (Bothell, 2017). It is assumed that attending takes 50 ms, the default value of process firing in ACT-R. Since visual attention is modeled as a fixed amount of time, any fit of the model to the data must be driven only by retrieval processes: the retrieval of lexical information or the retrieval of syntactic information, which are the only two retrieval processes considered in this paper.

The processes *press key* and *move visual attention* interact with the motor module and the visual module, respectively. *Press key* is modeled assuming the basic model of motor actions in ACT-R, which is inspired by the EPIC cognitive architecture (Bothell, 2017). It is assumed that readers have their fingers ready on the key to be pressed. In that case, the simple model of motor actions in ACT-R, followed here, postulates that it takes 150 ms to press the key. Crucially, during this time, the procedural system is

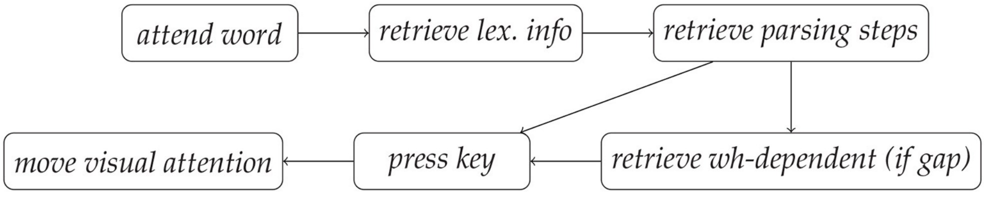


FIGURE 3 | Sequential model of reading on one word. Each box represents one process. Arrows show the order in which the processes fire. There are two arrows from *retrieve parsing steps* because *retrieve wh-dependent* is only triggered when a gap is postulated by the parser.

free to carry out any other actions in the sequential model. That means that moving visual attention can happen concurrently with key presses.

The processes *retrieve lex. info, retrieve parsing steps* and *retrieve wh-dependent* are the processes that depend on declarative memory. All processes take at least r amount of time each. Aside from that, they will also take some extra time: the amount of time needed to retrieve a chunk from declarative memory. All relevant equations to calculate retrieval time have been given in section 2. Let us repeat that the retrieval time is a function of activation of a retrieved chunk and modulated by two free parameters (23-a). Activation is calculated as the sum of base-level activation and spreading activation (23-b).

(23) a. $T_i = Fe^{-fA_i}$ (F, f – free parameters)
b. $A_i = B_i + S_i$

The base-level activation and spreading activation have been discussed in detail in section 2. Recall that these activations had several free parameters: decay d, weight W, maximum associative strength S. We set the first two parameters at their default value 0.5 and 1, respectively (see Anderson and Lebiere, 1998; Bothell, 2017). The maximum associative strength is set at 20 to ensure that associative strength is always positive (see Bothell, 2017). Furthermore, r, the time for the procedural system to fire a process, see (22), is set at 33 ms, as this was found in Dotlačil (accepted)[3] to be the median value for an ACT-R model that simulates reading in a self-paced reading experiment. Finally, the time component needed to calculate base-level activation is calculated in the same way for the retrieval of lexical information (words) and the retrieval of parsing steps. It is derived from the frequencies of words and parsing steps, based on the procedure summarized in **Appendix A**.

This leaves us with two parameters needed to estimate retrieval times from activations: F and f. These will be estimated with a Bayesian modeling procedure.

4.3.3. Bayesian Modeling

There are two parameters that we need to model to fit the reading model to the corpus data: F and f. We will estimate them using Bayesian techniques (see Dotlačil, 2018, Brasoveanu and Dotlačil, 2018, Brasoveanu and Dotlačil, 2019, Brasoveanu and Dotlačil, 2020; Rabe et al., 2021 for other examples of combining Bayesian modeling with ACT-R cognitive models; see Weaver, 2008; Dotlačil, 2018 for arguments why this is necessary).

We assume the structure of the model as shown in **Figure 4**. In this graph, the top layer represents priors, the bottom part is the likelihood. **ACT-R**(F;f) is the ACT-R cognitive model of reading described in the previous section. When run and supplied with F and f values, it outputs latencies per word. The latencies of the model are then evaluated against the data assuming the likelihood is a normal distribution (measured in milliseconds) with standard deviation 20 ms (the bottom part of the graph). The actual data that we try to model are mean reading times (mRT) per word in the self-paced reading corpus. We select the first two (out of 10) stories for the estimation of the parameters. In each story, there is an observable effect of speed-up as readers progress beyond the first few sentences. Since our model does not represent that, we decided to remove the first 10 sentences from each story. Furthermore, we model mRTs only starting at the second word and ending at the second to last word in each sentence since the first and last words tend to be outliers due to starting and wrap-up effects. Besides, the starting words are also outliers in our model (see also text footnote 6).

The following prior structure for the parameters is assumed:

- $F \sim Gamma(\alpha = 2, \beta = 10)$
- $f \sim Gamma(\alpha = 2, \beta = 4)$

Given these priors, the values in the range 0–1 are most likely but extremely low values are penalized. The priors for the parameters have the mean values of 0.2 and 0.5, respectively. These priors take into account previous findings that when F and f are estimated on language studies, including reading data, they are below 1 but usually not exceedingly small and F tends to be smaller than f (Brasoveanu and Dotlačil, 2018, 2020).

The estimation of parameters was done using PYMC3 and MCMC-sampling with 1,200 draws, 2 chains and 400 burn-in draws. The sampling chains converged as witnessed by the Rhat value (Rhat for F was 1.036; Rhat for f was 1.028).

4.3.4. Results

The mean, median and standard deviation values for the latency factor (F) and latency exponent (f) of the posterior distributions can be seen in **Table 1**.

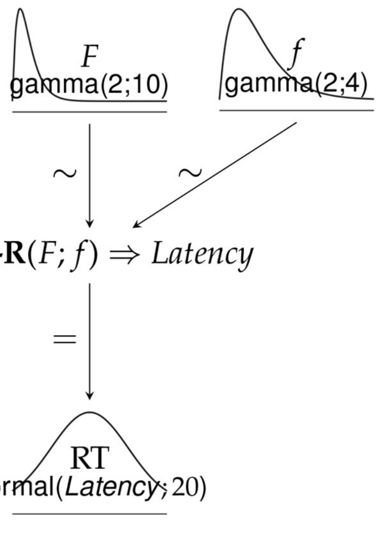


FIGURE 4 | Bayesian model for parameter estimation of Natural Stories Corpus.

TABLE 1 | Estimated parameter values.

	Mean	Median	Std
F	0.0139	0.0139	0.001
f	0.661	0.655	0.068

TABLE 2 | The linear model with Predictive RT as the only independent variable.

	Estimate	SE	*t*-Value	*p*-Value
Predictive RT	0.993	0.0024	415.5	$p < 0.0001$

TABLE 3 | The linear model with Intercept and Predictive RT.

	Estimate	SE	*t*-Value	*p*-Value
Intercept	248.4	12.7	19.57	$p < 0.0001$
Predicted RT	0.220	0.040	5.55	$p < 0.0001$

The mean and median values for F match the estimate in previous Bayesian + ACT-R reading models (Brasoveanu and Dotlačil, 2018, 2020). However, the estimate of f is greater than in previous reading studies. It is possible that this is because the previous reading studies did not take the retrieval of parsing steps into account, focusing only on lexical retrieval and that the previous studies mainly looked at experimental data, while this study models corpus data.

To further investigate the model, we check samples from its posterior distribution of predicted RTs (i.e., RTs that the reading model predicts using the posterior distribution of the fitted parameters). We expect that these should correlate with observed meanRTs. This is because the model simulates two steps in processing, namely, lexical retrieval and parsing. Lexical retrieval is affected by the activation of words, which depends on frequency and causes less frequent words take more time to retrieve than more frequent words (see **Appendix A** for the estimation of base-level activation based on frequency). Syntactic retrieval is affected by the activation of parsing steps, which is the sum of base-level activation and spreading activation. The base-level activation is related to frequency just like word activation and makes less frequent parsing steps take more time to retrieve (see **Appendix A**). Furthermore, if a reader is in a rare syntactic context (i.e., an uncommon syntactic construction), they are less likely to find parsing steps in the past that would provide a good match. This results in a decreased spreading activation, which again affects reading times. Finally, the parser models wh-dependency and retrieving wh-words will increase reading times when the wh-words are far away from the gap site, due to the decrease in their activation.

We now inspect the predictions of the model. First, we run a simple linear model with predicted RTs per word (i.e., RTs that the reading model predicts using the posterior distribution of the fitted parameters) as the independent variable and observed mean RTs as the dependent variable. We see in the summary of the linear model given in **Table 2** that the Maximum Likelihood Estimate (MLE) of predicted RT is very close to 1, i.e., in the best linear fit between the predicted and observed RT, the increase of 1 ms in predicted RTs corresponds to the increase of 1 ms in observed RTs. **Table 3** shows the fit of the intercept + predicted RT linear model. As we see, predicted RTs are a highly significant predictor for observed mean RTs.

TABLE 4 | A full linear model for RTs in the NSC.

	Estimate	SE	*t*-Value	*p*-Value
Intercept	258.5	17.2	15	$p < 0.0001$
Story	7.3	1.3	5.5	$p < 0.0001$
Zone	−3.9	0.87	−4.5	$p < 0.0001$
Position	−2	0.7	−3	0.003
Story:Zone	−3.3	1.34	−2.5	0.01
Zone:Position	1.65	0.73	2.25	0.02
Nchar	16.3	3.79	4.3	$p < 0.0001$
Log(Freq)	0.21	0.52	0.4	0.7
Nchar:log(Freq)	−0.68	0.22	−3.1	0.002
Log(Bigram)	0.25	0.63	0.4	0.7
Log(Trigram)	−0.88	0.48	−1.82	0.07
Predicted RT	0.15	0.04	3.66	0.0003

The finding in **Table 3** shows that our reading model can capture some aspects of self-paced reading data. However, we want to see that this modeling capability goes beyond what surface features of a text, i.e., position, word length or string frequencies, known to influence reading times, can account for. For this reason, we consider a more complex model, summarized in **Table 4**. The confounds we consider are the following: (i) Story (story 1 or story 2, the former being the reference level), (ii) Zone (the word position in its story, z-transformed), (iii) Position (the word position in its sentence, z-transformed), (iv) the interaction of Story × Zone, (v) the interaction of Zone × Position, (vi) Log(Freq) (log-unigram frequency), (vii) Nchar (the length of the word in number of characters, z-transformed), (viii) the interaction of Nchar × Log(Freq), (ix) Log(Bigram) (log bigram probability), (x) Log(Trigram) (log trigram probability). Frequencies and bigram and trigram probabilities are provided in the NSC. Most of the confounds that we input are considered when evaluating computational psycholinguistic models on corpus data (Demberg and Keller, 2008; Boston et al., 2011; Hale, 2014, among others). We see that even after adding the confounds, predicted RTs remain a significant predictor and the effect goes in the expected (positive) direction ($t = 3.66$, $p = 0.0003$). Thus, our parsing model captures aspects of reading data that are not captured by surface-like factors, e.g., string frequencies, position, number of characters and the interaction of those.[7]

To further inspect the predictions of our Bayesian + ACT-R model and the actual data, we split the predicted and observed data sets into deciles based on trigrams, word frequencies and the actual observed mean RTs. The graphical summaries per decile are given in **Figure 5**. For trigram probabilities and unigram frequencies, we see that the data predicted by the model follows the trend of the actual data and the mean predicted RT is generally close to the observed mean RT in each decile (with the

[7] It might seem surprising that the effect of log-frequency is not significant in **Table 4**. This is because predicted RTs correlate with frequency and because we also include the Nchar × Log(Freq) interaction. In a simpler model lacking the interaction, Log(Freq) is significant and goes in the expected direction.

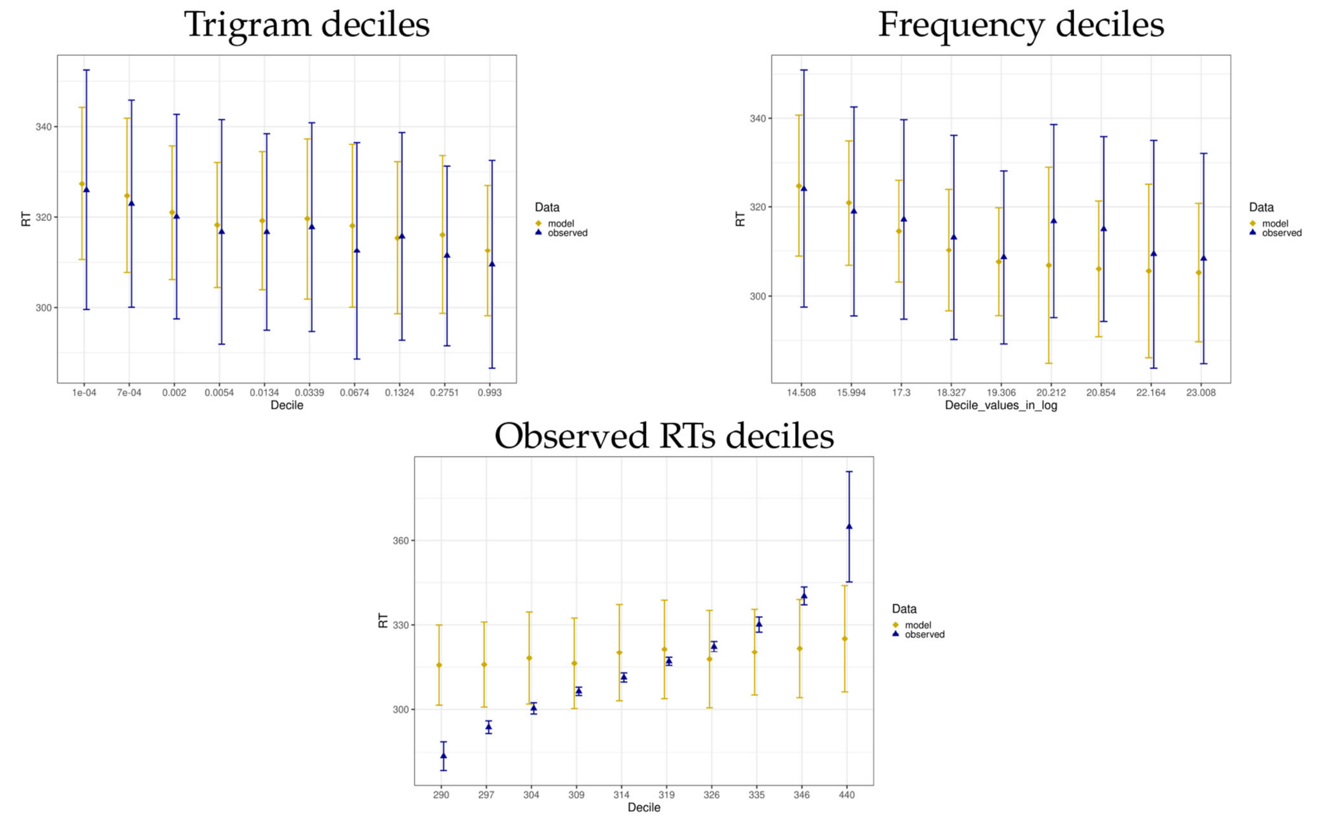


FIGURE 5 | Mean and standard deviation summaries of model and data split per trigram, frequency and observed mean RT deciles. The x-axis label shows the upper cut-off point per decile (given in log in case of Frequency). In case of Frequency, only 9 deciles are present. This is because a single word (*the*) spans the top two deciles.

slight divergence in the 6th and 7th decile of Frequency, for which the model assumes mean RTs faster by 10 and 9 ms). In case of the last graph, in which data are split by observed mean RT deciles, the model copies the linear trend of the data, i.e., predicted mean RTs increase per decile. This trend is also confirmed by a highly significant Pearson correlation between predicted mean RT and observed mean RT split by decile ($r = 0.88, p < 0.001$). However, compared to the actual data, the model has much less extreme values on both ends of the decile spectrum and as the result. While it captures the linear trend in the data, it overestimates RTs in low deciles and underestimates RTs in high deciles.

Finally, we compare the predictions of our model to another ACT-R model of reading, presented in Boston et al. (2011). The model of Boston et al. (2011) models the retrieval of dependencies using the assumptions of the ACT-R rational memory. In contrast to our work, Boston et al. (2011) do not model structure building, i.e., the knowledge of parsing steps, using the ACT-R memory.[8] For this reason, we would expect that the time predictions of our model remain a significant predictor when the predictions of Boston et al. (2011) are included in a linear model of the NSC reading data. To check this, we constructed time predictions of the ACT-R reading model of Boston et al. (2011) for the NSC sub-corpus that we used for testing (the first two stories).[9] We tested the ACT-R retrieval model of Boston et al. (2011) with various levels of beam-width k ($k = 1, 3, 9, 20, 50, 100$), where k specifies the number of syntactic parses built in parallel. It turned out that model predictions with low numbers of k ($k \leq 20$) did not show a significant effect on our NSC reading data. For $k = 50$ and $k = 100$, the model showed a very wide range of predicted reading times (from 50 to 5,000 ms). When we removed predictions beyond 2,000 ms, the model predictions were significant ($\beta = 0.005, t = 3.1$). Crucially, the predictions of our model, PREDICTED RT, were also significant ($\beta = 0.2, t = 4.1$). This supports the position that our model captures the properties of reading missing in an ACT-R model that only simulates the retrieval of dependencies using the ACT-R theory of memory.

4.4. Summary of the Results

We provided empirical evidence for the parsing model that is built on the assumptions of the rational theory of memory proposed in Anderson (1991) and embedded in ACT-R. Two

[8] See also section 5 for comparisons of our model to related works.

[9] We used the code available at https://conf.ling.cornell.edu/Marisa/. To generate predictions, we made use of the default English training corpus, Brown. We would like to thank an anonymous reviewer, Marisa Boston and John Hale for discussion and help.

types of evidence were collected. First, processing difficulties of garden-path phenomena correspond to activation drop of retrieved parsing steps. Second, the parsing model, combined with some basic assumptions about reading, has been used to model self-paced reading data from the Natural Stories Corpus. After fitting two parameters, the resulting model showed a highly significant correlation with observed reading times. The model was able to capture aspects of the reading data that were not captured by other, low-level factors like string frequencies, position or word length. We leave it open which particular aspects of the rational memory might play a dominant role in model fitting, in particular, which of base-level activation and spreading activation was crucial in our finding.

5. COMPARISON TO RELATED WORKS

5.1. Parsers in Computational Psycholinguistics

It is possible to split the computational psycholinguistic approaches to parsing into two types, experience-based theories and memory-based theories. In experience-based theories, it is studied how past experience with syntactic structures affect parsing, most often because of expectations readers form during sentence processing. A popular framework belonging to experience-based approaches is Surprisal Theory (Hale, 2001; Boston et al., 2008, 2011; Levy, 2008, 2011; Smith and Levy, 2013, among others). In memory-based theories, it is studied how the bottleneck of memory affects storage and retrieval during processing. Dependency Locality Theory is an example of a memory-based explanation of processing difficulties (Gibson, 1998), and so are theories studying the effect of integration and recall of information from parsing stacks (Van Schijndel and Schuler, 2013; Shain et al., 2016; Rasmussen and Schuler, 2018). Another memory-based theory is the activation-based approach to dependency resolution, often implemented in ACT-R (see Lewis and Vasishth, 2005; Lewis et al., 2006).

The two types of approaches offer different advantages. While experience-based theories can account for processing difficulties tied to construction frequency and local ambiguities (garden-path phenomena), memory-based approaches are used to capture locality effects. However, the integration of the two accounts into one framework is arguably still an open issue. In most accounts, two research lines are simply put together as two different and separated parts of a model (Demberg and Keller, 2008; Boston et al., 2011; Levy et al., 2013; Van Schijndel and Schuler, 2013).

In contrast to the just cited approaches, the current account builds a single analysis of experience-driven and memory-driven processing difficulties. It is assumed that both difficulties are driven by memory limitations in retrieval, as predicted by rational memory systems. The only difference is what is being retrieved: memory-driven processing difficulties arise when the memory system tries to recall a recently constructed phrase/element to satisfy dependency and encounters problems; experience-driven difficulties arise when *the same* memory system tries to recall a parsing step and encounters problems. The first type of difficulties has been well-investigated in computational psycholinguistics in general and in the sub-field of modeling using cognitive architectures like ACT-R in particular (see Lewis and Vasishth, 2005; Lewis et al., 2006; Dubey et al., 2008; Reitter et al., 2011; Engelmann et al., 2013; Engelmann, 2016; Vogelzang et al., 2017; Brasoveanu and Dotlačil, 2020). Crucially, the second type of difficulties has been investigated much less from this perspective. This paper can be seen as an attempt to enhance our understanding on this topic. In this respect, this paper advances current ACT-R analyses of reading, notably Lewis and Vasishth (2005), which do not generalize parsing, relying instead only on hand-coded rules for selected syntactic constructions. An account that offered one framework for both types of processing difficulties has been developed in Futrell and Levy (2017), which provides a computational-level analysis (in contrast to the algorithmic-level analysis developed here) and comes to the problem from the opposite direction. Futrell and Levy (2017) provides a single analysis to processing difficulties by expanding Surprisal Theory with an extra component (noisy-context) to capture memory-driven difficulties.

In works within cognitive architectures, a close affinity can be found between this account and the models of Reitter et al. (2011) and Hale (2014).

Unlike Reitter et al. (2011), the current account does not model production, but focuses on comprehension, and it does not study priming of syntactic rules. Furthermore, Reitter et al. (2011) developed a model to generate qualitative effects in priming, while this paper shows that, through the application of ACT-R models in a Bayesian framework, it is possible to model quantitative data patterns. In fact, the presented approach makes it possible to develop a model in which the reading profile of experience-driven processing difficulties quantitatively constrains the reading profile of memory-driven processing difficulties, since both phenomena are modeled in the same way and modulated by the same free parameters. This has also been assumed in this paper (e.g., the parser for the Natural Stories Corpus assumes the same model for retrieval of wh-dependency, lexical retrieval and the retrieval of parsing steps). However, a close investigation of the interaction of different cases of retrieval in the same model goes beyond the scope of this paper. See Dotlačil (accepted)[3] for more work in this direction.

Finally, Hale (2014), Chapters 7 and 8, derives experience-driven processing difficulties as a case of (failed, less likely) production compilation/cohesion. This position is not incompatible with the current account, in fact, it complements it. While this work studies the role of declarative memory on parsing, Hale (2014) focuses on the role of procedural memory on parsing. The latter position has arguably been investigated in much more detail in psycholinguistics and in ACT-R than the former position since the seminal works of Lewis (1993) and Lewis and Vasishth (2005). In this respect, the current proposal can be seen as breaking with this tradition. However, both types of memory are crucial for ACT-R as well as other cognitive architectures (see Anderson, 2007) and their interaction is needed to account for complex learning patterns (Lebiere, 1999; Taatgen and Anderson, 2002). It is likely that a highly non-trivial task, such as syntactic structure-building will benefit from investigations that do not limit its investigation to the procedural memory system.

5.2. Transition-Based Parsing in Computational (Psycho)linguistics

Transition-based parsers were a popular choice of parsers in computational linguistics, especially for dependency grammars (see Nivre et al., 2007; Zhang and Clark, 2008; Kübler et al., 2009). One advantage of transition-based parsers over graph-based parsing and grammar-based parsing is that they are fast, incremental and they allows for rich feature representations (Nivre, 2004; McDonald and Nivre, 2011). Transition-based parsers have also been applied to phrase-structure parsing (Kalt, 2004; Sagae and Lavie, 2005). The recent neural transition-based parsers for phrase-structure building have the F1 value around 95% on the PTB section 23 (Liu and Zhang, 2017; Kitaev and Klein, 2018). Transition-based parsers have also been used in computational psycholinguistics to model EEG data (Recurrent neural network grammars; Dyer et al., 2016; Hale et al., 2018) and reading data (Boston et al., 2008; Rasmussen and Schuler, 2018).[10]

While the high accuracy of the state-of-the-art transition-based parsing is encouraging, as it suggests that this line of parsing can eventually be used to create a very accurate parser, we should note that our parser is nowhere near this accuracy performance. When tested on the section 23 of the Penn Treebank, the parser shows Label Precision as 70.2, Label Recall as 72.4, F1 as 71.3. When we restrict attention to sentences of 40 words or less, as is common, Label Precision is 73.7, Label Recall is 75.9, and F1 is 74.8.[11]

There are arguably several reasons for the low performance. First, it has been found that one of the disadvantages of transition-based parsers when compared to another class of data-driven parsers, graph-based parsers, is that they get worse with increase in sentence length and increase in dependence, i.e., error propagation (McDonald and Nivre, 2011). Traditional transition-based parsers, including the parser in this paper, explore just one path. They have to greedily select what path they will follow and stick to it until the end of the sentence. Thus, early mistakes will propagate the error throughout the whole sentence. Better transition-based parsers mitigate this type of mistake through beam search or methods to recover from errors. While the adaptation of these methods could be investigated for psycholinguistics, we are not primarily interested in the best accuracy of the parser on the complex Penn Treebank sentences, but in parsing that is human-like. It is known that a human processor also shows error propagation in parsing, as witnessed by the fact that readers struggle to recover from garden path sentences the longer the wrong interpretation can be held (e.g., Frazier and Rayner, 1982). Thus, it is not a priori clear that error propagation should be avoided.

Another reason why we see a low accuracy is that the parser assumes a very straightforward relation between memory instances and a parsing step. A parsing step is simply stored in declarative memory.[12] This is in contrast to complex training methods commonly assumed in current neural parsers. Relatedly, current computational parsers assume a much richer feature system. They are enriched by vector space models representing lexical information and syntactic information is usually encapsulated in 200 or more features, while our parser has 19 features.

In any case, it might be worth pointing out that even though the accuracy of the parser is not very high, it suffices for the research presented in this paper. The chosen examples in section 4.2 are correctly constructed by the parser when they do not lead to garden path and the parser in section 4.3 was at the end of every step (word) corrected to match the gold standard provided in the corpus, ensuring that the constructed parse is correct.

The decision to have a simple feature model is driven by the fact that we want to first establish that this model of parsing can be useful in predicting reading times. For that, it is preferable to keep the model as comprehensible and simple as possible, otherwise, it would not be clear whether the results reported in section 4 are due to the parsing model or some confound we are not interested in (e.g., meaning similarity present in word vector spaces). For the same reason, we currently made use of the bottom-up parsing algorithm, even though there is a good argument to be made that the bottom-up parsing algorithm is not cognitively adequate. There are well-known issues with bottom-up parsing for psycholinguistics: it accumulates elements on the stack in right-branching structures, suffers from disconnectedness and has problems when tied to incremental interpretation (see Resnik, 1992; Crocker, 1999). We assumed the bottom-up parsing algorithm since it is arguably the most common parsing algorithm for transition-based phrase structure parsers and thus, it serves as a very good starting point. We leave it for the future to see whether other parsing algorithms, notably, left-corner parsers, can improve on the current modeling results.

6. CONCLUSION

This paper presented and tested a psycholinguistic parser that has been developed using insights from the rational theory of memory. It has been shown that the rational theory of memory

[10] While the mentioned works in computational psycholinguistics make use of transition-based parsing, they are not closely related to this work. The cited approaches, unlike the current account, do not construct the parsers inside a cognitive architecture and their goal is different than developing a single account for experience-based and memory-based processing difficulties based on the rational theory of memory.

[11] Label Precision is calculated as the number of correctly constructed constituents divided by the number of all constituents proposed by the parser. Label Recall is calculated as the number of correctly constructed constituents divided by the number of all constituents present in the gold standard. F1 is the harmonic mean of the two accuracy measures. For the calculation, only non-terminal constituents are used for accuracy (i.e., trivial constituents like $\langle a, DT\rangle$ are ignored so that the accuracy measures are not artificially inflated).

[12] The parser could be subsumed under a case of memory-based parsing, see Daelemans et al. (2004). However, unlike the past cases of memory-based parsing, which were inspired by memory structures to deliver the best accuracy on data-driven parsing, the current approach is inspired by memory structures to connect parsing to on-line behavioral measures. Such a link is not considered in the approach of Daelemans et al. (2004).

can be combined with transition-based parsing to produce a data-driven parser that can be embedded in the ACT-R cognitive architecture. The parser has been tested on garden-path sentences and it has been shown that the parser to a large extent predicts processing difficulties at correct disambiguation points. The parser has also been evaluated on on-line behavioral data from a self-paced reading corpus and it has been shown that the parser can be fit to data and model quantitative patterns in reading times.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article are available at https://github.com/jakdot/parsing-model-and-a-rational-theory-of-memory.

AUTHOR CONTRIBUTIONS

JD contributed to the development of the theory, coding, and modeling. PH contributed to coding and modeling.

FUNDING

The research reported in this paper was supported by the NWO grant VC.GW.17.122.

ACKNOWLEDGMENTS

The authors would like to thank the audience of the conference Rational Approaches to Language Science, and to Marisa Boston, Matthew Crocker, Vera Demberg, and the Frontiers reviewers.

REFERENCES

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cogn. Psychol.* 6, 451–474. doi: 10.1016/0010-0285(74)90021-8

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. (1991). Is human cognition adaptive? *Behav. Brain Sci.* 14, 471–517. doi: 10.1017/S0140525X00070801

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe*? New York, NY: Oxford University Press.

Anderson, J. R., Bothell, D., and Byrne, M. D. (2004). An integrated theory of the mind. *Psychol. Rev.* 111, 1036–1060. doi: 10.1037/0033-295X.111.4.1036

Anderson, J. R., and Lebiere, C. (1998). *The Atomic Components of Thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., and Reder, L. M. (1999). The Fan effect: new results and new theories. *J. Exp. Psychol. Gen.* 128, 186–197. doi: 10.1037/0096-3445.128.2.186

Bever, T. G. (1970). "The cognitive basis for linguistic structures," in *Cognition and the Development of Language*, ed J. Hayes (New York, NY: Wiley), 279–362.

Boston, M. F., Hale, J., Kliegl, R., Patil, U., and Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: an evaluation using the potsdam sentence corpus. *J. Eye Mov. Res.* 2, 1–12. doi: 10.16910/jemr.2.1.1

Boston, M. F., Hale, J. T., Vasishth, S., and Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Lang. Cogn. Process.* 26, 301–349. doi: 10.1080/01690965.2010.492228

Bothell, D. (2017). *ACT-R 7 Reference Manual*. Available online at: http://act-r.psy.cmu.edu/actr7.x/reference-manual.pdf

Brasoveanu, A., and Dotlačil, J. (2018). "An extensible framework for mechanistic processing models: from representational linguistic theories to quantitative model comparison," in *Proceedings of the 2018 International Conference on Cognitive Modelling*.

Brasoveanu, A., and Dotlačil, J. (2019). "Quantitative comparison for generative theories," in *Proceedings of the 2018 Berkeley Linguistic Society 44* (Berkeley, CA).

Brasoveanu, A., and Dotlačil, J. (2020). *Computational Cognitive Modeling and Linguistic Theory*. Language, Cognition, and Mind (LCAM) Series. Springer (Open Access). doi: 10.1007/978-3-030-31846-8

Coavoux, M., and Crabbé, B. (2017a). "Incremental discontinuous phrase structure parsing with the gap transition," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (Valencia: Association for Computational Linguistics), 1259–1270. doi: 10.18653/v1/E17-1118

Coavoux, M., and Crabbé, B. (2017b). "Multilingual lexicalized constituency parsing with word-level auxiliary tasks," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (Valencia), 331–336. doi: 10.18653/v1/E17-2053

Collins, M. (1997). "Three generative, lexicalized models for statistical parsing," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics* (Madrid), 16–23. doi: 10.3115/976909.979620

Crabbé, B. (2015). "Multilingual discriminative lexicalized phrase structure parsing," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Lisbon), 1847–1856. doi: 10.18653/v1/D15-1212

Crain, S., and Steedman, M. (1985). "On not being led up the garden path: the use of context by the psychological syntax processor," in *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*, eds D. Dowty, L. Karttunen, and A. Zwicky (Cambridge: Cambridge University Press), 320–358. doi: 10.1017/CBO9780511597855.011

Crocker, M. W. (1999). "Mechanisms for sentence processing," in *Language Processing*, eds S. Garrod and M. Pickering (London: Psychology Press Hove), 191–232.

Daelemans, W., Zavrel, J., Van Der Sloot, K., and Van den Bosch, A. (2004). *TiMBL: Tilburg Memory-Based Learner*. Tilburg: Tilburg University.

Demberg, V., and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109, 193–210. doi: 10.1016/j.cognition.2008.07.008

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Dotlačil, J. (2018). Building an ACT-R reader for eye-tracking corpus data. *Top. Cogn. Sci.* 10, 144–160. doi: 10.1111/tops.12315

Dubey, A., Keller, F., and Sturt, P. (2008). A probabilistic corpus-based model of syntactic parallelism. *Cognition* 109, 326–344. doi: 10.1016/j.cognition.2008.09.006

Dyer, C., Kuncoro, A., Ballesteros, M., and Smith, N. A. (2016). "Recurrent neural network grammars," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics* (San Diego, CA), 199–209. doi: 10.18653/v1/N16-1024

Engelmann, F. (2016). *Toward an integrated model of sentence processing in reading* (Ph.D. thesis), University of Potsdam, Potsdam, Germany.

Engelmann, F., Jäger, L. A., and Vasishth, S. (2019). The effect of prominence and cue association on retrieval processes: a computational account. *Cogn. Sci.* 43:e12800. doi: 10.1111/cogs.12800

Engelmann, F., Vasishth, S., Engbert, R., and Kliegl, R. (2013). A framework for modeling the interaction of syntactic processing and eye movement control. *Top. Cogn. Sci.* 5, 452–474. doi: 10.1111/tops.12026

Franke, M., and Jäger, G. (2016). Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Z. Sprachwiss.* 35, 3–44. doi: 10.1515/zfs-2016-0002

Frazier, L. (1978). *On comprehending sentences: syntactic parsing strategies* (Ph.D. thesis), University of Connecticut, Storrs, CT, United States.

Frazier, L., and Rayner, K. (1982). Making and correcting errors during sentence comprehension: eye movements in the analysis of structurally ambiguous sentences. *Cogn. Psychol.* 14, 178–210. doi: 10.1016/0010-0285(82)90008-1

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., et al. (2018). "The natural stories corpus," in *Proceedings of LREC 2018, Eleventh*

International Conference on Language Resources and Evaluation (Miyazaki), 76–82.

Futrell, R., and Levy, R. (2017). "Noisy-context surprisal as a human sentence processing cost model," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (Valencia), 688–698. doi: 10.18653/v1/E17-1065

Gibson, E. (1991). *A computational theory of human linguistic processing: memory limitations and processing breakdown* (Ph.D. thesis), Carnegie Mellon University, Pittsburgh, PA, United States.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/S0010-0277(98)00034-1

Hale, J. (2001). "A probabilistic Earley parser as a psycholinguistic model," in *Proceedings of the 2nd Meeting of the North American Asssociation for Computational Linguistics* (Stroudsburg, PA), 159–166. doi: 10.3115/1073336.1073357

Hale, J., Dyer, C., Kuncoro, A., and Brennan, J. R. (2018). "Finding syntax in human encephalography with beam search," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Melbourne, VIC). doi: 10.18653/v1/P18-1254

Hale, J. T. (2014). *Automaton Theories of Human Sentence Comprehension*. Stanford, CA: CSLI Publications.

Hart, B., and Risley, T. R. (1995). *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore, MD: Paul H. Brookes Publishing.

Jäger, L. A., Engelmann, F., and Vasishth, S. (2017). Similarity-based interference in sentence comprehension: literature review and bayesian meta-analysis. *J. Mem. Lang.* 94, 316–339. doi: 10.1016/j.jml.2017.01.004

Jäger, L. A., Mertzen, D., Van Dyke, J. A., and Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: a large-sample study. *J. Mem. Lang.* 111:104063. doi: 10.1016/j.jml.2019.104063

Kalt, T. (2004). "Induction of greedy controllers for deterministic treebank parsers," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (Barcelona).

Kitaev, N., and Klein, D. (2018). "Constituency parsing with a self-attentive encoder," in *Proceedings of the 56 meeting of the Association for Computational Linguistics* (Melbourne, VIC). doi: 10.18653/v1/P18-1249

Klein, D., and Manning, C. D. (2003). "A* parsing: fast exact viterbi parse selection," in *Proceedings of the Human Language Technology Conference and The North American Association for Computational Linguistics (HLT-NAACL)*, 119–126. doi: 10.3115/1073445.1073461

Kübler, S., McDonald, R., and Nivre, J. (2009). *Dependency Parsing*. Synthesis lectures on human language technologies. Morgan & Claypool Publishers.

Kush, D., Lidz, J., and Phillips, C. (2015). Relation-sensitive retrieval: evidence from bound variable pronouns. *J. Mem. Lang.* 82, 18–40. doi: 10.1016/j.jml.2015.02.003

Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., and Phillips, C. (2015). Agreement attraction in spanish comprehension. *J. Mem. Lang.* 82, 133–149. doi: 10.1016/j.jml.2015.02.002

Lebiere, C. (1999). The dynamics of cognition: an ACT-R model of cognitive arithmetic. *Kognitionswissenschaft* 8, 5–19. doi: 10.1007/s001970050071

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

Levy, R. (2011). "Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, OR), 1055–1065.

Levy, R., Fedorenko, E., and Gibson, E. (2013). The syntactic complexity of Russian relative clauses. *J. Mem. Lang.* 69, 461–495. doi: 10.1016/j.jml.2012.10.005

Lewis, R. (1993). *An architecturally-based theory of human sentence comprehension* (Ph.D. thesis), Carnegie Mellon University, Pittsburgh, PA, United States.

Lewis, R., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 1–45. doi: 10.1207/s15516709cog0000_25

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Liu, J., and Zhang, Y. (2017). In-order transition-based constituent parsing. *Trans. Assoc. Comput. Linguist.* 5, 413–424. doi: 10.1162/tacl_a_00070

Marcus, M. P. (1978). *A theory of syntactic recognition for natural language* (Ph.D. thesis), Massachusetts Institute of Technology, Cambridge, MA, United States.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: the PENN treebank. *Comput. Linguist.* 19, 313–330. doi: 10.21236/ADA273556

McDonald, R., and Nivre, J. (2011). Analyzing and integrating dependency parsers. *Comput. Linguist.* 37, 197–230. doi: 10.1162/coli_a_00039

Nicenboim, B., Vasishth, S., Engelmann, F., and Suckow, K. (2018). Exploratory and confirmatory analyses in sentence processing: a case study of number interference in german. *Cogn. Sci.* 42, 1075–1100. doi: 10.1111/cogs.12589

Nivre, J. (2004). "Incrementality in deterministic dependency parsing," in *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together* (Stroudsburg, PA), 50–57. doi: 10.3115/1613148.1613156

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., et al. (2007). Maltparser: a language-independent system for data-driven dependency parsing. *Nat. Lang. Eng.* 13, 95–135. doi: 10.1017/S1351324906004505

Oaksford, M., and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychol. Rev.* 101:608. doi: 10.1037/0033-295X.101.4.608

Oaksford, M., and Chater, N. (2007). *Bayesian Rationality: the Probabilistic Approach to Human Reasoning*. Oxford University Press.

Piantadosi, S. T., Tenenbaum, J. B., and Goodman, N. D. (2016). The logical primitives of thought: empirical foundations for compositional cognitive models. *Psychol. Rev.* 123, 392-424. doi: 10.1037/a0039980

Pritchett, B. L. (1992). *Grammatical Competence and Parsing Performance*. Chicago, IL: The University of Chicago Press.

Rabe, M. M., Paape, D., Vasishth, S., and Engbert, R. (2021). Dynamical cognitive modeling of syntactic processing and eye movement control in reading. *PsyArXiv*. doi: 10.31234/osf.io/w89zt

Rasmussen, N. E., and Schuler, W. (2018). Left-corner parsing with distributed associative memory produces surprisal and locality effects. *Cogn. Sci.* 42, 1009–1042. doi: 10.1111/cogs.12511

Reitter, D., Keller, F., and Moore, J. D. (2011). A computational cognitive model of syntactic priming. *Cogn. Sci.* 35, 587–637. doi: 10.1111/j.1551-6709.2010.01165.x

Resnik, P. (1992). "Left-corner parsing and psychological plausibility," in *Proceedings of the Fourteenth International Conference on Computational Linguistics* (Nantes). doi: 10.3115/992066.992098

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Comput. Linguist.* 27, 249–276. doi: 10.1162/089120101750300526

Sagae, K., and Lavie, A. (2005). "A classifier-based parser with linear run-time complexity," in *Proceedings of the Ninth International Workshop on Parsing Technology* (Vancouver, BC), 125–132. doi: 10.3115/1654494.1654507

Shain, C., Van Schijndel, M., Futrell, R., Gibson, E., and Schuler, W. (2016). "Memory access during incremental sentence processing causes reading time latency," in *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)* (Osaka), 49–58.

Smith, G., and Vasishth, S. (2020). A principled approach to feature selection in models of sentence processing. *Cogn. Sci.* 44:e12918. doi: 10.1111/cogs.12918

Smith, N. J., and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition* 128, 302–319. doi: 10.1016/j.cognition.2013.02.013

Taatgen, N. A., and Anderson, J. R. (2002). Why do children learn to say "broke"? A model of learning the past tense without feedback. *Cognition* 86, 123–155. doi: 10.1016/S0010-0277(02)00176-2

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279–1285. doi: 10.1126/science.1192788

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 33:407. doi: 10.1037/0278-7393.33.2.407

Van Schijndel, M., and Schuler, W. (2013). "An analysis of frequency-and memory-based processing costs," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Atlanta, GA), 95–105.

Vasishth, S., Nicenboim, B., Engelmann, F., and Burchert, F. (2019). Computational models of retrieval processes in sentence processing. *Trends Cogn. Sci.* 23, 968–982. doi: 10.1016/j.tics.2019.09.003

Villata, S., Tabor, W., and Franck, J. (2018). Encoding and retrieval interference in sentence comprehension: evidence from agreement. *Front. Psychol.* 9:2. doi: 10.3389/fpsyg.2018.00002

Vogelzang, M., Mills, A. C., Reitter, D., Van Rij, J., Hendriks, P., and Van Rijn, H. (2017). Toward cognitively constrained models of language processing: a review. *Front. Commun.* 2:11. doi: 10.3389/fcomm.2017.00011

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Weaver, R. (2008). Parameters, predictions, and evidence computational modeling: a statistical view informed by ACT-R. *Cogn. Sci.* 32, 1349–1375. doi: 10.1080/03640210802463724

Zhang, Y., and Clark, S. (2008). "A tale of two parsers: investigating and combining graph-based and transition-based dependency parsing," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (Honolulu, HI), 562–571. doi: 10.3115/1613715.1613784

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.

APPENDIX A: CALCULATE BASE-LEVEL ACTIVATION FROM WORD/RULE FREQUENCIES

We want to calculate B_i from frequency. d is a free parameter and can be ignored in this discussion.

$$B_i = \log\left(\sum_{k=1}^{n} t_k^{-d}\right) \qquad \text{(d- free parameter)} \tag{A1}$$

Consider a 15-year old speaker. How can we estimate how often a word/parsing step x was used in language interactions that the speaker participated in?

First, let's notice that we know the relative frequency of x. We collect that from the British National Corpus (for words) and from the Penn Treebank corpus (for parsing steps).

We know the lifetime of the speaker (15 years), so if we know the total number of words an average 15-year old speaker has been exposed to, we can easily calculate how many times x was used on average based on the frequency of x. A good approximation of the number of words a speaker is exposed to per year can be found in Hart and Risley (1995). Based on recordings of 42 families, Hart and Risley estimate that children comprehend between 10 million to 35 million words a year, depending to a large extent on the social class of the family, and this amount increases linearly with age. According to the study, a 15-year old has been exposed to anywhere between 50 and 175 million words total. For simplicity, the model will work with the mean of 112.5 million words as the total amount of words a 15-year old speaker has been exposed to. This is a conservative estimate as it ignores production and the linguistic exposure associated with mass media. Furthermore, we assume that each word is accompanied by one parsing step, so there are as many parsing steps as words (again, this is a simplification that should not harm modeling).

We now know how we get from frequency to the number of usages of x. Simplifying again, we assume that the usages, t_k above, are evenly spread during the life span.

The procedure described here was successfully used in translating frequencies to activations and ultimately reaction times in sentence production (Reitter et al., 2011), eye tracking reading times (Dotlačil, 2018) and reaction times in lexical decision tasks (Brasoveanu and Dotlačil, 2020).

ORIGINAL RESEARCH
published: 30 June 2021
doi: 10.3389/fpsyg.2021.662277

On the Correlation of Context-Aware Language Models With the Intelligibility of Polish Target Words to Czech Readers

Klára Jágrová[1]*, *Michael Hedderich*[1,2], *Marius Mosbach*[1,2], *Tania Avgustinova*[1,3] and *Dietrich Klakow*[1,2]

[1] *Collaborative Research Center 1102: Information Density and Linguistic Encoding, Saarland University, Saarbrücken, Germany,* [2] *Saarland Informatics Campus, Spoken Language Systems, Saarland University, Saarbrücken, Germany,* [3] *Language Science and Technology, Saarland University, Saarbrücken, Germany*

This contribution seeks to provide a rational probabilistic explanation for the intelligibility of words in a genetically related language that is unknown to the reader, a phenomenon referred to as intercomprehension. In this research domain, linguistic distance, among other factors, was proved to correlate well with the mutual intelligibility of individual words. However, the role of context for the intelligibility of target words in sentences was subject to very few studies. To address this, we analyze data from web-based experiments in which Czech (CS) respondents were asked to translate highly predictable target words at the final position of Polish sentences. We compare correlations of target word intelligibility with data from 3-g language models (LMs) to their correlations with data obtained from context-aware LMs. More specifically, we evaluate two context-aware LM architectures: Long Short-Term Memory (LSTMs) that can, theoretically, take infinitely long-distance dependencies into account and Transformer-based LMs which can access the whole input sequence at the same time. We investigate how their use of context affects surprisal and its correlation with intelligibility.

Keywords: intercomprehension, predictive context, Polish, Czech, context-aware language models, Long Short-Term Memory, transformer, surprisal

Edited by:
Hannah Rohde,
University of Edinburgh,
United Kingdom

Reviewed by:
Robert Möller,
University of Liège, Belgium
Raphael Berthele,
Université de Fribourg, Switzerland

Correspondence:
Klára Jágrová
kjagrova@coli.uni-saarland.de

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

Received: *31 January 2021*
Accepted: *20 May 2021*
Published: *30 June 2021*

Citation:
Jágrová K, Hedderich M, Mosbach M, Avgustinova T and Klakow D (2021) On the Correlation of Context-Aware Language Models With the Intelligibility of Polish Target Words to Czech Readers. Front. Psychol. 12:662277. doi: 10.3389/fpsyg.2021.662277

1. INTRODUCTION

In the research domain of intercomprehension, the intelligibility of stimuli has been, among other linguistic and extra-linguistic factors, traditionally explained by the linguistic distance of the stimulus toward a language in the linguistic repertoire of the reader, mostly the native language (L1) (e.g., Gooskens, 2007; Möller and Zeevaert, 2015; Golubović, 2016) or a combination of the L1 and other acquired languages (Vanhove, 2014; Vanhove and Berthele, 2015; Jágrová et al., 2017). It has been shown many times that the lower the measurable cross-lingual similarity or regularity of orthographic correspondences (Stenger et al., 2017) is, the more the languages are mutually intelligible in general. This applies to individual words in language pairs, too: The lower the linguistic (orthographic, phonetic, and morphological) distance between a concrete word pair, the more the words are expected to be comprehensible to the reader of the respective other related languages.

So far there have been only a few studies focusing on the role of context as an additional factor influencing the mutual intelligibility of target words. Muikku-Werner (2014) observed that the role of neighborhood density (number of available similar word forms that readers might consider suitable translation equivalents) decreases through context since the potential other options have to fit the syntactic frame. She also found that it appears easier for respondents to guess a frequent collocate of a word, once the other word is successfully recognized (Muikku-Werner, 2014, p. 105). In a study on the disambiguation of false friends with students of Slavic languages, Heinz (2009) points out that the amount of correctly understood context is crucial for the correct recognition of target words. He also refers to the negative role that context can play: Previous (correct) lexical decisions can be revised to formulate an utterance that respondents believe is reasonable.

Jágrová (2018) investigated the influence of divergent word order in Polish (PL) noun phrases (adjective-noun vs. noun-adjective) on their intelligibility to Czech (CS) readers, since the noun-adjective linearization is more typical in PL than in CS which is reflected in higher surprisal scores of the CS translations of the stimuli. She correlated the product of linguistic distance and 3-g language model (LM) surprisal ("overall difficulty") of the stimuli phrases to processing time and intelligibility and found a higher correlation than with linguistic distance only. This method of determining an overall difficulty consisting of distance and surprisal for individual words within sentences was also applied in Jágrová et al. (2019) in "an attempt to use LMs to describe the role of context in the stimuli and translations thereof" (Jágrová et al., 2019, p. 261), without claiming to present statistically sufficient data for the PL-to-CS scenario (12 sentences, 16 respondent pairs). There it was found that the calculated difficulty levels of the words within the stimuli did not always agree with the actual performance of the respondents. Contrary to the expectations of the authors, even cognates with very low linguistic distance or internationalisms that are identical in both languages were not always translated correctly, especially when they also had low corpus frequency and thus high surprisal scores. Respondents often considered these words unlikely or not fitting the context. In another study by Jágrová and Avgustinova (2019), data from a representative sample of stimuli sentences and respondents was collected in a web-based cloze translation experiment in the same language-reader-scenario. In the present study, we build upon the data from their experiment.

The language models applied in the studies by Jágrová (2018), Jágrová et al. (2019), and Jágrová and Avgustinova (2019) were all 3-g models. The principle according to which these models work is that they iterate through a training corpus and count all occurrences of any three subsequent words. When then applied to a sentence, they can help statistically assess the predictability of a word in relation to its two preceding words. In practice, however, the sentential context relevant for the intelligibility of a target word can be larger than only its two preceding words. Consequently, other types of statistical LMs might be better in capturing the role of semantic primes and concepts that allow for correct associations within the sentences.

To verify this hypothesis, we trained different context-aware LMs on the Czech National Corpus (Křen et al., 2016) and the PolEval 2018 language modeling corpus (Ogrodniczuk and Kobyliński, 2018). We applied these LMs to score the PL stimuli sentences used in the experiment by Jágrová and Avgustinova (2019) and on the closest CS translations thereof. We correlated the surprisal scores of the target words and the whole sentences with target word intelligibility and compared them to the correlations with 3-g surprisal from Jágrová and Avgustinova (2019). Although all correlations proved to be fairly low, we found slightly better results for the target word surprisals from the CS context-aware models. In individual examples, we found that the context-aware models appear to be better suitable to capture the predictability of semantic associations within the sentences, while 3-g models appear to be better representations of predictability caused by collocates directly preceding the target words.

This study is structured as follows. In section 2, we first explain how data from 3-g LMs were correlated with target word intelligibility in Jágrová and Avgustinova (2019). We then outline the hypothesis regarding the better performance of context-aware LMs in comparison to 3-g LMs in section 3 and explain their architectures in section 4. Next, we present the results from the context-aware LMs in section 5 and compare them with the correlations observed in Jágrová and Avgustinova (2019). Finally, we summarize the findings in the discussion in section 6.

2. PREVIOUS RESEARCH

In a previous study, using surprisal estimates from 3-g LMs, Jágrová and Avgustinova (2019) showed that predictability in context contributes to the intelligibility of target words in sentence-final position when compared to the intelligibility of the same words without context. They gathered data from web-based cloze translation experiments for highly predictable target words in 149 PL sentences.

The sentence stimuli presented in the experiment are translations of sentences published in a study by Block and Baldwin (2010) who tested a set of 500 constructed sentences in a cloze completion task. In addition to that, Block and Baldwin (2010) validated the predictability of the target words in their sentences in event-related potential(s) experiments. The study resulted in a dataset of 400 high-constraint, high cloze probability sentences. For the study of Jágrová and Avgustinova (2019), those sentences with the most predictable target words (90–99% cloze probability) were translated into PL and applied in cloze translation experiments. Sentences containing culturally specific context were omitted, which resulted in a set of 149 sentences. The translation into PL was provided by a linguist and professional translator who was instructed to keep the original target words in the last position in the sentences.

These 149 sentences were presented to CS respondents who were asked to guess and translate the PL target words into CS. After having filled out a sociodemographic survey and having provided a self-assessment of language skills, only those respondents were admitted to the experiment who did not indicate any prior knowledge of PL. The PL sentences were presented in seven blocks, each consisting of 17–24 sentences.

The order of the sentences within a block was randomized. Data of at least 30 respondents (mean age 25.3) were gathered for each target word. To make sure that respondents indeed read the sentential context, the experiment was designed in a way that respondents initially saw only the first word of the sentence and then were asked to click on it to make the next word appear. In this way, they clicked through the whole sentence till the last word (target word) appeared. After clicking on the target word, the window for entering the translation of the target word appeared. The time limit for entering the translation of the target word was set to 20–30 s, depending on the length of the sentence. The respondents were not informed that the target words are highly predictable.

To obtain a baseline for comparison, the PL target words were also presented without any context and in their base forms to other CS respondents as a cognate guessing task. The majority of the words were more comprehensible within the sentences (68.0% intelligibility) than if presented without context (49.7% intelligibility).

For instance, the PL target word *głosu* "voice [genitive]" in the PL sentence

(1) PL: *Że był wściekły, rozpoznała po tonie jego* ***głosu****.*
(CS: *Že byl vzteklý, poznala podle tónu jeho* ***hlasu****.*)
"That he was mad, she could tell by the tone of his **voice**."

was translated correctly into CS as a form of *hlas* "voice" more often in the predictive context (93.3%) than without context (26.7%). As shown in **Figure 1**, the predictability of the target word is, in this case, reflected well by the surprisal scores obtained from the 3-g LM, since PL *głosu* (CS *hlasu*) "voice [genitive]" is highly predictable after PL *tonie jego* (CS *tónu jeho*) "the tone of his."

The PL 3-g LM was trained on the PL part of InterCorp (Čermák and Rosen, 2012), and the CS LM was trained on the SYN2015 version of the Czech National Corpus (CNC, Křen et al., 2015). Kneser-Ney smoothing (Kneser and Ney, 1995) was applied on both LMs. The PL LM provides the information density profile of the stimuli sentences. To obtain the best possible representation of the comprehension process of a CS reader, the PL stimuli sentences were translated literally into CS before the CS LM was applied for scoring (for detail on the method of Vanhove, 2014; Jágrová and Avgustinova, 2019). The blue graph in **Figure 1** represents the surprisal of a PL sentence scored by the PL 3-g LM. The orange graph represents the closest literal CS translation of this PL sentence scored by the CS 3-g LM accordingly.

In other sentences, however, the predictability of the target word resulting in greater ease of understanding was not reflected by the surprisal scores from the PL 3-g LM. For instance, the PL target word *gwoździa* "nail [genitive]" in the sentence

(2) PL: *Aby zawiesić obraz Ted potrzebował młotka i* ***gwoździa****.*
(CS: *Aby zavěsil obraz, Ted potřeboval kladivo a* ***hřebík****.*)
"To hang the picture Ted needed a hammer and a **nail**." (Block and Baldwin, 2010)

was translated more often correctly as a form of CS *hřebík* "nail" in context (53.3%) than without context (3.03%). However, as shown in **Figure 2**, the 3-g LM displays a rise in surprisal at the target word position, which is a typical indication of high processing difficulty due to unexpectedness in context. This suggests that the predictability of the target word does not depend exclusively on the immediately preceding words, as could have been reflected by the 3-g LM. Instead, the better comprehension of the target seems to be connected to the correct identification of the concept of hanging a picture: PL *zawiesić* "to hang" is a cognate of CS *zavěsit*, the sentence-initial conjunction *aby* "to" as well as the noun *obraz* "picture" are identical in form and meaning in both languages, PL *potrzebował* "he needed" is a cognate of its CS translation *potřeboval*. PL *młotka* "hammer [genitive]" preceding the target word is a non-cognate to its CS translation equivalent *kladivo*. However, there might be a clue in the CS lexicon through the concept of *mlátit* "to hit" or *mlat* as in *sekeromlat* "threshel, stone axe," provided that the CS respondents successfully apply the regular PL:CS correspondence *ło:la/lá* in the stem.

Even though the context was helpful for the comprehension of targets in most of the sentences, the situation was reversed for some target words in context if compared to the condition without context. An analysis of the errors made by respondents revealed some systematic patterns, such as L1 interferences, inferences from other acquired languages, or perceived morphological mismatches. Also, priming by readers or association with a dominant but misleading concept in the sentence seems to have played a crucial role in the misinterpretations of some target words. For instance, the PL target word *dzień* (CS *den*) "day" in the sentence

(3) PL: *Dentysta zaleca myć zęby dwa razy na* ***dzień****.*
(CS: *Zubař doporučuje čistit si zuby dvakrát za* ***den***).
"The dentist recommends brushing your teeth twice a **day**." (Block and Baldwin, 2010)

was translated wrongly by some respondents as *dáseň* "gum." Not only are PL *dzień* and CS *dáseň* orthographically relatively similar [Levenshtein distance: 0.5 (Levenshtein, 1966), the mean pronunciation-based orthographic distance of the 149 target words is 42.6%], but also does the concept of the easily identifiable PL *dentysta* (CS *dentista* or *zubař*) "dentist" mislead respondents to an association of the target word with the dentist. The intelligibility of PL *dzień* for CS respondents was higher without context (80.0%) than in context (66.7%). The question is whether such effects can be predicted by an LM that would also take into account cross-lingual similarity. We explore this setting in section 4.6.

It has to be mentioned that sentence context is not equally easy to understand in all test sentences, some of the sentences contain non-cognates or false friends, while others do not. Also, the orthographic distance of cognates is different in each sentence. Admittedly, it is difficult to capture the whole complexity of intercomprehension in these translation experiments and to

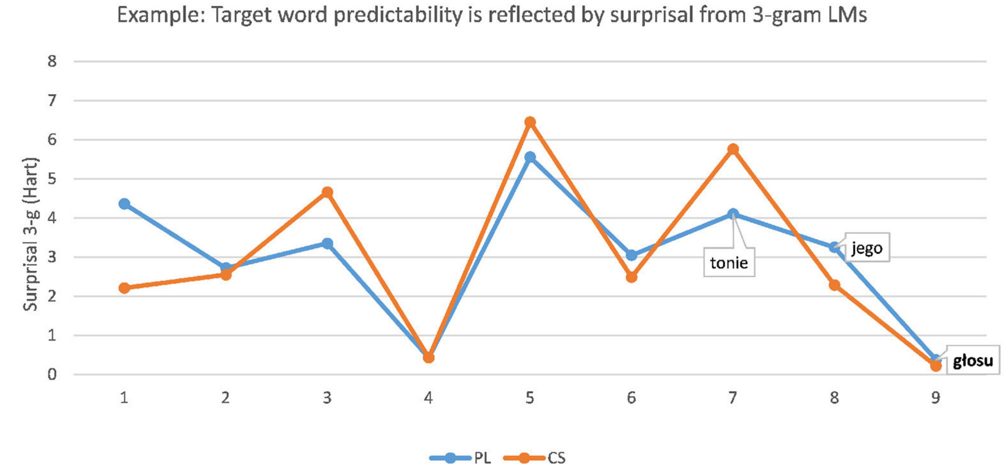


FIGURE 1 | Example: Predictability of PL target word *głosu* "voice [genitive]" is reflected well by the low surprisal score of the target obtained from the 3-g LM.

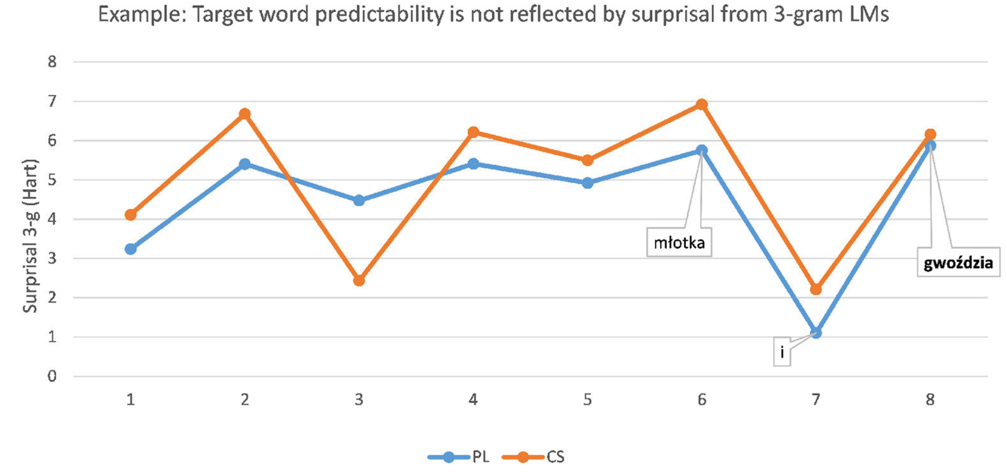


FIGURE 2 | Example: Predictability of PL target word *gwoździa* "nail [genitive]" is not reflected well by the 3-g language model (LM): surprisal curve rises at the target word.

control for a whole range of (linguistic) factors that come into play when the context is concerned.

In an ideal world, all words of which the stimuli sentences consist should have been tested for intelligibility separately to reliably assess how much of the context the respondents understand. Although it was not tested how intelligible the context is, it was approximated by measuring the linguistic distance (lexical and orthographic distance) of the stimuli sentences toward the closest CS translation in Jágrová and Avgustinova (2019). The distance of the target word and the total number of non-cognates per sentence were then added as variables into a multiple linear regression model and could, together with the sum of surprisal of the PL sentence, account for 49.6% of the variance in the data (Jágrová and Avgustinova, 2019, p. 15).

Jágrová and Avgustinova (2019) also found that besides high correlations with orthographic distance ($r = -0.772, p < 0.001$ without context and $r = -0.680, p < 0.001$ in predictive context), the correlation of intelligibility with surprisal depends on the lexical similarity of the target words. For the whole set of 149 sentences, the best correlation found was a fairly low one with the sum of surprisal of the whole PL sentence ($r = -0.215, p < 0.01$). When excluding sentences with target cognates (words with etymologically related translation equivalents in both languages) from the analysis, the correlation of intelligibility with the total surprisal of the PL sentence reaches $r = -0.411, p < 0.01$. Three-gram surprisal and intelligibility correlate best for sentences in which the target words are false friends ($r = -0.443, p < 0.01$), especially those that, despite their misleading character, allow for correct semantic associations with the correct translation. Even though all correlations turned out relatively low, predictability effects and associations seem to be more important for targets with high linguistic distance, especially for non-cognates and false friends (lexical distance), than for cognates with low linguistic distance. For instance, PL *drzewo* was translated more often correctly as CS *strom* "tree" (36.6%) or *rodokmen* "family tree" in the sentence

(4) PL: *Aby dowiedzieć się czegoś o swoich przodkach, narysowali genealogiczne* ***drzewo****.*
(CS: *Aby se dozvěděli něco o svých předcích, nakreslili genealogický* ***strom*** / ***rodokmen****.*)
"To learn about their ancestors they drew a family **tree**."
(Block and Baldwin, 2010)

than in the condition without context (0%). There it was frequently mistaken for its CS false friend *dřevo* "wood." Together with the partly identifiable context of this sentence [PL *dowiedzieć się* (CS *dozvědět se*) "to learn (about)"; PL *o swoich przodkach* (CS *o svých předcích*) "about their ancestors"], PL *drzewo* allows for a correct semantic association of wood and trees (Jágrová and Avgustinova, 2019, p. 11).

3. HYPOTHESIS

Since the 3-g LM used by Jágrová and Avgustinova (2019) cannot reflect the influence of contextual cues from any other position in the sentence than the two words immediately preceding the target word, we hypothesize that the intelligibility of highly predictable target words will have a stronger correlation with surprisal values obtained from language models which incorporate information from the entire sentence than with surprisal values from 3-g LMs.

4. METHODS

We build upon the study by Jágrová and Avgustinova (2019) and estimate the surprisal of target words in a given sentence by relying on language models that are capable of considering context beyond 3-g.

In recent years, two main approaches have dominated context-aware, neural LMs: Long Short-Term Memory (LSTM) and more recently Transformer. For both architectures, we investigate whether their use of sentence-level context affects surprisal and its correlation with intelligibility. We start by providing a brief recap on (neural) language modeling.

4.1. Language Modeling

Language models are machine learning models that are typically trained on text corpora and can predict the probability of a word given its context. As an example, an LM trained on a standard English corpus, given the start of the sentence *A small, green* would assign most likely the word *frog* a higher probability for continuing the sentence than the word *cow*. The probability for a target word, given its context, is obtained *via* a learned model that bases its predictions on occurrence statistics in the training corpus.

Most commonly, an LM predicts the probability of a word given the previous (left) context. Formally, for a sentence s consisting of words or tokens $w_1, ..., w_n$, an LM computes the probability $p(w_t|w_{t-1}, ..., w_0)$. The probability of a sentence can be obtained by factorizing the joint probability as a product of conditional probabilities, i.e., by applying the product rule of probabilities:

$$p(s) = \prod_{t=1}^{n} p(w_t|w_{t-1}, ..., w_0) \tag{1}$$

Traditionally, count-based n-gram models have been used for language modeling. In this case, the previous context is limited to $n-1$ words. A 3-g model, therefore, can only compute the probability of a word given its two predecessors, i.e., $p(w_t|w_{t-1}, w_{t-2})$. Increasing the value of n for count-based models is difficult due to factors like data sparsity (Jelinek and Mercer, 1980).

4.2. Long Short-Term Memory

Long Short-Term Memories (Hochreiter and Schmidhuber, 1997) are a form of Recurrent Neural Networks (RNNs) (Elman, 1990). They learn a parametric model of the distribution of words given their context. These machine learning models can handle sequences of input words of arbitrary length. This removes the hard limitation of history size n that n-gram models have. At each time-step t, the RNN obtains as input the previous word or token w_{t-1}. It then updates its internal state based on that input and its previous internal state. As output, at each time-step, the probability for the current word is given $p(w_t|w_{t-1}, ..., w_0)$.

While RNNs have in theory no limitation on sequence length, in practice, effects like vanishing gradients (Bengio et al., 1994) do limit the amount of previous words that are taken into consideration for the probability of the next token. LSTMs contain special components, such as cell states that improve the handling of such long-term dependencies. An in-depth discussion of the use of LSTMs for language modeling is given in Sundermeyer et al. (2012).

In this study, we build a four layer LSTM with embedding and hidden state sizes of 300. Dropout (Srivastava et al., 2014) of 0.1 is applied between the layers, and gradient clipping is performed with a gradient norm size of 1. As an optimizer, we use Adam (Kingma and Ba, 2015) with a learning rate of 2.5 * 10^{-4}.

4.3. Transformer

Originally proposed for the task of neural machine translation, Transformers (Vaswani et al., 2017) have recently shown strong empirical performance on various natural language processing tasks and have become the predominant architecture for many natural language processing tasks. Other than RNNs, such as LSTMs, Transformers typically do not contain any recurrence and hence have access to the whole input sequence at once *via* an attention mechanism. They can model $p(w_t|w_{t-1}, ..., w_0)$ while taking into consideration all previous context words $w_{t-1}, ..., w_0$ in equal measure. This allows them to make more efficient use of context. Given a large enough input size and positional encodings, Transformers have become the dominating architecture for neural language modeling (Al-Rfou et al., 2019; Dai et al., 2019).

In this study, we train two different Transformer based LMs: (1) a vanilla Transformer decoder with 16 hidden layers, learned positional encoding and a context size of 32 tokens (Al-Rfou et al., 2019) and (2) a 16-layer Transformer-XL decoder with relative positional encodings (Dai et al., 2019). The same gradient clipping and optimizer are used for the LSTM. We choose a context size of 32 tokens based on the sentence length statistics of the stimuli sentences.

4.4. Corpora

The PL LMs were trained on the PolEval 2018 language modeling corpus (Ogrodniczuk and Kobyliński, 2018). It contains 20 million sentences selected from PL Wikipedia, Internet forums, PL books, the National Corpus of Polish Przepiórkowski et al. (2012), and the Polish Parliament Corpus (Ogrodniczuk, 2012). We used the unsegmented version released by the PolEval organizers[1]. This corpus is larger than the Polish part of InterCorp (Čermák and Rosen, 2012) used by Jágrová and Avgustinova (2019).

The CS LMs were trained on the SYN v4 version of the Czech National Corpus (Křen et al., 2016), a collection of contemporary written CS containing ∼4.3 billion tokens. This is the same data as in the study by Jágrová and Avgustinova (2019).

We tokenized both corpora using byte-pair-encoding (Sennrich et al., 2016) and using the SentencePiece toolkit (Kudo and Richardson, 2018). More specifically, for each of the corpora, we automatically create a vocabulary containing the 32.000 most frequent subunits (so-called subwords) and then tokenize the training data as well as the stimuli sentences according to this vocabulary. Both LSTM and Transformer models use the same vocabulary. If a target word is tokenized into several subunits, the probability of the target word is the product of the probabilities of the subunits.

The PL stimuli sentences were scored with the LMs trained on the PL corpus. To obtain the surprisal scores for the CS versions of the sentences and hence to represent their understanding by the CS reader, both the closest CS translation (not necessarily grammatically correct) and a grammatically correct CS translation were scored by the LMs trained on the CS corpus. Models of both languages were used to find out if the surprisal of the stimulus (PL) or the language of the readers (CS) correlates better with target word intelligibility.

4.5. Language Model Performance

The performance of LMs is commonly measured in perplexity over the test corpus T. It is defined as

$$PPL(T) = 2^{-\sum_{w \in T} p(w) \log_e p(w)} \tag{2}$$

where w are all the words or subwords in T. The lower the perplexity of the LM, the better is the performance of the model on predicting the correct next token. The test perplexities for the CS and PL language models are given in **Tables 1**, **2**, respectively. For both languages, the Transformer model outperforms the LSTM and Transformer XL performs best. To the best of our knowledge, we reach a new state-of-the-art for language modeling on the PL corpus (Czapla et al., 2018).

TABLE 1 | The perplexity of the language models on the CS validation corpus.

Model	Subword PPL	Word PPL
LSTM	17.85	38.80
Transformer	15.59	32.67
TransformerXL	**13.94**	**28.35**

The lowest perplexity values are marked bold.

4.6. Toward a Model of the Reader

Following the previous study, we train the aforementioned LMs on PL and CS and then evaluate their surprisal on sentences in the same language. In addition, we also propose a model that is conceptually closer to the human participants. In this case, these are CS native speakers who read PL text. We, therefore, also use the CS Transformer LM to compute surprisal on PL sentences. The model should, e.g., have a low surprisal by the PL word *testamencie* "testament [locative]" as it is close to its CS translation *testamentu*. This is in contrast to the PL word *gwoździa* "nail [genitive]" where the equivalent in CS would be a form of *hřebík* which should result in a high surprisal for the model. It is important to note that this is possible since the surprisal of the Transformer model is computed on a subword or character level (as shown in section 4.4) and not exclusively on a word level. While the PL word *testamencie* will most likely be unknown to a CS LM, its subwords *te*, *sta*, *men*, and *cie* are part of the subword vocabulary of the model.

There are several PL characters with diacritics, e.g., *ą*, *ć*, and *ł*, that are not part of the CS alphabet and thus unknown to this LM. As an attempt to overcome this issue, such PL characters are mapped to CS characters that CS respondents assumed to be corresponding in a previous experiment. There, CS respondents were asked to read out PL stimuli including the unknown PL characters aloud and translate them (Jágrová, 2021). With the help of the transcripts of these recordings, it was possible to obtain statistics about how likely an unknown character was pronounced similar to a (seemingly) corresponding CS character. We use these insights to map certain PL characters to the CS alphabet. In the case of PL *ł*, for instance, the CS character *l* would also be the linguistically correct correspondence. However, while the linguistically correct CS correspondence for PL *ć* would be *t* (regular correspondence in infinitive endings, which a CS reader is not expected to be aware of), we map it to CS *č*.

4.7. Intelligibility

The intelligibility of the word is measured here as the percentage of correct translations provided by respondents for this word. For instance, if the PL word *dzień* "day" has intelligibility of 80%, it means that 80% of the CS respondents translated the word correctly. As for the scoring of responses of the participants, the

[1] http://2018.poleval.pl/index.php/tasks/.

TABLE 2 | The perplexity of the LMs on the PL validation corpus.

Model	Subword PPL	Word PPL
LSTM	49.83	125.5
Transformer	31.12	70.11
TransformerXL	**29.92**	**66.78**
ULMFiT-SP (Czapla et al., 2018)	–	117.67
ULMFiT-SP (Czapla et al., 2018)	–	95.0

The lowest perplexity values are marked bold.

experiment software automatically classified responses as correct or wrong, according to the previous definition. All responses were, however, manually checked afterward so that cases which were classified as wrong but where respondents had understood the stimulus, e.g., typos, missing letters at word end due to time restrictions, or synonyms, could be categorized as correct subsequently. Also, responses that were base forms of targets were counted as correct even if the target word was inflected. The wrong gender of verb forms was tolerated if the translation was otherwise correct, but the wrong tense was not accepted.

4.8. Predictors

We first perform a linear regression with surprisal as the main predictor in question and then add other predictors into a multiple linear regression model. Surprisal as a predictor variable is provided by the models in the unit *nat*. For each sentence and each model trained, we determine the surprisal of the target word as well as the surprisal of the whole sentence. Since higher surprisal is related to higher difficulty, higher surprisal should predict lower intelligibility of an item. If a word is segmented into subword units, then its surprisal is the product of the subword surprisals.

As a representation of the (dis-)similarity of the PL stimulus toward CS, a measure referred to as total pronunciation-based distance is determined for the whole sentence, the final 3-g, 2-g, and target word and examined for correlations with intelligibility. The distances are calculated automatically with the help of the *incom.py* toolbox (Mosbach et al., 2019) for each word. Distances of the 2-g, 3-g, and sentences are the mean distances of the individual words they consist of. For the calculation, two words are aligned by their consonants and vowels in a way that the cheapest alignment option is preferred. The alignment cost for every single PL:CS character pair can be defined when using the *incom.py* tool. For this purpose, a cost of 1 is charged for every different character. As illustrated in **Table 3**, the pronunciation-based distance differs from traditionally calculated Levenshtein distance (Levenshtein, 1966) in a way that it does not charge any costs for the alignment of such characters whose pronunciation should be obvious to the respondents: *y*:*i*, *i*:*y*, *ł*:*l*, *w*:*v*, *ż*:*ž* (PL:CS). The share of different characters is normalized by the alignment length of the word pair and given as a percentage. The more distant a PL word, the less it is expected to be intelligible to the CS respondents. The total pronunciation-based distance measure also incorporates lexical distance by assigning a distance of 100% to non-cognates.

TABLE 3 | Traditionally calculated Levenshtein distance vs. pronunciation-based distance.

Traditional Levenshtein distance										
	1	2	3	4	5	6	7	8	9	10
PL			s	i	ł	o	w	n	i	ę
CS	p	o	s	i	l	o	v	n	u	
Distance	1	1	0	0	0.5	0	1	0	1	1
Normalized distance	55%									
Pronunciation-based Levenshtein distance										
PL			s	i	ł	o	w	n	i	ę
CS	p	o	s	i	l	o	v	n	u	
Distance	1	1	0	0	0	0	0	0	1	1
Normalized distance	40%									

TABLE 4 | Correlations of the context-aware LMs with intelligibility (all sentences).

Model	Surprisal of	Correlation
Transformer CS	target word	$r = -0.247, p < 0.01$
LSTM CS	target word	$r = -0.240, p < 0.01$
TransformerXL CS	target word	$r = -0.223, p < 0.01$
3-g PL (Jágrová and Avgustinova, 2019)	sentence (sum)	$r = -0.215, p < 0.01$
Reader Model	target word	$r = -0.214, p < 0.01$
3-g CS (Jágrová and Avgustinova, 2019)	target word	$r = -0.191, p < 0.05$
3-g PL (Jágrová and Avgustinova, 2019)	target word	$r = -0.186, p < 0.05$
TransformerXL PL	target word	$r = -0.150, p > 0.05$
LSTM PL	target word	$r = -0.148, p > 0.05$
Transformer PL	target word	$r = -0.141, p > 0.05$

The total number and the percentage of non-cognates per sentence are determined as an additional separate predictor of lexical (dis-)similarity. For this purpose, non-cognates are PL words that, in the given context, do not have a CS translation equivalent with the same or a related root in terms of etymological origin. For instance, the sentence in example (2) contains one non-cognate, *gwoździa* "nail [genitive]." The other seven words of the sentence are cognates. If normalized by the number of words in the sentence, the percentage of non-cognates in this sentence is 12.5%. The more non-cognates a CS respondent encounters in a sentence, the less intelligible the sentence should be.

5. RESULTS

5.1. Regression Results

The correlations of target word intelligibility and surprisal from the context-aware LMs are listed in **Table 4** together with the correlations of the 3-g surprisal from the previous study for comparison. When considering the whole dataset of 149 sentences, the highest correlation of intelligibility and surprisal could be found for the target word surprisal from the CS

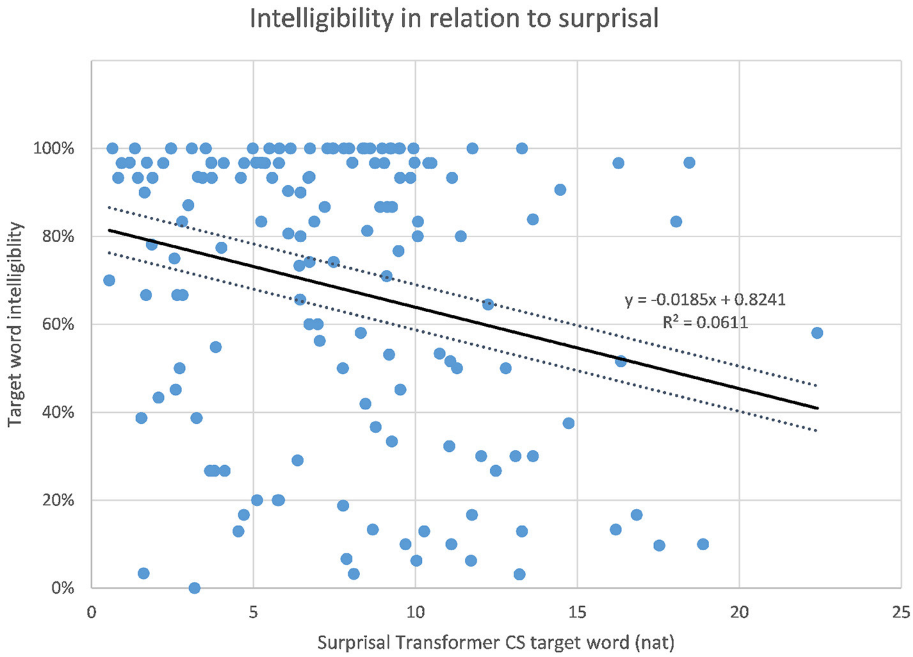


FIGURE 3 | Intelligibility of target words and surprisal from the CS Transformer model.

Transformer model ($r = -0.247, p < 0.01$, as shown in **Figure 3**), followed by the CS LSTM (**Figure 4**), and the CS Transformer XL (**Figure 5**). Thus, the surprisal from all three models correlates slightly stronger with intelligibility than the surprisal from the 3-g models in the previous study, which weakly confirms the hypothesis. No significant correlation could be found with the sum of surprisal per sentence or with the surprisal obtained from the PL versions of the context-aware models.

As pointed out in the introduction, the intelligibility scores have high (negative) correlations with linguistic distance. The best correlation found regarding distance is that of target word distance and intelligibility. Intelligibility and target word distance correlate with $r = -0.680, p < 0.01$ (Jágrová and Avgustinova, 2019, p. 15), as shown in **Figure 6**. As presented in **Figure 7**, the number of non-cognates per sentence as a measure of lexical distance also shows a significant negative correlation with target word intelligibility with $r = -0.507, p < 0.0001$, although the correlation is lower than that of target word distance.

5.2. Multiple Linear Regression Analyses

When we add the variables surprisal and linguistic distance into a multiple linear regression model, the best fitting model is one that consists of the pronunciation-based distance of the target word, the total number of non-cognates per sentence, and surprisal from the CS Transformer model. A regression equation was found [$F_{(3,\,145)} = 59.569, p < 0.0001$] with an adjusted R^2 of 0.543, $p < 0.001$. This is higher than the $R^2 = 0.496, p < 0.001$ of the model containing 3-g surprisal reported in Jágrová and Avgustinova (2019). The predicted intelligibility of the target word is equal to $1.209 - 0.648 * distance - 0.065 * NC - 0.023 * surpTransCS$, where *distance* (in %) is the pronunciation-based distance of the target word normalized by the alignment length of the word pair, *NC* is the number of non-cognates per sentence as a total number (not normalized by the number of words per sentence) and surprisal is measured in *nat* (*surpTransCS* is surprisal from the CS Transformer model). According to the model, the predicted intelligibility of a target word decreased by 0.648% for each % of the distance of the target. As of the model, target word intelligibility decreased by 6.5% for each non-cognate per sentence. For each *nat* of surprisal, target word intelligibility decreased by 2.3%. All three variables distance of the target, number of non-cognates per sentence, and surprisal from the CS Transformer model were significant predictors of target word intelligibility.

5.3. Illustrative Examples

For all example sentences mentioned so far, the surprisal scores from the 3-g LMs (target word surprisal PL and CS and sum of surprisal of the PL sentence) are compared to the target word surprisals from the best performing context-aware LMs in **Table 5**. Contrary to expectations, all models display relatively high coefficients of variance when it comes to target word surprisal of the whole data set, while the coefficient of variance of the 3-g surprisal of the PL sentence is less than half as high. If the LMs provided optimal representations of predictable target words, then target word surprisal would be rather constantly low and would not vary to a high degree.

Since it is interesting to observe whether the context-aware LMs perform better with sentences containing semantic associations or hyponymy outside of the final 3-g, which the 3-g LMs were not able to capture, we take a closer look at the results for the following sentences (also listed in **Table 5**):

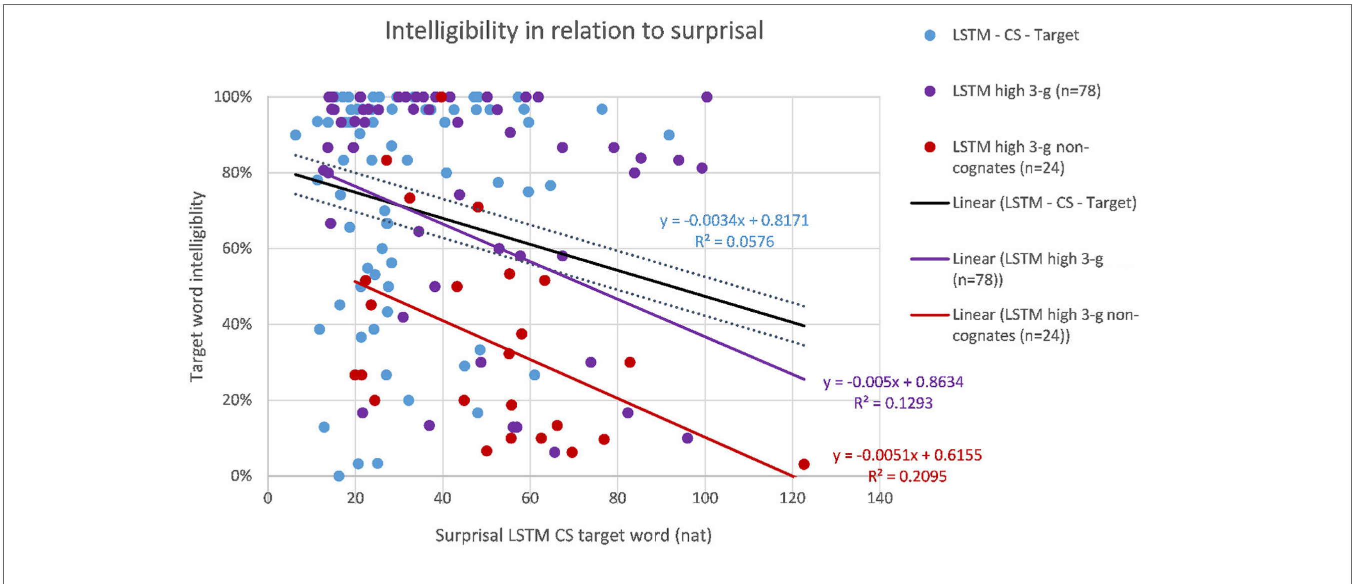


FIGURE 4 | Intelligibility of target words (including filtered subsets) and surprisal from the CS LSTM model.

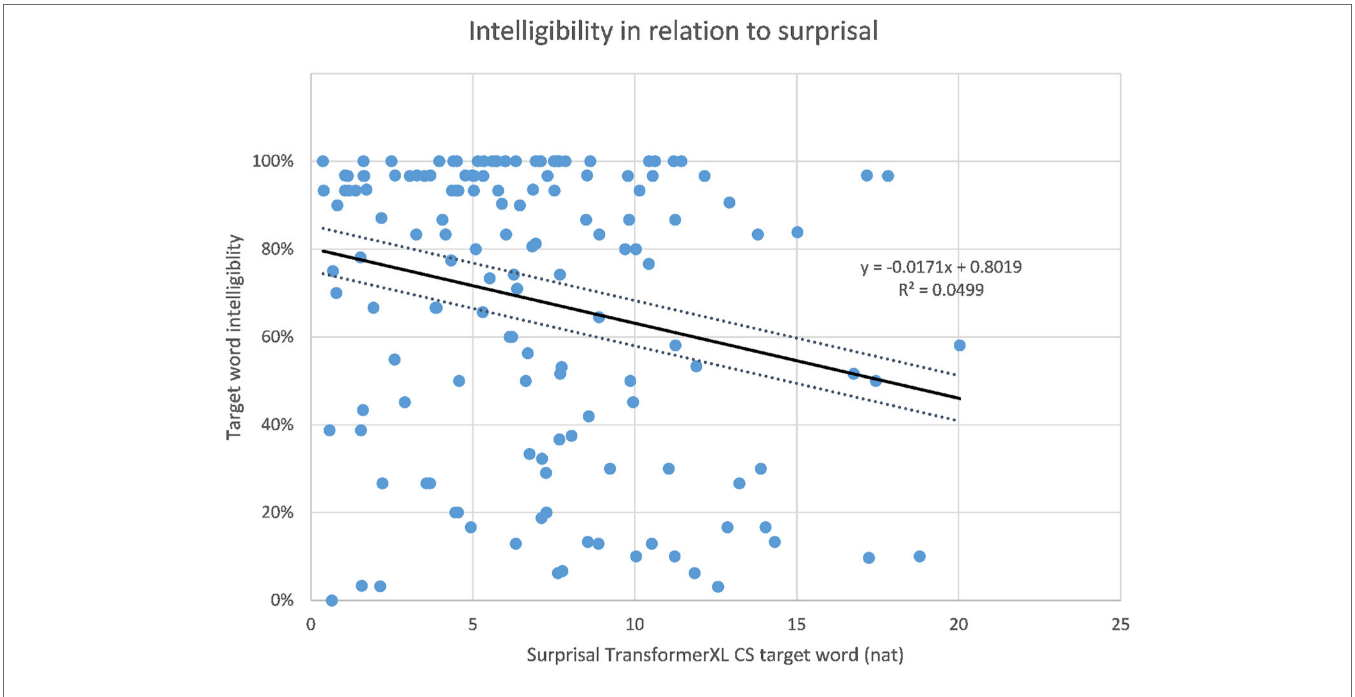


FIGURE 5 | Intelligibility of target words and surprisal from the CS Transformer XL model.

(5) PL: *Farmer spędził ranek dojąc swoje* ***krowy***.
(CS: *Farmář strávil ráno tím, že dojil svoje* ***krávy***.)
"The farmer spent[2] the morning milking his **cows**."
(Block and Baldwin, 2010)

(6) PL: *Ellen lubi poezję, malarstwo i inne formy* ***sztuki***.
(CS: *Ellen má ráda poezii, malířství a jiné formy* ***umění***.)
"Ellen enjoys poetry, painting, and other forms of **art**."
(Block and Baldwin, 2010)

(7) PL: *Sportowiec lubi chodzić na podnoszenie ciężarów na* ***siłownię***.
(CS: *Sportovec rád chodí na vzpírání do* ***posilovny***.)
"The sportsman likes to do weightlifting at the **gym**."
(Block and Baldwin, 2010)

The mean surprisal scores, their SEs, and coefficients of variance for all sentences ($n = 149$) are indicated at the bottom of **Table 5** for the different models. All surprisal scores below the mean of the whole dataset (i.e., low surprisal) are marked in bold font. Note that the surprisal

[2] The original stimulus as of Block and Baldwin (2010) uses *spend*.

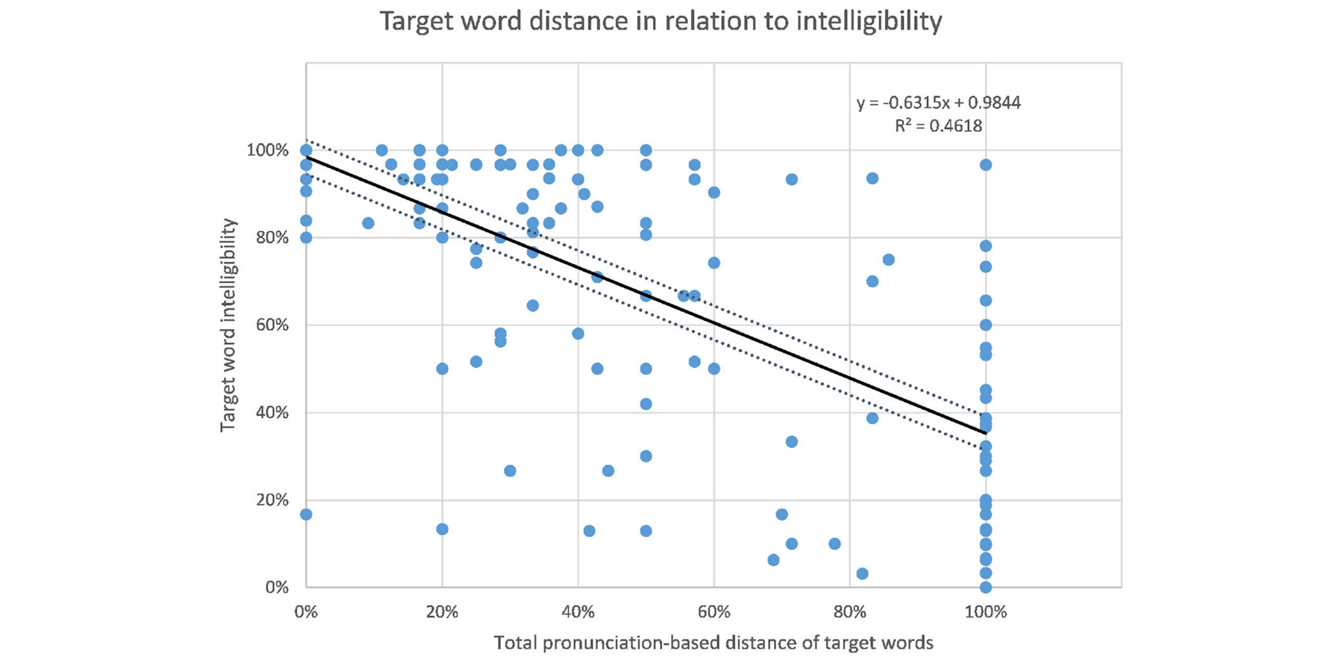


FIGURE 6 | Relation of target word intelligibility and target word distance.

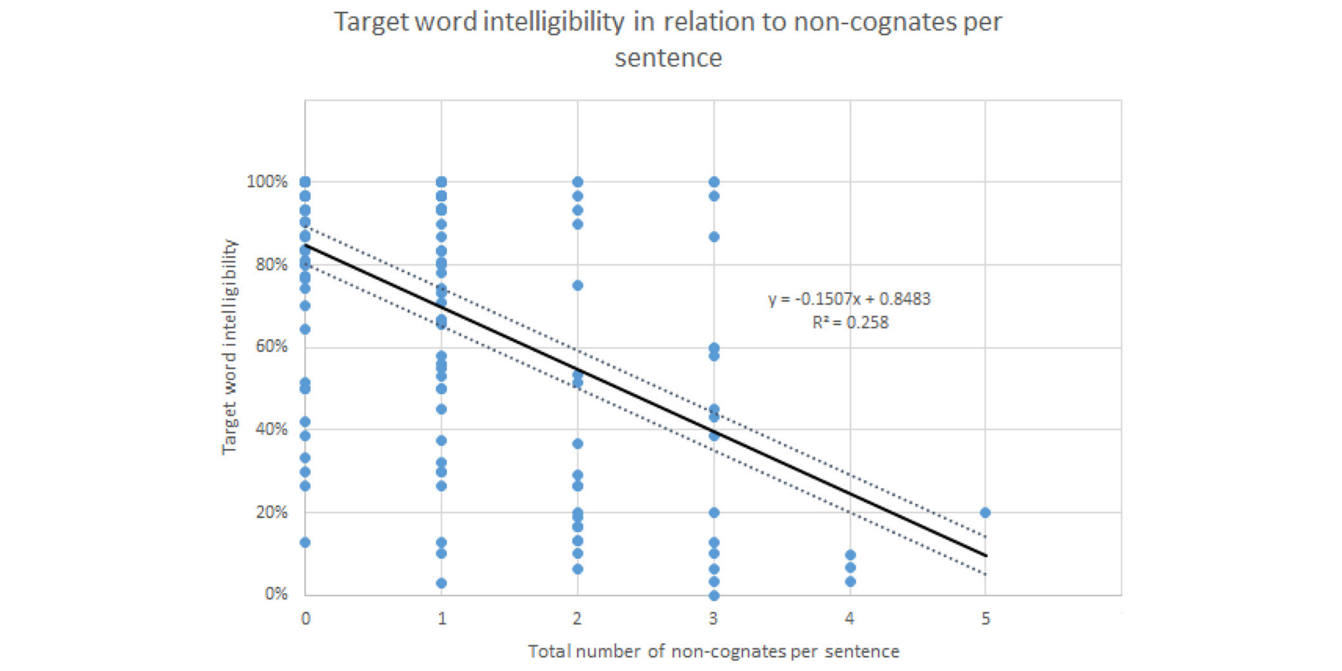


FIGURE 7 | Relation of target word intelligibility and the number of non-cognates per sentence.

for the 3-g models is given in *Hart* (log base 10) while our models use the unit *nat* (log base *e*). While they are not directly comparable, their correlations with intelligibility and the difference to the means for the same models can be compared.

As mentioned earlier, the predictability of the target word *głosu* in example (1) was already reflected well by the 3-g LMs and is also reflected well by the surprisal from the Transformer and Transformer XL model, but surprisingly not by the LSTM. Also, all models assigned a low surprisal to the target *dzień*

TABLE 5 | Surprisal scores: 3-g vs. context-aware LMs of example sentences 1–7 (surprisal below the mean of the whole dataset is marked bold).

Example	Target	3-g PL	3-g CS	3-g PL sentence	LSTM CS	Trans CS	TransXL CS
1	*głosu*	**038**	**0.22**	26.76	40.47	**0.82**	**0.40**
2	*gwoździa*	5.88	6.16	36.19	55.28	10.75	11.90
3	*dzień*	**0.75**	**3.63**	**21.78**	**27.17**	**2.81**	**3.88**
4	*drzewo*	**1.61**	**0.44**	29.28	**21.38**	8.77	7.67
5	*krowy*	3.86	4.05	30.42	**23.03**	**3.70**	**1.06**
6	*sztuki*	3.52	**2.34**	29.03	**22.84**	**3.83**	**2.59**
7	*siłownię*	4.15	5.58	26.52	57.75	8.31	11.25
Mean surp all		3.14	3.85	24.74	38.97	7.59	6.94
SE all		1.76	1.96	5.91	22.73	4.34	4.25
CV (%)		56.12	50.96	23.89	58.34	57.12	61.29

in example (3), suggesting greater ease of cognitive processing, although its intelligibility was lower in context than without any context. We can observe that the predictability of the target words in examples (5) and (6) is better reflected by the context-aware LMs when compared to the 3-g LMs since their target word surprisals are considerably below average. This suggests that the context-aware models can capture the implication (farmer and cows) in example (5) or the relation of art with poetry and painting in example (6). In the case of example (5) it is likely that the high surprisal score of the 3-g LMs is due to the low corpus frequency of the present participle form *dojąc* "milking" (as opposed to the more frequent infinitive *doić* "to milk"). In this study, respondents can in the first place rely on target word similarity: PL *krowy* and CS *krávy* "cows" are cognates with a pronunciation-based distance of only 20%. In example (6), however, a correct response can be based only on expectations, since the target word *sztuki* "art [genitive]" is a non-cognate to CS *umění*. The remaining sentence in example (6) consists of cognates and should thus be understandable. A possible inference might be drawn through *štyk* as it occurs in the CS compound and Germanism *majstrštyk* "masterpiece" (or through knowledge of German) which might in addition to the context evoke an association with the concept of art and hence lead the respondent toward a correct understanding of the target.

However, all of the models assigned a relatively high surprisal to the target word *siłownię* in example (7) and *gwoździa* "nail [genitive]" in example (2). In example (2), this might be because PL *gwoździa* and forms of its CS translation equivalent *hřebík* have very low corpus frequencies in general. It could have been expected that the occurrence of the words for *hanging* and *picture* might lead to the predictability of the context-aware models and hence lower surprisal of *hammer and nail*, but, judging from the surprisal scores, these concepts most likely do not co-occur often enough in the training corpora. As for what could be expected regarding the transformation of the target word *gwoździa* by the reader model (section 4.6), the model transforms PL *ź* into CS *ž*, resulting in *gwoždzia*, which is then scored by the CS model. Since this string of characters is rather unusual in CS, it is no surprise that the surprisal score from this model is rather high. Accordingly, PL *ł* in the preceding collocate *młotka* "hammer [genitive]" is transformed into CS *l* and not into the linguistically correctly corresponding root *mlat* or *mlát*, so that *mlotka* is scored by the model. Since this is a non-word in CS, it is unlikely to lead to a lower surprisal of *gwoždzia*.

Despite its high surprisal score, the target word *siłownię* "gym" was translated more often correctly as a form of CS *posilovna* in context (58.1%) than without context (30.3%). The whole sentence should be more or less understandable for the CS respondents: PL *sportowiec* "sportsman" is an orthographically relatively close cognate to CS *sportovec*, PL *lubi* "likes" can be inferred from CS *líbit* (*se*) "to like [reflexive]," PL *chodzić* "to go" through CS *chodit*, the preposition *na* is, in this case, identical in form and meaning in both languages, PL *podnoszenie* "lifting" can be segmented into the prefix *pod* "under," which is again identical in both languages, and *noszenie* which is related to CS *nošení* "carrying." The only problem here could be in PL *ciężarów* "weights [genitive plural]": Although it contains the Pan-Slavic root *cięż*, which linguistically corresponds to the CS root *těž*, a non-linguist respondent cannot be expected to know of the applicable regular cross-lingual correspondence of *cię:tě* (PL:CS). While CS uses the term *vzpírání* "weightlifting," PL uses the noun phrase *podnoszenie ciężarów* (literally *lifting of weights*) in which *podnoszenie* is post-modified with the genitive plural *ciężarów*. Hence, it is not expected that *ciężarów* is understood, but this might not have a negative influence on the overall understanding of the topic of the sentence, which seems to be help understand the target word. However, this also means that while the final CS 3-g contains the whole concept of weightlifting, only *weights* [genitive plural] is part of the final 3-g in PL. It appears as if the correct understanding of the target *siłownię* is supported by correct identification of the concept of sports and the PL keyword *sportowiec* (CS *sportovec*) "sportsman" at the sentence onset, which can result in associative priming. However, it also appears as if neither of the context-aware LMs performed better than the 3-g LMs in reflecting the predictability of the target word.

A relatively low surprisal was assigned to the PL target word *drzewo* in example (4) by the 3-g LMs and by the CS LSTM model, but not by the Transformer and Transformer XL. PL *drzewo* "tree" is a frequent collocate of PL *genealogiczne* "genealogical" just as CS *strom* "tree" is a frequent collocate of CS *genealogický* "genealogical." In this particular example in which the directly preceding word is a frequent collocate, the 3-g LMs and the LSTM reflect predictability of the target word better than the Transformer LMs.

5.4. Controlling for Local Context

We filtered the original dataset ($n = 149$) for sentences for which the 3-g LMs did not reflect predictability of the target word, i.e., sentences with 3-g surprisals above the mean (≥3.2 *Hart* for PL; ≥3.9 *Hart* for CS, cf. **Table 5**). When we correlate the target word surprisals from the CS context-aware LMs for these sentences ($n = 78$) with the intelligibility of the target words, the correlation of surprisal from the CS LSTM model proves to be higher than the best correlation for the whole set of sentences [$r(78) = -0.35, p < 0.05$]. It has to be noted that for the same subset, the correlation did not improve with surprisal from the

CS Transformer and the CS Transformer XL. This might suggest that the LSTMs perform somewhat better for such sentences in which the 3-g LMs failed. However, since the difference in correlations is still rather small, this effect could also be due to the lower number of data points and hence the lower number of outliers. Since Jágrová and Avgustinova (2019) found that the lexical distance of the targets is crucial and 3-g surprisal correlates better with the intelligibility of those target words that are not cognates, we also filtered the 78 sentences again for sentences with target words that are not cognates ($n = 24$) and obtained a better correlation with intelligibility and surprisal from the CS LSTM [$r_{(24)} = -0.457, p < 0.05$]. The correlations of both filtered subsets are displayed together with the whole dataset for the LSTMs in **Figure 4**.

While the reader model introduced in section 4.6 did not improve correlation, one can still observe examples in which a change in surprisal on the subword level corresponds to what one would also expect from a CS reader. In **Figure 8** this is visualized in an example for the CS and PL locative forms of the word *testament* "testament." On the CS version of the word, the model trained on CS text has a decreasing surprisal. For the last subword *mentu*, the surprisal is low given the previous subwords *te* and *sta*. On the PL version of the word, the surprisal also decreases for the first three subwords *te*, *sta*, and *men* as these are shared between CS and PL. For the last subword *cie*, the surprisal increases, however, as this is not the expected ending of this word in CS. We hypothesize that a similar reaction would be evoked in a CS reader. The segmentation into these units can be explained by the fact that *-cie* is a frequent string of characters at the end of CS nominative singular forms of feminine internationalisms, e.g., *policie* "police," *byrokracie* "bureaucracy," *Francie* "France."

6. DISCUSSION

We investigated whether surprisal obtained from context-aware LMs correlates better with the intelligibility of highly predictable PL target words to CS readers than surprisal obtained from 3-g LMs in a previous experiment. To this end, we trained seven context-aware LMs on large corpora of PL and CS and scored the stimuli and their CS translations with these models. The surprisal values represent the (un-)predictability of words or their (sub-)sequences in relation to the context.

In general, the correlations of intelligibility and surprisal scores obtained from the context-aware models are slightly higher than the correlations with surprisal from the 3-g LMs. It has to be noted that the differences between these correlations are rather small and the correlations themselves are very low. The highest correlation of intelligibility and surprisal from the LSTMs does not exceed a coefficient of $r = -0.46$, $p < 0.05$ in a number of selected sentences with lexically distant target words, which means that surprisal as an indicator of the predictability of words in context cannot explain more than 21% of the variance in the underlying data. Hence, it has to be noted that target word predictability in context appears to be only one of many other stimulus-related factors (linguistic distance, neighborhood density, associations, interferences from other acquired languages, and divergent grammatical gender) influencing the intelligibility of words in closely related languages in general, not to mention the many possible respondent-related factors that were not elaborated on in this study. Surprisal as a representation of predictability in context does not reach the level of the correlations with the linguistic distance that was many times demonstrated in previous research (e.g., Gooskens, 2007; Vanhove, 2014; Möller and Zeevaert, 2015; Vanhove and Berthele, 2015; Golubović, 2016; Jágrová et al., 2017; Stenger et al., 2017).

In the examples, it appears that the context-aware LMs perform better than 3-g LMs particularly in such sentences, where the helpful part allowing for an association with the correct translation lies outside of the window of two words preceding the target word, i.e., at another position in the sentence than the final 3-g, which is at least in some cases reflected in the lower surprisal scores of the highly predictable target words. However, the 3-g LMs and the LSTMs appear to represent predictability of direct collocates of the target words in the examples better than the context-aware LMs that take more context into account. Nevertheless, the examples discussed in this study were chosen to shed light on the possible processes in the first place and one should not generalize and draw conclusions as to the whole dataset. Also, it has to be noted that the LMs were trained on written language and that human performance in these experiments might be much more influenced by everyday language, which could explain why at least some of the models failed in example (5) and all models failed in example (2) since there might not be many texts about farming or handcraft in the corpora.

We found that the reader model (section 4.6), designed to observe whether cross-lingual similarities can be taken into account with such a type of language model, was only to a certain extent able to predict the greater difficulty of unexpected sequences. This outcome is open for interpretation. It is possible that this model of the reader does not perform ideally, since it also aligns incorrect correspondences, such as *ć:č* (PL:CS) based on interpretations of the respondents. Consequently, when the CS respondent, for instance, encounters the PL infinitive form *bawić* (*się*) "to play," the model can approximate that the CS reader will interpret the verb as the noun *bavič* "entertainer," which is, of course, considered a wrong response in the experiment. The reader model will thus calculate the predictability of *bavič* in the sentence according to the CS model and not the predictability of the correct CS translation of the verb equivalent *hrát* (*si*) "to play." Nevertheless, it was demonstrated how such a cross-lingual model could work to support a linguistically reasonable model of the reader. Improved modeling of the reader with regard to cross-lingual similarity, also taking linguistic distance into account, could be an interesting avenue for future work. Moreover, predicting the effects of misleading dominant concepts in sentences or interference not only from the L1 of the reader but also from other acquired languages, remains a topic for future research in the field of intercomprehension.

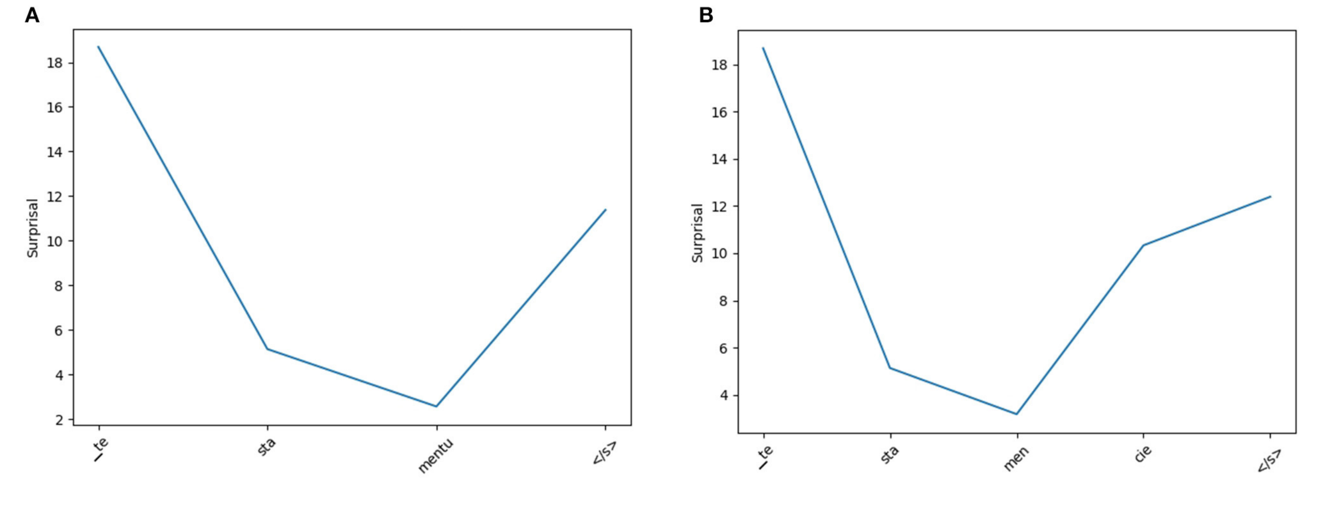


FIGURE 8 | Subword level surprisal of the CS reader model when applied on CS *testamentu* **(A)** and PL *testamencie* **(B)** (both "testament [locative]"). From the perspective of a CS reader, the model displays a rise in surprisal at the unexpected subword *cie* (PL) as opposed to the CS subword units.

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.

AUTHOR CONTRIBUTIONS

KJ initiated this study, gathered and supplied the data from intercomprehension experiments, developed the research question, and performed the analysis of the results (correlations, comparisons, and linguistic interpretations of these). MM and MH developed the language modeling part and ran the technical experiments. TA and DK advised on the project. All authors contributed to the writing.

FUNDING

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 232722074—SFB 1102.

ACKNOWLEDGMENTS

We wish to thank the three reviewers for their constructive and helpful comments on this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.662277/full#supplementary-material

REFERENCES

Al-Rfou, R., Choe, D., Constant, N., Guo, M., and Jones, L. (2019). "Character-level language modeling with deeper self-attention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33 (Palo Alto, CA: AAAI Press), 3159–3166. doi: 10.1609/aaai.v33i01.33013159

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* 5, 157–166. doi: 10.1109/72.279181

Block, C., and Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: behavioral and neural validation using event-related potentials. *Behav. Res. Methods* 42, 665–670. doi: 10.3758/BRM.42.3.665

Čermák, F., and Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *Int. J. Corpus Linguist.* 17, 411–427. doi: 10.1075/ijcl.17.3.05cer

Czapla, P., Howard, J., and Kardas, M. (2018). "Universal language model fine-tuning with subword tokenization for Polish," in *PolEval 2018 Workshop Proceedings* (Warsaw).

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. (2019). "Transformer-XL: attentive language models beyond a fixed-length context," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence: Association for Computational Linguistics), 2978–2988. doi: 10.18653/v1/P19-1285

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1207/s15516709cog1402_1

Golubović, J. (2016). *Mutual Intelligibility in the Slavic Language Area*. Groningen: Rijksuniversiteit Groningen.

Gooskens, C. (2007). The contribution of linguistic factors to the intelligibility of closely related languages. *J. Multiling. Multicult. Dev.* 28, 445–467. doi: 10.2167/jmmd511.0

Heinz, C. (2009). Semantische Disambiguierung von false friends in slavischen L3: die Rolle des Kontexts. *Z. Slawistik* 54, 145–166. doi: 10.1524/slaw.2009.0013

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Jágrová, K. (2018). Processing effort of Polish NPs for Czech readers-A+N vs. N+A. Canonical and non-canonical structures in Polish. *Stud. Linguist. Methodol.* 12, 123–143. Available online at: http://www.coli.uni-saarland.de/~tania/CANONICAL_PL_preprint.pdf

Jágrová, K. (2021). *Reading Polish with Czech eyes. Distance and surprisal in qualitative, quantitative and error analyses of mutual intelligibility* (Ph.D. thesis), Saarland University, Saarbrücken, Germany.

Jágrová, K., and Avgustinova, T. (2019). "Intelligibility of highly predictable Polish target words in sentences presented to Czech readers," in *Proceedings of CICLing: International Conference on Intelligent Text Processing and Computational Linguistics* (La Rochelle).

Jágrová, K., Avgustinova, T., Stenger, I., and Fischer, A. (2019). Language models, surprisal and fantasy in Slavic intercomprehension. *Comput. Speech Lang*. 53, 242–275. doi: 10.1016/j.csl.2018.04.005

Jágrová, K., Stenger, I., and Avgustinova, T. (2017). Polski nadal nieskomplikowany? Interkomprehensionsexperimente mit Nominalphrasen [Is Polish still uncomplicated? Intercomprehension experiments with noun phrases]. Polnisch in Deutschland. *Z. Bundesverein. Polnischlehrkr*. 5, 20–37. Available online at: http://polnischunterricht.de/wp-content/uploads/2018/02/www_gazeta_2017.pdf

Jelinek, F., and Mercer, R. L. (1980). "Interpolated estimation of Markov source parameters from sparse data," in *Proceedings, Workshop on Pattern Recognition in Practice*, eds E. S. Gelsema and L. N. Kanal (Amsterdam), 381–397.

Kingma, D. P., and Ba, J. (2015). "Adam: a method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, eds Y. Bengio and Y. LeCun (San Diego, CA).

Kneser, R., and Ney, H. (1995). "Improved backing-off for m-gram language modeling," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1 (Detroit, MI: IEEE), 181–184. doi: 10.1109/ICASSP.1995.479394

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., et al. (2015). *Syn2015: reprezentativní korpus psané češtiny*. Prague: Ústav Českého narodního korpusu FF UK. Available online at: http://www.korpus.cz

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., et al. (2016). *SYN v4: Large Corpus of Written Czech*. Prague: LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kudo, T., and Richardson, J. (2018). "SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Brussels: Association for Computational Linguistics), 66–71. doi: 10.18653/v1/D18-2012

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10, 707–710.

Möller, R., and Zeevaert, L. (2015). Investigating word recognition in intercomprehension: methods and findings. *Linguistics* 53, 313–352. doi: 10.1515/ling-2015-0006

Mosbach, M., Stenger, I., Avgustinova, T., and Klakow, D. (2019). "incom.py–A toolbox for calculating linguistic distances and asymmetries between related languages," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)* (Varna), 810–818. doi: 10.26615/978-954-452-056-4_094

Muikku-Werner, P. (2014). Co-text and receptive multilingualism-Finnish students comprehending Estonian. Eesti ja soome-ugri keeleteaduse ajakiri. *J. Eston. Finno Ugric Linguist*. 5, 99–113. doi: 10.12697/jeful.2014.5.3.05

Ogrodniczuk, M. (2012). "The Polish sejm corpus," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (Istanbul: European Language Resources Association [ELRA]), 2219–2223.

Ogrodniczuk, M., and Kobyliński, L. (Eds.). (2018). *Proceedings of the PolEval 2018 Workshop* (Warsaw: Institute of Computer Science, Polish Academy of Sciences).

Przepiórkowski, A., Bańko, M., Górski, R. L., and Lewandowska-Tomaszczyk, B. (2012). *Narodowy korpus jezyka polskiego*. Warsaw: Wydawnictwo Naukowe PWN.

Sennrich, R., Haddow, B., and Birch, A. (2016). "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Berlin: Association for Computational Linguistics), 1715–1725. doi: 10.18653/v1/P16-1162

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res*. 15, 1929–1958. Available online at: https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm_campaign=buffer&utm_content=buffer79b43&utm_medium=social&utm_source=twitter.com

Stenger, I., Avgustinova, T., and Marti, R. (2017). "Levenshtein distance and word adaptation surprisal as methods of measuring mutual intelligibility in reading comprehension of Slavic languages," in *Computational Linguistics and Intellectual Technologies: International Conference 'Dialogue 2017' Proceedings*, Vol. 16 (Moscow), 304–317.

Sundermeyer, M., Schlüter, R., and Ney, H. (2012). "LSTM neural networks for language modeling," in *Thirteenth Annual Conference of the International Speech Communication Association* (Portland).

Vanhove, J. (2014). *Receptive multilingualism across the lifespan* (Ph.D. thesis), Université de Fribourg, Fribourg, Switzerland.

Vanhove, J., and Berthele, R. (2015). Item-related determinants of cognate guessing in multilinguals. *Crosslinguist. Influence Crosslinguist. Interact. Multiling. Lang. Learn*. 95:118. Available online at: https://core.ac.uk/download/pdf/43669306.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*, Vol. 30, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (Long Beach, CA: Curran Associates, Inc.), 5998–6008.

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: 12 July 2021
doi: 10.3389/fpsyg.2021.648200

Cross-Linguistic Trade-Offs and Causal Relationships Between Cues to Grammatical Subject and Object, and the Problem of Efficiency-Related Explanations

*Natalia Levshina**

Neurobiology of Language Department, Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands

Cross-linguistic studies focus on inverse correlations (trade-offs) between linguistic variables that reflect different cues to linguistic meanings. For example, if a language has no case marking, it is likely to rely on word order as a cue for identification of grammatical roles. Such inverse correlations are interpreted as manifestations of language users' tendency to use language efficiently. The present study argues that this interpretation is problematic. Linguistic variables, such as the presence of case, or flexibility of word order, are aggregate properties, which do not represent the use of linguistic cues in context directly. Still, such variables can be useful for circumscribing the potential role of communicative efficiency in language evolution, if we move from cross-linguistic trade-offs to multivariate causal networks. This idea is illustrated by a case study of linguistic variables related to four types of Subject and Object cues: case marking, rigid word order of Subject and Object, tight semantics and verb-medial order. The variables are obtained from online language corpora in thirty languages, annotated with the Universal Dependencies. The causal model suggests that the relationships between the variables can be explained predominantly by sociolinguistic factors, leaving little space for a potential impact of efficient linguistic behavior.

Keywords: efficiency, trade-offs, causal networks, subject, object

OPEN ACCESS

Edited by:
Gerhard Jäger,
University of Tübingen, Germany

Reviewed by:
Agnieszka Konopka,
University of Aberdeen,
United Kingdom
T. Florian Jaeger,
University of Rochester, United States

****Correspondence:***
Natalia Levshina
natalia.levshina@mpi.nl

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

Received: *31 December 2020*
Accepted: *15 June 2021*
Published: *12 July 2021*

Citation:
Levshina N (2021)
Cross-Linguistic Trade-Offs
and Causal Relationships Between
Cues to Grammatical Subject
and Object, and the Problem
of Efficiency-Related Explanations.
Front. Psychol. 12:648200.
doi: 10.3389/fpsyg.2021.648200

SOME PROBLEMS WITH EFFICIENT TRADE-OFFS

In recent years there have been quite a few cross-linguistic studies that investigate trade-offs between different communicative or cognitive costs. It is often claimed that these trade-offs are explained by the need to support efficient communication. For example, Kemp et al. (2018) argue that lexical systems of kinship words or color terms demonstrate a trade-off between cognitive costs (number of rules needed to describe a system) and communicative costs (divergence between the probability distributions of the speaker and the addressee). Coupé et al. (2019) find a trade-off between information rate and speech rate, which, on the one hand, saves language users from cognitive overload, and helps to save time, on the other hand.

Similarly, Koplenig et al. (2017) demonstrate a trade-off between information conveyed by word order and word structure, represented by information-theoretic measures and based on corpus data from almost 1,000 languages. Isolating languages, such as Mandarin Chinese, have high scores on information conveyed by word order, but low scores on information carried by word structure. In contrast, polysynthetic languages, such as Ojibwa and Greenlandic Inuktitut, have high word structure scores, but low word order scores. Koplenig et al. (2017) interpret this correlation as an efficient trade-off: Language users can dispense with morphological marking when word order provides sufficient information about the message.

A more specific trade-off is related to the expression of grammatical subject. Berdicevskis et al. (2020) provide typological data showing that languages that have subject indexing (verbal affixes and clitics) more frequently allow for omission of subject pronouns, although this trend is not supported in Eurasia. They also use corpora of East Slavic languages to show that that the absence of person indexation in past tense encourages speakers to encode accessible subject referents by independent pronouns significantly more often (note that this tendency is also observed in some other Slavic languages, where person is always marked). The results are interpreted in terms of efficiency: Information should be conveyed linguistically, but redundancy is undesirable.

Inverse correlations between different linguistic variables have enjoyed considerable attention in research on linguistic complexity. For example, Fenk-Oczlon and Fenk (2008) argue for the following trade-offs between different language subsystems:

- Phonological complexity (e.g., large phonemic inventory, complex syllable structure, and high number of syllable types) vs. morphological complexity (e.g., high number of morphemes per word and low number of monosyllabic words);
- Morphological complexity (see above) vs. semantic complexity (polysemy and synonymy);
- Semantic complexity (see above) vs. word order complexity (e.g., flexible word order, which has low predictability and implies that language users have to learn many additional stylistic rules).[1]

As an illustration, compare English and Russian. English has a higher number of syllable types, shorter words with fewer morphemes, higher lexical and grammatical ambiguity and rigid word order. In contrast, Russian has fewer syllable types, longer words with more morphemes, lower ambiguity and more flexible word order. At least some of these trade-offs can be interpreted in terms of efficiency. The trade-off between phonological and morphological complexity is in accordance with Menzerath–Altmann's law (Altmann, 1980), which predicts an inverse correlation between word length and syllable length. Stave et al. (2021) argue that this trade-off is efficient: it allows language users to save costs needed for working memory and planning. The trade-off between semantic and word-order complexity can be explained by the fact that ambiguous words rely on their context for assignment of lexico-semantic and grammatical properties (cf. Piantadosi et al., 2012; Hawkins, 2019).

An assumption behind these and similar claims is that language users tend to avoid both linguistic overspecification and underspecification when expressing certain information. This tendency can be interpreted as rational and efficient behavior. So, one might expect that different types of linguistic cues that express similar information will be negatively correlated. And the other way round, negative correlations could be interpreted as a sign of efficient behavior.

These assumptions are not as self-evident as they may seem, however. First of all, aggregate variables, such as the presence of case marking or flexible word order in a language, do not take into account the joint distribution of cues in usage contexts. While this lack of information may be irrelevant for languages with categorical values on linguistic variables (e.g., total lack of case marking vs. obligatory case marking without case syncretism; or perfectly rigid vs. completely random word order), this creates problems for languages with in-between values, such as optional or differential case marking, or a dominant but not exclusive word order. In fact, these are the majority of languages (e.g., Sinnemäki, 2014a; Levshina, 2019). In this case, there is a possibility of one clause containing two or zero cues, which means overspecification or underspecification, respectively. A trade-off at the aggregate level can mask these uses. Therefore, not all inverse correlations between linguistic variables representing different cues can be interpreted as a sign of efficient behavior.

Second, an inverse correlation between two linguistic variables can disappear or become weaker if we control for a third variable (e.g., Levshina, 2020a). Most importantly, we need to control for the role of accessibility of information from context in a broad sense (that is, including linguistic context, situational, and encyclopedic information), which itself is in a trade-off relationship with the amount of linguistic coding required. This trade-off has been observed in studies of phonological reduction (Jurafsky et al., 2001; Aylett and Turk, 2004; Cohen Priva, 2008; Seyfarth, 2014; Jaeger and Buz, 2017; Hall et al., 2018). In the lexicon, there is a correlation between predictability (defined in different ways) and word length (Zipf, 1965[1935]; Manin, 2006; Piantadosi et al., 2011; Mahowald et al., 2013). The length of referential expressions is known to depend on their accessibility (Ariel, 1990), which is determined by common ground (Clark and Wilkes-Gibbs, 1986). As for morphosyntactic coding asymmetries and splits, it is well known that more predictable grammatical meanings are expressed by shorter forms (including zero) than less predictable ones (e.g., Jäger, 2007; Kurumada and Jaeger, 2015; Kurumada and Grimm, 2019; Haspelmath, 2021). Lemke et al. (2021) demonstrate that fragments (i.e., incomplete sentential structures) encoding events known from everyday scripts and scenarios are perceived as more natural than fragments encoding unpredictable events. See more examples in Hawkins (2004), Jaeger and Tily (2011), and Gibson et al. (2019). That is, if some meaning is highly predictable from context or in general, it is efficient to use no overt cues

[1] Notably, Fenk-Oczlon and Fenk disagree on what makes word order more or less complex. This is symptomatic of complexity research, with many possible definitions.

at all[2]. For example, it is known that the subject of canonical imperatives does not have to be overtly expressed in the vast majority of the world's languages, especially if the addressee is singular (Aikhenvald, 2010). If some meaning is difficult to retrieve, it may be perfectly efficient to use multiple cues. For instance, the use of resumptive pronouns, as in Hebrew and Cantonese, in certain types of relative clauses can be efficient because it makes processing easier in structurally more complex environments (Hawkins, 2004). Another case is clitic doubling in some high-contact varieties, such as languages of the Balkan Sprachbund, which means that some objects are expressed twice[3]. According to Wiemer and Hansen (2012: 127), it helps "speakers in multilingual settings of a primarily oral culture . . . to achieve the most reliable degree of mutual intelligibility." So, a negative correlation between *linguistic* cues does not tell us much about efficiency if other factors, such as predictability and ease of processing, are not controlled for.

Moreover, the use of linguistic cues is multifunctional. For example, in addition to helping to identify main grammatical roles, constituent order can also allow language users to manage information structure, to facilitate production by putting accessible elements first (e.g., Bock and Warren, 1985; Ferreira and Yoshita, 2003), to maximize early access to semantic and grammatical structure (Hawkins, 2004), to save memory costs by minimizing dependency distances or syntactic domains (Hawkins, 2004; Ferrer-i-Cancho, 2006; Liu, 2008; Futrell et al., 2015), and so on. There is also a claim (Maurits, 2011) that constituent orders that frequently occur in the world's languages make information density more uniform, avoiding peaks and troughs (Jaeger, 2006; Levy and Jaeger, 2007). This means that the overall communicative efficiency of a certain language system depends on multiple parameters, which need to be taken into account.

In addition, language users' communicative preferences are not the only factor that shapes language structure. An important role is played by analogy (Haspelmath, 2014) and by diverse frequency effects (Bybee, 2010). In addition, many language changes are attributed to sociolinguistic factors. Under normal circumstances, for example, languages tend to accumulate morphological complexity (Dahl, 2004), but an increase in the proportion of adult L2 speakers and population size can lead to simplification and loss of inflectional morphology (McWhorter, 2011). Cross-linguistic studies reveal inverse relationships between morphological complexity and population size (Lupyan and Dale, 2010) and proportion of L2 speakers (Bentz and Winter, 2013). Fenk-Oczlon and Pilz (2021) find that languages with more speakers tend to have larger phoneme inventories, shorter words in number of syllables and a higher number of words per clause, among other things[4]. It therefore does not necessarily follow that changes in language structure should be attributed solely to the pressure for communicative efficiency, i.e., the balance between robust information transfer and articulation and processing costs, which rational language users try to achieve.

It is also important to keep in mind that transfer of information between the speaker and the addressee takes place in a noisy channel (Shannon, 1948 ; Gibson et al., 2019). This means that a message from Speaker to Addressee can be corrupted on the way – due to external noise, or due to production and processing errors. Therefore, there is a possibility that not all cues to a particular meaning or function are recovered from the signal. Producing only one cue to express a certain meaning may not be enough. In fact, typologists find redundancy at all linguistic levels (Hengeveld and Leufkens, 2018).

It is not surprising then that not all potential trade-offs are detected in actual linguistic data. For example, Sinnemäki (2008) finds significant inverse correlations between rigid word order and the presence of case marking of the core arguments in a representative sample of languages (also see below), but no correlation between word order and verb agreement, or verb agreement and case marking. Moreover, different cues may work in synergy. As an illustration, consider verbal and visual cues in communication. One would believe that processing one modality should be at the cost of the other. However, Holler et al. (2018) demonstrate that interlocutors respond faster to questions that have an accompanying manual and/or head gesture, than to questions without such visual components. According to Holler and Levinson (2019), multimodal information is easier to process than unimodal information (at least, for neurotypical speakers) thanks to synergy effects and creation of Gestalts.

To summarize, trade-offs, or inverse correlations, between linguistic variables related to different cues do not automatically imply efficiency as a driving force of language use and change, and the other way round.

I will illustrate these considerations by a case study of linguistic cues that help language users understand "who did what to whom." There are multiple cues that help to infer this information: case marking, verb agreement, word order, and semantics. Languages differ in how they employ these cues. For example, Hungarian has case marking, agreement, but flexible word order (Pleh and MacWhinney, 1997), while others rely mostly on rigid word order, such as Present-Day English or Mandarin Chinese.

In this article, I will focus on four types of cues, which will be obtained from corpora in thirty languages, annotated with the Universal Dependencies (Zeman et al., 2020). The cues are as follows:

- Case marking, measured as Mutual Information between grammatical role and case;
- Semantic tightness, measured as Mutual Information between role and lexeme (lemma);
- Rigid word order, measured as 1 minus entropy of Subject and Object order;

[2] I thank Mira Ariel (p.c.) for sharing this idea.

[3] Thanks to Björn Wiemer (p.c.) for making me aware of this interesting feature.

[4] Fenk-Oczlon and Pilz attribute the inverse correlation between word length and population words to a general increase in frequency of words when population increases, such that more frequent words will undergo formal reduction, according to Zipf's law of abbreviation. But it is not clear how the higher frequency in the entire population would affect predictability of a word for individual speakers, who only communicate within their social networks. A more plausible explanation, in my view, is an increase in phonological inventories due to borrowings, which would allow for more monosyllabic words.

- The proportion of clauses with verb-middle order, which is claimed to facilitate processing in a noisy channel (Gibson et al., 2013).

The role of these cues is discussed in section "Cues for Identification of Subject and Object." The previous studies of these cues in typology focused mostly on binary trade-offs, such as rigid word order vs. case marking (Sinnemäki, 2014b), and case morphology vs. verb-medial order (Sinnemäki, 2010). Other cues and their relationships have received less attention, however, the present study is the first attempt to examine all four cues systematically with the help of quantitative measures and corpus data, which are presented in section "Data and Variables."

Using pairwise correlations, I will show that the relationships are quite complex (see section "A Correlational Analysis of Cross-Linguistic Data"). Not all these cues are correlated, and not all correlations are negative. There is a robust negative correlation, however, between rigid word order and case marking. Next, I will move from binary correlations to causal networks in section "A Causal Analysis of Subject and Object Cues" (cf. Blasi and Roberts, 2017). Causal networks are more informative, because they allow us to identify directional relationships between different variables. There are some studies that employ diverse types of causal inference for different types of linguistic questions (e.g., Moscoso del Prado Martín, 2014; Baayen et al., 2016; Blasi, 2018; Dellert, 2019), but the approach has not yet become mainstream. In this article, I explore how causal inference based on synchronic corpus data can be used in token-based functional typology (Levshina, 2019). This type of corpus-based approach complements recent miniature language learning experiments that investigate the links between communicative efficiency (and other learning biases) and different linguistic cues to the same linguistic meaning (e.g., Culbertson et al., 2012; Fedzechkina et al., 2016; Kanwal et al., 2017; Kurumada and Grimm, 2019; Fedzechkina and Jaeger, 2020, to name just a few). Corpora are a valuable source because they represent language produced in naturalistic settings by real language users. I will demonstrate that some of the corpus-based results converge with previous experimental results (in particular, Fedzechkina et al., 2016; Fedzechkina and Jaeger, 2020), which shows that causal analysis can be added as a useful tool for studying linguistic cues across languages. I interpret the resulting causal network, discussing a possible diachronic scenario, which involves extralinguistic factors, such as the number of adult L2 learners. I argue that the potential for efficient and rational behavior playing a role in this scenario is quite limited.

CUES FOR IDENTIFICATION OF SUBJECT AND OBJECT

Formal Marking

This section describes different cues which can help to communicate "who did what to whom." One type of cues is formal marking, most importantly, case marking and agreement (indexing). Some languages have consistent case marking on either the subject, the object, or both. For example, Lithuanian nouns, with the exception of some loan words, have distinct nominative and accusative case forms in all declension types. Some languages have differential marking, when A or P are marked in some situations, and not marked in others. For example, in Spanish, only animate and specific objects are marked, while other objects are unmarked (see more examples in Aissen, 2003). There are also case systems in which the distinctions between the Nominative and the Accusative forms are made only in some lexical classes, while the forms are identical in others, e.g., inanimate masculine nouns in Russian, e.g., *stol-Ø* "table.NOM/ACC", or neuter nouns in Latin, e.g., *bell-um* "war-NOM/ACC".

In some languages, the marking is probabilistic. An example is Korean (Lee, 2009), where the object markers are more or less likely depending on animacy, definiteness, person, heaviness of the object and other factors. Often, variation is contextual. For example, the Japanese object marker is used more frequently when the role configurations are not typical, e.g., when it is a thief who arrests a policeman, and not the other way round (Kurumada and Jaeger, 2015).

Both in probabilistic and categorical differential marking systems, there is a negative correlation between the presence of the case marker and predictability or accessibility of the role given the semantic and other properties of the nominal phrase. This correlation can be explained by efficiency considerations and rational behavior (e.g., Jäger, 2007; Levshina, 2021).

The arguments can also be marked on the verb. This is called agreement, or indexing. Subject indexing is popular across languages, e.g., German *er komm-**t*** "he comes". As for object indexing, it is less frequently obligatory. The reason is that the relevant grammatical elements usually do not advance further down the cline of grammaticalization and do not become obligatory agreement markers, as it very often happens with subject agreement. Typically, object markers remain at the stage of differential object indexing (Haig, 2018). Their use or omission depends on diverse semantic and pragmatic factors, which are similar to the ones relevant for differential case marking. For example, in Maltese, the index is always present if the object is pronominal and given, and is always absent if it is new and non-specific. In the remaining situations, there is variation (Just and Čéplö, in press)[5]. This means that the use of differential object indexing is efficient.

Word Order Cues

Fixed word order can also help the addressee to understand who did what to whom. It is used as a compensatory strategy in languages without case marking (Sapir, 1921). The position of the verb can be another cue. It is claimed that it is easier to assign the roles when the verb occurs between the subject and the object:

> *[V]erb position is the particular vehicle which most conveniently enables these basic grammatical relations to be expressed by means of word order: the subject occurs to the immediate left, and the object to the immediate right of*

[5] These results are for sentences with canonical (i.e., VO) order. When the order is non-canonical, the object index is always present.

the verb. I.e., the verb acts as an anchor (Hawkins, 1986, 48–49).

In experiments that involve gestural communication, participants prefer SOV when trying to convey a transitive event (Goldin-Meadow et al., 2008; Gibson et al., 2013; Hall et al., 2013). However, when an event is reversible, i.e., both participants can be Subject or Object, such as "The mother hugs the boy" and "The boy hugs the mother", users tend to use SVO more often than when the role assignment is clear (Hall et al., 2013). Notably, some participants in Gibson et al. (2013) used some sort of *ad hoc* "spatial marking" that helps to distinguish between Subject and Object. For example, they used one hand to designate Subject and the other to represent Object, or gestured Subject in one location in space and Object in another. In the presence of such marking, they used the SVO order less frequently. Thus, SVO is used more often in the absence of any – formal or semantic – cues.

How can one explain these findings? Gibson et al. (2013) argue that verb-medial order is more robust to the presence of noise as far as conveying the roles of subject and object are concerned. If the addressee fails to recognize one of the nouns before the verb, he or she will be unable to decide if the noun is a subject or an object. For example, if instead of *The mother the boy hugs*, he only hears, *The mother hugs*, it will be difficult to interpret the role of the argument in the absence of the second nominal phrase, if there are no other cues. But if one noun is before the verb and one is after the verb, then the noise is less disruptive. If the argument that the addressee discerns is before the verb, e.g., *The mother hugs*, it can be identified as the subject. If the noun is after the verb, e.g., *Hugs the boy*, then it should be the object.

At the same time, Hall et al. (2015) show that pantomime comprehenders interpret SOV sequences robustly as subject-first, for both reversible and non-reversible events. This means that the role of ambiguity avoidance is probably less important than previously assumed (cf. Wasow, 2015). It may be that the preference for SVO in production has to do with avoidance of two semantically similar elements in close proximity. In linguistics, one speaks of the *horror aequi* principle, which describes the tendency to avoid placing formally, structurally or semantically similar units close to one another (cf. Ferreira and Firato, 2002; Rohdenburg, 2003; Walter and Jaeger, 2008). In phonology, this constraint is known as the Obligatory Contour Principle (Leben, 1973). By using the SVO order, the signers may avoid interference based on semantic similarity of Subject and Object.

Semantic and Pragmatic Properties of the Arguments

Semantics of the arguments can provide strong cues for assigning the roles. For example, one can expect that it is a dog who bites a man, a hunter who kills a bear, a journalist who interviews a politician, and not the other way round.

There are also strong associations between roles and more abstract referential features, such as animacy, definiteness, discourse status, etc. According to cross-linguistic spoken corpus data, if an argument is human, 1st or 2nd person, definite or discourse-given, it is more likely to be Subject than Object. If an argument is non-human, 3rd person, indefinite or new, it is more likely to be Object than Subject.

Languages differ in how flexible they are with restrictions in the expression of Subject and Object. For example, Lummi (Straits Salish, British Columbia) does not allow the person of the subject argument to be lower on the person scale than the person of a non-subject argument. For example, if the subject in a potential active sentence is 3rd person and the object is 1st or 2nd person, then passivization is obligatory. In English, active sentences of this kind are possible, although there is a tendency to use passive more often in those cases (Bresnan et al., 2001).

A comparison of the associations between grammatical roles and semantics in English and German was performed by Hawkins (1986: 121–127, 1995) and extended cross-linguistically by Müller-Gotama (1994). For instance, Present-Day English has fewer semantic restrictions on the subject and object than Old English or German. Consider several examples below.

(1) a. Locative: *This tent sleeps four.*
b. Temporal: *2020 witnessed a spread of the highly infectious coronavirus disease.*
c. Source: *The roof leaks water.*

This suggests that subjects in English are less semantically restricted than subjects in German and Russian, in which these sentences would sound unnatural or incorrect (see also Plank, 1984). We can also say that English is a "loose-fit" language, while German, as well as Russian, Korean and Turkish, are "tight-fit" languages. A corpus-based study of thirty languages showed that the tightness rankings can be reproduced with the help of Mutual Information between grammatical roles and lexemes (Levshina, 2020b) – a method also used in the present article.

Correlations and Causal Links From Previous Studies

Some correlations between the variables are already known from the previous studies. In particular, there is an inverse correlation between argument marking and rigid word order (Sapir, 1921; Sinnemäki, 2014b). Also, Greenberg's (1966b) Universal 41 says: "If in a language the verb follows both the nominal subject and nominal object as the dominant order, the language almost always has a case system." This means that verb-final order is associated with case marking, while verb-medial order is associated with lack of case marking (Sinnemäki, 2010).

Hawkins (1986) wrote about a positive correlation between verb-finalness and semantic tightness, which has been confirmed empirically (Levshina, 2020b). Moreover, he predicted a positive correlation between case marking and semantic tightness. Verb-final languages should be semantically tight and have case marking because an early incorrect assignment of roles would result in re-analysis, which has high cognitive costs.

As for the causal relationships, we know much less. Some diachronic accounts suggest that word order can determine case marking, according to the principle *post hoc ergo propter hoc*. According to Kiparsky (1996), the shift to VO began in Old English. It happened before the case system collapsed, and also before the loss of subject-verb agreement. Bauer (2009) demonstrates that that the change to VO and rigid word order in Late and Vulgar Latin was before the loss of inflection, which happened later in Romance.

There is also some support of this hypothesis in experimental linguistics. Fedzechkina et al. (2016) had their participants learn a miniature artificial language. The languages contained optional case marking on the object. Some languages had fixed constituent order, and some had flexible order. Learners of the fixed order language produced case marking significantly less often than learners of the flexible order language. In addition, a follow-up study by Fedzechkina and Jaeger (2020) demonstrates that the loss of marking in a fixed-order artificial language is observed only when case production requires additional effort, which indicates that the learners' behavior is motivated by communicative efficiency and not by other considerations.

In the study presented below, I will investigate the correlational and causal relationships between four variables: case marking, rigid word order, verb-medial order and semantic tightness. These variables will be estimated with the help of corpus data, which are described below.

DATA AND VARIABLES

Corpus Data

Available cross-linguistic syntactically annotated collections, such as the Universal Dependencies corpora (Zeman et al., 2020), are too small for the purposes of the present study because one cue type, namely, semantic tightness, requires distributional information about the frequencies of individual lexemes as Subject and Object. This is why I used freely downloadable web-based corpora from the Leipzig Corpora Collection (Goldhahn et al., 2012). These corpora contain collections of randomized sentences in diverse languages. The language sample consists of thirty languages (see **Table 1**). For each language, I took one million sentences representing online news (categories "news" and "newscrawl"). The choice of languages and the sample size were determined by the availability of language models in the UDPipe annotation toolkit, which was used to tokenize, lemmatize and annotate the sentences morphologically and syntactically (Straka and Straková, 2017). The processing was performed with the help of the R package *udpipe* (Wijffels, 2020). Importantly, the models provide uniform parts-of-speech tags and dependency relations (Universal Dependencies), which allows us to compare the data in different languages.

This annotation was used to extract all nominal subjects and objects. Here and below by subjects I mean only subjects of transitive clauses. Intransitive clauses were not taken into account. Pronominal arguments were excluded for the sake of comparability. Some languages are pro-drop, and it would be technically impossible and linguistically incorrect to recover the "missing" pronouns.

Of course, using automatic annotation is risky. Additional checks were performed in order to make sure that the subjects and objects are identified correctly. Moreover, another study (Levshina, 2020a) compared several word order and case marking scores based on the online news corpora and the training corpora in the UD collection. It revealed very strong positive correlations between the scores based on these two data sources, which can serve as an indication that the data are reliable.

TABLE 1 | Languages in this study.

Language	Genus	Family	UD model
Arabic	Semitic	Afro-Asiatic	arabic-padt-ud-2.4
Bulgarian	Slavic	Indo-European	bulgarian-btb-ud-2.4
Croatian	Slavic	Indo-European	croatian-set-ud-2.4
Czech	Slavic	Indo-European	czech-pdt-ud-2.4
Danish	Germanic	Indo-European	danish-ddt-ud-2.4
Dutch	Germanic	Indo-European	dutch-alpino-ud-2.4
English	Germanic	Indo-European	english-ewt-ud-2.4
Estonian	Finnic	Uralic	estonian-edt-ud-2.4
Finnish	Finnic	Uralic	finnish-tdt-ud-2.4
French	Romance	Indo-European	french-gsd-ud-2.4
German	Germanic	Indo-European	german-gsd-ud-2.4
Greek (modern)	Greek	Indo-European	greek-gdt-ud-2.4
Hindi	Indic	Indo-European	hindi-hdtb-ud-2.4
Hungarian	Ugric	Uralic	hungarian-szeged-ud-2.4
Indonesian	Malayo-Sumbawan	Austronesian	indonesian-gsd-ud-2.4
Italian	Romance	Indo-European	italian-isdt-ud-2.4
Japanese	Japanese	Japanese	japanese-gsd-ud-2.4
Korean	Korean	Korean	korean-gsd-ud-2.4
Latvian	Baltic	Indo-European	latvian-lvtb-ud-2.4
Lithuanian	Baltic	Indo-European	lithuanian-hse-ud-2.4
Persian	Iranian	Indo-European	persian-seraji-ud-2.4
Portuguese	Romance	Indo-European	portuguese-bosque-ud-2.4
Romanian	Romance	Indo-European	romanian-rrt-ud-2.4
Russian	Slavic	Indo-European	russian-syntagrus-ud-2.4
Slovenian	Slavic	Indo-European	slovenian-ssj-ud-2.4
Spanish	Romance	Indo-European	spanish-gsd-ud-2.4
Swedish	Germanic	Indo-European	swedish-talbanken-ud-2.4
Tamil	Southern Dravidian	Dravidian	tamil-ttb-ud-2.4
Turkish	Turkic	Altaic	turkish-imst-ud-2.4
Vietnamese	Viet-Muong	Austro-Asiatic	vietnamese-vtb-ud-2.4

Variables

Case Marking

Case marking is represented here as Mutual Information between Role (Subject or Object) and Case (depending on the case inventory in a particular language). In comparison with traditional classifications, such as the number of morphological cases in a language, this method can determine more precisely the amount of information obtained about Role through observing Case in language use. This is particularly important for languages with differential case marking. For example, in Russian some nouns have different forms in the Nominative and Accusative (e.g., *devočk-a* "girl-NOM" and *devočk-u* "girl-ACC"), while some nouns have identical forms (e.g., *stol* "table" or *myš* "mouse"). Similarly, as already mentioned, Korean has variable marking on Subject and Object with complex probabilistic rules (Lee, 2009). In some languages, like Finnish and Estonian, the same morphological cases (e.g., Nominative and Partitive) can express both Subject and Object under certain conditions. The question is then, how frequently do the Subject and Object forms help the addressee to infer the grammatical role of a noun? In order to answer this question, we need a quantitative corpus-based approach.

The frequencies of Role-Case combinations were determined in the following way. In some languages, the roles are marked by adpositions or case particles marking the roles that are treated as separate words by the Universal Dependencies, e.g., the preposition *a* in Spanish. In this case, I simply counted the number of Subjects and Objects with and without these markers, which are marked with the dependency "case." **Table 2** displays the counts for Spanish.

If a language has a special Subject form, which cannot be used to represent Object, I counted in three Cases (rows in the table): strictly the Subject form, the Object form and the ambiguous form, which usually has zero marking. For example, Hindi has three Cases under this approach: absolutive (with zero marking), ergative (only transitive subjects) and accusative (only transitive objects). **Table 3** represents the counts for Hindi. A similar situation is in Japanese and Korean, which have Subject-only particles, Object-only particles, and unmarked forms.

In order to obtain the counts of morphological cases, I used two approaches: automatic and manual. The automatic method was used in simple case systems. I compared the case wordforms with the corresponding lemmas, which represent the Nominative (Subject) case. This is how I obtained the counts for Object forms in several languages. In more complex situations, I analyzed manually samples of 200 Subjects or Objects (or 500, if the system was relatively simple to analyze) with the help of dictionaries, and obtained the counts by extrapolating the frequencies from the sample. This procedure was used in those languages in which automatic comparison of case wordforms with lemmas was problematic because of the presence of other morphemes, e.g., definite articles or possessive suffixes, as in Arabic, Bulgarian, Finnish or Hungarian. **Table 4** displays the extrapolated counts for Finnish. It has Nominative (no marking), Genitive and Partitive cases that are used with Subject and Object. Subjects can be expressed by the zero Nominative and occasionally by Partitive and Genitive forms, while Objects can have no marking, or be in the Partitive or Genitive form with case suffixes.

TABLE 2 | Frequencies of case forms in Spanish.

Case	Subject	Object
Zero marking	126,736	569,252
Preposition *a*	0	55,442

TABLE 3 | Frequencies of case forms in Hindi.

Case	Subject	Object
Absolutive (zero marking)	46,241	363,647
Ergative	61,512	0
Accusative	0	92,510

TABLE 4 | Frequencies of case forms in Finnish (extrapolated).

Case	Subject	Object
Nominative (zero marking)	132,631	94,077
Genitive + Partitive	9,562	386,268

Note that in order to perform the automatic comparison and facilitate the manual annotation, I took only non-plural and non-dual forms in all languages, so that the formal variation based on number could be excluded. I do not expect this restriction to influence the results strongly because plural forms are less frequent than singular ones (Greenberg, 1966a).

German was treated in a special way because the carriers of case information are the articles, pronouns and adjectives, e.g., the nominative form *der Tisch* "the table" is contrasted with the accusative form *den Tisch*. This contrast is only available for masculine nouns. I inferred the number of marked forms by computing the number of masculine singular nouns in the role of Subject and Object, which are modified by determiners or adjectives. Feminine and neuter nouns, as well as the masculine ones without determiners or adjectives, were treated as having ambiguous forms.

Next, for each Case-by-Role frequency table, I computed Mutual Information (MI) between Case and Role:

$$I\ (Case;\ Role) = \sum_{i,j} p\ (case_i,\ role_j)\ log_2 \frac{p\ (case_i,\ role_j)}{p\ (case_i)\ p(role_j)}$$

Finally, in languages without any Subject or Object markers (that is, Danish, Dutch, English, Indonesian, Swedish, and Vietnamese), the MI scores were set to 0. Note that in some case-free languages, e.g., in French, a tiny fraction of objects are marked with a preposition. These are objects representing unspecified quantity, e.g., *Je voudrais de l'eau* "I would like some water."

The MI scores are displayed in **Figure 1**. The languages at the bottom have no or very limited case marking (English, Indonesian, the Romance languages and Vietnamese), while the languages at the top have extensive marking, which contributes substantially to discriminating between Subject and Object (e.g., the Baltic languages and Hungarian). Lithuanian, the Indo-European language that has preserved most of the ancient nominal morphology, has the highest distinctiveness. Most Slavic languages, Hindi, Persian, and Turkish and other languages with differential marking are in the middle, as expected. The low score of Spanish, which has differential object marking, as well, is somewhat surprising. The reason may be that animate specific

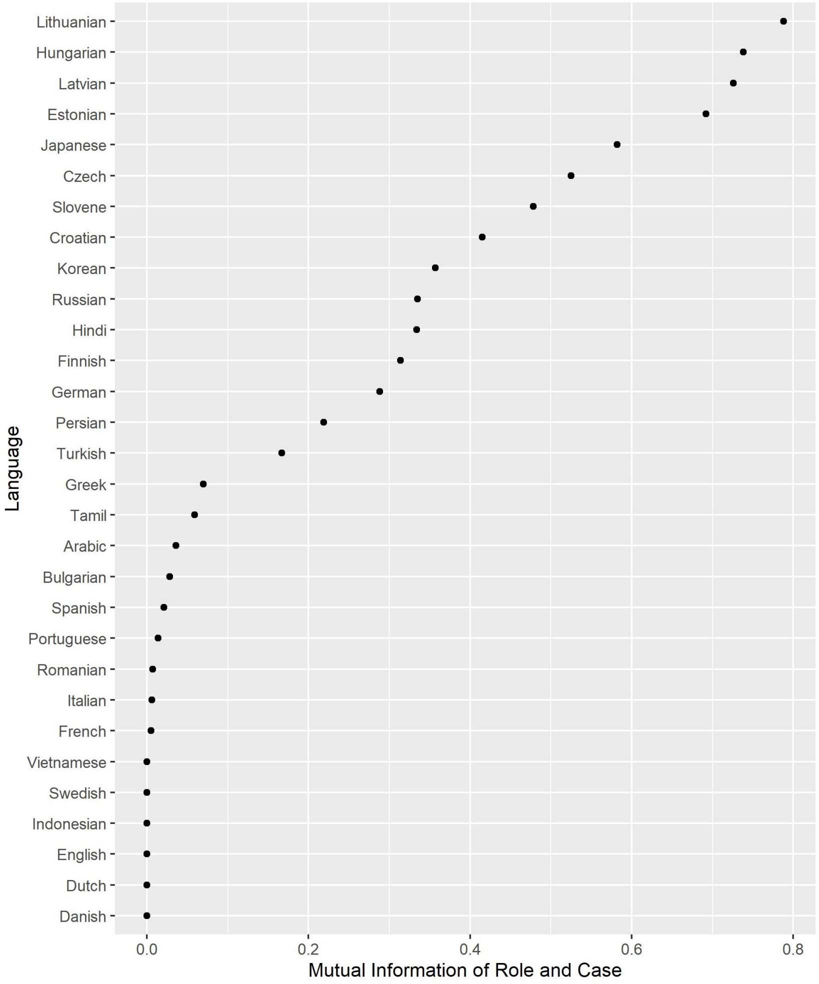


FIGURE 1 | Case marking (Mutual Information between Role and Case).

objects, which are marked with the preposition a, are much rarer than other nominal phrases (see **Table 2**).

Agreement markers are not investigated in this article. There are several reasons. First, it is difficult to quantify how much they help to distinguish between Subject and Object. Second, previous research has shown that subject agreement is not significantly correlated with other cues, such as word order or case marking (Sinnemäki, 2008). At the same time, it has been found that object agreement is not observed when both other cues are present simultaneously in a language. At the moment, my sample of languages does not allow me to test the role of object agreement statistically. I leave that to future research.

Semantic Tightness

As a proxy for semantic tightness, I computed Mutual Information between Role and individual lexemes. For this purpose, I extracted frequencies of common nouns as Subject and Object from the corpora. Examples are displayed in **Table 5**. Usually, human nouns tend to be biased toward the role of Subject (e.g., *hunter*), while inanimate nouns more frequently occur in the object role (e.g., *t-shirt* and *street*). The stronger these biases, the higher the MI score and therefore the tighter the semantic fit. The MI scores are shown in **Figure 2**.

TABLE 5 | A fragment of the Lexeme – Role matrix for English.

Lexeme (lemma)	Transitive subject	Object
hunter	40	22
street	34	466
t-shirt	3	118

The tightest languages are Hindi, Korean, Russian, Hungarian, and Japanese. This supports previous accounts (see section "Semantic and Pragmatic Properties of the Arguments"). Among the loosest languages are English and Indonesian, which are also well known as semantically loose. It is surprising that Turkish is the loosest language in the sample, although if we also take into account more grammatical roles (such as intransitive subjects and obliques), it becomes relatively tight (Levshina, 2020b).

An important issue in language comparison is what to count as a word (Haspelmath, 2011). For example, in English, the phrase *art history* consists of two words, but its German equivalent

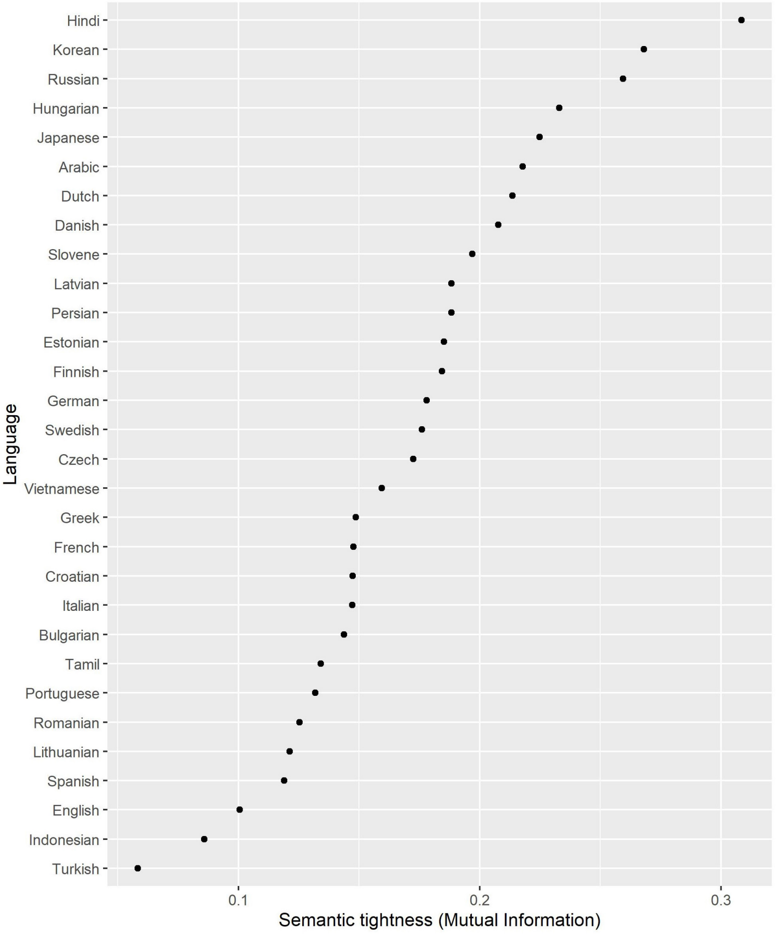


FIGURE 2 | Semantic tightness (Mutual Information between Role and Lexeme).

Kunstgeschichte is only one word. In order to counterbalance the influence of orthographic conventions, I also computed the scores treating multiword units like *art history* as one lexeme, based on the Universal Dependencies "compound", "fixed" and "flat". In the subsequent correlational and causal analyses, this variable, however, did not perform differently from the first one. This is why the analyses presented below are based only on lemmas of single orthographic words (but see Levshina, 2020b).

Rigid Word Order

The next type of information reflects if rigid word order can be a reliable cue of the syntactic roles. In order to compute it, I used anti-entropy, which is 1 minus Shannon entropy of the order of Subject and Object. The formula for computing entropy of orders SO and OS is as follows:

$$H = -\sum_{i=1}^{n} P\ (Order_i) * \ log\ P(Order_i)$$

where P (Order_i) stands for the probability of SO or OS. The probabilities were computed as simple proportions of each word order in the corpora. More on this approach can be found in Levshina (2019).

If either Subject is always before Object or the other way round, i.e., P (SO) = 1 and P (OS) = 0, or P (SO) = 0 and P (OS) = 1, the entropy value is minimal (H = 0) and therefore the rigidity score is maximal: 1 – H = 1 – 0 = 1. If both orders have equal probabilities, i.e., P (SO) = P (OS) = 0.5, then the entropy value is maximal (H = 1) and the rigidity score is minimal: 1 – 1 = 0. The rigidity scores are displayed in **Figure 3**.

The Baltic, Finno-Ugric and most Slavic languages, as expected, have the lowest rigidity scores, allowing for word order flexibility. In contrast, English, French, Indonesian have the most rigid order, followed by the Scandinavian and other Romance languages and Vietnamese. Interestingly, Korean and Japanese do not display much variability, although it is assumed that they have flexible order of Subject and Object.

Verb-Medial Order

The fourth and final variable considered in this study is "verb-medialness," which shows how frequently the head verb occurs between the subject and the object. The procedure was as follows. I computed the number of clauses in the corpora (only finite

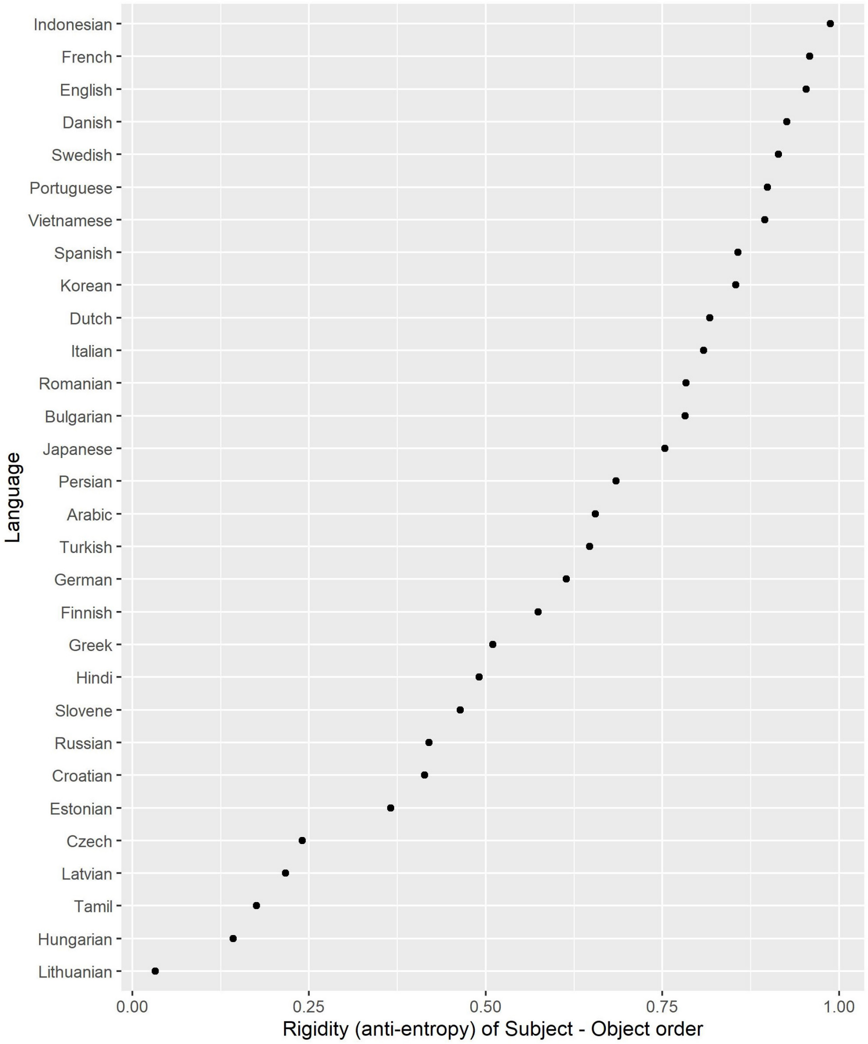


FIGURE 3 | Rigidity of Subject – Object order (1 – entropy).

main and subordinate clauses with a lexical verbal predicate were considered), which had overt Subject and Object, and a lexical head verb. Next, I computed the proportion of all clauses where the verb is between Subject and Object (in either order). The scores based on the UD corpora and the online news corpora are displayed in **Figure 4**. One can see a gap between the typical SOV languages (Japanese, Tamil, Korean, Hindi, and Turkish) with the lowest scores and all the rest. Indonesian, English and French are nearly always verb-medial.

A CORRELATIONAL ANALYSIS OF CROSS-LINGUISTIC DATA

The Problem of Dependent Observations

Computing correlations between the variables in this case study is not straightforward because the dataset contains dependent observations. Many languages come from the same family or even genus. In order to address this issue, I used a combination of sampling and permutation. I followed Dryer's (1992) approach relying on genera as the main taxonomic level. In 1,000 simulations, I sampled only one language from each genus and computed the Spearman's rank-based correlation coefficients for each sample. These coefficients were then averaged for each pair of variables. The Spearman method was used because some of the relationships displayed small non-linearity, but Pearson's product-moment coefficients, as well as Kendall's coefficients, reveal similar results.

In order to perform the null hypothesis significance testing, I computed and logged the test statistic for the original pairs of scores in every simulation. I also ran 1,000 permutations, in which the original scores of the second variable were randomly reshuffled. The permutation scores represented the distribution of the test statistic under the null hypothesis. Next, I counted the number of cases out of 1,000 permutations where the permuted scores were equal to or more extreme than the original test

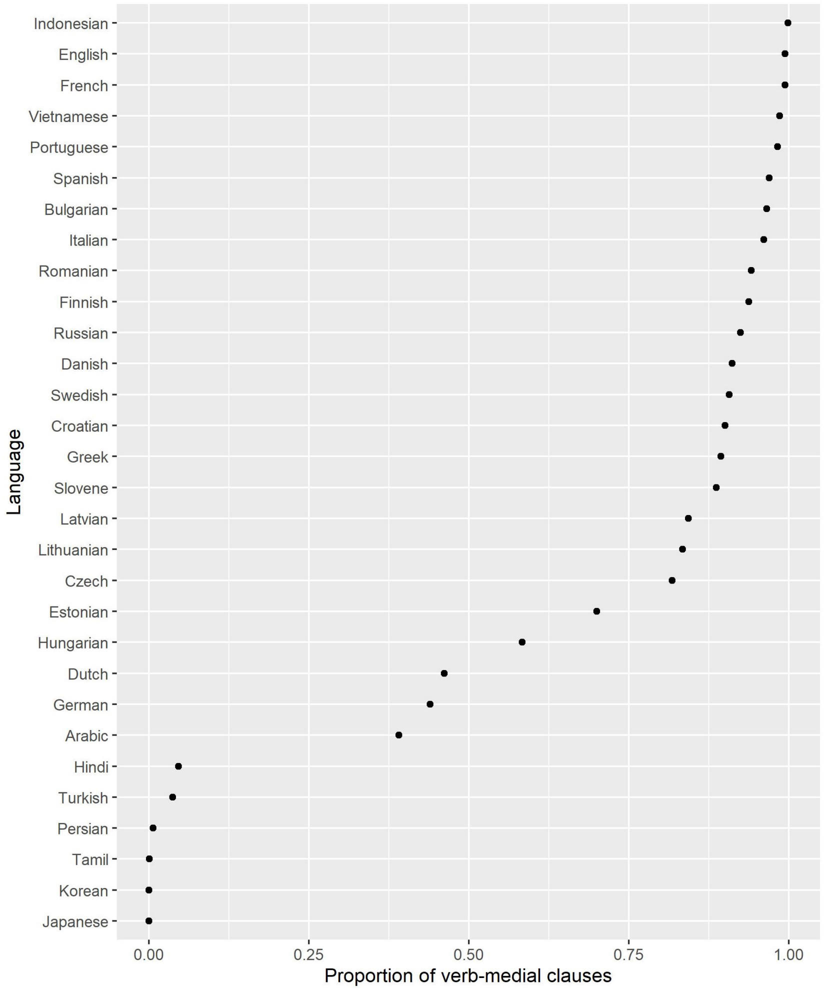


FIGURE 4 | Proportion of verb-medial clauses.

statistics based on the unpermuted data. These proportions served as *p*-values. The *p*-values were then averaged across the 1,000 samplings from the genera.

Results of Correlational Analyses

The Spearman correlation coefficients are displayed in **Figure 5**. The 95% confidence intervals around the average values can be found in **Appendix Table 1**. The simple (non-partial) pairwise correlations are represented by bold labels at the top of the squares. The strongest negative correlation is between case marking and rigid order of Subject and Object. The correlation is negative and significant (ρ = −0.67, *p* = 0.004). This means that distinctiveness of case marking increases with word order flexibility and decreases with word order rigidity. Next follows a positive correlation between case marking and tight semantics (ρ = 0.49, *p* = 0.043). From this we can conclude that semantically tight languages tend to have more informative case marking than semantically loose ones. The negative correlation between case marking and proportion of verbs located medially, between Subject and Object (ρ = −0.47, *p* = 0.042), means that languages without distinctive case marking tend to have SVO. There is also a negative correlation between semantic tightness and the proportion of verbs in the middle (ρ = −0.44, *p* = 0.047). This suggests that semantically loose languages are usually verb-medial, whereas semantically tight ones are usually verb-final (the only language in the sample with partly verb-initial order is Arabic). The remaining correlations are not significant.

If we compute partial correlations, which represent the relationships between variables X and Y taking into account all other variables, as in multiple regression, the direction of the significant correlations is mostly similar, as one can see from the coefficients represented by dark-gray labels in italics in **Figure 5**. The 95% confidence intervals around the average coefficients can be found in **Appendix Table 1**. The correlations between rigid order and case marking, and between tight semantics and case marking change very little, but the correlations between the proportions of verbs in the middle and the other variables become much weaker. In this case, only the correlation between rigid word order and

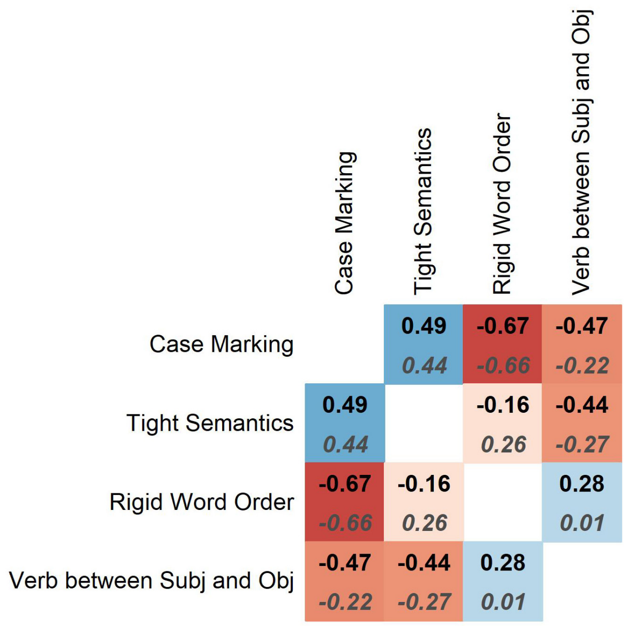


FIGURE 5 | Spearman's correlation coefficients between pairs of variables, averaged across 1,000 simulations. Top: simple pairwise coefficients. Bottom: partial coefficients.

case marking is statistically significant at the level of 0.05 ($p = 0.012$).

To summarize, we see that not all correlations are negative (though all significant partial correlations are): the correlation between semantic tightness and case marking is positive, for example. Also, not all variables are correlated (although this can be due to the relatively small sample size). It is also remarkable that case marking is the most strongly correlated with the other variables.

A CAUSAL ANALYSIS OF SUBJECT AND OBJECT CUES

Motivation for Causal Analysis

Hypotheses about causal mechanisms can be performed with the help of experiments, by manipulating the variables of interest while carefully controlling for possible confounding effects. If diachronic data are available, causal relationships can be discovered with the help of a Granger-causality analysis (Moscoso del Prado Martín, 2014). Here I will use statistical methods to identify causal relationships using the synchronic observational data. In this case, causal analysis is based on tests of conditional independence of one variable X from another variable Y, given another variable (Z) or variables. Independence between X and Y means, informally speaking, that we do not know more about the value of X if we know the value of Y, and the other way round. For example, if we know that it will rain today, this information will not help us to guess the exchange rate of euro to British pound sterling. Conditional independence means that we cannot say anything more about X if we know Y, given Z. For example, if we take children's heights and their vocabulary size, we are likely to find a positive correlation. But if we control for age, this correlation will disappear. In this case, there are several scenarios of causal relationships. For example, the relationship between X and Y can be a so-called fork X ← Z → Y, which means that Z is the common cause for both X and Y. This can be illustrated by the above-mentioned example with age as the common cause of height and vocabulary size. A linguistic example is lexical borrowing into different languages from English. If we take two unrelated languages, e.g., Japanese and Telugu, and compare their vocabularies, we will find that they overlap to some extent due to shared loanwords. But if we control for English loans, the languages will become independent (Dellert, 2019: 69). This is not the only possibility when X and Y are conditionally independent given Z. The relationships can also represent a causal chain, X → Z → Y or X ← Z ← Y, where all the influence from X to Y or from Y to X is mediated by Z. For example, there is a dependency between Modern English and Old English, but it is mediated by Middle English. More variables are needed in order to distinguish between forks and different kinds of chains.

Consider now the opposite scenario: X and Y are independent in the absence of Z, but become dependent if we control for Z. In this case, the variables are likely to form a so-called collider, or v-structure: X → Z ← Y. To give a very basic example, we can assume that the amount of talent (X) and amount of luck (Y) are independent. We can also assume that they both contribute to success (Z). If we control for success, talent and luck will become dependent. That is, if we know how successful one is, and the amount of talent, we can figure out the amount of luck. For instance, if someone has achieved a lot, but has no talent, people will say that he or she has been very lucky. And if someone is obviously talented, but remains an underdog, then bad luck is to blame.

There are many different algorithms for causal inference. Here I use the FCI (Fast Causal Inference) algorithm, which is preferred in the situations when we are not sure if the assumption of causal sufficiency is met. This means that we could miss some other variables that represent common causes for two or more variables in the data (Spirtes et al., 2000; Dellert, 2019: 80). In other words, FCI allows latent variables. In our case, potential latent variables can be sociolinguistic ones, such as intensity of language contact or population size (e.g., Trudgill, 2011; see also section "A Possible Diachronic Scenario"). The relevance of different sociolinguistic variables for grammar, however, is not fully understood yet (Sinnemäki and Di Garbo, 2018).

FCI also allows unmeasured selection variables, which determine whether or not a measured unit (here: a language) is included in the data sample. They represent selection bias. In our case, this can be the fact that all languages in the sample are written languages with a large number of speakers. Also, these languages are spoken in Eurasia only.

The result of a FCI algorithm is a Partial Ancestral Graph (PAG), where causal relationships are represented as edges between nodes (here: linguistic cues). Different types of edges are possible. When a relationship is directional, it is represented as an arrow: X → Y. If variables X and Y have a common latent cause, the edge will be bidirectional: X ↔ Y. Undirected edges (X – Y) suggest the presence of selection variables. In addition, there can be edges X ∘→ Y, X ∘– Y and X ∘–∘Y, where the circle represents uncertainty: it stands for either an arrowhead, or a tail.

The FCI algorithm runs as follows. The first step is to identify the undirected complete graph, or the skeleton. The algorithm used here is stable in the sense that the result does not depend on the order of variables in the data, cf. Colombo and Maathuis (2014). All edges of this skeleton are of the form X ∘–∘Y. This means that they are undetermined, or not oriented. Next, v-structures are identified using conditional independence tests, and superfluous edges are removed if a conditional independence is found. Finally, the v-structures are oriented again, and all possible undetermined edge marks ∘ are eliminated using the orientation rules in Zhang (2008). See more details in Dellert, (2019: 80–85).

The causal analysis was performed with the help of the FCI algorithm implemented in the *pcalg* package in R (Kalish et al., 2012; R Core Team, 2020). The rank-transformed variables were used instead of the original ones, to ensure the compatibility with the correlational analyses.

Due to the presence of dependent observations the causal analysis was repeated 1,000 times on subsets of the data, where one language was picked randomly from every genus. In each iteration, the algorithm returned an asymmetric adjacency matrix with information about the edges from X to Y and from Y to X represented by number codes. The presence of every edge was tested with the significance level of 0.05. Every matrix was logged, and the different types of edges were counted and analyzed, as will be shown below.

A Causal Network

The causal graph based on the FCI algorithm is displayed in **Figure 6**. The thickness of the edges corresponds to their frequency in 1,000 simulations, during which languages were randomly sampled from the genera. All links that have passed the significance test in at least one simulation are displayed in the causal network. One can see that some links are missing, which means the corresponding nodes are conditionally independent in all iterations at α = 0.05. In every simulation, the FCI algorithm computes maximal *p*-values for all conditional independence tests performed on every edge. If it is less than 0.05, the nodes are treated as conditionally dependent, and there exists a connection between them. The average *p*-values and their minimum and maximum values in the 1,000 simulations are displayed in **Table 6**.

There are four edges which pass the conditional independence test at least once. The links are between case marking and word order, between case marking and verb-medialness, between case marking and semantic tightness, and between verb-medialness and semantic tightness. The causal network also represents two types of links which emerged during the simulation. Most links are so-called unoriented edges of the type X ∘–∘Y, which means that no direction could be identified. Each end of such an edge could be an arrowhead or a tail. This can happen due to lack of v-structures, or colliders, in the sample. The most frequent link of this type is between rigid order and case marking. It occurred in 650 out of 1,000 iterations. Next comes the link between verb-medialness and semantic tightness with 59 occurrences. Finally, the link between tight semantics and case marking was observed only six times.

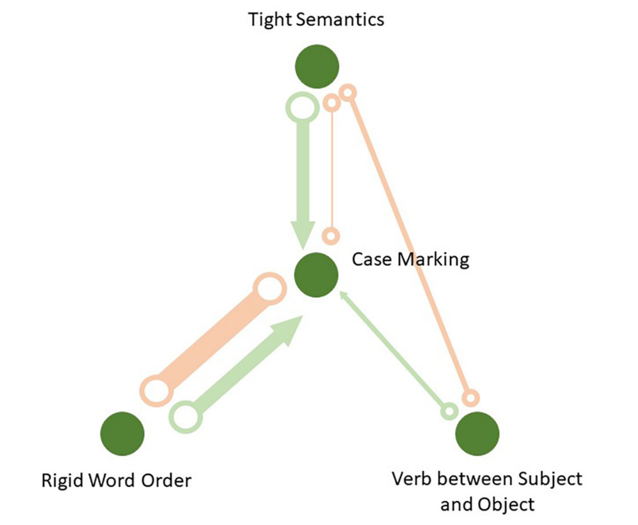


FIGURE 6 | Causal network based on the FCI algorithm: Thickness of the edges reflects their frequencies in 1,000 samples.

In addition, there were several partially directional edges of the type X ∘→Y. This means that there is no certainty whether the relationship is X →Y or it is bidirectional, X ↔Y. Recall that bidirectional edges suggest the presence of a common latent cause. Importantly, all of these edges have their arrowheads pointed to case marking. This means that case marking is more likely to be influenced by the other variables than the other way round. The most frequent edge of this type is the one from rigid word order to case marking with 344 occurrences in 1,000 simulations. It is followed by the edge from tight semantics to case marking with 314 occurrences, and finally by the link from verb-medialness to case marking, which occurred 30 times only. The edge between verb-medialness and semantic tightness does not have any partially directed links.

These results contain a lot of uncertainty. More data are apparently needed. Still, we can draw some conclusions. First of all, case marking is in the center of the graph. Second, we see that all partially directed edges lead to case marking, and none from case marking to the other cues. This suggests that formal marking is probably the most sensitive to other parameters' influence.

Also, the total number of edges of any type between tight word order and case marking was 994 out of possible 1,000. It was present in almost all iterations. This means that the causal link between word order and case marking has by far the strongest support. However, we also see that there are some chances of a causal relationship from word order to case marking, and no partially or fully directed edges in the opposite direction.

TABLE 6 | Mean *p*-values of the edges in FCI.

	Case marking	Tight semantics	Rigid order	Verb middle
Case marking		0.099 (0.002, 0.392)	0.011 (0.001, 0.068)	0.122 (0.027, 0.346)
Tight semantics	0.099 (0.002, 0.392)		0.564 (0.109, 1)	0.128 (0.021, 0.895)
Rigid order	0.011 (0.001, 0.068)	0.564 (0.109, 1)		0.319 (0.058, 0.750)
Verb middle	0.122 (0.027, 0.346)	0.128 (0.021, 0.895)	0.319 (0.058, 0.750)	

In parentheses: minimum and maximum values.

The evidence for the link between tight semantics and case is weaker. The total number of edges between tight semantics and case marking was 320. Nearly all of them are partially directed. Therefore, a unidirectional effect of case marking on tight semantics is less likely than the reverse effect. There were 59 non-directed edges between tight semantics and verb-medialness. The total number of edges from verb-medialness to case marking was only 30, the smallest value. All these links were partially directed.

A Possible Diachronic Scenario

How can we interpret these correlations and causal links? A tentative historical scenario could be as follows. Under normal circumstances, languages tend to accumulate complexity (Dahl, 2004), which explains why languages are vastly redundant (Hengeveld and Leufkens, 2018). Tight semantics and rich case morphology can be among those complexities. Mature and complex languages can also have complex contextual rules for choosing SO or OS for managing information flow, which makes the unconditional entropy of Subject and Object order high. All these complexities are not a problem for child L1 learners and are transmitted faithfully from one generation to another. Also, these languages can retain verb-final order, which was arguably the order in the ancestral language (Gell-Mann and Ruhlen, 2011).

Now imagine that due to increasing language contact the number of adult learners of this language increases. What would the consequences be like? We can expect the following changes.

First, evidence from artificial language learning experiments suggests that adults are better at learning word sequences that are produced by rules, while children are better at memorizing sequences without any underlying rules (Nowak and Baggio, 2017). Although there is evidence that adults tend to probability-match free variation in an input language under certain conditions more than children do (Hudson Kam and Newport, 2005)[6], experiments with artificial languages show that input languages exhibiting free variation become increasingly regular, revealing a strong bias toward regularity in adult learners during language diffusion (Smith and Wonnacott, 2010). Moreover, it is important to emphasize that variation in word order is hardly ever free. On the contrary, it is constrained by individual constructions and stylistic and information-management considerations. It is then possible that a rigid order of Subject and Object, which represents a simple generalization, is easier for adult L2 learning than a so-called flexible order with many local rules[7]. Adults will learn patterns that can be captured by a few simple rules. As for L1 speakers in language contact situations, there is evidence that they prefer more rigid word order if they are immersed in another language. For example, Namboodiripad (2017) and Namboodiripad et al. (2019) show that increased language contact with English leads to a greater preference for canonical order (SOV) in Korean and Malayalam speakers. So we can expect the order to become more rigid in a language contact situation.

Second, the associations between roles and lexemes or semantic classes can become looser due to the cognitive limitations of adult learners. Acquisition of the role – semantics associations, and which constructions to use if some combinations are not allowed (e.g., passives), is difficult. Also, growth and increasing diversity of a language community can cause greater variability in the role – referent mappings[8]. Since L2 learners can subconsciously transfer their mother tongue features to the target language (Siegel, 2008: Ch. 5), this can increase the pool of variants in the expression of grammatical roles, which makes the associations between the roles and semantics looser.

Third, the verb can shift to the middle position due to increased noise in L2 communication. Following the hypothesis in Gibson et al. (2013), the verb-medial order is more robust for information transmission in a noisy channel. One can consider L2 communication noisier than L1 communication. In fact, if we look at high-contact pidgins and creoles represented in the Atlas of Pidgin and Creole Language Structures, we will find that 71 of 76 languages (93%) have SVO, with 63 languages (83%) relying on this word order as the exclusive or dominant pattern (Huber and the APiCS Consortium, 2013). According to Bentz and Christiansen (2010), the increase of L2 learners of (Vulgar) Latin as *lingua franca* of the expanding Roman Empire provided an important pressure toward the Romance SVO without case marking and the reduction of word order flexibility. It is also possible that the high proportion of L2 speakers is responsible for the predominant SVO in the three most widespread languages: Chinese, English, and Spanish. Bentz and Christiansen explain this development by production pressures. In particular, they

[6] This effect is restricted by different factors. In particular, it is observed when the free variation is between only two alternatives, and when adults reproduce already familiar input. When producing new utterances, adults fall back on their bias toward regularization (Wonnacott and Newport, 2005).

[7] Note that languages with so-called free word order tend to have strong preferences with regard to the pragmatic role of the elements. For example, some polysynthetic languages tend to put newsworthy constituents first. This could also be an easy rule to acquire. This pattern is characteristic of languages with a full set of substantive bound pronouns referring to all core arguments attached to the verb in a rigid order, so the full noun phrases act like appositives to the pronominal affixes (Mithun, 1987). The question is then what serves as the main cues for Subject and Object – bound morphemes or nouns.

[8] I thank Laura Becker (p.c.) for this idea.

claim that it is easier to assign the case to the object when the verb comes first.

We also see a weak causal link between the position of the verb and semantic tightness. According to Hawkins (1986), semantic tightness helps to avoid reanalysis in verb-final clauses and thus to avoid extra effort (see also Levshina, 2020b). This can be seen as a manifestation of efficiency. The model does not show which of the variables influences which one. It may be that tight semantics allows for verb-finalness, or verb-finalness leads to semantic tightness. More research is needed to understand this relationship.

Finally, case morphology represents another source of complexity, which L2 learners can be tempted to get rid of. In the causal network, we saw that there are some chances that the directional relationships are in fact bidirectional, which is usually due to latent common causes. It seems that the presence of L2 learners and similar sociolinguistic variables can be such a common cause.

In addition, the changes toward more rigid word order, semantic looseness and verb-medial order create favorable conditions for the language to lose case marking. Semantic looseness leads to more abstract semantics of the case forms, which do not contribute much beyond the syntactic relationships. Since the forms do not express much beyond what is already conveyed by word order, it would be rational and efficient to save articulatory and processing effort by not using case marking. The role of production effort in loss of case marking has been demonstrated in Fedzechkina and Jaeger's (2020) experiment involving adult learners of an artificial language, so it is a valid factor. That said, it is important to emphasize that the loss of marking as a way of saving effort can happen only after appropriate conditions have been created.

CONCLUSION

This case study investigated the relationships between different cues that help the addressee to assign the grammatical roles of Subject and Object in a transitive clause. The cues included case marking, tight association between lexemes and roles (semantic tightness), rigid order of Subject and Object, and the position of the verb between Subject and Object. The measures that reflect the prominence of these cues were obtained from corpora in thirty languages.

The results of the correlation analyses demonstrated that some cues were negatively correlated, and some were not. By far the strongest correlation is the inverse correlation between case marking and rigid order of Subject and Object. This correlation has been discussed in numerous previous accounts (e.g., Sapir, 1921; Sinnemäki, 2014b; Fedzechkina et al., 2016; Fedzechkina and Jaeger, 2020). Importantly, the correlation between word order rigidity and case marking distinctiveness is not influenced by the presence or absence of the other variables. Therefore, the relationship between word order and case marking is robust, which means that the previous studies that focused only on this pair of cues are valid.

The other correlations are also in accordance with the previous studies. Semantic tightness and case marking display a strong positive correlation: the more information is provided by the lexemes (semantics), the more distinctive are the case forms in a language. This supports Hawkins (1986) ideas about tight-fit and loose-fit languages, where semantic tightness is associated with case marking. The analysis also revealed an expected negative correlation between verb-medialness and semantic tightness (Hawkins, 1986; Levshina, 2020b). Moreover, languages with the verb between Subject and Object usually have no case marking (cf. Sinnemäki, 2010), and tend to have rigid word order. Verb-final languages can have flexible word order and usually have case marking. This ties in well with the results of the gesture experiment in Gibson et al. (2013), who found a correlation between verb-finalness and the use of spatial marking of the core arguments.

The results of the correlational analysis are in accordance with previous grammar-based and experimental studies, which means that corpus-based variables can be used successfully to represent the linguistic cues. At the same time, only rigid word order and case marking have a significant partial correlation when the other variables are taken into account. This finding requires further research on a larger sample of languages. Also, the results indicate that case marking is more strongly correlated with the other cues than any other variable – a fact that has not been previously reported.

The causal analysis based on the Fast Causal Inference algorithm showed that case marking is the variable that is the most likely to be affected by the other variables. The most probable causal link is found between rigid word order and case marking, with greater probability of the directional relationship from word order to case marking than the other way round. This supports the previous observations based on the history of English and the Romance languages (see section "Correlations and Causal Links From Previous Studies"), saying that fixation of word order and transition toward SVO triggered the loss of case marking. It also provides empirical evidence for the reasoning in Koplenig et al. (2017) about the directionality of this relationship. Importantly, it converges with the experimental results in Fedzechkina et al. (2016) and Fedzechkina and Jaeger (2020), which point in a similar direction. Also, cross-linguistic evidence (Sommer and Levshina, 2021) demonstrates that word order plays an important role in differential case marking of core arguments. The use of a case marker is more likely when the word order in a clause is different from the dominant one, supporting the experimental results in Fedzechkina and Jaeger (2020) and Tal et al. (2020). This effect is found in quite a few languages from all over the world, including Dazaga (Saharan), Gurindji Kriol (mixed), Kakua (Cacua-Nukak), Sheko (Afro-Asiatic), and Udihe (Altaic). Case markers are often used on topicalized objects in left dislocation (Iemmolo, 2010), but also in other situations. The function of case marking is to override the addressee's expectations about the grammatical role of the argument and/or about the topic of the clause (cf. Diessel, 2019: Ch. 11).

At the same time, we do not find conclusive evidence that word order flexibility or rigidity is determined by the presence or absence of case. This goes against Sapir's hypothesis, who wrote about the historical change in English, "[a]s the inflected forms of English became scantier, as the syntactic relations were more and more inadequately expressed by the forms of the words themselves, position in the sentence gradually took over functions originally foreign to it" (Sapir, 1921: 166). Although some languages are known to use word order freezing (i.e., choosing the dominant word order) in ambiguous contexts, in particular, when the case forms are not informative enough (Jakobson, 1971), this effect is relatively weak in real language use (see Berdicevskis and Piperski, 2020 on Russian and German), so it is unlikely to have a major impact on language change.

Moreover, the causal analysis shows some probability that case marking can be affected by semantic tightness. We also find some weak evidence that the position of the verb can affect case marking, as well. In addition, there is a possibility of an undirected causal link between the degree of semantic tightness and the position of the verb in a sentence.

To summarize, the study shows that not all grammatical cues to subject and object are negatively correlated, as one would expect if one assumed that efficiency is directly reflected in relationships between aggregate typological variables. Still, there is a possibility that the trade-off between rigid word order and case marking is a manifestation of efficient behavior, and so is the weak correlation between tight semantics and the (final) position of the verb, where tight semantics helps to avoid costly reanalysis. The first claim is in fact supported by convergent evidence from artificial language learning experiments (Fedzechkina et al., 2016; Fedzechkina and Jaeger, 2020). Indeed, adult L2 learners avoid case marking in the presence of fixed word order. However, as was argued above, this manifestation of efficiency is only possible under certain conditions, which depend on the growing proportion of L2 users and possibly population size. Since the Subject and Object cues seem to be mostly influenced by the sociolinguistic factors, this leaves little space for potential manifestations of communicative efficiency.

A proper test of efficient behavior would require context-sensitive information about the joint distribution of linguistic cues, which also takes into account their diverse functions in discourse. This is difficult to do at the moment due to technical reasons, such as data sparseness and lack of reliable morphological annotation. Still, this article shows that a causal analysis of aggregate linguistic variables can be used to circumscribe the potential effects of communicative efficiency in language evolution. These results need further support from typological and experimental data, as well as from corpora representing other languages and registers.

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://github.com/levshina/SubjectObjectCues.

AUTHOR CONTRIBUTIONS

NL was responsible for the design of the study, data collection, statistical analysis, and interpretation.

FUNDING

The research in this manuscript was funded by the Netherlands Organization for Scientific Research (NWO) under Gravitation grant Language in Interaction, grant number 024.001.006.

ACKNOWLEDGMENTS

The author is extremely grateful to Johannes Dellert for sharing his vast expertise in theory and methods of causal analysis. This manuscript has greatly benefited from constructive, very detailed and insightful suggestions and comments from Florian Jaeger, Agnieszka Konopka and the editor Gerhard Jäger. The author thanks the audience at the Neurobiology of Language department of the MPI Nijmegen, and colleagues from the University of Tübingen for useful and friendly feedback on this project. All remaining errors are the author's own responsibility.

REFERENCES

Aikhenvald, A. Y. (2010). *Imperatives and Commands*. Oxford: Oxford University Press.

Aissen, J. (2003). Differential object marking: iconicity vs. economy. *Nat. Lang. Linguistic Theory* 21, 435–483. doi: 10.1023/A:1024109008573

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika* 2, 1–10.

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. London: Routledge.

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

Baayen, R. H., Milin, P., and Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology* 30, 1174–1220. doi: 10.1080/02687038.2016.1147767

Bauer, B. M. (2009). "Word order," in *New Perspectives on Historical Latin Syntax: Vol 1: Syntax of the Sentence*, eds P. Baldi and P. Cuzzolin (Berlin: Mouton de Gruyter), 241–316.

Bentz, Ch, and Christiansen, M. H. (2010). "Linguistic adaptation at work? The change of word order and case system from Latin to the Romance languages," in *Proceedings of the 8th International Conference on the Evolution of Language*, eds A. D. M. Smith, M. Schouwstra, B. de Boer, and K. Smith (Singapore: World Scientific), 26–33. doi: 10.1142/9789814295222_0004

Bentz, C., and Winter, B. (2013). Languages with more second language learners tend to lose nominal case. *Lang. Dynam. Change* 3, 1–27. doi: 10.1163/22105832-13030105

Berdicevskis, A., and Piperski, A. (2020). "Corpus evidence for word order freezing in Russian and German," in *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, (Spain), 26–33.

Berdicevskis, A., Schmidtke-Bode, K., and Seržant, I. (2020). "Subjects tend to be coded only once: corpus-based and grammar-based evidence for an efficiency-driven trade-off," in *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories (Düsseldoft: ACL)*, 79–92. doi: 10.18653/v1/2020.tlt-1.8

Blasi, D. E. (2018). *Linguistic Diversity Through Data. Ph, D. Thesis*. Leipzig: University of Leipzig.

Blasi, D. E., and Roberts, S. G. (2017). "Beyond binary dependencies in language structure," in *Dependencies in Language*, ed. N. J. Enfield (Berlin: Language Science Press), 117–128.

Bock, J. K., and Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition* 21, 47–67. doi: 10.1016/0010-0277(85)90023-X

Bresnan, J., Dingare, Sh, and Manning, Ch.D (2001). "Soft constraints mirror hard constraints: voice and person in English and Lummi," in *Proceedings of the LFG01 Conference*, eds M. Butt and T. Holloway King (Stanford: CSLI publications), 13–32.

Bybee, J. L. (2010). *Language, Usage, and Cognition*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511750526

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

Cohen Priva, U. (2008). "Using information content to predict phone deletion," in *Proceedings of the 27th West Coast Conference on Formal Linguistics*, eds N. Abner, J. Bishop, and M. A. Somerville (Cascadilla Proceedings Project), 90–98.

Colombo, D., and Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Machine Learn. Res.* 15, 3741–3782.

Coupé, Ch, Oh, Y. M., Dediu, D., and Pellegrino, F. (2019). Different languages, similar encoding efficiency: comparable information rates across the human communication niche. *Sci. Adv,* 5:eeaw2594. doi: 10.1126/sciadv.aaw2594

Culbertson, J., Smolensky, P., and Legendre, G. (2012). Learning biases predict a word order universal. *Cognition* 122, 306–329. doi: 10.1016/j.cognition.2011.10.017

Dahl, Ö (2004). *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: John Benjamins. doi: 10.1075/slcs.71

Dellert, J. (2019). *Information-Theoretic Causal Inference of Lexical Flow*. Berlin: Language Science Press.

Diessel, H. (2019). *The Grammar Network: How Linguistic Structure is Shaped by Language Use*. Cambridge: Cambridge University Press. doi: 10.1017/9781108671040

Dryer, M. (1992). The Greenbergian word order correlations. *Language* 68, 81–138. doi: 10.1353/lan.1992.0028

Fedzechkina, M., and Jaeger, T. F. (2020). Production efficiency can cause grammatical change: learners deviate from the input to better balance efficiency against robust message transmission. *Cognition* 196:104115. doi: 10.1016/j.cognition.2019.104115

Fedzechkina, M., Newport, E. L., and Jaeger, T. F. (2016). Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. *Cognit. Sci.* 41, 416–446. doi: 10.1111/cogs.12346

Fenk-Oczlon, G., and Fenk, A. (2008). "Complexity trade-offs between the subsystems of language," in *Language Complexity: Typology, Contact, Change*, eds M. Miestamo, K. Sinnemäki, and F. Karlsson (Amsterdam: John Benjamins), 43–65. doi: 10.1075/slcs.94.05fen

Fenk-Oczlon, G., and Pilz, J. (2021). Linguistic complexity: relationships between phoneme inventory size, syllable complexity, word and clause length, and population size. *Front. Commun.* 6:626032. doi: 10.3389/fcomm.2021.626032

Ferrer-i-Cancho, R. (2006). Why do syntactic links not cross? *Europhys. Lett.* 76, 1228–1234. doi: 10.1209/epl/i2006-10406-0

Ferreira, V. S., and Firato, C. E. (2002). Proactive interference effects on sentence production. *Psychon. Bull. Rev.* 9, 795–800. doi: 10.3758/BF03196337

Ferreira, V. S., and Yoshita, H. (2003). Given-new ordering effects on the production of scrambled sentences in Japanese. *J. Psycholing. Res.* 32, 669–692. doi: 10.1023/A:1026146332132

Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *PNAS* 112, 10336–10341. doi: 10.1073/pnas.1502134112

Gell-Mann, M., and Ruhlen, M. (2011). The origin and evolution of word order. *PNAS* 108, 17290–17295. doi: 10.1073/pnas.1113716108

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., and Saxe, R. (2013). A noisy-channel account of crosslinguistic word order variation. *Psychol. Sci.* 24, 1079–1088. doi: 10.1177/0956797612463705

Gibson, E., Futrell, R., Piantadosi, S., Dautriche, I., Mahowald, K., Bergen, L., et al. (2019). How efficiency shapes human language. *Trends Cognit. Sci.* 23, 389–407. doi: 10.1016/j.tics.2019.02.003

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

Goldin-Meadow, S., So, W. C., Özyürek, A., and Mylander, C. (2008). The natural order of events: how speakers of different languages represent events nonverbally. *Proc. Natl. Acad. Sci. U.S.A.* 105, 9163–9168. doi: 10.1073/pnas.0710060105

Greenberg, J. H. (1966a). *Language Universals, With Special Reference to Feature Hierarchies*. The Hague: Mouton.

Greenberg, J. H. ed (1966b). "Some universals of grammar with particular reference to the order of meaningful elements," in *Universals of Grammar*, (Cambridge, MA: MIT Press), 73–113.

Haig, J. (2018). The grammaticalization of object pronouns: why differential object indexing is an attractor state. *Linguistics* 56, 781–818. doi: 10.1515/ling-2018-0011

Hall, M. L., Mayberry, R. I., and Ferreira, V. S. (2013). Cognitive constraints on constituent order: evidence from elicited pantomime. *Cognition* 129, 1–17. doi: 10.1016/j.cognition.2013.05.004

Hall, M. L., Ahn, D. Y., Mayberry, R. I., and Ferreira, V. S. (2015). Production and comprehension show divergent constituent order preferences: evidence from elicited pantomime. *J. Memory Lang.* 81, 16–33. doi: 10.1016/j.jml.2014.12.003

Hall, K. C., Hume, E., Jaeger, T. F., and Wedel, A. (2018). The role of predictability in shaping phonological patterns. *Linguistics Vanguard* 4:20170027. doi: 10.1515/lingvan-2017-0027

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45, 31–80. doi: 10.1515/flin.2011.002

Haspelmath, M. (2014). "On system pressure competing with economic motivation," in *Competing Motivations in Grammar and Usage*, eds B. MacWhinney, A. Malchukov, and E. Moravcsik (Oxford: Oxford University Press), 197–208. doi: 10.1093/acprof:oso/9780198709848.003.0012

Haspelmath, M. (2021). Explaining grammatical coding asymmetries: Form-frequency correspondences and predictability. *J. Linguistics*. 1–29. doi: 10.1017/S0022226720000535

Hawkins, J. A. (1986). *A Comparative Typology of English and German*. London: Unifying the contrasts.

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199252695.001.0001

Hawkins, J. A. (2019). Word-external properties in a typology of modern English: a comparison with German. *Eng. Lang. Linguist.* 23, 701–727. doi: 10.1017/S1360674318000060

Hengeveld, K., and Leufkens, S. (2018). Transparent and non-transparent languages. *Folia Linguistica* 52, 139–175. doi: 10.1515/flin-2018-0003

Holler, J., Kendrick, K. H., and Levinson, S. C. (2018). Processing language in face-to-face conversation: questions with gestures get faster responses. *Psychonom. Bull. Rev.* 25, 1900–1908. doi: 10.3758/s13423-017-1363-z

Holler, J., and Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends Cognit. Sci.* 23, 639–652. doi: 10.1016/j.tics.2019.05.006

Huber, M., and the APiCS Consortium. (2013). "Order of subject, object, and verb," in *The Atlas of Pidgin and Creole Language Structures*, eds S. M. Michaelis, P. H. Maurer, M. Haspelmath, and M. Huber (Oxford: Oxford University Press), 1–5.

Hudson Kam, C., and Newport, E. L. (2005). Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. *Lang. Learn. Dev.* 1, 151–195. doi: 10.1207/s15473341lld0102_3

Iemmolo, G. (2010). Topicality and differential object marking. evidence from romance and beyond. *Stud. Lang.* 34, 239–272. doi: 10.1075/sl.34.2.01iem

Jäger, G. (2007). Evolutionary game theory and typology. a case study. *Language* 83, 74–109. doi: 10.1353/lan.2007.0020

Jaeger, T. F. (2006). *Redundancy and Syntactic Reduction in Spontaneous Speech. Ph, D. Thesis*. Stanford, CA: Stanford University.

Jaeger, T. F., and Buz, E. (2017). "Signal reduction and linguistic encoding," in *The Handbook of Psycholinguistics*, eds E. M. Fernández and H. S. M. I. T. H. CAIRNS (Hoboken, NJ: John Wiley & Sons), 38–81. doi: 10.1002/9781118829516.ch3

Jaeger, T. F., and Tily, H. (2011). On language 'utility': processing complexity and communicative efficiency. *Wiley Int. Rev. Cognit. Sci.* 2, 323–335. doi: 10.1002/wcs.126

Jakobson, R. (1971). *Selected Writings. Vol. II. Word and Language*. Berlin: De Gruyter Mouton. doi: 10.1515/9783110873269

Jurafsky, D., Bell, A., Gregory, M. L., and Raymond, W. D. (2001). "Probabilistic relations between words: evidence from reduction in lexical production," in *Frequency and the Emergence of Linguistic Structure*, eds J. L. Bybee and P. Hopper (Amsterdam: John Benjamins), 229–254. doi: 10.1075/tsl.45.13jur

Just, E., and Čéplö, S. (in press). "Differential object indexing in maltese: a corpus based pilot study," in *Proceedings of the 7th International Conference on Maltese Linguistics*, (Poland).

Kalish, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *J. Statist. Softw.* 47, 1–26. doi: 10.18637/jss.v047.i11

Kanwal, J., Smith, K., Culbertson, J., and Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: language users optimise a miniature lexicon for efficient communication. *Cognition* 165, 45–52. doi: 10.1016/j.cognition.2017.05.001

Kemp, C., Xu, Y., and Regier, T. (2018). Semantic typology and efficient communication. *Ann. Rev. Linguistics* 4, 109–128. doi: 10.1146/annurev-linguistics-011817-045406

Kiparsky, P. (1996). "The shift to head-initial VP in Germanic," in *Studies in Comparative Germanic Syntax II*, eds H. Thráinsson, S. D. Epstein, and S. Peter (Dordrecht: Kluwer), 140–179. doi: 10.1007/978-94-010-9806-9_6

Koplenig, A., Meyer, P., Wolfer, S., and Müller-Spitzer, C. (2017). The statistical trade-off between word order and word structurelarge-scale evidence for the principle of least effort. *PLoS ONE* 12:e0173614. doi: 10.1371/journal.pone.0173614

Kurumada, Ch, and Grimm, S. (2019). Predictability of meaning in grammatical encoding: optional plural marking. *Cognition* 191:103953. doi: 10.1016/j.cognition.2019.04.022

Kurumada, Ch, and Jaeger, T. F. (2015). Communicative efficiency in language production: optional case-marking in Japanese. *J. Memory Lang.* 83, 152–178. doi: 10.1016/j.jml.2015.03.003

Leben, W. (1973). *Suprasegmental Phonology. PhD Dissertation*. Cambridge, MA: MIT.

Lee, H. (2009). "Quantitative variation in korean case ellipsis: implications for case theory," in *Differential Subject Marking*, eds H. de Hoop and P. de Swart (Dordrecht: Springer), 41–61. doi: 10.1007/978-1-4020-6497-5_3

Lemke, R., Schäfer, L., and Reich, I. (2021). Modeling the predictive potential of extralinguistic context with script knowledge: the case of fragments. *PLoS One* 16:e0246255. doi: 10.1371/journal.pone.0246255

Levshina, N. (2019). Token-based typology and word order entropy. *Linguistic Typol.* 23, 533–572. doi: 10.1515/lingty-2019-0025

Levshina, N. (2020a). Efficient trade-offs as explanations in functional linguistics: some problems and an alternative proposal. *Revista da Abralin* 19, 50–78. doi: 10.25189/rabralin.v19i3.1728

Levshina, N. (2020b). "How tight is your language? A semantic typology based on mutual information," in *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories (Düsseldorf: ACL)*, 70–78. doi: 10.18653/v1/2020.tlt-1.7

Levshina, N. (2021). Communicative efficiency and differential case marking: a reverse-engineering approach. *Linguistics Vanguard* 7:20190087. doi: 10.1515/lingvan-2019-0087

Levy, R., and Jaeger, T. F. (2007). "Speakers optimize information density through syntactic reduction," in *Advances in Neural Information Processing Systems (NIPS)*, Vol. 19, eds B. Schlökopf, J. Platt, and T. H. Hoffman (Cambridge, MA: MIT Press), 849–856.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *J. Cognit. Sci.* 9, 159–191. doi: 10.17791/jcs.2008.9.2.159

Lupyan, G., and Dale, R. (2010). Language structure is partly determined by social structure. *PLoS One* 5:e8559. doi: 10.1371/journal.pone.0008559

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

Manin, D. Y. (2006). Experiments on predictability of word in context and information rate in natural language. *J. Inform. Proc.* 6, 229–236.

Maurits, L. (2011). *Representation, Information Theory and Basic Word Order. Ph, D. Thesis*. Adelaide: University of Adelaide.

McWhorter, J. (2011). *Linguistic simplicity and complexity: Why do languages undress?* Berlin: de Gruyter Mouton. doi: 10.1515/9781934078402

Mithun, M. (1987). "Is basic word order universal?," in *Coherence and Grounding in Discourse: Outcome of a Symposium, Eugene, Oregon, June 1984*, ed. R. S. Tomlin (Amsterdam: John Benjamins), 281–328. doi: 10.1075/tsl.11.14mit

Moscoso del Prado Martín, F. (2014). "Grammatical change begins within the word: causal modeling of the co-evolution of icelandic morphology and syntax," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2657–2662.

Müller-Gotama, F. (1994). *Grammatical Relations: A Cross-Linguistic Perspective on Their Syntax and Semantics*. Berlin: Mouton de Gruyter. doi: 10.1515/9783110887334

Namboodiripad, S. (2017). *An Experimental Approach to Variation and Variability in Constituent Order. Ph, D. Thesis*. UC San Diego.

Namboodiripad, S., Kim, D., and Kim, G. (2019). "English dominant Korean speakers show reduced flexibility in constituent order," in *Proceedings of the Fifty-third Annual Meeting of the Chicago Linguistic Society*, eds D. Edmiston, M. Ermolaeva, E. Hakgüder, et al. (Chicago: Chicago Linguistic Society), 247–260.

Nowak, I., and Baggio, G. (2017). Developmental constraints on learning artificial grammars with fixed, flexible and free word order. *Front. Psychol.* 8:1816. doi: 10.3389/fpsyg.2017.01816

Piantadosi, S. T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. *PNAS* 108, 3526–3529. doi: 10.1073/pnas.1012551108

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

Plank, F. (1984). Verbs and objects in semantic agreement: minor differences between english and German might that might suggest a major one. *J. Semant.* 3, 305–360. doi: 10.1093/jos/3.4.305

Pleh, Cs, and MacWhinney, B. (1997). Double agreement: role identification in hungarian. *Lang. Cognit. Proc.* 12, 67–102. doi: 10.1080/016909697386916

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rohdenburg, G. (2003). "Horror aequi and cognitive complexity as factors determining the use of interrogative clause linkers," in *Determinants of Grammatical Variation in English*, eds G. Rohdenburg and B. Mondorf (Berlin: Mouton de Gruyter), 205–250. doi: 10.1515/9783110900019.205

Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. New York, NY: Harcourt.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition* 133, 140–155. doi: 10.1016/j.cognition.2014.06.013

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Techn. J.* 27, 379–423 and 623–656. doi: 10.1002/j.1538-7305.1948.tb01338.x

Siegel, J. (2008). *The Emergence of Pidgin and Creole Languages*. Oxford: Oxford University Press.

Sinnemäki, K. (2008). "Complexity trade-offs in core argument marking," in *Language Complexity: Typology, Contact, Change*, eds M. Miestamo, K. Sinnemäki, and F. Karlsson (Amsterdam: John Benjamins), 67–88. doi: 10.1075/slcs.94.06sin

Sinnemäki, K. (2010). Word order in zero-marking languages. *Stud. Lang.* 34, 869–912. doi: 10.1075/sl.34.4.04sin

Sinnemäki, K. (2014a). A typological perspective on differential object marking. *Linguistics* 52, 218–313. doi: 10.1515/ling-2013-0063

Sinnemäki, K. (2014b). "Complexity trade-offs: a case study," in *Measuring Grammatical Complexity*, eds F. J. Newmeyer and L. B. Preston (Oxford: Oxford University Press), 179–201. doi: 10.1093/acprof:oso/9780199685301.003.0009

Sinnemäki, K., and Di Garbo, F. (2018). Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: a typological study of verbal and nominal complexity. *Front. Psychol.* 9:1141. doi: 10.3389/fpsyg.2018.01141

Smith, K., and Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition* 116, 444–449. doi: 10.1016/j.cognition.2010.06.004

Sommer, N., and Levshina, N. (2021). *Cross-Linguistic Differential and Optional Marking Database (Version v1.0.0)*.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*, 2nd Edn. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/1754.001.0001

Stave, M., Paschen, L., Pellegrino, F., and Seifart, F. (2021). Optimization of morpheme length: a cross-linguistic assessment of Zipf's and Menzerath's laws. *Linguistics Vanguard* 7:20190076. doi: 10.1515/lingvan-2019-0076

Straka, M., and Straková, J. (2017). "Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UD-pipe," in *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, (Canada). doi: 10.18653/v1/K17-3009

Tal, Sh, Smith, K., Culbertson, J., Grossman, E., and Arnon, I. (2020). The impact of information structure on the emergence of differential object marking: an experimental study. *PsyArXiv* [preprint] doi: 10.31234/osf.io/759gm

Trudgill, P. (2011). *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

Walter, M. A., and Jaeger, T. F. (2008). "Constraints on optional that: a strong word form OCP effect," in *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, eds R. L. Edwards, P. J. Midtlyng, C. L. Sprague, and K. G. Stensrud (Chicago, IL: CLS), 505–519.

Wasow, T. (2015). "Ambiguity avoidance is overrated," in *Ambiguity: Language and Communication*, ed. S. Winkler (Berlin: De Gruyter Mouton), 29–47.

Wiemer, B., and Hansen, B. (2012). "Assessing the range of contact-induced grammaticalization in Slavonic," in *Grammatical Replication and Borrowability in Language Contact*, eds B. Wiemer, B. Wälchli, and B. Hansen (Berlin: De Gruyter Mouton), 67–155. doi: 10.1515/9783110271973

Wijffels, J. (2020). *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the UDPipe NLP Toolkit. R package version 0.8.4-1. 2020*. Available online at: https://CRAN.R-project.org/package=udpipe (accessed June 24, 2021).

Wonnacott, E., and Newport, E. L. (2005). "Novelty and regularization: the effect of novel instances on rule formation. in BUCLD 29," in *Proceedings of the 29th Annual Boston University Conference on Language Development*, eds A. Brugos, M. R. Clark-Cotton, and S. Ha (Somerville, MA: Cascadilla Press).

Zeman, D., Nivre, J., Abrams, M., et al. (2020). *Universal Dependencies 2.6. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University*. http://hdl.handle.net/11234/1-3226.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* 172, 1873–1896. doi: 10.1016/j.artint.2008.08.001

Zipf, G. (1965[1935]). *The Psychobiology of Language: An Introduction to Dynamic Philology*. Cambridge, MA: M.I.T. Press.

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

TABLE A1 | 95% confidence intervals around Spearman's correlation coefficients based on 1,000 simulations.

	Case marking	Tight semantics	Rigid word order
Tight semantics	0.484, 0.499 *0.434, 0.448*		
Rigid word order	−0.670, −0.664 *−0.663,−0.657*	−0.162, −0.149 *0.257, 0.264*	
Verb between Subject and Object	−0.480, −0.469 *−0.229,−0.218*	−0.446, −0.440 *−0.278,−0.266*	0.269, 0.281 *0.011, 0.017*

Upper numbers, non-partial correlations. Lower numbers, italics: partial correlations.

frontiers in Psychology

ORIGINAL RESEARCH
published: 19 July 2021
doi: 10.3389/fpsyg.2021.682132

Phylogenetic Typology

Gerhard Jäger and Johannes Wahle*

Department of Linguistics, University of Tübingen, Tübingen, Germany

In this article we propose a novel method to estimate the frequency distribution of linguistic variables while controlling for statistical non-independence due to shared ancestry. Unlike previous approaches, our technique uses all available data, from language families large and small as well as from isolates, while controlling for different degrees of relatedness on a continuous scale estimated from the data. Our approach involves three steps: First, distributions of phylogenies are inferred from lexical data. Second, these phylogenies are used as part of a statistical model to estimate transition rates between parameter states. Finally, the long-term equilibrium of the resulting Markov process is computed. As a case study, we investigate a series of potential word-order correlations across the languages of the world.

Keywords: typology, phylogenetics, Bayesian inference, word-order, language universals

OPEN ACCESS

Edited by:
Xiaofei Lu,
The Pennsylvania State University (PSU), United States

Reviewed by:
Claudia Marzi,
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC), Italy
Anna Feldman,
Montclair State University, United States

****Correspondence:***
Gerhard Jäger
gerhard.jaeger@uni-tuebingen.de

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

Received: *17 March 2021*
Accepted: *18 June 2021*
Published: *19 July 2021*

Citation:
Jäger G and Wahle J (2021) Phylogenetic Typology. Front. Psychol. 12:682132. doi: 10.3389/fpsyg.2021.682132

1. INTRODUCTION

One of the central research topics of linguistic typology concerns the distribution of structural properties across the languages of the world. Typologists are concerned with describing these distributions, understanding their determinants and identifying possible distributional dependencies between different linguistic features. Greenbergian language universals (Greenberg, 1963) provide prototypical examples of typological generalizations. Absolute universals[1] describe the distribution of a single feature, while implicative universals[2] state a dependency between different features. In subsequent work (such as Dryer, 1992), the quest for implicative universals was generalized to the study of *correlations* between features.

Validating such kind of findings requires statistical techniques, and the quest for suitable methods has been a research topic for the last 30 years. A major obstacle is the fact that languages are not independent samples—pairwise similarities may be the result of common descent or language contact. As the common statistical tests presuppose independence of samples, they are not readily applicable to cross-linguistic data.

One way to mitigate this effect—pioneered by Bell (1978), Dryer (1989), and Perkins (1989)—is to control for genealogy and areal effects when sampling. In the simplest case, only one language is sampled per genealogical unit, and statistical effects are applied to different macro-areas independently. More recent work often uses more sophisticated techniques such as repeated stratified random sampling (e.g., Blasi et al., 2016). Another approach currently gaining traction is the usage of (generalized) mixed-effects models (Breslow and Clayton, 1993), where genealogical units such as families or genera, as well as linguistic areas, are random effects see, e.g., Atkinson (2011), Bentz and Winter (2013), and Jaeger et al. (2011) for applications to typology.

[1] For example, Greenberg's Universal 1 "In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object".

[2] For instance Universal 3 "Languages with dominant VSO order are always prepositional".

In a seminal paper, Maslova (2000) proposes an entirely different conceptual take on the problems of typological generalizations and typological sampling. Briefly put, if languages of type A (e.g., nominative-accusative marking) are more frequent than languages of type B (e.g., ergative-absolutive marking), this may be due to three different reasons: (1) diachronic shifts B→A are more likely than shifts A→B; (2) proto-languages of type A diversified stronger than those of type B, and the daughter languages mostly preserve their ancestor's type, and (3) proto-world, or the proto-languages at relevant prehistoric population bottlenecks, happened to be of type A, and this asymmetry is maintained due to diachronic inertia. Only the first type of reason is potentially linguistically interesting and amenable to a cognitive or functional explanation. Reasons of category (2) or (3) reflect contingent accidents. Stratified sampling controls for biases due to (2), but it is hard to factor out (1) from (3) on the basis of synchronic data. Maslova suggests that the theory of Markov processes offers a principled solution. If it is possible to estimate the diachronic transition probabilities, and if one assumes that language change has the Markov property (i.e., is memoryless), one can compute the long-term equilibrium probability of this Markov process. This equilibrium distribution should be used as the basis to identify linguistically meaningful distributional universals.

Maslova and Nikitina (2007) make proposals how to implement this research program involving the systematic gauging of the distribution of the features in question within language families.

Bickel (2011, 2013) introduces the *Family Bias Theory* as a statistical technique to detect biased distributions of feature values across languages of different lineages while controlling for statistical non-independence. Briefly put, the method assesses the tendency for biased distributions within families on the basis of large families, and extrapolates the results to small families and isolates.

In this article we will propose an implementation of Maslova's program that makes use of algorithmic techniques from computational biology, especially the *phylogenetic comparative method*. A technically similar approach has been pursued by Dunn et al. (2011), where it was confined to individual language families. Here we will propose an extension of their method that uses data from several language families and isolates simultaneously. Unlike the above-mentioned approaches, our method makes use of the entire phylogenetic structures of language families including branch lengths—to be estimated from lexical data—, and it treats large families, small families, and isolates completely alike. Also, our method is formulated as a generative model in the statistical sense. This affords the usage of standard techniques from Bayesian statistics such as inferring posterior uncertainty of latent parameter values, predictive checks via simulations, and statistical model comparison.

The model will be exemplified with a study of the potential correlations between eight word-order features from the *World Atlas of Language Structure* (Dryer and Haspelmath, 2013) that were also used by Dunn et al. (2011).

2. STATISTICAL ANALYSIS

2.1. Continuous Time Markov Processes

Following Maslova (2000), we assume that the diachronic change of typological features follows a continuous time Markov process (abbreviated as CTMC, for continuous time Markov chain). Briefly put, this means that a language is always in one of finitely many different states. Spontaneous mutations can occur at any point in time. If a mutation occurs, the language switches to some other state according to some probability distribution. This process has the Markov property, i.e., the mutation probabilities—both the chance of a mutation occurring, and the probabilities of the mutation targets—only depend on the state the language is in, not on its history.

Mathematically, the properties of such a process can be succinctly expressed in a single $n \times n$-matrix Q, where n is the number of states. The entries along the diagonal are non-positive and all other entries non-negative. Each row sums up to 0. The waiting time until the next mutation, when being in state i, is an exponentially distributed random variable with rate parameter $-q_{ii}$. If a mutation occurs while being in state i, the probability of a mutation $i \to j$ is proportional to q_{ij}.

The probability of a language ending up in state j after a time interval t when being in state i at the beginning of the interval (regardless of the number and type of mutations happening during the interval) is p_{ij}, where

$$P = e^{tQ}.$$

(The number e is the base of the natural logarithm). According to the theory of Markov processes[4], if each state can be reached from each other state in a finite amount of time, there is a unique equilibrium distribution π. Regardless of the initial distribution, the probability of a language being in state i after time t converges to π_i when t grows to infinity. Also, the proportion of time a language spent in state i during a time interval t converges to π_i when t grows to infinity. According to Maslova, it is this equilibrium distribution π that affords linguistic insights and that therefore should be identified in distributional typology.

2.2. Phylogenetic Markov Chains

Different languages are not samples from independent runs of such a CTMC. Rather, their properties are correlated due to common descent, which can be represented by a *language phylogeny*. A phylogeny is a family tree of related languages with the common ancestor at the root and the extant (or documented) languages at the leaves. Unlike the family trees used in classical historical linguistics, branches of a phylogeny have a length, i.e., a positive real number that is proportional to the time interval the branch represents. According to the model used here, when a language splits into two daughter languages, those initially have

[3] In the context of the present study, this is taken to be a possibly simplifying assumption that is part of the statistical approach taken. Whether or not it is empirically true would require a more careful discussion than what is possible here. Our currently best guess is that the assumption holds true provided a sufficiently fine-grained notion of "state" is adopted. For the coarse-grained states taken from WALS, such as VO or AN, the assumption is arguably too simplistic.

[4] See, e.g., Grimmett and Stirzaker (2001).

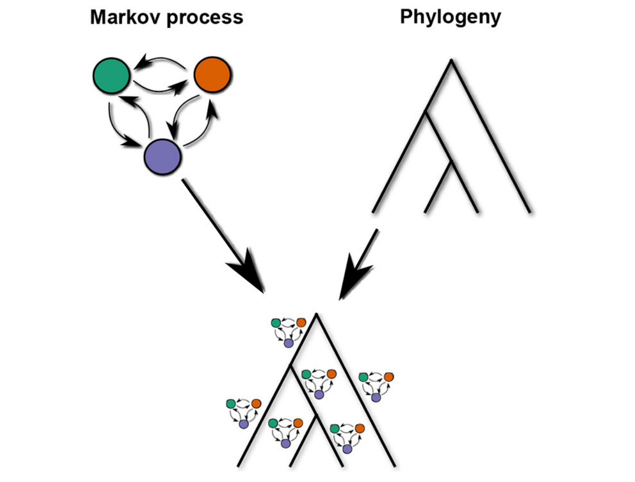


FIGURE 1 | Schematic structure of the phylogenetic CTMC model. Independent but identical instances of a CTMC run on the branches of a phylogeny.

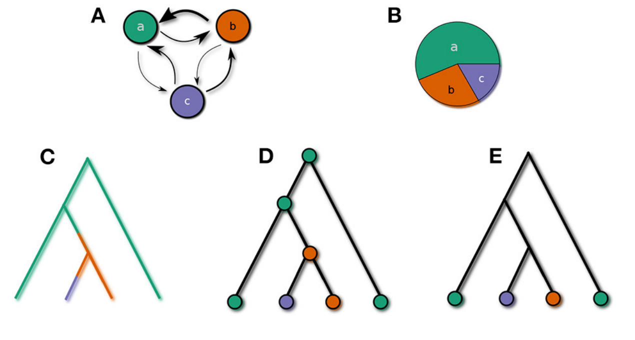


FIGURE 2 | (A) CTMC, **(B)** equilibrium distribution, **(C)** fully specified history of a phylogenetic Markov chain, **(D)** Marginalizing over events at branches, **(E)** marginalizing over states at internal nodes.

the same state but then evolve independently according to the same CTMC. This is schematically illustrated in **Figure 1**.

Let us illustrate this with an example. Suppose the feature in question has three possible values, *a*, *b*, and *c*. The *Q*-matrix characterizing the CTMC is given in (1).

$$Q = \begin{pmatrix} -3 & 2 & 1 \\ 5 & -6 & 1 \\ 2 & 3 & -5 \end{pmatrix} \tag{1}$$

The equilibrium distribution[5] π for this CTMC is

$$\begin{aligned} \pi &= (9/16, 13/48, 1/6) \\ &\approx (0.56, 0.27, 0.17) \end{aligned} \tag{2}$$

The transition rates and the equilibrium distribution are illustrated in the upper panels of **Figure 2**.

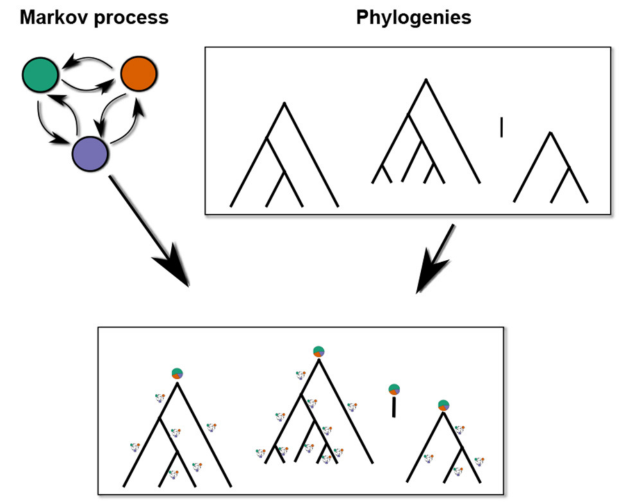


FIGURE 3 | Phylogenetic Markov CTMC with a collection of phylogenies.

A complete history of a run of this CTMC along the branches of a phylogeny is shown in **Figure 2C**. If the transition rates and branch lengths are known, the likelihood of this history, conditional on the state at the root, can be computed. To completely specify the likelihood of the history, one needs the probability distribution over states at the root of the tree—i.e., at the proto-language. In this paper we assume that proto-languages of known language families are the result of a long time of language change. If nothing about this history is known, the distribution of states at the proto-language is therefore virtually identical to the equilibrium distribution[6].

The precise points in time where state transitions occur are usually unknown though. We can specify an infinite set of possible histories which only agree on the states of the nodes of the tree (illustrated in **Figure 2D**). The marginal likelihood of this set is the product of the conditional likelihood of the bottom node of each branch given the top node and the length of each branch, multiplied with the equilibrium probability of the root state.

Under normal circumstances only the states of the extant languages, i.e., at the leaves of the tree, are known. The marginal likelihood of all histories agreeing only in the states at the leaves can be determined by summing over all possible distribution of states at internal nodes (illustrated in **Figure 2E**). This quantity can be computed efficiently via a recursive bottom-up procedure known as Felsenstein's (1981) *pruning algorithm*.

This can easily be extended to sets of phylogenies (e.g., a collection of phylogenies for different language families; schematically illustrated in **Figure 3**). Language isolates are degenerate phylogenies with only one leave that is also the root. The likelihood of the state of an isolate is thus the equilibrium probability of its state.

[5] This is the left null vector of Q, normalized such that it sums to 1.

[6] It is possible to test for a given feature distribution and phylogeny whether the data support this hypothesis. We leave this issue for future work.

Under the assumption that the distributions in different language families are independent, the total likelihood of such a collection of phylogenies is the product of the individual tree likelihoods.

Under realistic conditions, the precise phylogeny of a language family is never known. Rather, it is possible to infer a probability distribution over phylogenies using Bayesian inference and, e.g., lexical data. In such a scenario, the *expected likelihood* for a language family is the averaged likelihood over this distribution of trees.

If only the phylogeny and the states at the leaves are known, statistical inference can be used to determine the transition rates (and thereby also the equilibrium distribution). Bayesian inference, that is used in this study, requires to specify prior assumptions over the transition rates and results in a posterior distribution over these rates.

In the remainder of this paper, this program is illustrated with a case on word order features and their potential correlations.

2.3. Word Order Features

The typical order of major syntactic constituents in declarative sentences of a language, and the distribution of these properties across the languages of the world, has occupied the attention of typologists continuously since the work of Greenberg (1963) (see, e.g., Lehmann, 1973; Vennemann, 1974; Hawkins, 1983; Dryer, 1992, among many others). There is a widespread consensus that certain word-order features are typologically correlated. For instance, languages with verb-object order tend to be prepositional while object-verb languages are predominantly postpositional. Other putative correlations, like the one between verb-object order and adjective-noun order, are more controversial.

The study in Dunn et al. (2011) undermined this entire research program. They considered eight word-order features and four major language families. For each pair of features and each family, they conducted a statistical test whether the feature pair is correlated in that family, using Bayesian phylogenetic inference. Surprisingly, they found virtually no agreement across language families. From this finding they conclude that the dynamics of change of word-order features is lineage specific; so the search for universals is void.

We will take up this problem and will consider the same eight word order features, which are taken from the *World Atlas of Language Structures* (WALS; Dryer and Haspelmath, 2013). For each of the 28 feature pairs, we will test two hypotheses:

1. All lineages (language families and isolates) share the parameters of a CTMC governing the evolution of these features (vs. Each lineage has its own CTMC parameters), and
2. If all lineages share CTMC parameters, the two features are correlated.

For each of the eight features considered, only the values "head-dependent" vs. "dependent-head" are considered. Languages that do not fall in either category are treated as "missing value". These features and the corresponding values are listed in **Table 1**.

TABLE 1 | Word order features.

Feature	Value 1	Value 2
VS	Verb-subject	Subject-verb
VO	Verb-object	Object-verb
PN	Adposition-noun	Noun-adposition
NG	Noun-genitive	Genitive-noun
NA	Noun-adjective	Adjective-noun
ND	Noun-demonstrative	Demonstrative-noun
NNum	Noun-numeral	Numeral-noun
NRc	Noun-relative clause	Relative clause-noun

2.4. Obtaining Language Phylogenies

Applying the phylogenetic Markov chain model to typological data requires phylogenies of the languages involved. In this section we describe how these phylogenies were obtained.

In Jäger (2018), a method is described how to extract binary characters out of the lexical data from the *Automated Similarity Judgment Program* (ASJP v. 18; Wichmann et al., 2018). These characters are suitable to use for Bayesian phylogenetic inference.

The processing pipeline described in Jäger (2018) is briefly recapitulated here. The ASJP data contains word lists from more than 7,000 languages and dialects, covering the translations of 40 core concepts. All entries are given in a uniform phonetic transcription.

In a first step, mutual string similarities are computed using pairwise sequence alignment along the lines of Jäger (2013). From these similarities, pairwise language distances are computed. These two measures are used to group the words for a each concept into cluster approximating *cognate classes*. Each such cluster defines a binary character, with value 1 for languages containing an element of the cluster in its word list, 0 if the entries for the same concepts all belong to different clusters, and undefined if there is no entry for that concept.

An additional class of binary characters is obtained from the Cartesian product of the 40 concepts and the 41 sound classes used in ASJP. A language has entry 1 for character "concept c/sound class s" if one of the entries for concept "c" contains at least one occurrence of sound class "s," 0 if none of the entries for "c" contain "s," and undefined if there is no entry for that concept.

In Jäger (2018), it is demonstrated that phylogenetic inference based on these characters is quite precise. For this assessment, the expert classification from Glottolog (Hammarström et al., 2018) is used as gold standard.

For the present study, we identified a total of 1,626 of languages for which WALS contains information about at least one word-order feature and the data from Jäger (2018) contain characters. These languages comprise 175 lineages according to the Glottolog classification, including 81 isolates[7]. The geographic distribution of this sample is shown in **Figure 4**.

Here, we used the cognate classes occurring within the language sample, as well as the concept/sound class characters

[7] A language is called an isolate here if our 1,626-language sample contains no other language belonging to the same Glottolog family.

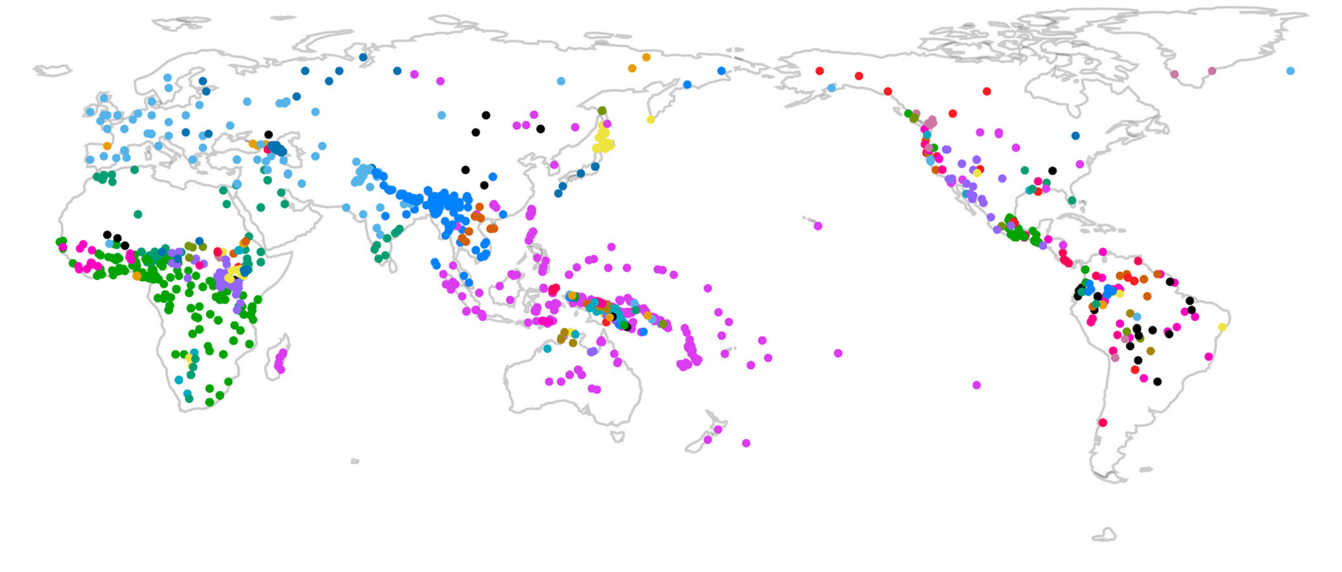


FIGURE 4 | Geographic distribution of the sample of languages used. Colors indicate Glottolog classification.

as input for Bayesian phylogenetic inference. For each language family, a posterior tree sample was inferred using the Glottolog classification as constraint trees[8]. For each family, we sampled 1,000 phylogenies from the posterior distribution for further processing.

2.5. Generative Models

To study the co-evolution of two potentially correlated word-order features, we assume a four-state CTMC for each pair of such features—one state for each combination of values. We assume that all twelve state transitions are *a priori* possible, including simultaneous changes of both feature values[9]. The structure of the CTMC is schematically shown in **Figure 5** for the feature pair VO/NA.

As pointed out above, Dunn et al. (2011) argue that the transition rates between the states of word-order features follow lineage-specific dynamics. To test this assumption (Hypothesis 1 above), we fitted two models for each feature pair:

- a **universal model** where all lineages follow the same CTMC with universally identical transition rates, and
- a **lineage-specific model** where each lineage has its own set of transition rate parameters.

These two model structures are illustrated in **Figure 6**.

For all models we chose a log-normal distribution with parameters $\mu = 0$ and $\sigma = 1$ as prior for all rate parameters.

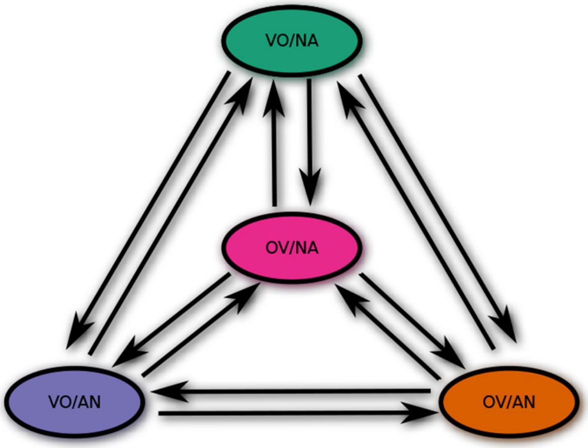


FIGURE 5 | CTMC for a possibly correlated feature pair.

We will determine via statistical model comparison for each feature pair which of the two models fits the data better.

2.6. Prior Predictive Sampling

In a first step, we performed prior predictive sampling for both model types. This means that we simulated artificial datasets that were drawn from the prior distributions, and then compared them along several dimensions with the empirical data. This step is a useful heuristics to assess whether the chosen model and the chosen prior distributions are in principle capable to adequately model the data under investigation.

We identified three statistics to summarize the properties of these artificial data and compare them with the empirically observed data. For this purpose we represented each language

[8] For this purpose we used the software *MrBayes* (Ronquist and Huelsenbeck, 2003) for families with at least three members and *RevBayes* (Höhna et al., 2016) for two-language families.

[9] Dunn et al. (2011), following the general methodology of Pagel and Meade (2006), exclude this possibility. We believe, however, that this possibility should not be excluded *a priori* because it is conceivable that during a major reorganization of the grammar of a language, several features change their values at once.

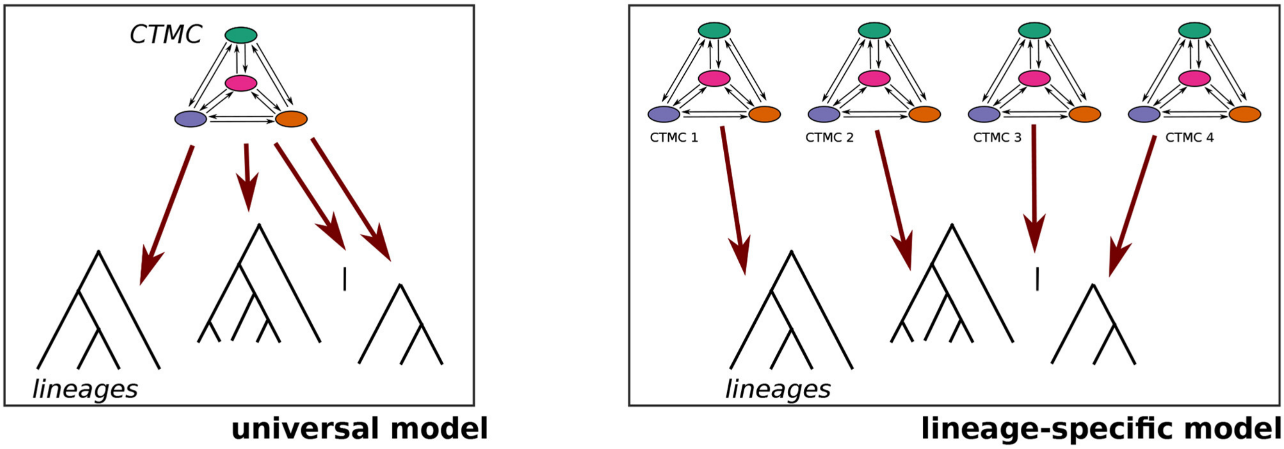


FIGURE 6 | Universal vs. lineage-specific model.

as a probability vector over the four possible state. Let Y be the data matrix with languages as rows and states as columns, and n the number of languages, and F the set of lineages, where each lineage is a set of languages. The statistics used are:

- the **total variance**:

$$\frac{1}{n}\sum_i\left(\sum_l Y_{l,i}^2 - \left(\sum_l Y_{l,i}\right)^2\right)$$

- the **mean lineage-wise variance**:

$$\frac{1}{|F|}\sum_k \frac{1}{|F_k|}\sum_i\left(\sum_{l\in F_k} Y_{l,i}^2 - \left(\sum_{l\in F_k} Y_{l,i}\right)^2\right)$$

- the **cross-family variance**, i.e., the total variance between the centroids for each lineage:

$$\sum_i\left(\frac{1}{|F|}\sum_k\left(\frac{1}{|F_k|}\sum_{l\in F_k} Y_{l,i}\right)^2 - \left(\frac{1}{|F|}\sum_k \frac{1}{|F_k|}\sum_{l\in F_k} Y_{l,i}\right)^2\right)$$

In **Figure 7**, the distribution of these statistics for the 28 feature pairs for the empirical data are compared with the prior distributions for the universal model (top panels) and the lineage-specific model (bottom panels). For each model, we conducted 1,000 simulation runs.

From visual inspection it is easy to see that for the universal model, the empirically observed values fall squarely within the range of the prior distributions. For the lineage-specific model, the observed variances are generally lower than expected under the prior distribution. This is especially obvious with regard to the cross-family variance, which is much lower for the empirical data than predicted by the model.

Following the suggestion of a reviewer, we performed a Mann-Whitney U-test for each configuration to test the hypothesis that empirical and simulated data come from the same distribution. The results (shown inside the plots in **Figure 7**) confirm the visual inspection. For the total variance and the cross-lineage variance and the universal model, the hypothesis of equal distributions cannot be rejected, while the empirical distribution differs significantly from the simulated data for the other four configurations.

2.7. Model Fitting

Both models were fitted for each of the 28 feature pairs. Computations were performed using the programming language *Julia* and Brian J. Smith's package *Mamba* (https://github.com/brian-j-smith/Mamba.jl) for Bayesian inference. We extended *Mamba* by functionality to handle phylogenetic CTMC models.

Posterior samples were obtained via Slice sampling (Neal, 2003). Averaging over the prior of phylogenies was achieved by randomly picking one phylogeny from the prior (see section 2.4) in each MCMC step. Posterior sampling was stopped when the *potential scale reduction factor* (PSRF; Gelman and Rubin, 1992) was ≤ 1.1 for all parameters.

2.8. Posterior Predictive Sampling

To test the fit of the models to the data, we performed *posterior predictive sampling* for all fitted models. This means that for each model, we randomly picked 1,000 samples from the posterior distribution and used it to simulate artificial datasets. The three statistics used above for prior predictive sampling were computed for each simulation. The results are shown in the **Supplementary Material**.

With regard to total variance, we find that the empirical value falls outside the 95% highest posterior interval for three out of 28 feature pairs (VO-NRc, PN-NRc, and NA-ND), where the model overestimates the total variance. The lineage-specific model overestimates the total variance for 10 feature pairs.

Since three outliers out of 28 is within the expected range for a 95% interval, we can conclude that the universal model

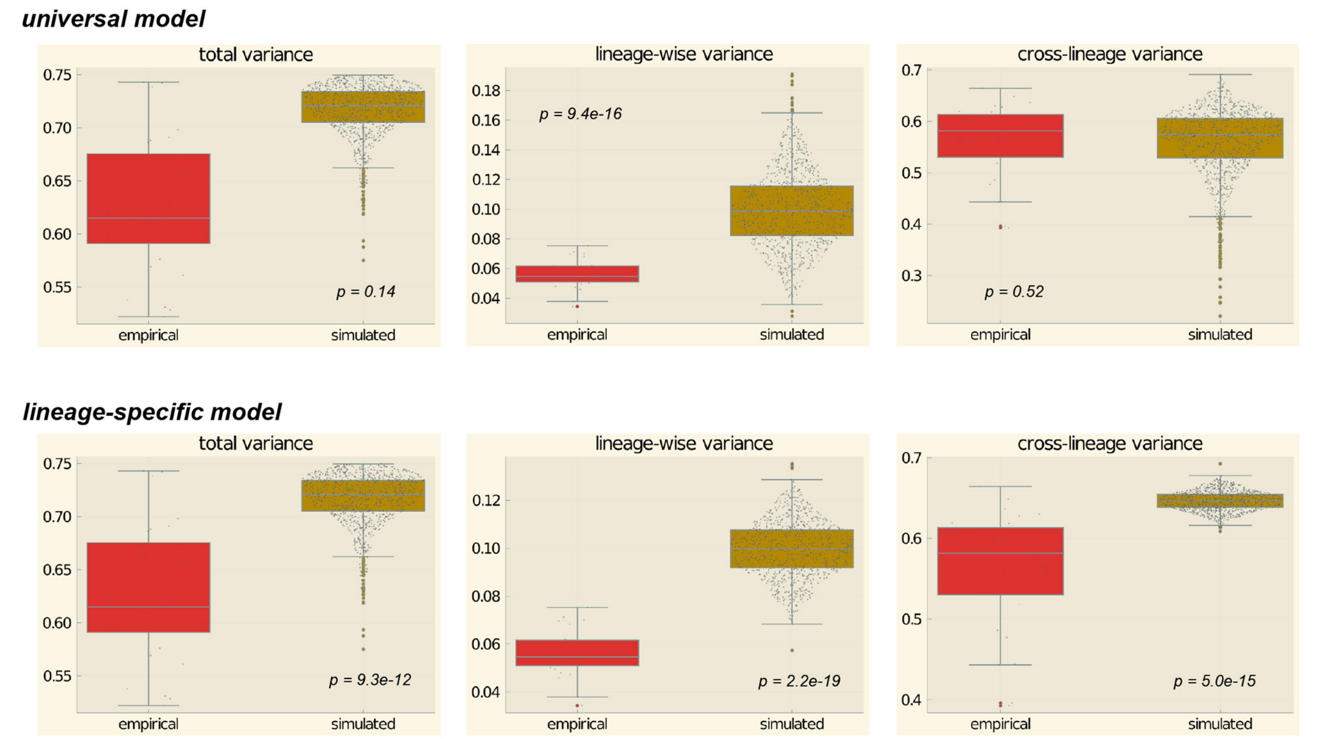


FIGURE 7 | Prior predictive simulations. The *p*-values are the result of a Mann-Whitney *U*-test whether empirical and simulated values come from the same distribution.

generally predicts the right amount of cross-linguistic variance. The lineage-specific model overestimates this quantity.

For cross-linguistic variance, the empirical value falls outside the HPD (95% highest posterior density interval) for 14 pairs for the universal model and for 21 pairs for the lineage-specific model. So both models tend to overestimate this variable. This might be due to the fact that phylogenetic CTMC models disregard the effect of language contact, which arguably reduced within-family variance.

The cross-family variance falls into the universal model's HPD for all pairs, but only for two pairs (VO-NA, VO-NNum) for the lineage-specific model. Briefly put, the universal model gets this quantity right while the lineage-specific model massively overestimates it.

2.9. Bayesian Model Comparison

As a next step we performed statistical model comparison between the universal and the lineage-specific model. Briefly put, model comparison estimates how well models will serve to predict unseen data that are generated by the same process as the observed data, and compares the predictive performances. Everything else being equal, the model with the better predictive performance can be considered a better explanation for the observed data.

Since there is no general consensus about the best method to compare Bayesian models (see, e.g., Vehtari and Ojanen, 2012 for an overview), we applied two techniques.

The *marginal likelihood* of the data under a Bayesian model is the expected likelihood of the data y weighted by the prior probability of the model parameters θ.

$$D(y|M) = \int_{\theta} p(y|\theta)p(\theta|M)d\theta$$

The Bayes factor between two models M_1 and M_2 is the ratio of their marginal densities:

$$BF = \frac{D(y|M_1)}{D(y|M_2)}$$

To estimate the marginal densities, we used **bridge sampling** (cf. Gronau et al., 2017). For our implementation we depended strongly on the *R*-package *bridgesampling* (Gronau et al., 2020). The logarithmically transformed Bayes factors between the universal model ($\approx M_1$) and the lineage-specific model ($\approx M_2$) are shown for each feature pair in **Table 2**.

All log-Bayes factors are positive, i.e., favor the universal over the lineage-specific model.

According to the widely used criteria by Jeffreys (1998), a Bayes factor of $\geq$ 100, which corresponds to a logarithmic Bayes factor of 4.6, is considered as decisive evidence. So except for the feature pair VO-NNum, this test provides decisive evidence in favor of the universal model.

Unlike frequentist hypothesis testing, Bayesian model comparison does not require a correction for multiple testing.

TABLE 2 | log-Bayes factor between universal and lineage-specific model.

Feature pair	(log) Bayes factor	Cumulative posterior probability
VS-VO	72.9	0.000
VS-NG	65.9	0.000
PN-NG	64.5	0.000
VO-PN	56.8	0.000
VS-PN	54.4	0.000
VO-NG	41.3	0.000
VS-NRc	36.5	1.11e-16
NA-ND	32.1	1.18e-14
VS-NNum	31.1	4.19e-14
VS-NA	30.8	8.57e-14
NG-NRc	28.0	8.09e-13
VO-NRc	27.7	1.79e-12
VS-ND	27.0	3.67e-12
PN-NRc	25.6	1.12e-11
NA-NRc	22.1	2.63e-10
NG-ND	19.0	5.98e-9
NG-NA	18.8	1.29e-8
ND-NNum	15.6	1.76e-7
PN-NA	15.2	4.38e-7
NA-NNum	8.8	0.000147
PN-ND	8.7	0.000319
ND-NRc	7.3	0.00101
NG-NNum	6.7	0.00223
VO-ND	6.4	0.00393
NNum-NRc	5.3	0.00892
PN-NNum	5.1	0.0152
VO-NA	5.0	0.0218
VO-NNum	2.4	0.102

The last row gives the upper limit of the posterior probability that for at least one feature-pair up to this line the lineage-specific model is true.

Still, since 28 different hypotheses are tested simultaneously here, the question arises how confident we can be that a given subset of the hypotheses are true. Assuming the uninformative prior that the universal and the lineage-specific model are equally likely *a priori*, the posterior probability of the universal model being true given that one of the two models is true, is the logistic transformation of the log-Bayes factor. Let us call this quantity p_i^u for feature pair *i*. We assume that feature-pairs are sorted in descending order according to their Bayes factor, as in **Table 2**. The posterior probability of the lineage-specific model is $p_i^l = 1 - p_i^u$. The quantity $p_{1\cdots k}^l$ is the cumulative probability that the lineage-specific model is true for at least one feature pair *i* with $1 \leq i \leq k$[10].

Since the hypotheses for the individual feature pairs are not mutually independent, it is not possible to compute this probability. However, according to the *Bonferroni inequality*, it holds that

$$p_{1\cdots k}^l \leq \sum_{1\leq i\leq k} p_k^l$$

The right-hand side of this inequality provides an upper limit for the left-hand side. This upper limit is shown in the third column of **Table 2**. For all but the feature pair VO-NNum, this probability is < 0.05. We conclude that this line of reasoning also confirms that the data strongly support the universal over the lineage-specific hypothesis for all feature pairs except VO-NNum.

As alternative approach to model comparison, we conducted **Pareto-smoothed cross-validation** (Vehtari et al., 2017) using the *R*-package *loo* (Vehtari et al., 2020).

Leave-one-out cross-validation means to loop through all data points y_i and compute the quantity

$$\log p(y_i|y_{-i}) = \int_\theta p(y_i|\theta)p(\theta|y_{-i})d\theta$$

Here, y_{-i} denotes the collection of all datapoints $\neq y_i$. Since this amounts to fitting a posterior distribution as often as there are datapoints, this is computationally not feasible in most cases (including the present case study). The quantity

$$\text{elpd} = \sum_i \log p(y_i|y_{-i}),$$

the *expected log pointwise predictive density*, is a good measure of how well a model predicts unseen data and can be used to compare models.

Since computing the elpd amounts to fitting a posterior distribution for each datapoint, the method is not feasible though in most cases (including the present case study). Pareto-smoothed leave-one-out cross-validation is a technique to estimate elpd from the posterior distribution of the entire dataset.

However, this algorithm depends on the assumption that individual datapoints are mutually *conditionally independent*, i.e.,

$$p(y|\theta) = \prod_i p(y_i|\theta).$$

This is evidently not the case for phylogenetic CTMC models if we treat each language as a datapoint[11]. However, conditional independence does hold between lineages both in the universal and the lineage-dependent model. Pareto-smoothed leave-one-out cross-validation can be therefore be performed if entire lineages are treated as datapoints.

The difference in elpd, i.e., elpd of universal model minus elpd of lineage-specific model, are shown in **Table 3**. For all feature pairs, the elpd is higher for the universal than for the lineage-specific model.

[10] This amounts to the Holm-Bonferroni correction (Holm, 1979), but we use it here to compute an upper limit for the posterior probability rather than for the expected α level.

[11] To see why, consider an extreme case where the phylogeny consists of two leaves with an infinitesimally small co-phenetic distance, and the equilibrium distribution over states is the uniform distribution. Then $p(y_1 = y_2 = s_1) \approx p(y_1 = s_1) = p(y_2 = s_1) < p(y_1 = s_1)p(y_2 = s_1)$.

TABLE 3 | Differences in elpd.

Feature pair	△ elpd
VS-VO	79.7
PN-NG	75.9
VS-NG	72.6
VO-PN	65.3
VS-PN	61.4
VO-NG	48.3
VS-NRc	45.5
NA-ND	37.3
NG-NRc	36.7
VS-NA	35.3
VO-NRc	34.6
PN-NRc	32.7
VS-NNum	28.7
NA-NRc	27.9
VS-ND	26.1
NG-NA	21.1
PN-ND	19.2
PN-NA	18.0
NG-ND	16.1
VO-ND	13.6
ND-NNum	12.5
NG-NNum	12.3
ND-NRc	7.5
PN-NNum	6.6
NA-NNum	4.2
NNum-NRc	3.7
VO-NNum	3.5
VO-NA	1.1

To summarize, for all feature pairs except VO-NNum, different methods of model comparison agree that the universal model provides a better fit to the data than the lineage-specific model. For VO-NNum, the evidence is more equivocal, but it is also slightly favors the universal model.

From this we conclude that there is no sufficient empirical evidence for the assumption of lineage specificity in the evolution of correlated word-order features. Dunn et al.'s (2011) finding to the contrary is based on a much smaller sample of 301 languages from just four families, and it omits an explicit model comparison.

2.10. Feature Correlations

Let us know turn to the second hypothesis mentioned in section 2.3, repeated here. For each feature pair, we will probe the question:

> If all lineages share CTMC parameters, the two features are correlated.

To operationalize correlation, we define the feature value "dependent precedes head" as 0 and "head precedes dependent" as 1. For a given feature pair, this defines a 2 × 2 contingency table with posterior equilibrium probabilities for each value combination. They are displayed in **Figure 8**. In each diagram, the x-axis represents the first feature and the y-axis the second feature. The size of the circles at the corners of the unit square indicate the equilibrium probability of the corresponding value combination. Blurred edges of the circles represent posterior uncertainty.

The diagrams also show the posterior distribution of regression lines indicating the direction and strength of the association between the two features[12]. Perhaps surprisingly, for some feature pairs the association is negative.

The *correlation* between two features binary f_1, f_2 in the strict mathematical sense, also called the *Phi coefficient*, is

$$\frac{\mathrm{cov}(f_1,f_2)}{\sqrt{\mathrm{var}(f_1)\mathrm{var}(f_2)}}$$
$$= \frac{p_{00}p_{11} - p_{10}p_{01}}{\sqrt{(p_{00}+p_{01})(p_{10}+p_{11})(p_{00}+p_{10})(p_{10}+p_{11})}}$$

and ranges from -1 (perfect negative relationship) to 1 (perfect positive relationship), with 0 indicating no relationship.

The median posterior correlations and the corresponding HPD interval given in **Table 4** and shown in **Figure 9**.

How reliable are these estimates? The Bayes factor between the hypotheses "correlation $\neq$ 0" and "correlation $=$ 0" can be determined via the Savage-Dickey method (Dickey and Lientz, 1970; see also Wagenmakers et al., 2010). We used the *R*-package *LRO.utilities* (https://github.com/LudvigOlsen/LRO.utilities/) to carry out the computations. The log-Bayes factors for the individual feature pairs are shown in **Table 5**.

Using the same method as in section 2.9, we can conclude with 95% confidence that there is a non-zero correlation for 13 feature pairs: VO-PN, VS-VO, VS-NG, PN-NG, NA-NNum, ND-NNum, VO-NG, VO-NRc, NA-NRc, NA-ND, NNum-NRc, PN-NRc. For all these pairs, the correlation coefficient is credibly positive (meaning the 95% HPD interval is entirely positive). There is not sufficient evidence that there is a negative correlation for any feature pair. For the four feature pairs where the HPD interval for the correlation coefficient is entirely negative (VO-NA, VO-NNum, VS-NNum, PN-NA), the log-Bayes factors in favor of a non-zero correlation (1.60, 2.01, 2.47, 3.17) are too small to merit a definite conclusion.

Conversely, for no feature pair is the Bayes factor in favor of a zero-correlation large enough to infer the absence of a correlation.

3. DISCUSSION

3.1. Equilibrium Analysis vs. Language Sampling

Maslova (2000) argues that the frequency distribution of typological feature values may be biased by accidents of history, and that the equilibrium distribution of the underlying Markov process more accurately reflects the effects of the cognitive

[12] Intercept and slope of the regression lines are $\frac{p_{01}}{p_{00}+p_{01}}$ and $\frac{p_{11}}{p_{10}+p_{11}} - \frac{p_{01}}{p_{00}+p_{01}}$, respectively.

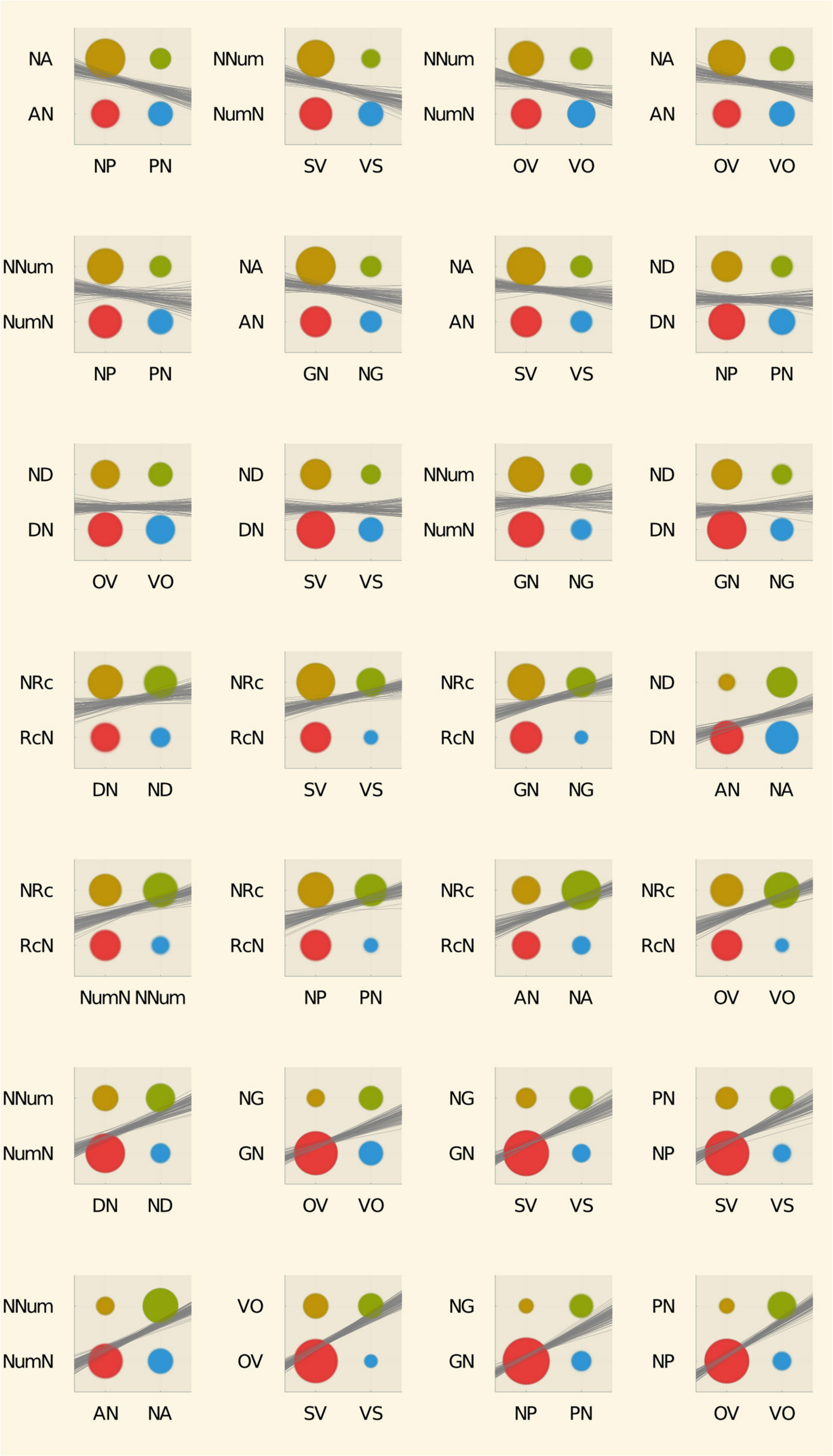


FIGURE 8 | Posterior equilibrium probabilities and linear regression.

TABLE 4 | Correlation coefficients for feature pairs: median and 95% HPD interval.

Feature pair	Median	HPD
VO-PN	0.64	(0.53, 0.75)
PN-NG	0.55	(0.41, 0.67)
VS-VO	0.49	(0.38, 0.60)
NA-NNum	0.47	(0.34, 0.59)
VS-PN	0.45	(0.32, 0.58)
VS-NG	0.45	(0.32, 0.58)
VO-NG	0.41	(0.27, 0.53)
ND-NNum	0.38	(0.26, 0.50)
VO-NRc	0.38	(0.24, 0.50)
NA-NRc	0.37	(0.23, 0.51)
PN-NRc	0.28	(0.14, 0.42)
NNum-NRc	0.28	(0.13, 0.42)
NA-ND	0.27	(0.15, 0.39)
NG-NRc	0.24	(0.09, 0.38)
VS-NRc	0.19	(0.05, 0.32)
ND-NRc	0.17	(0.00, 0.32)
NG-ND	0.06	(−0.06, 0.20)
NG-NNum	0.05	(−0.09, 0.19)
VS-ND	−0.00	(−0.13, 0.14)
VO-ND	−0.01	(−0.13, 0.12)
PN-ND	−0.01	(−0.15, 0.11)
VS-NA	−0.09	(−0.22, 0.05)
NG-NA	−0.12	(−0.24, 0.02)
PN-NNum	−0.12	(−0.27, 0.05)
VO-NA	−0.17	(−0.30, −0.04)
VO-NNum	−0.19	(−0.32, −0.05)
VS-NNum	−0.20	(−0.33, −0.06)
PN-NA	−0.21	(−0.34, −0.08)

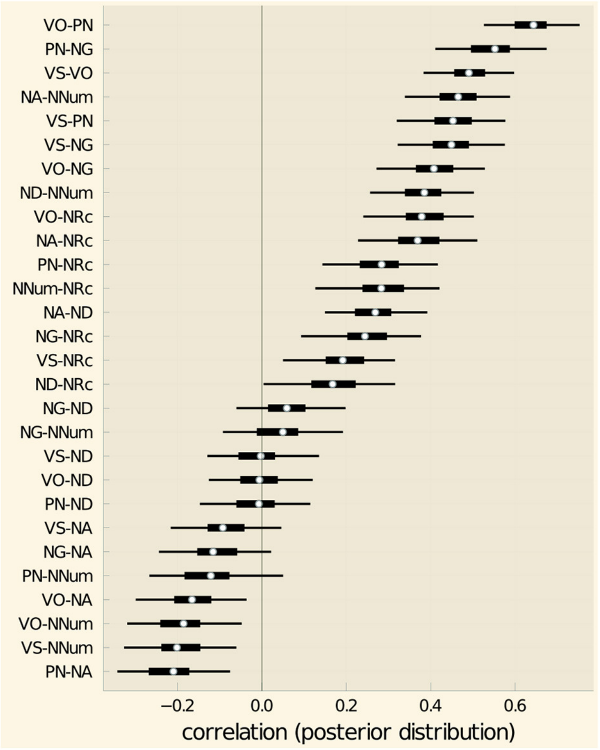


FIGURE 9 | Correlation coefficients for feature pairs. White dots indicate the median, thick lines the 50% and thin lines the 95% HPD intervals.

and functional forces. Inspection of our results reveals that the difference between raw frequencies and equilibrium probabilities can be quite substantial. In **Table 6**, the relative frequencies, the equilibrium frequencies and the 95% HPD intervals for the four values of the feature combination "verb-object/adposition-noun" are shown.

We also computed the *stratified frequencies*, i.e., the weighted means where each language is weighted by the inverse of the size of its Glottolog lineage. As a result, each lineage has the same cumulative weight.

The same information is displayed in **Figure 10**. It can be discerned that uniformly head-initial languages (VO-AdpN) are over-represented among the languages of the world in comparison to the equilibrium distribution while uniformly head-final languages (OV-NAdp) are underrepresented. The stratified frequencies come very close to the equilibrium distribution though. This discrepancies are arguably due to the fact that head-initial languages are predominant in several large families while head-final languages are quite frequent among small families and isolates.

This example suggests that our approach effectively achieves something similar than stratified sampling, namely discounting the impact of large families and give more weight to small families and isolates. A more detailed study of the relationship between stratified sampling and equilibrium analysis is a topic for future research.

3.2. Universal vs. Lineage-Specific Models

The findings from section 2.9 clearly demonstrate that the universal model provides a better fit of the data than the lineage-specific model. This raises the question why Dunn et al. (2011) came to the opposite conclusion. There are several relevant considerations. First, these authors did not directly test a universal model. Rather, they fitted two lineage-specific models for each feature pair—one where the features evolve independently and one where the mutation rates of one feature may depend on the state of the other feature. They then compute the Bayes factor between these models for each family separately and conclude that the patterns of Bayes factors vary wildly between families. So essentially it is tested whether the pattern of feature correlations is identical across families.

In this paper, we explored slightly different hypotheses. We tested whether the data support a model where all lineages following the same dynamics with the same parameters (where a correlation between features is possible), or whether they support different parameters (each admitting a correlation between features). Having the same model across lineages implies an identical correlation structure, but it also implies many other things, such as identical equilibrium distributions, identical rate of change etc.

To pick an example, Dunn et al. (2011) found evidence for a correlation between NA and NRc for Austronesian and Indo-European but not for Bantu and Uto-Aztecan. This seems to speak against a universal model. However, inspection of our data reveals that the feature value "relative clause precedes noun" only occurs in 1.8% of all Austronesian and 13.8% of all Indo-European languages, and it does not occur at all in Bantu or Uto-Aztecan. The universal model correctly predicts that the observed frequency distributions will be similar across lineages (as demonstrated by the low cross-family variance in the prior predictive simulations discussed in section 2.6). The lineage-specific model cannot account for this kind of cross-family similarities. More generally, our approach to test the relative merits of a universal versus lineage-specific dynamics regarding word-order features takes more sources of information into account than just correlation patterns. This more inclusive view clearly supports the universal model.

TABLE 5 | log-Bayes factor between "correlation ≠ 0" and "correlation = 0".

Feature pair	(log) Bayes factor	Cumulative posterior probability
VO-PN	19.25	**4.37e-9**
VS-VO	15.82	**1.39e-7**
VS-NG	14.12	**8.78e-7**
PN-NG	12.07	**6.62e-6**
VS-PN	12.03	**1.26e-5**
NA-NNum	10.93	**3.05e-5**
ND-NNum	9.98	**7.68e-5**
VO-NG	8.85	**0.00022**
VO-NRc	8.15	**0.000509**
NA-NRc	7.43	**0.0011**
NA-ND	4.72	**0.00995**
NNum-NRc	4.07	**0.0267**
PN-NRc	3.83	**0.0479**
PN-NA	3.17	0.0884
NG-NRc	2.64	0.155
VS-NNum	2.47	0.233
VO-NNum	2.01	0.351
VS-NRc	1.93	0.478
VO-NA	1.60	0.646
ND-NRc	0.55	1.000
NG-NA	−0.03	1.000
PN-NNum	−0.43	1.000
VS-NA	−0.51	1.000
NG-ND	−1.17	1.000
NG-NNum	−1.32	1.000
PN-ND	−1.52	1.000
VO-ND	−1.56	1.000
VS-ND	−1.64	1.000

The last row gives the upper limit of the posterior probability that for at least one feature-pair up to this line correlation = 0. Cumulative posterior probabilites < 0.05 are shown in bold.

3.3. Word-Order Correlations

The 13 feature pairs identified in section 2.10 for which there is credible evidence for a correlation are shown in **Figure 11**, where connecting lines indicate credible evidence for a correlation.

The four features correlated with VO are exactly those among the features considered here that were identified by Dryer (1992) as "verb patterners," i.e., for which he found evidence for a correlation with verb-object order. These are verb-subject, noun-genitive, adposition-noun and noun-relative clause. It is perhaps noteworthy that like Dryer, we did not find credible evidence for a correlation between verb-object order and noun-adjective order, even though such a connection has repeatedly been hypothesized, e.g., by Lehmann (1973) and Vennemann (1974), and, more recently, by Ferrer-i-Cancho and Liu (2014).

Besides Dryer's verb patterns, we found a group of three mutually correlated features, noun-numeral, noun-adjective and noun-demonstrative. Two of them, noun-numeral and

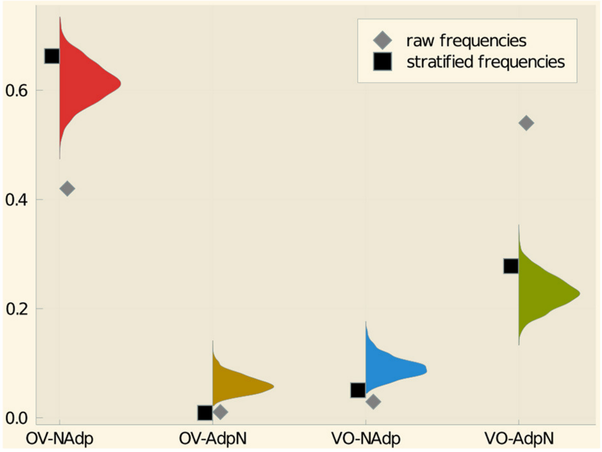


FIGURE 10 | Relative frequencies, stratified frequencies, and posterior probabilities of the four value combinations of VO-PN.

TABLE 6 | Relative frequencies, stratified frequencies, and posterior probabilities of the four value combinations of VO-PN.

Values	Relative frequencies	Stratified relative frequencies	Equilibrium (median)	HPD
OV-NAdp	0.420	0.663	0.614	(0.540, 0.685)
OV-AdpN	0.011	0.009	0.060	(0.032, 0.096)
VO-NAdp	0.030	0.051	0.091	(0.054, 0.134)
VO-AdpN	0.540	0.278	0.230	(0.175, 0.293)

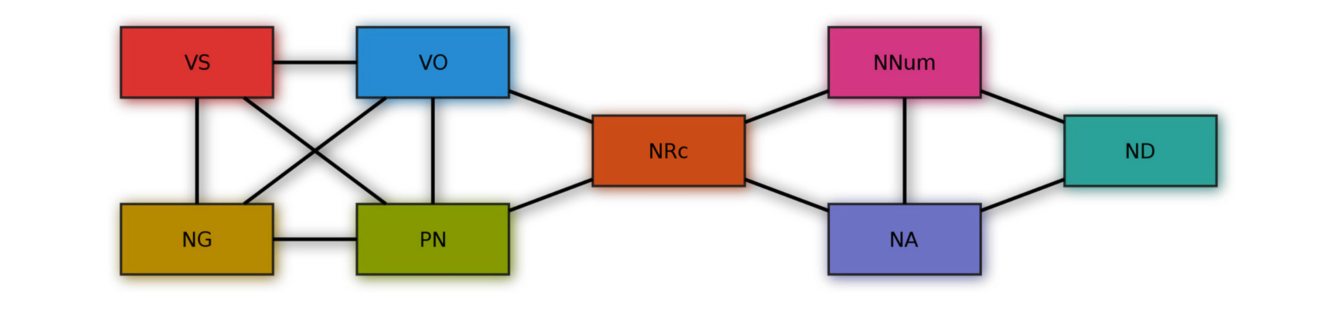


FIGURE 11 | Feature-pairs with credible evidence for a correlation.

noun-adjective, are also correlated with noun-relative clause. These correlations have received less attention in the typological literature. The findings are not very surprising though, given that all these features pertain to the ordering of noun-phrase material relative to the head noun.

4. CONCLUSION

In this article we demonstrated that the modeling of typological feature distributions in terms of phylogenetic continuous-time Markov chains—inspired Maslova's (2000) theoretical work as well as by research within the framework of the biological comparative method such as Pagel and Meade (2006) and Dunn et al. (2011)—has several advantages for typology. It allows to use all data, from families large and small as well as from isolate languages. The method controls for non-independence due to common descent. Couched in a Bayesian framework, it affords standard techniques for model checking and model comparison as well as quantification of the uncertainty in inference. We do see it as essential though that this kind of study uses data from a variety of lineages since individual families generally do not display evidence for all the possible diachronic transitions required to estimate transition rates reliably. Working with forests rather than single trees, i.e., with trees or tree distributions for several families and also including isolates as elementary trees is a suitable way to achieve this goal[13].

To demonstrate the viability of this method, we chose a re-assessment of the issue broad up by Dunn et al. (2011): Are word-order correlations lineage specific or universal? Using a collection of 1,626 languages from 175 lineages (94 families and 81 isolates), we found conclusive evidence that a universal model provides a much better fit to the word-order data from WALS than a lineage-specific model. Furthermore we found statistical evidence for a correlation for 13 word-order features (out of 28 considered), which largely confirm the findings of traditional typological research.

There is a variety of open issues for future research. Maslova (2000) also discusses the possibility that the current distribution of feature value represents traces of proto-world or some later bottleneck language, which would bias the estimation of the equilibrium distribution. In the present paper this option was disregarded. It is possible to address this question using Bayesian model comparison.

By design, phylogenetic models only capture vertical transmission. The effects of language contact and areal tendencies are systematically ignored. In future work, this could be remedied by including areal and spatial random effects into the model.

Statistical research in other disciplines involving stratified data suggest that the binary alternative between a lineage-specific and a universal model might be ill-posed. Both approaches can be integrated within *hierarchical models* (see, e.g., Gelman et al., 2014; McElreath, 2016) where between-group variance is as small as possible but as large as the data require. Due to the high number of parameters involved, fitting such models, however, poses a considerable computational challenge.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: http://asjp.clld.org/ (Automated Similarity Judgment Program); http://glottolog.org (Glottolog); http://wals.info/ (World Atlas of Language Structure); https://doi.org/10.17605/OSF.IO/CUFV7 (accompanying data for Jäger, 2018, Global-scale phylogenetic linguistic inference from lexical resources, Scientific Data). The code used for this study can be found at https://github.com/gerhardJaeger/phylogeneticTypology.

AUTHOR CONTRIBUTIONS

GJ conducted the study and wrote up the article. JW programmed the software used and gave technical support. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the DFG Centre for Advanced Studies in the Humanities Words, Bones, Genes, Tools (DFG-KFG 2237) and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement 834050). We acknowledge support by Open Access Publishing Fund of University of Tübingen.

[13] Verkerk et al. (2021) use a similar approach but utilize a universal tree encompassing all lineages.

ACKNOWLEDGMENTS

We thank Michael Franke for helpful input regarding Bayesian model comparison, and two reviewers for their feedback.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.682132/full#supplementary-material

REFERENCES

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332, 346–349. doi: 10.1126/science.1199295

Bell, A. (1978). "Language sampling," in *Universals of Human Language I: Method and Theory*, eds J. H. Greenberg, C. A. Ferguson, and E. A. Moravcsik (Stanford, CA: Stanford University Press), 125–156.

Bentz, C., and Winter, B. (2013). Languages with more second language learners tend to lose nominal case. *Lang. Dyn. Change* 3, 1–27. doi: 10.1163/22105832-13030105

Bickel, B. (2011). Statistical modeling of language universals. *Linguist. Typol*. 15, 401–413. doi: 10.1515/lity.2011.027

Bickel, B. (2013). "Distributional biases in language families," in *Language Typology and Historical Contingency: In Honor of Johanna Nichols*, eds B. Bickel, L. A. Grenoble, D. A. Peterson, and A. Timberlake (Amsterdam: John Benjamins), 415–444. doi: 10.1075/tsl.104.19bic

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., and Christiansen, M. H. (2016). Sound-meaning association biases evidenced across thousands of languages. *Proc. Natl. Acad. Sci. U.S.A*. 113, 10818–10823. doi: 10.1073/pnas.1605782113

Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc*. 88, 9–25. doi: 10.1080/01621459.1993.10594284

Dickey, J. M., and Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *Ann. Math. Stat*. 41, 214–226. doi: 10.1214/aoms/1177697203

Dryer, M. S. (1989). Large linguistic areas and language sampling. *Stud. Lang*. 13, 257–292. doi: 10.1075/sl.13.2.03dry

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language* 68, 81–138. doi: 10.1353/lan.1992.0028

Dryer, M. S., and Haspelmath, M. (2013). *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. Available online at: http://wals.info/

Dunn, M., Greenhill, S. J., Levinson, S., and Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473, 79–82. doi: 10.1038/nature09923

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol*. 17, 368–376. doi: 10.1007/BF01734359

Ferrer-i-Cancho, R., and Liu, H. (2014). The risks of mixing dependency lengths from sequences of different length. *Glottotheory* 143–155. doi: 10.1515/glot-2014-0014

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. Boca Raton, FL: CRC Press. doi: 10.1201/b16018

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci*. 7, 457–472. doi: 10.1214/ss/1177011136

Greenberg, J. (ed.). (1963). "Some universals of grammar with special reference to the order of meaningful elements," in *Universals of Language* (Cambridge, MA: MIT Press), 73–113.

Grimmett, G. and Stirzaker, D. (2001). *Probability and Random Processes*. Oxford: Oxford University Press.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., et al. (2017). A tutorial on bridge sampling. *J. Math. Psychol*. 81, 80–97. doi: 10.1016/j.jmp.2017.09.005

Gronau, Q. F., Singmann, H., and Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *J. Stat. Softw*. 92, 1–29. doi: 10.18637/jss.v092.i10

Hammarström, H., Forkel, R., and Haspelmath, M. (2018). *Glottolog 3.3.2. Jena: Max Planck Institute for the Science of Human History*. Available online at: http://glottolog.org (accessed April 8, 2021).

Hawkins, J. A. (1983). *Word Order Universals*. London: Academic Press.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., et al. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol*. 65, 726–736. doi: 10.1093/sysbio/syw021

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat*. 6, 65–70.

Jaeger, T. F., Graff, P., Croft, W., and Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguist. Typol*. 15, 281–319. doi: 10.1515/lity.2011.021

Jäger, G. (2013). Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Lang. Dyn. Change* 3, 245–291. doi: 10.1163/22105832-13030204

Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Sci. Data* 5:180189. doi: 10.1038/sdata.2018.189

Jeffreys, H. (1998). *The Theory of Probability*. Oxford: Oxford University Press.

Lehmann, W. P. (1973). A structural principle of language and its implications. *Language* 9, 47–66. doi: 10.2307/412102

Maslova, E. (2000). A dynamic approach to the verification of distributional universals. *Linguist. Typol*. 4, 307–333. doi: 10.1515/lity.2000.4.3.307

Maslova, E., and Nikitina, E. (2007). *Stochastic Universals and Dynamics of Cross-Linguistic Distributions: the Case of Alignment Types*. Stanford University.

McElreath, R. (2016). *Statistical Rethinking. A Bayesian Course with Examples in R and Stan*. Boca Raton, FL: CRC Press.

Neal, R. M. (2003). Slice sampling. *Ann. Stat*. 31, 705–741. doi: 10.1214/aos/1056562461

Pagel, M., and Meade, A. (2006). Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am. Nat*. 167, 808–825. doi: 10.1086/503444

Perkins, R. D. (1989). Statistical techniques for determining language sample size. *Stud. Lang*. 13, 293–315. doi: 10.1075/sl.13.2.04per

Ronquist, F., and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574. doi: 10.1093/bioinformatics/btg180

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Brkner, P.-C., Paananen, T., et al. (2020). *loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models*. R Package Version 2.4.1. Available online at: https://mc-stan.org/loo/. (accessed April 8, 2021).

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput*. 27, 1413–1432. doi: 10.1007/s11222-016-9696-4

Vehtari, A., and Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Stat. Surveys* 6, 142–228. doi: 10.1214/12-SS102

Vennemann, T. (1974). Theoretical word order studies: results and problems. *Papiere Linguist*. 7, 5–25.

Verkerk, A., Haynie, H., Gray, R., Greenhill, S., Shcherbakova, O., and Skirgrd, H. (2021). "Revisiting typological universals with Grambank," in *Paper Presented at the DGfS Annual Meeting* (Freiburg).

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., and Grasman, R. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the Savage-Dickey method. *Cogn. Psychol*. 60, 158–189. doi: 10.1016/j.cogpsych.2009.12.001

Wichmann, S., Holman, E. W., and Brown, C. H. (2018). *The ASJP Database (version 18)*. Available online at: http://asjp.clld.org/ (accessed April 8, 2021).

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 Psychology

ORIGINAL RESEARCH
published: 22 July 2021
doi: 10.3389/fpsyg.2021.662125

Predictable Words Are More Likely to Be Omitted in Fragments–Evidence From Production Data

Robin Lemke[1,2]*, Ingo Reich[1,2], Lisa Schäfer[1,2] and Heiner Drenhaus[1,3]

[1] Collaborative Research Center 1102, Saarland University, Saarbrücken, Germany, [2] Department of Modern German Linguistics, Saarland University, Saarbrücken, Germany, [3] Department of Language Science and Technology, Saarland University, Saarbrücken, Germany

Instead of a full sentence like *Bring me to the university* (uttered by the passenger to a taxi driver) speakers often use fragments like *To the university* to get their message across. So far there is no comprehensive and empirically supported account of why and under which circumstances speakers sometimes prefer a fragment over the corresponding full sentence. We propose an information-theoretic account to model this choice: A speaker chooses the encoding that distributes information most uniformly across the utterance in order to make the most efficient use of the hearer's processing resources (Uniform Information Density, Levy and Jaeger, 2007). Since processing effort is related to the predictability of words (Hale, 2001) our account predicts two effects of word probability on omissions: First, omitting predictable words (which are more easily processed), avoids underutilizing processing resources. Second, inserting words before very unpredictable words distributes otherwise excessively high processing effort more uniformly. We test these predictions with a production study that supports both of these predictions. Our study makes two main contributions: First we develop an empirically motivated and supported account of fragment usage. Second, we extend previous evidence for information-theoretic processing constraints on language in two ways: We find predictability effects on omissions driven by extralinguistic context, whereas previous research mostly focused on effects of local linguistic context. Furthermore, we show that omissions of content words are also subject to information-theoretic well-formedness considerations. Previously, this has been shown mostly for the omission of function words.

Keywords: information theory, fragments, ellipsis, script knowledge, surprisal

Edited by:
Hannah Rohde,
University of Edinburgh,
United Kingdom

Reviewed by:
Leon Bergen,
University of California, San Diego,
United States
Cassandra L. Jacobs,
University of Wisconsin-Madison,
United States

****Correspondence:***
Robin Lemke
robin.lemke@uni-saarland.de

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

Received: *31 January 2021*
Accepted: *18 May 2021*
Published: *22 July 2021*

Citation:
Lemke R, Reich I, Schäfer L and Drenhaus H (2021) Predictable Words Are More Likely to Be Omitted in Fragments–Evidence From Production Data. Front. Psychol. 12:662125. doi: 10.3389/fpsyg.2021.662125

1. INTRODUCTION

In order to communicate a message to a hearer, speakers have to select a particular utterance from a set of utterances that can be used to convey this message in the utterance situation. Besides utterances that contain different word forms or syntactic constructions, speakers can often resort to a subsentential utterance like (1-a) instead of a full sentence like (1-b). Despite their reduced form, given an appropriate context, such subsentential utterances are interpreted as roughly meaning-equivalent to their fully sentential counterparts.

(1) [A pedestrian approaches a taxi at the train station and says:]
   a. To the university, please.
   b. Bring me to the university, please.

Subsentential utterances, or fragments[1] (Morgan, 1973), have been discussed extensively in the theoretical literature from a syntactic perspective, in particular with respect to the question of whether they are a genuinely nonsentential output of syntax (Ginzburg and Sag, 2000; Barton and Progovac, 2005; Culicover and Jackendoff, 2005; Stainton, 2006), or derived by ellipsis from regular sentences (Merchant, 2004; Reich, 2007; Weir, 2014).

Only a few studies have looked into the questions of why speakers use fragments at all, and under which circumstances they prefer them over the corresponding full sentence. In the theoretical literature, the grammaticality of omissions has been related to information structure, in particular to the notions of focus and givenness (Merchant, 2004; Reich, 2007; Weir, 2014; Ott and Struckmeier, 2016; Griffiths, 2019). Leaving aside conceptual differences between these accounts, overall they agree on the prediction that only material that is given in an information-structural sense (Schwarzschild, 1999) can be omitted and that words that belong to the focus (see e.g., Rooth, 1992) must be realized. Such information structure-based accounts however explain only why fragments can or cannot be used under particular conditions, but not why they are (not) used when they are licensed by grammar.

The sparse evidence available with respect to the actual usage of fragments suggests that the choice between a fragment and a sentence is driven by the general tendency to maximize communicative efficiency: Speakers adapt the form of the utterance to properties of the situation and the hearer. This idea has been formalized in information-theoretic (Levy and Jaeger, 2007; Levy, 2008) and game-theoretic frameworks (Franke, 2009; Frank and Goodman, 2012). Bergen and Goodman (2015) combine a game-theoretic model of rational communication with a noisy channel model, which in sum predicts that the choice between a fragment and a complete sentence is based on a trade-off between the cost for producing an utterance and the risk of not being understood correctly. Even though the account is promising, it is only illustrated with a highly simplified example of a question-answer pair. Bergen and Goodman (2015) do not apply it to more realistic communication situations which involve more diverse utterances, potentially communicated messages and predictability effects drive by extralinguistic context. Lemke et al. (2021) in turn explain the choice as adaptation to the processing resources of the hearer. They argue that predictable utterances, which require less processing effort (Hale, 2001; Demberg and Keller, 2008; Levy, 2008), are more often reduced in order to use the hearer's processing resources efficiently.

Both Bergen and Goodman (2015) and Lemke et al. (2021) provide an explanation for when speakers prefer to reduce their utterance more strongly and consequently to produce a fragment rather than a full sentence, but they do not make clear predictions about why speakers prefer *particular* fragments if a sentence can be reduced in different ways. For instance, in the taxi example (1), it seems more natural for the passenger to say *to the university* than *me the university*, even though both of these fragments reduce the utterance to a similar extent. While Lemke et al. (2021) just show that the reduction of predictable utterances is more acceptable, Bergen and Goodman (2015) include a cost term in their model that penalizes utterances that are effortful to produce. Since Bergen and Goodman (2015) derive a preference for fragments from this cost term, it is most likely intended to be affected by the length of an utterance, but they do not make this explicit or discuss other sources of production effort, like a cost for retrieving unpredictable words (Ferreira and Dell, 2000). In the absence of more specific accounts of why particular words are more likely to be omitted, the general tendency to densify predictable utterances or those produced in the absence of noise cannot fully explain speakers' production choices.

In this article we propose an information-theoretic account of fragment usage, according to which the predictability of utterances and words therein determines (i) whether speakers choose a sentence or a fragment to get their message across, and, in the latter case, (ii) which words are omitted in fragments. We hypothesize that this choice is driven by the tendency to distribute processing effort uniformly across the utterance (Fenk and Fenk, 1980; Levy and Jaeger, 2007). Since the effort required to process a word is inversely proportional to its probability (Hale, 2001; Levy, 2008), our account makes two specific predictions: First, likely words are preferably omitted, and second, words that increase the likelihood of otherwise unlikely following words are preferably inserted.

An information-theoretic approach is particularly promising for two reasons: First, Lemke et al. (2021) found that fragments are overall more strongly preferred in predictive contexts. This finding is in line with our account, because in predictive contexts the words within an utterance are overall more likely and thus also more likely to be omitted. However, Lemke et al. (2021) did not look into the more fine-grained predictions of the information-theoretic account on the word level. Second, information-theoretic constraints have been shown to explain the distribution of omissions particularly on the word level, such as those of complementizers, pronouns, articles and case markers (Levy and Jaeger, 2007; Frank and Jaeger, 2008; Jaeger, 2010; Asr and Demberg, 2015; Kurumada and Jaeger, 2015; Norcliffe and Jaeger, 2016; Lemke et al., 2017). Even though most of these studies focused on semantically relatively empty function words,[2] information-theoretic constraints make similar predictions on

[1] In the theoretical literature there is no agreement on a definition of the notions *nonsentential utterance* or *fragment*. Researchers diverge in particular with respect to the question of which elliptical utterances are categorized as fragments and with respect to the presence or absence of an explicit antecedent. As for the first question, in this article we restrict ourselves to DP fragments, which are analyzed as fragments by all accounts of fragments. As for the second question, some researchers (see e.g., Merchant, 2004; Barton and Progovac, 2005) do not distinguish between fragments that occur discourse-initially and those that have a linguistic antecedent, like short answers (but cf. Klein, 1993; Reich, 2011). We avoid this debate by investigating only uncontroversial discourse-initial fragments.

[2] But cf. Kravtchenko (2014), Schäfer (2021).

the omission of content words. Therefore, our research extends previous evidence for information-theoretic considerations on the speaker's preferences on optional omissions to content words and larger phrases.[3]

Testing the predictions of information-theoretic constraints on optional omissions in general requires (i) a corpus that contains instances of the relevant omissions and the corresponding full forms (in our case, complete sentences), (ii) knowing *which* expressions have been omitted, and (iii) a method to estimate the probability of both the omitted and the realized expressions in the data. Logistic regressions can then show whether the predictability of a target expression given the material surrounding it has a significant effect on the likelihood of the omission of the target expression. The application of this procedure to fragments however is difficult due to three properties of fragments that inhibit probability estimation with standard language models. We address these challenges by collecting a data set with a production task that allows us to investigate the predictions of our account.

1. It is often unclear which lexical item has been omitted in fragments. For instance, in the taxi example, a hearer who processes the fragment *to the university* in (1-a) can interpret it as *bring me to the university*, or *I'd like to go to the university*, and which of these reconstructions is assumed obviously affects the words' probability estimates. In order to reduce this ambiguity, we preprocess our data set so that that omissions can be relatively unambiguously resolved.
2. Fragments often occur discourse initially, so that the likelihood of utterances and words therein is determined by extralinguistic context, which language cannot take into account, because this information is not contained in standard text corpora. We quantify effects of extralinguistic context by eliciting the utterances in our data set with script-based context stories, which are based on probabilistic event chains extracted from DeScript (Wanzare et al., 2016), a freely available crowd-sourced corpus of script knowledge.
3. Levy and Jaeger (2007) observe a circularity issue that concerns probability estimations on elliptical data: If predictable expressions are particularly often omitted, they will appear to be rare in a corpus, or at least not as frequent as their probability would suggest, just *because* of their high ratio of omission. We propose a new approach to estimate the probability of each word in our data that is not vulnerable to the circularity issue. Our method relies on a version of the data set that does not contain omissions, so probability estimation is not affected by a word's actual omission rate.

Our study is the first empirical investigation of why speakers choose (not) to omit particular words in fragments, and consequently, in which situations they prefer to use a fragment rather than a full sentence. From the broader perspective of information-theoretic research on language, we extend previous evidence for information-theoretic processing constraints in two ways. First, we provide evidence for such effects on content words, and second, we find that not only local linguistic context, but also extralinguistic context drives predictability effects. From a methodological perspective, our probability estimation method circumvents the circularity issue observed by Levy and Jaeger (2007) and provides an approach to quantifying by-word probability in the presence of omissions.

The article is structured as follows: Section 2 sketches our information-theoretic account and its central predictions on fragment usage. Section 3 presents the production experiment and section 4 summarizes our main results and contributions and their relevance for related theories of probability effects on optional omission and reduction.

2. AN INFORMATION-THEORETIC ACCOUNT OF FRAGMENT USAGE

Information-theoretic processing constraints have been shown to explain the distribution of a wide range of reduction phenomena. Their application ranges from phonological reduction (Bell et al., 2003, 2009; Aylett and Turk, 2004; Tily et al., 2009; Demberg et al., 2012; Kuperman and Bresnan, 2012; Seyfarth, 2014; Pate and Goldwater, 2015; Brandt et al., 2017, 2018; Malisz et al., 2018) to morphological effects on contraction (Frank and Jaeger, 2008) and case marker omission (Kurumada and Jaeger, 2015; Norcliffe and Jaeger, 2016) to pronominalization (Tily and Piantadosi, 2009), and, what is most closely related to omissions in fragments, optional omissions of various types of function words (Levy and Jaeger, 2007; Jaeger, 2010; Asr and Demberg, 2015; Lemke et al., 2017) and preverbal subjects (Kravtchenko, 2014; Schäfer, 2021).

The central idea of information-theoretic accounts of omission phenomena is that speakers use omissions in order to optimize their utterance with respect to properties of the situation and the hearer. Information-theoretic approaches model this as the *channel capacity* in the sense of Shannon (1948), i.e., the maximum amount of information that can be transmitted across a channel with a limited capacity. *Information*, or *surprisal* (Hale, 2001) is defined probabilistically as $-\log_2 p(word \mid context)$, i.e., the negative logarithm of a word's likelihood to appear in a given context. The less likely a word is, the more information it conveys. Since Hale (2001), this notion of information has been related to processing effort: The more information a word conveys, the more processing effort it requires (see also Hale, 2006; Demberg and Keller, 2008; Levy, 2008). Given the link between information and processing effort, we interpret channel capacity as an upper bound to the cognitive resources of the hearer. If this upper bound is exceeded, the hearer is not

[3] In the literature there is no consensus on which expressions are classified as fragments. For instance, Barton and Progovac (2005) discuss the omission of articles and prepositions in an article on the syntax of fragments, whereas e.g., Merchant (2004) takes into account mostly bare DPs and PPs. In this article we avoid this issue by testing only discourse-initial bare DP fragments, which are classified as fragments by all theories thereof.

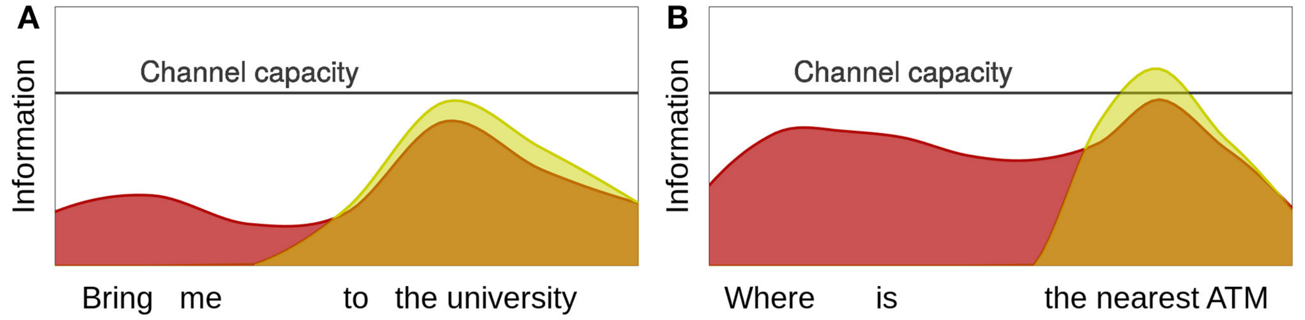


FIGURE 1 | Hypothetical ID profiles for the predictable **(A)** and the unpredictable **(B)** sentence (red) and fragment (yellow) examples in the taxi example. In the case of the predictable utterance the fragment is more well-formed, because it fragment avoids the trough in the profile caused by *bring me*. In the case of the unpredictable utterance the sentence is more well-formed, because the fragment causes a peak in the profile that exceeds channel capacity.

able to successfully process an input, whereas under-utilizing channel capacity results in inefficient communication.[4] Taken together, this predicts that speakers adapt their utterance so as to communicate at a rate close to, but not exceeding, channel capacity. Information maxima that exceed channel capacity shall be avoided, just like information minima that do not make use of the full cognitive resources available to the hearer.

UID makes two main predictions with respect to omissions in fragments. Words that are more likely in context are more likely to be omitted in order to **avoid local information minima** which result in the underutilization of the hearer's processing resources and appear as *troughs* in the information density (ID) profile, as the left facet of **Figure 1** shows. In the taxi example in (1), it is very likely that the pedestrian approaching the vehicle wants to be brought somewhere, hence the words *bring me* are highly predictable and convey only little information. In contrast, the destination is less predictable in this context, hence the information on *the university* is higher. Such information minima are inefficient and can be smoothed by omitting these words. In situations where the structure resulting from this omission is a fragment, UID hence predicts that speakers prefer this fragment over the corresponding full sentence.

In contrast, words that precede unpredictable words are more likely to be realized in order to **avoid local information maxima**, which exceed the hearer's processing resources and appear as *peaks* in the ID profile. Inserting optional words often increases the predictability of following ones, because this restricts the range of possible successors (Hale, 2001; Levy, 2008). Consequently, inserting these words can reduce the information maximum on following words.[5] For instance, in the taxi example the pedestrian could ask for the nearest ATM instead of asking for a ride. If asking for a direction is less likely, the words *where is* will be more informative and hence less likely to underutilize the hearer's processing resources. Furthermore, if *the nearest ATM* is less likely to be a potential destination than the university, it will be more likely to yield a peak in the ID profile that exceeds channel capacity. Inserting *where is* in turn might increase the likelihood of locations that are asked for frequently, like a subway station, a bus stop or an ATM, and thus smooth the peak on *the nearest ATM* that occurs in the fragment. Hence, in case of this example, a speaker should prefer to produce a complete sentence rather than a fragment.

An important property of UID is that the omission or insertion of optimal is limited to variation between "the bounds defined by grammar" (Jaeger, 2010, p. 25): Omissions which are ruled out by grammar will not be preferred even if they distribute information more uniformly across the signal. For instance, Schäfer (2021) finds UID effects on the omission of preverbal subjects in German text messages, however, in more prototypically written text types in her corpus there is not a single instance of this construction. With respect to fragments, this predicts that omissions are restricted to those that are available in the language and text type under investigation.

Determining whether a specific omission contributes to the optimization of an utterance given these predictions in principle would require knowing channel capacity. Only information maxima which exceed channel capacity are to be avoided. In practice however, channel capacity is necessarily unknown, since the amount of processing resources that are available to the hearer depends on properties of the situation (Engonopoulos et al., 2013; Häuser et al., 2019) and of the individual hearer (Pate and Goldwater, 2015). Interlocutors must therefore

[4] This view differs from the discussion of channel capacity in related work (Levy and Jaeger, 2007; Jaeger, 2010), which emphasizes on the role of noise in determining channel capacity. Shannon (1948) shows that if the channel is noisy, the transmission rate can be increased up to channel capacity without increasing the noise ratio. Attempts to increase the transmission rate beyond channel capacity however reduces the actual transmissions rate because of an increased noise rate. The noise-based and the processing effort-based interpretations of channel capacity do not contradict each other, and with respect to our study, their predictions are identical. However, we tentatively assume that on the word level our processing account is more appropriate. Falsely perceiving a phoneme like /p/ as /b/ due to noise is relatively likely, whereas perceiving a word like *Mary* as *John* under normal communication conditions is much more unlikely, since they differ in a larger set of phonemes.

[5] Whether inserting words can only increase or also reduce the likelihood depends on the method used for surprisal estimation. Following the approach proposed by Hale (2001) words can only reduce the surprisal of following material, but the entropy-based method by Levy (2008) predicts that the insertion of preceding material can also increase the surprisal of a target word.

infer channel capacity. This assumption is also psychologically plausible, because we can never precisely know the amount of processing resources that is available to a hearer. In consequence, our hypothesis pertains even if channel capacity is unknown: Predictable words are more likely to yield an information minimum, which is smoothed by omission, and unpredictable ones are more likely to yield an information maximum, which can be reduced by the insertion of preceding material.

3. PRODUCTION STUDY

We use a production task to collect a data set that is suitable for the investigation of the predictions of the information-theoretic account: In order to avoid troughs and peaks in the ID profile, speakers prefer to omit predictable words and to insert additional redundancy before unpredictable words. Such a data set must (i) contain both instances of such omissions and of the corresponding full forms, (ii) allow for the quantification of predictability effects driven by extralinguistic context, and (iii) it must allow for the unambiguous reconstruction of the omitted material, because the way in which omissions are reconstructed affects the estimation of individual words' surprisal. In order to control for extralinguistic context, we elicited our data set with 24 script-based stories, as we describe at detail in sections 3.1, 3.2. In section 3.3 we discuss how we pre-processed the data in order to ensure a relatively unambiguous reconstruction of omitted material. Section 3.4 describes our surprisal estimation methods and section 3.5 the statistical analysis of the data.

3.1. Materials

In order to control and quantify effects of extralinguistic context, we used 24 stories like (2) to elicit participants' responses. We conducted the study in German and translate materials presented here for convenience. Participants were asked to produce the most likely utterance to be produced by the specified person in the situation described in the story. For each story, we collected a total of 100 responses. Since all of these responses are produced in the same context, this approach allows us to quantify effects of extralinguistic context on the likelihood of a response and the words therein.

(2) Annika and Jenny want to cook pasta. Annika has put a pot with water on the stove. Then she has turned the stove on. After a few minutes, the water has started to boil. Now Annika says to Jenny:

Stories like (2) might in principle trigger different expectations in different subjects, depending on their experiences and world knowledge. In order to minimize such effects, we based our stories on scripts, i.e., knowledge about the stereotypical time-course of everyday activities that is represented by partially ordered sequences of events (Schank and Abelson, 1977). For instance, the script about cooking pasta that underlies (2) contains events like *put a pot on the stove*, *turn the stove on* and *wait for the water to boil*, which most of the time appear in this order. Psycholinguistic studies have shown that script events prime upcoming events within the same script (see e.g., Bower et al., 1979; McKoon and Ratcliff, 1986; Millis et al., 1990; van den Broek, 1994; van der Meer et al., 2002; Nuthmann and Van Der Meer, 2005; Bicknell et al., 2010; Delogu et al., 2018), hence we expect that our context stories trigger expectations about what happens next and consequently determine which utterance is produced. For instance, in our example in (2), a request to pour the pasta into the pot or to give the speaker the pasta seems intuitively likely, whereas a question about ingredients of the sauce might be less likely.

We based our materials on event chains extracted from DeScript (Wanzare et al., 2016), a crowd-sourced corpus of script knowledge, in order to rely on empirically founded script representations rather than on our own intuitions. DeScript is a publicly available resource that contains about 100 descriptions of the stereotypical time-course of 40 everyday activities which differ in their complexity, the degree of variation and conventionalization (e.g., *flying on an airplane*, *making scrambled eggs* or *taking a bath*). We used a semi-automatic approach for extracting event chains from the corpus, i.e., sequences of events that are likely to follow each other.[6] Following Manshadi et al. (2008), we defined an event as the finite verb and its nominal complement, e.g., `put pot` in (2). After dependency-parsing the corpus with the Stanford parser (Klein and Manning, 2003) included in the Python Natural Language Toolkit (Loper and Bird, 2002), we extracted these event representations from it. We estimated the likelihood of an event given the previous one with bigram language models trained on the manually preprocessed data for each script with the SRILM toolkit (Stolcke, 2002). We then extracted sequences of three events that were most likely to follow each other and used these event chains to construct our materials. The first sentence in each item introduces the script (cooking pasta), and the next three ones elaborate the event chain (`put pot`, `turn on stove`, `boil water`). In context on this event chain, we expect a relatively high ratio of utterances referring to the most likely event to follow in the event chain, i.e., that of pouring the pasta into the pot.

3.2. Data Collection

The study was conducted using the LimeSurvey presentation software (LimeSurvey GmbH, 2012). The 24 stimuli were distributed across two lists (12 per list), mixed with 8 unrelated fillers that resembled our context stories and presented in individually fully randomized order. We recruited 200 self-reported native speakers of German on the crowdsourcing platform Clickworker, half of which were assigned to each of the lists. Each participant received €2 for participation. All participants agreed to the collection and aggregated or anonymized publication of their responses by participating in the study. We did not collect any personal data like participants' names, IP addresses or IDs on the Clickworker platform, whose collection would require additional data safety measures.

Subjects were asked to provide the most natural utterance to be produced by the specified person in the situation described by the context story. In the instructions we asked subjects

[6] The stories were originally used in the rating study in Lemke et al. (2021), we refer the reader there for more details on this procedure.

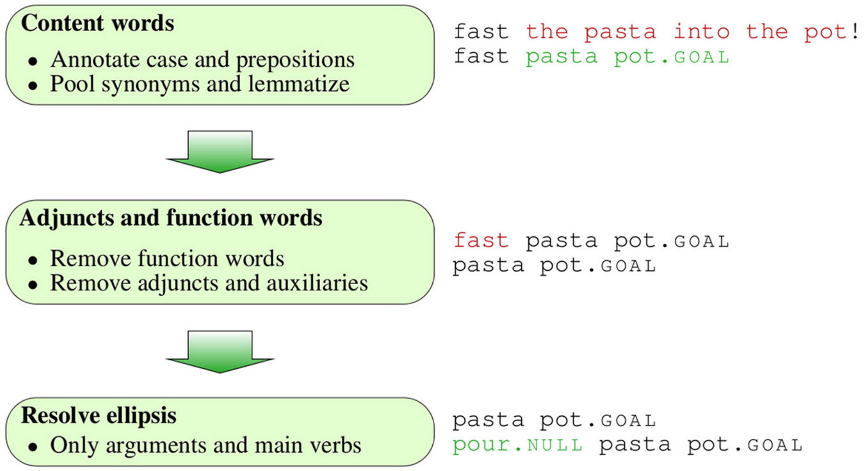

Content words
• Annotate case and prepositions
• Pool synonyms and lemmatize
fast the pasta into the pot!
fast pasta pot.GOAL
Adjuncts and function words
• Remove function words
• Remove adjuncts and auxiliaries
fast pasta pot.GOAL
pasta pot.GOAL
Resolve ellipsis
• Only arguments and main verbs
pasta pot.GOAL
pour.NULL pasta pot.GOAL

FIGURE 2 | Overview of the preprocessing procedure at the example of the fragment *Schnell die Nudeln in den Topf* "The pasta into the pot, quickly!" in (3), which we illustrate at the English translation of the example. First, the NP *the pasta* and the PP *into the pot* are merged to single expressions and the information conveyed by the function words annotated on the noun. Then the adverb *fast* is removed and finally the missing verb is reconstructed.

to produce *complete sentences*. Initially, we planned to collect two data sets, one without omissions, that would be used for surprisal estimation, and one with ellipses. Since subjects however produced omissions (up to 60% of all grammatically required words in a script were omitted, see also **Figure 3**) despite having been told not to do so, we used this data set for both surprisal estimation and for analysis. This might raise the concern that the ratio of omissions might be lower than if the task would be totally unconstrained, i.e., if we asked for *any* utterance that participants perceived as likely. In order to address this, we collected a second data set consisting of 50 responses for each item following the same procedure, but asking subjects to provide the most likely *utterance* in this context. The overall rate of omissions was slightly higher in the data set collected by asking for "utterances" (25.74%) than in the data set collected by asking for "sentences" (23.79%). However, a linear mixed effects regression that compared the omissions rates for each of the items between both data sets shows that omissions are not significantly more frequent when asking for "utterances" rather than for "sentences" ($\chi^2 = 0.8, p > .3$). Therefore, we used the initially collected data set, which was twice as large, for analysis.

3.3. Production Data Preprocessing

Our preprocessing procedure had two goals. First, we standardized lexical items in order to facilitate unambiguous and homogeneous ellipsis resolution and to facilitate surprisal estimation, and second, we adapted our data to requirements of the statistical analysis with logistic regressions. **Figure 2** provides an overview of the procedure for the fragment utterance in (3). The main steps, which we describe in detail in what follows, consisted in annotating information conveyed by function words and removing these, lemmatizing the remaining words, pooling synonyms to a single lemma, removing optional words, and finally manually resolving pronouns and ellipses.

(3) Schnell die Nudeln in den Topf!
fast the pasta into the.ACC pot
'The pasta into the pot, quickly!'

We first annotated the information conveyed by prepositions and articles as tags like ACC for accusative case on the corresponding nouns and subsequently removed the function words from the data set. This step accounts for the assumption that UID explains only *grammatical* variation. Since the omission of articles is ungrammatical in standard German (Reich, 2017) and that of prepositions from PPs highly degraded (Merchant et al., 2013; Lemke, 2017), their omission appears to be blocked in German for reasons which are independent from UID. Otherwise, our logistic regression analysis, which predicts the omission of a word from information-theoretic measures, might predict that the omission of a particular article or preposition is preferred even though it is ungrammatical. For prepositions we annotated the preposition as a tag on the noun, whereas for articles we annotated only distinctive case marking encoded on the article.[7] Annotating prepositions and case on the noun ensures that the complete phrase is treated as a single unit in the regression analysis and that the information conveyed by the removed word is preserved. Since it is an important cue toward the omitted

[7] For instance masculine singular DPs distinguish four cases in German, whereas feminine singular and plural DPs have a partially syncretic paradgim (i).

(i) a. der Mann / des Mannes / dem Mann / den Mann
the.NOM man / the.GEN man / the.DAT man / the.ACC man
b. die Frau / der Frau / der Frau / die Frau
the.NOM woman / the.GEN woman / the.DAT woman / the.ACC woman
c. die Leute / der Leute / den Leuten / die Leute
the.NOM people / the.GEN people / the.DAT people / the.ACC people

material, for instance, encountering an accusative DP reduces the possibility of encountering another one within the same sentence (Levy, 2008).

The next step consisted in pooling synonym content words, i.e., nouns and verbs, to a single lemma. By synonyms, we understand words that refer to the same object in a single scenario: For instance, verbs like *schütten* 'to pour' and *reintun* 'to put inside' are no synonyms in general, but we pooled them in the pasta scenario when they were used to refer to the action of pouring something into the pot. The same holds for *the water* and *the pot* in utterances asking the hearer to put something inside the pot with the boiling water: In this context it is impossible to put something into the pot without putting it into the water and vice versa. In contrast, we did not merge categorially different order items like *hamburger*, *nachos* or *fries* in ordering scenarios, because they are different items at the moment in which the utterance occurs and their usage results in different outcomes of the situation. We used the most frequently occurring lexicalization in the data set as the label for an object or action in the script. Since we estimated surprisal for the data of each scenario separately, possible duplicate labels in the data for different scripts have no effect on the surprisal estimates.

Merging synonyms to a single label is necessary for two reasons. First, it facilitates the resolution of omissions: If an omitted verb in a fragment like (2) can be resolved either as *schütten* 'to pour' or *reintun* 'to put inside', the decision between either of these verbs would be arbitrary, but if there is only one option after pooling, resolution becomes unambiguous. Second, we use the pre-processed structures for surprisal estimation, and the presence of various synonym lemmas in the data would split the total probability mass of e.g., an action to occur among these alternatives.[8] A further advantage of pooling is that it reduces the lexicon size in the data for each scenario and thus allows to estimate word probabilities more accurately.

We then removed all optional words from our data set, specifically adjectives, adverbials and adverbs, but also modal verbs and particles. This ensures that our data set contains only those expressions, whose absence indicates that they have been omitted. Since the data set must contain both omissions and the corresponding complete counterparts, including e.g., locative adjuncts in our analysis would imply that locative adjuncts have been omitted in all sentences that do not contain such. However, leaving predictable adjuncts implicit is not an omission that results in fragments, and hence not the type of omission that we are concerned with in this article.[9]

Finally, we resolved the reference of pronouns and reconstructed ellipses in our data. Resolving ellipsis is a prerequisite for modeling whether the words that UID predicts to be omitted are really more often omitted in the production data. We added those expressions that are minimally required in a full sentence, i.e., missing verbs and their arguments. Since we inserted the corresponding labels after pooling synonyms, ellipsis resolution was straightforward. For instance, in the case of a fragment like *The pasta into the pot!*, after pooling there is only a single verb *pour* that can be inserted to enrich the fragment to a full sentence. In what follows we refer to the annotated data set resulting from this procedure, that contains both words that were actually produced and those words that were omitted and reconstructed as the *enriched* data set. Based on this corpus, our regression analyses test for each word within this data set whether our information-theoretic predictors significantly determine its omission in the original data.

3.4. Surprisal Estimation

We investigate effects of three measures of surprisal: (i) *unigram surprisal*, (ii) *context-dependent surprisal* that takes into account preceding linguistic material within the utterance and (iii) *surprisal reduction*, which quantifies how much inserting a word reduces the surprisal of the following word. In our data set, unigram surprisal models the likelihood of a word to appear given a particular extralinguistic context, since we estimated it individually for each script. Our measure of context-dependent surprisal is similar to the approach to surprisal by Hale (2001), but it is robust with respect to the circularity issue that results from estimating surprisal on elliptical data. We use these first two measures to investigate whether, as our account predicts, predictable words are more often omitted. Our third predictor, surprisal reduction, allows us to investigate the second prediction of UID, i.e., whether words are more likely to be realized when they reduce the surprisal of following material. Previous research on function words (e.g., Levy and Jaeger, 2007; Jaeger, 2010; Lemke et al., 2017) addressed this question by estimating the surprisal of the word following the target word on a modified corpus, from which all instances of the target word, e.g., relative pronouns or articles, had been omitted. This approach is not applicable to fragments though, because in principle all parts of speech can be omitted in fragments. Therefore, we developed a measure of surprisal reduction that quantifies to what extent inserting or omitting a target word w_i before its successor w_{i+1} reduces the surprisal of w_{i+1} in a particular context.

3.4.1. Unigram Surprisal

We estimate the *unigram surprisal* of each word in the preprocessed data with unigram language models with Good-Turing discount on the preprocessed data that we trained using the SRILM toolkit (Stolcke, 2002). We trained an individual language model on the data for each script separately, because this allows us to interpret surprisal as conditioned on the script-based situation, i.e., on the extralinguistic context (4). For instance, it will show how likely a word like `pasta` is at a particular position in an utterance produced given the pasta script, without taking potentially preceding words into account.

(4) $$S(w_i) = -\log_2 p(w_i \mid context_{extraling.}).$$

[8] An anonymous reviewer pointed out that much of the previous research has focused on word probabilities, so that the surprisal difference between frequent and infrequent synonyms would affect the predictions of the theory. This is in principle correct, but we decided to pool synonyms because otherwise the unambiguous resolution of omissions would be impossible.

[9] In principle, it would be interesting to look into the implicit or explicit realization of adjuncts from a UID perspective. We would expect that highly predictable adverbials are more often omitted, too.

3.4.2. Context-Dependent Surprisal

We use a novel method to estimate *context-dependent surprisal*, which takes into account preceding words in addition to extralinguistic context. In previous research, effects of linguistic context on surprisal were often measured with bigram or higher order *n*-gram models, which return a word's likelihood given the previous $n - 1$ words. Currently there are more advanced language modeling techniques that take into account larger parts of linguistic context (Iyer and Ostendorf, 1996; Oualil et al., 2016a,b; Singh et al., 2016; Grave et al., 2017; Khandelwal et al., 2018; Devlin et al., 2019). However, training even those advanced models on corpus data brings along a circularity issue observed by Levy and Jaeger (2007, p. 852): If predictable words are omitted more often than unpredictable ones, their corpus frequency is not proportional to their predictability. This problem could be addressed by training the model on the enriched data set, i.e., after ellipsis resolution, but this option results in further issues. For instance, consider the case of the fragment (5-a), which is derived from the sentence (5-b) by omitting the NP `pasta`. An *n*-gram model trained on the complete sentence would estimate the surprisal of `pot.GOAL` as *p*(*pot*.GOAL | *pour pasta*), but this is psychologically implausible: Since the word `pasta` is not included in the actual linguistic context, it cannot affect the likelihood of `pot.GOAL`.

(5) a. `pour pot.GOAL` Fragment
b. `pour pasta pot.GOAL` Sentence

Therefore, we estimate context-dependent surprisal (and surprisal reduction, see below) with a method based on the approach by Hale (2001), who derives surprisal from the work done by a fully parallel parser. The parser rejects all parses that are compatible with the input before but not after processing a word, and the processing effort for that word is proportional to the probability mass of the discarded parses: The larger the total probability mass of the rejected parses is, the higher is the surprisal of this word. Hale (2001) calculates the surprisal of a word w_i as the log ratio between the prefix probability α, i.e., the total probability mass of the parses compatible with an input, before and after processing w_i, as shown in equation 1.

$$S(w_i) = \log_2 \frac{\alpha_{\text{i-1}}}{\alpha_{\text{i}}} \tag{1}$$

The application of this approach to a data set requires to know the set of possible parses, i.e., the possible structures in a language, and their respective likelihood. Hale (2001) uses a probabilistic context-free grammar (PCFG) to obtain both the set of possible parses and to calculate their probabilities. He does not discuss fragments, but in principle fragments like `pour pot.GOAL` could be accounted for by the rule in (6-a), whose likelihood can be estimated from a corpus. However, this would raise a circularity issue which is similar to the one discussed above. If speakers often omit objects like *the pasta* in such fragments, the rule corresponding to the complete structure (6-b) will have a lower probability than (6-a) and the NP consequently be assigned a high surprisal in this context, rather than the low one that motivates its frequent omission.

(6) a. S → V PP
b. S → V NP PP

The first main difference between Hale's approach and ours is that rather than using a PCFG to estimate the likelihood of structures, we assume that the set of possible complete structures is equal to the set of complete structures in our enriched production data set, i.e., the pre-processed data set after the reconstruction of omissions. Since this set is finite, it is straightforward to determine each complete structure's probability. The second main difference to Hale's method concerns the question of which complete structures are excluded by an input. Hale (2001) rules out all parses that are not identical to the input up to the currently processed word. For instance, a fragment like (7-a) is compatible with the complete structure in (7-b), but not with (7-c), because it does not start with the word *pour*.

(7) a. The pasta.
b. The pasta is ready.
c. Pour the pasta into the water.

However, the fragment in (7-a) can be derived from both (7-b,c) by omission. Therefore, we do not require the input and the parse to be identical to be included in the set of parses that are compatible with the input, but for the input to be potentially derived by omissions from the parse. More technically, we allow for an arbitrary number of omissions to occur before, between and after all words in the our enriched representations when checking whether the current input is compatible with a particular parse.

In what follows, we illustrate how our approach allows us to estimate the surprisal of omitted and realized words at the case of the fragment `pour pot.GOAL` in (5-a), for which we assume the underlying complete structure in (5-b). For expository purposes, we assume the hypothetical probability distribution over complete structures in (8), but the approach works identically for the actual production data.

(8) a. `pour pasta pot.GOAL` p = 0.75
b. `pour salt pot.GOAL` p = 0.2
c. `set table` p = 0.03
d. `pour onion pan.GOAL` p = 0.02

Given this probability distribution, the surprisal of `pour` at the onset of the utterance can be estimated just like Hale (2001) proposes. Before any input is processed, no parse is excluded, hence the prefix probability $\alpha_{onset} = 1$. Processing the word `pour` rules out (8-c), because it is the only complete structure that does not contain the word `pour`, so that $\alpha_{pour} = 0.97$. The surprisal of `pour` at the utterance onset is then calculated as shown in equation 2.

$$S(pour|onset) = \log_2 \frac{\alpha_{\text{onset}}}{\alpha_{\text{pour}}} = \log_2 \frac{1}{0.97} = 0.04\ bits \tag{2}$$

Similarly, the surprisal of the omitted word `pasta` given `pour` is equivalent to the ratio of the cumulated probability mass of all parses that contain the word `pour`, i.e., (7-a,b,d) and those which contain the word pour followed by pasta, i.e., (8-a). Since $\alpha_{i-1} = 0.97$ and $\alpha_i = 0.75$, the context-dependent surprisal of pasta is calculated as shown in equation 3. Note that this surprisal estimate is not affected by the actual omission of the word pasta, because the prefix probabilities are calculated based on the complete structures in (8) alone. Therefore, it is not affected by the circularity issue discussed above and can be used as a predictor of omission in our statistical analysis.

$$S(pasta|pour) = \log_2 \frac{\alpha_{\text{pour}}}{\alpha_{\text{pasta}}} = \log_2 \frac{0.97}{0.75} = 0.37 \; bits \quad (3)$$

In order to calculate the surprisal of pot.GOAL in (5-a), we compare the probability mass of all parses that contain pour, i.e., (8-a,b,d) and the probability mass of the parses that contain pot.GOAL somewhere after pour (8-a,d). Since pasta has been omitted, the current input pour pot.GOAL can be derived from both of the complete structures in (8-a,d) by omission. Again, the surprisal of pot.GOAL is calculated by applying Hale's formula, as shown in equation (9). In this case the surprisal estimate is affected by the omission of the word pasta that could precede the target word pot.GOAL. This is desirable, because it would not be psychologically realistic to assume that a hearer who processes the reduced utterance relies on words that have been omitted to estimate the surprisal of following ones.

$$(9) \qquad S(pot.\text{GOAL}|pour) = \log_2 \frac{\alpha_{\text{pour}}}{\alpha_{\text{pot.GOAL}}} = \log_2 \frac{0.97}{0.77} = 0.33 \; bits$$

Taken together, our approach avoids the circularity issue caused by omissions of frequent words in the data used for surprisal estimation because it relies on nonelliptical representations for calculating the prefix probabilities. It is also psychologically realistic because it quantifies the work done by the parser incrementally and omitted words in the context of a target word cannot affect the target word's probability.[10]

3.4.3. Surprisal Reduction

Our last measure is *surprisal reduction*, which quantifies how much inserting w_i reduces the surprisal of w_{i+1}. Whereas, context-dependent surprisal quantifies the processing effort of a w_i itself and thus allows us to investigate whether predictable words are more often omitted, surprisal reduction can show us whether the degree to which inserting a word w_i reduces the surprisal of the following word w_{i+1} also constrains the likelihood of the omission of w_i. Some of the previous studies investigating UID effects on reduction (e.g Levy and Jaeger, 2007; Frank and Jaeger, 2008) used the *n*-gram surprisal of w_{i+1} to investigate this prediction, but in case of our study this was not reasonable: UID does not predict arbitrary insertions before unpredictable words, but that insertions are only useful when they reduce the surprisal of unpredictable words.[11] For this purpose, we calculate the ratio between the prefix probability at w_{i+1} if w_i has been realized and the prefix probability at w_{i+1} if w_i has been omitted (10).

$$(10) \qquad S\,reduction(w_i, w_{i+1}) = \log_2 \frac{\alpha_{i+1}}{\alpha_{i,i+1}}$$

Again, we illustrate this idea at the simplified pasta script by quantifying how much inserting pasta before pot.GOAL in a fragment pour pot.GOAL reduces the surprisal of pot.GOAL as compared to the omission of pasta. In this case, the probability mass of all parses that contain the words put and pot.GOAL in this order, with potentially intervening material (i.e., 7a,b), is compared to the ratio of those parses that additionally contain pasta between these words (8-a). Since $\alpha_i = 0.95$ and $\alpha_{i,i+1} = 0.75$, inserting pasta reduces the surprisal of pot.GOAL by $\log_2(0.95/0.75) = 0.34$ bits.

3.5. Results

3.5.1. Data Set Statistics

The preprocessed data set comprises a total of 2.409 sentences consisting of 6.816 primitive expressions ("words"). 1.052 (15.43%) of these words had been omitted in the original data set. As **Figure 3** shows, scripts differ to a large extent with respect to the ratio of words that were omitted. For instance, in the train script 62.3% of the words were omitted, whereas there are no omissions at all in the cooking scrambled eggs script.

The low ratio of omission in some scripts raises the question of whether this variation occurs due to properties of the situation which might override the predictions of our information-theoretic account, or whether our account predicts such variation. For instance, sentences might be perceived as more polite than fragments, so that in situations where politeness matters there might be a preference for full sentences which is the result of information-theoretic considerations. In contrast, the responses collect for a script might differ between scripts in their degree of variation. If there are only few different words in the data for a scenario, and/or the probability distribution over these words is skewed, i.e., some words are much more likely than others, an average word in that data will be more predictable. Since we expect that a word's probability predicts the likelihood of its omission, a varying omission rate between scenarios could result from different probability distributions over words.

We test this hypothesis by investigating whether the ratio of omission is higher in scripts with a higher degree of variation between words. For this purpose, we estimate the entropy in the probability distribution over words in the enriched data set for each script after preprocessing and ellipsis resolution. Following (Shannon, 1948), the entropy, which quantifies the degree of uncertainty about the outcome of a random variable, is defined

[10] Note that our approach is not technically identical to the definition of surprisal in the literature, because the surprisal values assigned to the words that follow w_{i-1} do not necessarily sum up to 1. The reason for this is that individual parses can contribute to the probability mass of not only a single word w_i, but also to that of a word w_j, which appears after w_i in the parse, provided that w_i has been omitted. Even though our approach loses this mathematical property of the original definition of surprisal in the literature, the probability estimate that we propose is in line with the insight by Hale (2001) that processing effort is proportional to the probability mass of the parses that are compatible with an input.

[11] This concern does not apply to the studies cited here. Levy and Jaeger (2007) looked into effects of additional processing effort due to syntactic surprisal and Frank and Jaeger (2008) investigated contractions, so their study is not affected by the issues concerning omissions.

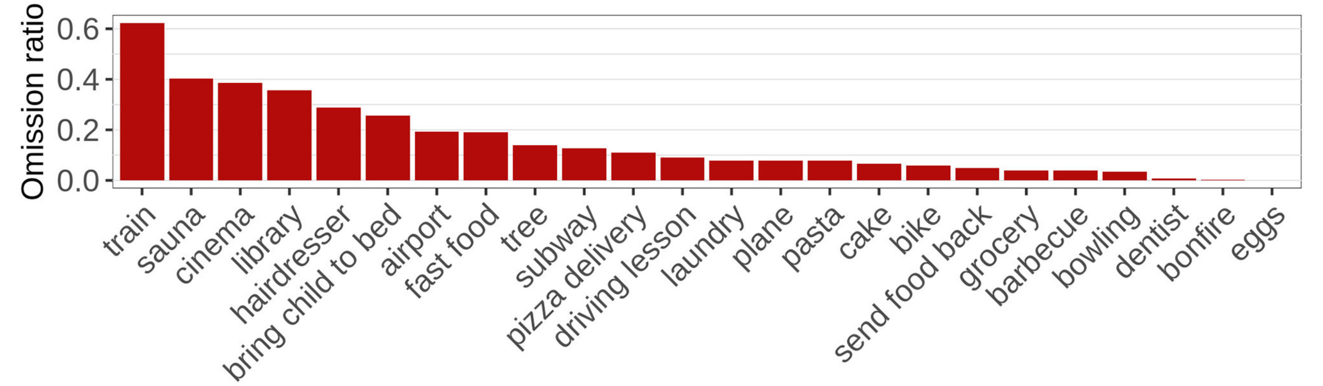


FIGURE 3 | Ratio of omission across scripts.

as shown in equation 4. It equals 0 when there is no variation in the data, i.e., when there is only one possible word, which has a probability of 1, and it is maximal when all words in the data are equally likely. Furthermore, it increases as the number of different words in the data grows. **Figure 4** suggests that the entropy in the data for a script is indeed related to the rate of omissions: The omission rate seems to be higher in scripts with a low entropy. This is confirmed by a linear regression (R Core Team, 2019) which shows that entropy has a significant effect on the ratio of omission ($F_{(1)} = 12.49, p < 0.01$).[12]

$$H = -K \sum_{i=1}^{n} p_i \log p_i \qquad (4)$$

3.5.2. Statistical Methodology

We analyzed the data with mixed effects logistic regressions (`lme4`, Bates et al., 2015) in R (R Core Team, 2019). Our regressions predict the actual omission of the words in the enriched data set from the three surprisal measures that we introduced in section 3.4. We investigate the effects of these predictors individually with three separate analyses. Even though it would have been desirable to test for effects of these predictions in a single analysis in order to tease apart effects of linguistic and extralinguistic context on predictability, **Table 1** shows that the measures are correlated with each other, and context-dependent surprisal is particularly strongly correlated with the other two measures.

Therefore, we first conduct two analyses that test for effects of unigram and context-dependent surprisal on the complete data set. In a third analysis we take into account unigram surprisal and surprisal reduction. This last analysis investigates only non-final words, since the last word in an utterance lacks a successor, whose predictability its insertion or omission could affect. In all analyses we conducted model comparisons with likelihood ratio tests computed with the `anova` function in R and maintained only those effects in the model that significantly improved model fit.

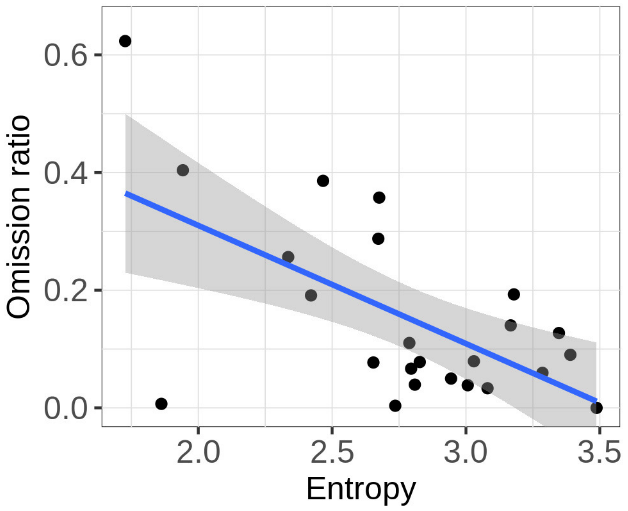


FIGURE 4 | Ratio of omission across scripts as a function of the entropy in the probability distribution over words in the data for that script. Each data point represents one script.

TABLE 1 | Correlations between surprisal predictors.

Predictors	r^2	*t*-value	*p*-value
Unigram, context	0.65	70.06	< 0.001
Unigram, reduction	0.48	37.99	< 0.001
Context, reduction	0.62	54.0	< 0.001

3.5.3. Avoid Troughs: Unigram Surprisal and Context-Dependent Surprisal

Figure 5 shows how the omitted and the realized words are distributed across the range of unigram surprisal (left facet) and context-dependent surprisal (right facet). For both predictors,

[12] We also tested for a potential effect of raw lexicon size, which also predicts the rate of omission ($F_{(1)} = 6.18, p < 0.05$). However, if entropy is included in the model, the effect of lexicon size is no longer significant ($F_{(1)} = 6.18, p < 0.05$).

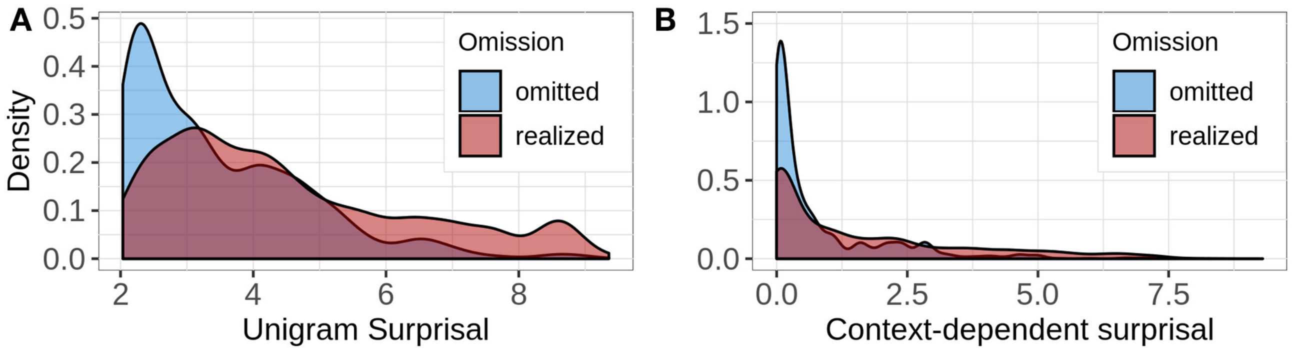


FIGURE 5 | The density plots illustrate the distribution of words that were originally omitted and those originally realized across the surprisal ranges. The left facet **(A)** shows the distribution for unigram surprisal and the right facet **(B)** shows the distribution for context-dependent surprisal.

TABLE 2 | Fixed effects in the final GLMM investigating the effect of UnigramSurprisal on Omission.

Predictor	Estimate	SE	χ^2	*p*-value
UnigramSurprisal	–0.337	0.117	7.39	< 0.01

TABLE 3 | Fixed effects in the final GLMM investigating the effect of ContextSurprisal on Omission.

Predictor	Estimate	SE	χ^2	*p*-value
ContextSurprisal	–0.28	0.126	4.86	< 0.05

TABLE 4 | Fixed effects in the final GLMM investigating effects of both UnigramSurprisal and SurprisalReduction.

Predictor	Estimate	SE	χ^2	*p*-value
UnigramSurprisal	–0.151	0.046	10.39	< 0.01
SurprisalReduction	–0.349	0.07	27.03	< 0.001

the originally omitted words appear toward the lower end of the scale, whereas the realized words seem to have a higher surprisal on average. The distribution for context-dependent surprisal is highly skewed, because in our highly standardized data set sometimes a single word fully disambiguates between two utterances and consequently, all words that appear later in the utterance have a surprisal of 0. For instance, in case of our simplified example in (8) above, put salt pot.GOAL is the only utterance that contains the word salt, so that pot.GOAL has a probability of 1 and a surprisal of 0 in this context.

The models in the analyses of unigram surprisal[13] and context-dependent surprisal[14] contained by-script random intercepts and slopes for surprisal and by-subject random intercepts. **Tables 2**, **3** summarize the final models for the analyses. Both of the analyses reveal significant main effects of the respective predictor, which support our hypothesis that predictable words are more likely to be omitted. Unlike we expected though, the effect for unigram surprisal ($\chi^2 = 7.39, p < 0.01$) is stronger than that of context-dependent surprisal ($\chi^2 = 4.86, p < 0.05$). In principle we would expect the opposite pattern, since previous research on information-theoretic constraints on omissions has found robust predictability effects driven by linguistic context. In the case of our data, the relatively large number of words that are assigned a context-dependent surprisal of 0 and that were nevertheless realized might account for this pattern. Even though these words are fully predictable in our enriched data set, they are not necessarily equally predictable to an actual speaker, for instance because of different lexicalizations which we merged during preprocessing. Furthermore, our data set contains only those utterances that appeared to be the most likely ones by at least one of our participants, but not utterances that everybody considers to be relatively unlikely. An actual speaker however must reserve some probability mass to those utterances as well, so she would not assign as many words a surprisal of 0 as our account does, and consequently choose not to omit some of these words. Therefore, we expect to find stronger effects of context-dependent surprisal in case of larger and more diverse data sets.

3.5.4. Avoid Peaks: Surprisal Reduction

The analysis that includes surprisal reduction and unigram surprisal was conducted on a subset of the data that contained those non-final words that were not followed by an ellipsis (55.51% of the total data). Only these words have a successor which is not omitted and whose surprisal might be constrained by the omission or realization of the preceding word. The full model contained main effects of both predictors as well as their interaction, as well as random intercepts for subjects and scripts.[15] The final model contains only the main effects, both of which support our predictions (see **Table 4**). The effect of unigram surprisal ($\chi^2 = 10.39, p < 0.01$) replicates the effect found in the analysis of the full data set. The effect of surprisal

[13] Ellipsis ∼ UnigramS + (1+UnigramS | Script) + (1 | Subject).

[14] Ellipsis ∼ ContextS + (1+ContextS | Script) + (1 | Subject).

[15] Ellipsis ∼ UnigramS * SReduction (1 | Script) + (1 | Subject).

reduction ($\chi^2 = 27.03, p < 0.001$) shows that, as we expected, words that reduce the surprisal of the following word more strongly are more likely to be realized. There is no significant interaction between both predictors ($\chi^2 = 0.01,\ p > 0.9$).

4. DISCUSSION

In this article we propose an information-theoretic account as an answer to the previously underexplored question of why speakers use fragments, when they prefer a fragment over the corresponding complete sentence, and if they do so, which fragment is ultimately selected. The empirical predictions of our account are supported by the results of a production task: First, speakers tend toward omitting words that are predictable in context in order to make a more efficient use of the hearer's cognitive resources. Second, speakers tend toward inserting words that could be omitted but that increase the predictability of the following word. This reduces peaks in the information density profile of the utterance, which would be otherwise likely to exceed the resources available to the hearer.

Our study provides the first systematic investigation of why speakers use fragments. Previous research on fragments investigated their syntactic properties and licensing conditions, and pursued almost exclusively theoretical approaches. As we observed in the introduction, information structure-based syntactic accounts of fragments (Merchant, 2004; Reich, 2007; Weir, 2014; Ott and Struckmeier, 2016; Griffiths, 2019) explain under which circumstances fragments can be used, but not why speakers choose to produce a fragment or a complete sentence. Our information-theoretic account provides a potential solution to this issue: Speakers prefer to use fragments when the omission of words that are obligatory in full sentences (like finite verbs and their arguments), which results in fragments, optimizes the form of the utterance with respect to the processing resources which are available to the hearer.

Our results are partially in line with other theories of probability-driven reduction of linguistic expressions, like the availability-based production theory (Ferreira and Dell, 2000) or a source coding account (Zipf, 1935; Pate and Goldwater, 2015). Even though not all of these theories have been applied to fragments, they make predictions on the distribution of omissions that can result in fragments. However, none of these theories covers the complete empirical picture, that is, the preferences to omit predictable words and to insert redundancy when this increases the likelihood of following unpredictable words.

Availability-based production (e.g., Bock, 1987; Ferreira and Dell, 2000) explains some predictability effects with speaker-centered production difficulties: Retrieving words from memory is effortful, and the retrieval of unpredictable words requires more effort, i.e., time. Since speakers intend to avoid disfluencies which result from the effortful retrieval of unpredictable words, they insert optional words before unpredictable material in order to keep speech fluent. Therefore, availability-based production predicts the insertion of optional words before unpredictable words, but not that predictable words are more likely to be omitted.

The opposite holds for a source coding account, which takes into account only properties of the *source* in the model of communication assumed by Shannon (1948), i.e., the frequency of expressions. A system that assigns a unique utterance to each message is more efficient if it assigns longer utterances to rare meanings and reserves the shortest encodings for the most frequent meanings. This predicts more likely utterances to be more often reduced, as Lemke et al. (2021) and we show, but not that speakers *insert* redundancy to *reduce* processing effort.

The main difference between the predictions of the game-theoretic account in Bergen and Goodman (2015) and our UID-based account is that according to game-theoretic accounts there is no upper bound to the densification of utterances, like channel capacity. In practice, if game-theoretic models are applied to larger and more diverse data sets, they might indirectly predict a similar effect though: Fragments like (11-a) can communicate different meanings, and if a particular meaning is more likely in a situation, like (11-b) as compared to (11-c) in our taxi scenario, the game-theoretic approach also predicts that speakers use the fragment to refer to the predictable utterance. Due to the high prior probability of (11-b) given (11-a), the hearer will choose this interpretation of the fragment and the speaker will in turn rather produce a complete sentence if she wishes to communicate the more unlikely message in (11-c). Empirically comparing game-theoretic models of fragment usage to our UID-based account would require a more precise formulation of how to derive the set of alternative utterances and messages to be considered in a situation and how the cost of producing an utterance is derived.

(11) a. To the university, please.
b. Bring me to the university, please.
c. Explain me the way to the university, please.

Our UID-based account predicts both the omission of predictable words and the insertion of additional redundancy before unpredictable words that the analysis of our production data revealed. Other theories that have been proposed in the literature to account for other optional omission phenomena, like availability-based production and source coding, or specifically applied to fragments, like the noisy channel model by Bergen and Goodman (2015), can only explain part of the data, but not the complete empirical pattern.

From the broader perspective of information-theoretic research on the choice between linguistic encodings, our study extends previous evidence for predictability effects on optional omissions in two ways. First, we present evidence for such effects on content words like nouns and verbs, whereas previous work focused on semantically relatively empty closed-class function words. This requires a modified approach to surprisal estimation, since *n*-gram models trained on regular corpus data suffer from a circularity issue and removing all target words from the corpus [following Levy and Jaeger (2007)] is not possible for content words. Second, we find predictability effects based on extralinguistic context on omissions. Most of the previous studies only took local linguistic context into account. Instead, we provide evidence for effects of script knowledge on predictability and consequently omissions.

Even though our study was relatively resource-intensive due to the large amount of manual preprocessing, there are several ways in which it could be extended in future research. First, we observed that effects of linguistic context are not as strong as we expected in our data set, and this might be in part due to the high degree of standardization and the amount of utterances per script, which is relatively small as compared to corpora like those used in previous studies on UID effects on omissions. Future research that relying more strongly on automatized procedures might be able to process larger data sets in a similar way and yield more fine-grained results. Second, our data set is probably a close approximation to hearers' expectations about what will be said in the situations described in our context stories. However, this expectation might differ to some extent from expectations developed by hearers in such situations, because we asked only for a single most likely utterance provided by each participant. There might be overall less likely, and yet salient, utterances, which hearers assign a probability, but which is not reflected in our data. A possible solution for this issue would consist in asking participants to provide a series of utterances and to specify the relative likelihood of each one, be it in terms of absolute probability or by ranks. Third, since we are interested in the usage of fragments, we deliberately preprocessed our data so that it contained only words whose omission would result in a fragment or that are obligatory in standard sentences, like main verbs and their arguments. As we noted above, in principle UID also predicts that likely adjuncts, e.g., temporal or local adverbials, will be omitted when they are predictable. This issue could be empirically investigated with a method that is similar to ours, but it would require an even more extensive preprocessing approach in order to neatly reconstruct all implicit adverbials.

Taken together, our research makes three main contributions: First, we propose an information-theoretic account as an answer to the question of why and when fragments are used. The two central predictions of our account are supported by our production study: Predictable words are more often omitted and additional redundancy is inserted in order to reduce the processing effort of following words. Second, we extend previous evidence for information-theoretic processing constraints on linguistic encoding choices, and third, our methodological approach might be also applied to other omission phenomena.

DATA AVAILABILITY STATEMENT

The original contributions generated for the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of the Deutsche Gesellschaft für Sprachwissenschaft (German Society for Language Science). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

RL conducted the experiment, supervised the annotation of the production data, developed the surprisal estimation methods, conducted the statistical analyses, and wrote the initial version of the article. IR acquired the project funding and conceptualized the overarching research program and goals. LS created software for pre-processing the corpus on which the experimental materials are based. HD and IR managed and supervised the research activities. IR, LS, and HD critically revised previous versions of the article. All authors contributed to the article and approved the submitted version.

FUNDING

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG)-Projektnummer 232722074-SFB 1102/Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)-Project-ID 232722074-CRC 1102.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.662125/full#supplementary-material

REFERENCES

Asr, F. T., and Demberg, V. (2015). "Uniform information density at the level of discourse relations: Negation markers and discourse donnective omission," in *Proceedings of the 11th International Conference on Computational Semantics*, eds M. Purver, M. Sadrzadeh, and M. Stonees (London: Association for Computational Linguistics), 118–128.

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

Barton, E., and Progovac, L. (2005). "Nonsententials in Minimalism," in *Ellipsis and Nonsentential Speech*, eds R. Elugardo and R. J. Stainton (Dordrecht: Springer), 71–93.

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

Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *J. Memory Lang.* 60, 92–111. doi: 10.1016/j.jml.2008.06.003

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *J. Acoust. Soc. Am.* 113, 1001–1024. doi: 10.1121/1.1534836

Bergen, L., and Goodman, N. D. (2015). The strategic use of noise in pragmatic reasoning. *Top. Cogn. Sci.* 7, 336–350. doi: 10.1111/tops.12144

Bicknell, K., Elman, J. L., Hare, M., McRae, K., and Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *J. Mem. Lang.* 63, 489–505. doi: 10.1016/j.jml.2010.08.004

Bock, K. (1987). An effect of the accessibility of word forms on sentence structures. *J. Mem. Lang.* 26, 119–137. doi: 10.1016/0749-596X(87)90120-3

Bower, G. H., Black, J. B., and Turner, T. J. (1979). Scripts in memory for text. *Cogn. Psychol.* 11, 177–220. doi: 10.1016/0010-0285(79)90009-4

Brandt, E., Zimmerer, F., Andreeva, B., and Möbius, B. (2017). "Mel-cepstral distortion of German vowels in different information density contexts," in *Interspeech 2017* (Stockholm: ISCA), 2993–2997.

Brandt, E., Zimmerer, F., Andreeva, B., and Möbius, B. (2018). "Impact of prosodic structure and information density on dynamic formant trajectories in German," in *9th International Conference on Speech Prosody 2018*, (Hyderabad: ISCA), 119–123.

Culicover, P., and Jackendoff, R. (2005). *Simpler Syntax*. New York, NY: Oxford University Press.

Delogu, F., Drenhaus, H., and Crocker, M. W. (2018). On the predictability of event boundaries in discourse: an ERP investigation. *Mem. Cogn.* 46, 315–325. doi: 10.3758/s13421-017-0766-4

Demberg, V., and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109, 193–210. doi: 10.1016/j.cognition.2008.07.008

Demberg, V., Sayeed, A. B., Gorinski, P. J., and Engonopoulos, N. (2012). "Syntactic surprisal affects spoken word duration in conversational contexts," in *Proceedings of EMNLP-CoNNL 2012* Jeyu Island.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [cs]*.

Engonopoulos, N., Sayeed, A., and Demberg, V. (2013). "Language and cognitive load in a dual task environment," in *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, eds M. Knauff, M. Pauen, N. Sebanz and I. Wachsmuth (Berlin: Cognitive Science Society, 2249–2254.

Fenk, A., and Fenk, G. (1980). Konstanz im Kurzzeitgedächtnis–Konstanz im sprachlichen Informationsfluß. *Z. Experiment. Angew. Psychol.* 27, 400–414.

Ferreira, V. S., and Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cogn. Psychol.* 40, 296–340. doi: 10.1006/cogp.1999.0730

Frank, A. F., and Jaeger, T. F. (2008). Speaking rationally: Uniform Information Density as an optimal strategy for language production. *Proc. Ann. Meet. Cogn. Sci. Soc.* 30, 939–944.

Frank, M. C., and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science* 336, 998–998. doi: 10.1126/science.1218633

Franke, M. (2009). *Signal to Act: Game Theory in Pragmatics*. Ph.D. thesis, Universiteit van Amsterdam.

Ginzburg, J., and Sag, I. A. (2000). *Interrogative Investigations: The Form, Meaning, and Use of English Interrogatives*. Stanford, CA: CSLI Publications.

Grave, E., Cisse, M., and Joulin, A. (2017). "Unbounded cache model for online language modeling with open vocabulary," in *Advances in Neural Information Processing Systems 30*, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Long Beach, CA: Curran Associates, Inc.), 6042–6052.

Griffiths, J. (2019). A Q-based approach to clausal ellipsis: deriving the preposition stranding and island sensitivity generalisations without movement. *Glossa* 4:12. doi: 10.5334/gjgl.653

Hale, J. (2001). "A probabilistic Earley parser as a psycholinguistic model," in *Proceedings of NAACL (Vol. 2)* (Pittsburgh, PA), 159–166.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cogn. Sci.* 30, 643–672. doi: 10.1207/s15516709cog0000_64

Häuser, K. I., Demberg, V., and Kray, J. (2019). Effects of aging and dual-task demands on the comprehension of less expected sentence continuations: evidence from pupillometry. *Front. Psychol.* 10:709. doi: 10.3389/fpsyg.2019.00709

Iyer, R., and Ostendorf, M. (1996). "Modeling long distance dependence in language: topic mixtures vs. dynamic cache models," in *Proceedings of Fourth International Conference on Spoken Language Processing. ICSLP '96*, Vol. 1, (Philadelphia, PA: IEEE), 236–239.

Jaeger, T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cogn. Psychol.* 61, 23–62. doi: 10.1016/j.cogpsych.2010.02.002

Khandelwal, U., He, H., Qi, P., and Jurafsky, D. (2018). "Sharp nearby, fuzzy far away: how neural language models use context," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, (Vol. 1, Long Papers)*, (Melbourne, SA: Association for Computational Linguistics), 284–294.

Klein, D., and Manning, C. D. (2003). "Accurate unlexicalized parsing," in *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, (Sapporo: Association for Computational Linguistics), 423–430.

Klein, W. (1993). "Ellipse," in *Syntax. An International Handbook of Contemporary Research*, eds J. Jacobs, A. von Stechow, W. Sternefeld, T. and Venneman (Berlin, NY: de Gruyter), 763–799.

Kravtchenko, E. (2014). Predictability and syntactic production: Evidence from subject omission Russian. *Proc. Annu. Meet. Cogn. Sci. Soc.* Quebec City, QC. 36, 785–790. doi: 10.1515/9783110095869.1.12.763

Kuperman, V., and Bresnan, J. (2012). The effects of construction probability on word durations during spontaneous incremental sentence production. *J. Mem. Lang.* 66, 588–611. doi: 10.1016/j.jml.2012.04.003

Kurumada, C., and Jaeger, T. F. (2015). Communicative efficiency in language production: optional case-marking in Japanese. *J. Mem. Lang.* 83, 152–178. doi: 10.1016/j.jml.2015.03.003

Lemke, R. (2017). "Sentential or not?–An experimental study on the syntax of fragments," in *Proceedings of Linguistic Evidence 2016*, eds S. Featherston, R. Hörnig, R. Steinberg, B. Umbreit, and J. Walli (Tübingen: University of Tübingen, online publication system).

Lemke, R., Horch, E., and Reich, I. (2017). "Optimal encoding!–Information Theory constrains article omission in newspaper headlines," in *Proceedings of the 15th Conference of the {E}uropean Chapter of the Association for Computational Linguistics: Vol. 2, Short Papers* (Valencia), 131–135.

Lemke, R., Schäfer, L., and Reich, I. (2021). Modeling the predictive potential of extralinguistic context with script knowledege: the case of fragments. *PLoS ONE* 16:e0246255. doi: 10.1371/journal.pone.0246255

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

Levy, R. P., and Jaeger, T. F. (2007). "Speakers optimize information density through syntactic reduction," in *Advances in Neural Information Processing*, Vol. 19, eds B. Schölkopf, J. Platt, and T. Hoffman (Cambridge, MA: MIT Press), 849–856.

LimeSurvey GmbH (2012). *LimeSurvey*: An Open Source Survey Tool Hamburg.

Loper, E., and Bird, S. (2002). "NLTK: the Natural Language Toolkit," in *ETMTNLP '02 Proceedings of the ACL-02 Workshop on Effective tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Vol. 1 (Philadelphia, PA: Association for Computational Linguistics), 63–70.

Malisz, Z., Brandt, E., Möbius, B., Oh, Y. M., and Andreeva, B. (2018). Dimensions of segmental variability: interaction of prosody and surprisal in six languages. *Front. Commun.* 3:25. doi: 10.3389/fcomm.2018.00025

Manshadi, M., Swanson, R., and Gordon, A. S. (2008). "Learning a probabilistic model of event sequences from internet weblog stories," in *Proceedings of the Twenty-First International FLAIRS Conference* Coconut Grove, FL.

McKoon, G., and Ratcliff, R. (1986). Inferences about predictable events. *J. Exp. Psychol. Learn. Mem. Cogn.* 12, 82–91. doi: 10.1037/0278-7393.12.1.82

Merchant, J. (2004). Fragments and ellipsis. *Linguist. Philos.* 27, 661–738. doi: 10.1007/s10988-005-7378-3

Merchant, J., Frazier, L., Weskott, T., and Clifton, C. (2013). "Fragment answers to questions. a case of inaudible syntax," in *Brevity*, ed L. Goldstein (Oxford: Oxford University Press), 21–35.

Millis, K., Morgan, D., and Graesser, A. (1990). The influence of knowledge-based inferences on the reading time of expository text. *Psychol. Learn. Motiv.* 25, 197–212. doi: 10.1016/S0079-7421(08)60256-X

Morgan, J. (1973). "Sentence fragments and the notion 'sentence'," in *Issues in Linguistics. Papers in Honor of Henry and Renée Kahane*, eds B. B. Kachru, R. Lees, Y. Malkiel, A. Pietrangeli, and S. Saporta (Urbana, IL: University of Illionois Press), 719–751.

Norcliffe, E., and Jaeger, T. F. (2016). Predicting head-marking variability in Yucatec Maya relative clause production. *Lang. Cogn.* 8, 167–205. doi: 10.1017/langcog.2014.39

Nuthmann, A., and Van Der Meer, E. (2005). Time's arrow and pupillary response. *Psychophysiology* 42, 306–317.

Ott, D., and Struckmeier, V. (2016). "Deletion in clausal ellipsis: remnants in the middle field," in *UPenn Working Papers in Linguistics* 22.

Oualil, Y., Greenberg, C., Singh, M., and Klakow, D. (2016a). "Sequential recurrent neural networks for language modeling," in *Interspeech 2016* (Austin, TX), 3509–3513.

Oualil, Y., Singh, M., Greenberg, C., and Klakow, D. (2016b). "Long-short range context neural network for language models," in *EMLP 2016*, (Austin, TX), 1473–1481.

Pate, J. K., and Goldwater, S. (2015). Talkers account for listener and channel characteristics to communicate efficiently. *J. Mem. Lang.* 78, 1–17. doi: 10.1016/j.jml.2014.10.003

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna.

Reich, I. (2007). "Toward a uniform analysis of short answers and gapping," in *On Information Structure, Meaning and Form*, eds K. Schwabe and D. Winkler (Amsterdam: John Benjamins), 467–484.

Reich, I. (2011). "Ellipsis," in *Semantics: An International Handbook of Natural Language Meaning*, eds K. von Heusinger, C. Maienborn, and P. Portner (Berlin, New Yor: Mouton de Gruyter), 1849–1874.

Reich, I. (2017). On the omission of articles and copulae in German newspaper headlines. *Linguist. Variat.* 17, 186–204. doi: 10.1075/lv.14017.rei

Rooth, M. (1992). A theory of focus interpretation. *Nat. Lang. Semant.* 1, 75–116.

Schäfer, L. (2021). Topic drop in German: Empirical support for an information-theoretic account to a long-known omission phenomenon. *Zeitschrift für Sprachwissenschaft*. doi: 10.1515/zfs-2021-2024. [Epub ahead of print].

Schank, R., and Abelson, R. (1977). *Scripts, Plans, Goals, and Understanding: An Enquiry Into Human Knowledge Structures.* Hillsdale, NJ: Erlbaum.

Schwarzschild, R. (1999). Givenness, AvoidF and other constraints on the placement of accent. *Nat. Lang. Semant.* 7, 141–177. doi: 10.1023/A:1008370902407

Seyfarth, S. (2014). Word informativity influences acoustic duration: effects of contextual predictability on lexical representation. *Cognition* 133, 140–155. doi: 10.1016/j.cognition.2014.06.013

Shannon, C. (1948). A mathematical theory of communications. *Bell Syst. Techn. J.* 27, 623–656. doi: 10.1002/j.1538-7305.1948.tb00917.x

Singh, M., Greenberg, C., and Klakow, D. (2016). "The custom decay language model for long range dependencies," in *Text, Speech, and Dialogue*, eds P. Sojka, A. Horák, I. Kopeček, and K. Pala (Cham: Springer), 343–351.

Stainton, R. J. (2006). "4. Neither fragments nor ellipsis," in *Linguistik Aktuell/Linguistics Today*, Vol. 93, eds L. Progovac, K. Paesani, E. Casielles, and E. Barton (Amsterdam: John Benjamins Publishing Company), 93–116.

Stolcke, A. (2002). "SRILM–an extensible language modeling toolkit," in *Proceedings International Conference Spoken Language Processing*, (Denver, CO).

Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., and Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Lang. Cogn.* 1, 147–165. doi: 10.1515/LANGCOG.2009.008

Tily, H., and Piantadosi, S. (2009). "Refer efficiently: use less informative expressions for more predictable meanings," in *Proceedings of the Workshop on the Production of Referring Expressions: Bridging the Gap Between Computational and Empirical Approaches to Reference*, (Amsterdam).

van den Broek, P. (1994). "Comprehension and memory of narrative texts," in *Handbook of Psycholinguistics*, ed M. A. Gernsbacher (San Diego, CA: Academic Press), 539–588.

van der Meer, E., Beyer, R., Heinze, B., and Badel, I. (2002). Temporal order relations in language comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 770–779. doi: 10.1037/0278-7393.28.4.770

Wanzare, L. D. A., Zarcone, A., Thater, S., and Pinkal, M. (2016). "DeScript: a crowdsourced corpus for the acquisition of high-quality script knowledge," in *Proceedings of LREC 2016*, (Portoroz), 3494–3501.

Weir, A. (2014). *Fragments and Clausal Ellipsis*. Ph.D. thesis, University of Massachusetts Amherst.

Zipf, G. K. (1935). *The Psycho-Biology of Language*. Houghton, MI; Mifflin; Oxford.

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 Psychology

ORIGINAL RESEARCH
published: 12 August 2021
doi: 10.3389/fpsyg.2021.652664

Hierarchical Inference in Sound Change: Words, Sounds, and Frequency of Use

*Vsevolod Kapatsinski**

Department of Linguistics, University of Oregon, Eugene, OR, United States

This paper aims examines the role of hierarchical inference in sound change. Through hierarchical inference, a language learner can distribute credit for a pronunciation between the intended phone and the larger units in which it is embedded, such as triphones, morphemes, words and larger syntactic constructions and collocations. In this way, hierarchical inference resolves the longstanding debate about the unit of sound change: it is not necessary for change to affect only sounds, or only words. Instead, both can be assigned their proper amount of credit for a particular pronunciation of a phone. Hierarchical inference is shown to generate novel predictions for the emergence of stable variation. Under standard assumptions about linguistic generalization, it also generates a counterintuitive prediction of a U-shaped frequency effect in an advanced articulatorily-motivated sound change. Once the change has progressed far enough for the phone to become associated with the reduced pronunciation, novel words will be more reduced than existing words that, for any reason, have become associated with the unreduced variant. Avoiding this prediction requires learners to not consider novel words to be representative of the experienced lexicon. Instead, learners should generalize to novel words from other words that are likely to exhibit similar behavior: rare words, and the words that occur in similar contexts. Directions for future work are outlined.

Keywords: hierarchical inference, sound change, lexical diffusion, frequency effects, usage-based phonology

OPEN ACCESS

Edited by:
Rory Turnbull,
Newcastle University, United Kingdom

Reviewed by:
Ingo Plag,
Heinrich Heine University of Düsseldorf, Germany
Adam Chong,
Queen Mary University of London, United Kingdom

****Correspondence:***
Vsevolod Kapatsinski
vkapatsi@uoregon.edu

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

Received: *12 January 2021*
Accepted: *08 July 2021*
Published: *12 August 2021*

Citation:
Kapatsinski V (2021) Hierarchical Inference in Sound Change: Words, Sounds, and Frequency of Use. Front. Psychol. 12:652664. doi: 10.3389/fpsyg.2021.652664

INTRODUCTION

Research on sound change has been characterized by a tension between the fact that changes affect specific sounds in phonological contexts, and the fact that changes progress faster in some words and expressions than in others. For example, a final post-consonantal /t/ is likely to be deleted in American English, compared to other comparable sounds like /k/ or /p/. At the same time, this deletion is more likely in a frequent word like *most* than in an infrequent word like *mast* (Bybee, 2002). These facts appear to be in conflict because approaches to sound change tend to assume that there is a particular unit of change, which is either the sound – in approaches growing out of the Neogrammarian tradition (Osthoff and Brugmann, 1878; Labov, 1981) – or the word, in the dialectological / lexical diffusion tradition where every word has its own history (Schuchardt, 1885; Mowrey and Pagliuca, 1995).

For example, generative grammatical theory (Chomsky and Halle, 1965), and allied approaches in psycholinguistics (Levelt, 1989; Levelt et al., 1999) have suggested that the long-term representations of words are composed of a small set of discrete segments (whether phones, features or syllables). In this architecture, words are not directly associated with specific pronunciations, and therefore the pronunciation of a segment is not lexically specific. As a result, only two types of sound change are possible – a phonetically abrupt deletion, insertion or substitution of a segment in the lexical representation of a particular word, or a continuous drift in the pronunciation of a particular segment that happens across all instances of the segment in a particular phonological environment, no matter what word it is embedded in Labov (1981). This theory has difficulty explaining how words can influence the pronunciation of a segment in a gradient manner (Bybee, 2002). For example, the durations of frequent words are shorter than the durations of homophonous infrequent words (Gahl, 2008). On the opposite end of the spectrum is Mowrey and Pagliuca's (1995) proposal that words are holistic motor programs specifying the timing and intensity of nerve impulses to muscles controlling articulator movement. This approach allows for each word to have its own history, and for lexical representations to change continuously rather than in discrete jumps (Mowrey and Pagliuca, 1995; Bybee, 2001, 2002). However, it has the converse problem of being unable to explain why a word's pronunciation does not change uniformly, i.e., why certain sounds are affected more than others.

Pierrehumbert (2002) unifies the segmental and lexical views of sound change by suggesting that the language system maintains representations of segmental categories, which are implemented as sets of exemplars, but that each exemplar of a segment is tagged with the word in which it occurred. In production, the selection of a segment exemplar is then driven both by the identity of the segment and the identity of the word: both are tags available to cue an exemplar in production. A related idea is the approach to reduction proposed by Browman and Goldstein (1989) within Articulatory Phonology, where gestures are units of change but the timing and magnitude of a gesture can be lexically specific.

The present paper combines this idea with rational probabilistic inference (Xu and Tenenbaum, 2007; Feldman et al., 2009; Perfors et al., 2011; Kleinschmidt and Jaeger, 2015; O'Donnell, 2015; Harmon et al., 2021). If both the identity of a segment and the identity of the word that contains it influence the pronunciation of a segment in a lexical context, then a rational language learner would use hierarchical inference to allocate credit for a particular pronunciation between the two influencers. This paper explores the consequences of this assumption for articulatorily-motivated sound change.

I focus on articulatorily-motivated changes because the role of inference in such changes has been underexplored. In the other major type of change, analogical change, a role for inference is relatively uncontroversial (e.g., Bybee, 2001). In analogical changes, words (or other stored forms) that exemplify a minority grammatical pattern succumb to analogical pressure from the rest of the lexicon. Low-frequency words succumb to this pressure more readily than high-frequency words (Phillips, 1984, 2001; Lieberman et al., 2007; Todd et al., 2019). This is exactly what is to be expected from hierarchical inference. Because the learner has little evidence for the behavior of a rare word being idiosyncratic, such a word is likely to be mistakenly inferred to behave like a typical word (of the same type).

In contrast to analogical changes, articulatorily-motivated changes start in frequent words (Schuchardt, 1885; Fidelholtz, 1975; Hooper, 1976; Phillips, 1984, 2001; Mowrey and Pagliuca, 1987, 1995; Bybee, 2001). These are words with which the speaker has had the most practice. A change that targets a frequent word or phrase (like *going to* reducing to a nasal schwa in some contexts) cannot be due to the learner receiving insufficient evidence for the original, conservative pronunciation. Instead, these changes appear to be due to streamlining of articulation of a word or phrase with extensive practice. This conclusion is supported by the reductive character of such changes, which invariably involve temporal and/or substantive reduction of articulations, or smoothing out of the transitions between articulatory targets (Mowrey and Pagliuca, 1987, 1995; Browman and Goldstein, 1989; Bybee, 2001; Kapatsinski et al., 2020).

Most research on articulatorily-motivated sound change has not considered inference to play a role in this process. This would be appropriate if the progression of articulatorily-motivated changes were entirely mechanical, rather than partly governed by the conventions of the speech community. That is, if you could perfectly predict the degree of reduction in a context from the phonetics of the context – the articulatory routine being automatized – and the amount of practice that speakers had with it.

However, it is clear that this is not a tenable assumption. For example, Coetzee and Pater (2011) show that the rate of reducing /t/ or /d/ at the ends of words like *most* is affected by the following phonological context in different ways across varieties of English. This means that the rate of t/d reduction in a particular context needs to be learned as part of acquiring a particular variety of English. Certain segments are more likely to be reduced than others in a particular language with probability of reduction varying between languages (e.g., /k/ is reduced in Indonesian, /t/ in English, and /s/ in Spanish; Cohen Priva, 2017). Furthermore, the same segment in the same context can be reduced in different ways in different language varieties. For example, where Americans flap, many Brits would produce a glottal stop. Thus, a speaker needs to learn what to reduce, how to reduce, and when / in what contexts to reduce in part from exposure to what is done in their community.

As discussed above, articulatorily-motivated reductions are often particularly advanced in specific segments or gestures. For example, [t] is often reduced to the point of being deleted in *massed*, *mast* or *most* (Bybee, 2002; Coetzee and Pater, 2011) but [k] in *mask* or *musk* is not equally reduced. At the same time, such changes are also affected by the identities of the words in which the segment is embedded. Furthermore, some of this lexical conditioning is idiosyncratic, rather than attributable to word frequency, suggesting that the effects of word identity on pronunciation choices also need to be learned from exposure to the ambient language variety (Pierrehumbert, 2002; Wolf, 2011). For example, Zuraw (2016) mentions that the verb *to text* shows a particularly high rate of final [t] deletion. Because both segments and words affect pronunciation choices, a rational learner would use hierarchical inference to infer how much responsibility for a particular pronunciation rests at the lexical level.

Contribution of This Paper

In this paper, I consider how automatization of articulation interacts with learning processes by which the listener infers when and what to reduce. The principal innovation of the present paper, in the context of the literature on sound change, is to model this learning process. In the proposed model, learning is understood as rational probabilistic inference. That is, the listener infers the likely combination of causes that resulted in a particular observed pronunciation. Crucially, this inference process is argued to be hierarchical in nature (Xu and Tenenbaum, 2007; Feldman et al., 2009; Perfors et al., 2011; Kleinschmidt and Jaeger, 2015; O'Donnell, 2015; Harmon et al., 2021).

As noted above, since the 1870s, research on sound change has been dominated by a debate between the Neogrammarian doctrine of regular sound change, in which the change affects all instances of a phonological structure at once ("sounds change"; Osthoff and Brugmann, 1878) and the doctrine of lexical diffusion, in which words change one by one, so that a sound change diffuses gradually through the lexicon ("words change"; Schuchardt, 1885). Hierarchical inference allows the proposed model to capture the insight that the answer is *both*. That is, the likelihood of producing a particular phone in a particular context is determined *both* by the phoneme it instantiates, and by the larger units in which it is embedded (Pierrehumbert, 2002). For example, even though a /t/, in the right phonological context, is generally very likely to be realized as a flap in American English, this likelihood is somewhat lower when the /t/ is embedded in the formal word *emitter*.

The model described here captures this effect of lexical identity on the choice of an articulatory target for a sublexical unit. It is intended as the simplest possible model incorporating hierarchical inference into a theory of sound change. The model is easily extendable to incorporate additional levels in the linguistic hierarchy as influences on pronunciation, such as phonological units above the segment, morphemes, or collocations, all of which influence pronunciation (Mowrey and Pagliuca, 1995). Speakers and speaker groups can also be incorporated as an additional random effect specifying knowledge of sociolinguistic variation to account for speakers' ability to produce or imitate more than one dialect (e.g., Vaughn and Kendall, 2019).

A classic problem in sound change is why it does not always happen, even though the seeds for it are ever present (termed the *actuation problem* by Weinreich et al., 1968). Inference appears to play a crucial role in actuation. For a sound change to take off, an innovative pronunciation needs to be reproduced, both by the same speaker and by the speakers s/he talks to. Inference of the causes of the pronunciation appears to play an important role in this process. Specifically, experimental research has demonstrated unconscious imitation of phonetic detail, which shows how innovative productions can influence both future productions by the same speaker and those of their interlocutors (Goldinger, 1998). However, the extent and even direction of this influence can be affected by the listener's perception of the reason for which the speaker produced the word in a novel way, or in an unfamiliar context. For example, when the speaker is perceived to not be a fully competent speaker of the language, or to be a carrier of a stigmatized dialect, the listener is less likely to imitate the production (Babel, 2012; see also Bannard et al., 2013; Oláh and Király, 2019). The speaker is also less likely to reuse a pronunciation that has received a negative evaluation by an interlocutor (Buz et al., 2016). The listener's evaluation of a production, and therefore the spread of a change that originates in production, is thus influenced by a process of inference that identifies the production's cause.

The aspect of actuation I focus on here is diffusion of an innovative pronunciation through the lexicon, rather than through the community of speakers. In this context, it is important for a listener who considers adopting a speaker's pronunciation to know how far to generalize from the experienced examples. For example, observing *butter* produced with a flap, the listener might think that this is the way that the speaker pronounced *butter*, the way they pronounce the phoneme /t/, the way they pronounce an intervocalic /t/, etc. Depending on the structure(s) to which credit for the new pronunciation is assigned, a listener who decides to adopt the speaker's innovation might confine it to the particular word in which it was observed, or generalize it to a larger subset of the vocabulary (see Xu and Tenenbaum, 2007, for the equivalent problem in generalizing a wordform to a specific Dalmatian, all Dalmatians or all dogs). Nielsen (2011) has shown that unconscious imitation generalizes beyond the experienced word to other instances of the same phone, and even other phones sharing phonological features with it. In order to know how far to generalize a pronunciation, the listener needs to infer what caused the speaker to produce it. It appears that not only do listeners make inferences about why a speaker pronounced a certain segment in a certain way (see also Marslen-Wilson et al., 1995; Kraljic et al., 2008), this inference also influences their likelihood of reproducing the pronunciation.

I show that hierarchical inference provides a novel perspective on the puzzling phenomenon of stable variation. Sometimes, the diffusion of an innovative pronunciation variant through the lexicon stalls, resulting in stable lexically specific variation. A classic example is *-ing* vs. *-in'* in English, which has been stable for decades. (Labov, 1989; Abramowicz, 2007; Gardiner and Nagy, 2017). Stable variation presents a challenge to exemplar-theoretic models of sound change (e.g., Pierrehumbert, 2001) because a consistent leniting bias should eliminate the conservative variant (Abramowicz, 2007). The proposed model accounts for how variation can remain stable, even if one of the variants is already statistically dominant, and articulatory pressures always favor the dominant variant. The proposed model is unique in making clear predictions about the conditions under which stable variation is likely to emerge, and the level at which variation is likely to stabilize (Sections "Inference of a Random Effect of Lexical Identity: Lexicalization, Polarization, Stable Variation and a U-Shaped Frequency Effect" – "Stable Variation Depends on the Frequency Distribution and Its Effect on Reduction").

An important question begged by suggesting that the language learner takes words to be samples from a classified lexicon is whether the learner expect words s/he encounters in the future to be like the words she has already encountered? Or does s/he think that the words s/he is about to encounter might differ systematically from words s/he already knows (see Navarro et al., 2013, for the latter in learning non-linguistic categories)? In particular, if frequent words systematically differ from rare words, does the learner catch onto this fact, extrapolating that newly encountered (and therefore presumably rare) words are likely not to be like the frequent words s/he already knows (see also Baayen, 1993; Barth and Kapatsinski, 2018; Pierrehumbert and Granell, 2018)? This hypothesis is compatible with the widely adopted assumption that the grammar is primarily for dealing with novel inputs, with known words largely retrieved from memory (e.g., Bybee, 2001; Albright and Hayes, 2003; Kapatsinski, 2010a,b, 2018a). If the grammar is there primarily to deal with novel inputs, then it would be rational for the learner to base their knowledge of how to deal with novel inputs on experience with rare/novel inputs. Alternatively, learners may simply learn how known words and phones are pronounced without inferring anything about the relationship between word frequency and pronunciation. I take this to be the standard assumption in usage-based linguistics (e.g., Bybee, 2001: 12). The proposed model allocates the most likely amount of credit for a pronunciation to each of its *conceivable* causes, where causes are conceivable if they are considered by the listener. From this perspective, the question raised in the preceding paragraph reduces to whether conceivable causes of reduction likely include frequency of use. I will show that this is necessary for a monotonic relationship between frequency and reduction to be maintained after the reduced variant becomes dominant in the lexicon (Section "If Novel Words Are Thought to be Like Rare Words, Frequency Effect Will Stay Monotonic").

Relations to Other Work

The proposed model views language acquisition as a combination of automatization of production and rational probabilistic inference. Automatization is often discussed in work on sound change (Mowrey and Pagliuca, 1995; Pierrehumbert, 2001) as well as on the effects of experience on production (Tomaschek et al., 2018). Probabilistic inference is extensively explored in work on acquiring language from perceptual input (Xu and Tenenbaum, 2007; Feldman et al., 2009; Perfors et al., 2011; Kleinschmidt and Jaeger, 2015; O'Donnell, 2015). However, the interaction of the two mechanisms and its implications for the structure of language have remained unexplored.

Hierarchical inference conceptualizes sublexical units as classes of words sharing a particular chunk, and words are conceptualized as classes of utterances. This view of the nature of hierarchies aligns with the usage-based view of linguistic representations in considering linguistic units to be categories of experienced utterances (Bybee, 1985, 2001; Edwards et al., 2004), rather than building blocks out of which larger units are composed. For example, there is nothing in the proposed model that demands that an utterance be exhaustively parsed into morphemes. Whatever morphemes affect pronunciation choices are simply attributes shared by a class of words. Words sharing the morpheme *-ado* in Spanish are a class in the same way that Latinate words are a class in English. Even though the former are all similar in the same way, and the latter share no more than a family resemblance, both can affect pronunciation choices (e.g., lenition of [d] and stress placement, respectively). Despite the 'hierarchical' in the name, hierarchical inference does not require classes to form a strict hierarchy. In fact, hierarchical inference is compatible with any model of linguistic categorization that results in associable categories that share members. For example, the structural descriptions of rules in Albright and Hayes (2003) can also be considered word classes and are potentially subject to hierarchical inference. In Albright and Hayes (2003), rules associated with the same change can be nested, so that a more specific rule like "0→ed after a voiceless fricative" can co-exist with a more general rule like "0→ed after any consonant". It is therefore possible for a learner to use hierarchical inference to allocate credit for a particular instance of *-ed* surfacing after a voiceless fricative across rules that enact the same change (see O'Donnell, 2015, for a model that does this in morphosyntax). However, the closest work to the present proposal in the literature is Pierrehumbert's (2002) hybrid exemplar/generative model of sound change.

Pierrehumbert (2002) proposed that the speaker stores tokens of phones, and tags them with the identities of the words in which they occurred (as well as other contextual characteristics). From the present perspective, these tags define partially overlapping classes of pronunciation exemplars. Again, a strict hierarchy is unnecessary: the class of segment exemplars tagged with the word *cattish* and the class of exemplars tagged with /t/ can co-exist in the model even though not all /t/ exemplars occur in *cat* and even though exemplars tagged with *cattish* also include exemplars of other sounds. Selection of a pronunciation variant in producing a word is then biased to some extent by the identity of the word. The model proposed here builds on Pierrehumbert's model by incorporating an inference mechanism, which infers the contribution of a particular class/tag to a particular pronunciation of a segment. This inference determines how much the tag should influence the pronunciation of the segment in the speaker's subsequent production. In addition, by treating word identity and phonological context as independent influences on variant choice, the proposed model can account for cases in which reduced variants surface in phonological contexts that otherwise disfavor them. For example, Shport et al. (2018) show that American English speakers flap the /t/ in *whatever* even though flapping is otherwise illegal inside a word before a stressed vowel. By treating words as a random effect, the proposed model predicts such cases to be fairly rare and restricted to frequent words that are likely to be reduced and can resist the pull of the rest of the lexicon to regress to the mean but, crucially, does not predict them to be impossible. In addition, the present model generates stable variation and makes predictions about when it is likely to emerge.

THE MODEL

The most basic version of the model thus consists of the following parts:

(1) there are two pronunciation variants, reduced and unreduced;
(2) every time a word is used, the likelihood of the reduced variant of the phone being used in that word is incremented; as a result, reduction advances further in frequent words than in rare ones; and
(3) when a learner is exposed to the language, s/he learns not only an overall probability for each variant but also how variant probabilities are affected by lexical context.

In other words, the model proposes that the child learns how often a certain phone is pronounced a certain way and that some words are pronounced exceptionally. This kind of word-specific phonetic learning appears to be necessary because lexical frequency does not account for all between-word variability in phone pronunciation; a residue of exceptionality remains after frequency is accounted for Pierrehumbert (2002); Wolf (2011); Zuraw (2016).

The model assumes that the inference process is functionally equivalent to hierarchical regression. Below, it is implemented specifically as a logistic regression because of the first assumption above, the existence of alternative production targets associated with a phoneme in context such as an intervocalic /t/, which can be realized as a flap or a stop in American English. However, most reductive processes can also be conceived of as phonetically gradient rather than categorical (e.g., De Jong, 1998, for flapping; Bybee, 2002, for t/d deletion). Fortunately, the same predictions would be made by the present model if reduction were assumed to be continuous. We would simply replace the logistic link function with the identity link function of linear regression. Nothing hinges on the choice of the logistic linking function below.

The model was implemented in R (R Core Team, 2020) and is available at https://osf.io/qt6x4/. For ecological validity, I elected to simulate real sublexica that might be affected by a sound change. I considered two sublexica that are on the opposite ends of a productivity continuum: a large sublexicon with many rare words and a low maximum token frequency, and a small sublexicon with few rare words and a high maximum token frequency. The first sublexicon is the set of words with an intervocalic /t/ or /d/, followed by an unstressed vowel. The second sublexicon is the set of words beginning with eth (/ð/). Words in the first set constitute words in which the /t/ or /d/ is eligible to be flapped regardless of the broader context in American English (e.g., Herd et al., 2010). Words in the second set are eligible to undergo stopping in some dialects (e.g., Drummond, 2018), though this is not the full set of words eligible for stopping. However, our aim here is not to model these specific changes, but rather to ensure that the results of modeling are robust across sublexica that are maximally distinct in type frequency and the token/type ratio, which are the only characteristics of words that the model can see. Where noted, these sublexica are modified by excluding the most frequent words, those with frequency above 300, to explore the influence of these lexical leaders of change on its progression.

The first generation was seeded with one of two sublexica. The first sublexicon was the full sample of words eligible for flapping from the Switchboard Corpus (Godfrey et al., 1992). All words with a flapping context in the CMU Pronouncing Dictionary (Weide, 1995) were included ($N = 762$). These words had a stressed vowel followed by a /t/ or /d/ followed by an unstressed vowel. Each word occurred in the input with the frequency with which it occurred in the corpus, which followed the highly skewed Zipfian distribution (Zipf, 1935): 236 words were hapax legomena, occurring in the input only once; the most frequent word, *little*, occurred 2793 times.

The second sublexicon is the set of English words that start with /ð/. This set has far lower type frequency (only 24 distinct wordforms are found in Switchboard). It is also not Zipfian-distributed because it includes several very frequent words (*the*, *this*, *they*, *than*, *then,* etc.) and a relatively small number of rare words (*theirselves, theirself, thereabouts* and *thereof* are the only hapax legomena found in Switchboard). The frequent words in this sublexicon are also far more frequent than the frequent words in the flap sublexicon. In these respects, it is representative of a change that affects or is triggered by an unproductive sublexical unit, and therefore can be seen to lie on the opposite end of the continuum of productivity from the flap sublexicon (Baayen, 1993; Bybee, 1995). In principle, any other lexicon can be substituted: the predictions below are a necessary consequence of hierarchical inference and a highly skewed frequency distribution.

The log odds of reduction were seeded as in (1), with b_0 set to either −1 or −3 on the logit scale in the simulations below (0.27 or 0.05 on the probability scale), the magnitude of the frequency effect b_{Freq} set to 0.02 or 0.0002. The effects of these manipulations are discussed below, but it is worth noting that the values allow the change to progress slowly enough for lexical diffusion to be observed, and to progress rather than sputtering out. A substantially higher b_{Freq} can make almost all words have ceiling rates of reduction, while a substantially lower one can make them all reduce at the same rate. A substantially lower b_0 can lead the change to sputter out rather than progressing, and a higher b_0 means that the change has already affected most of the lexicon. The random effect of word was set as a random distribution with a mean of 0 and standard deviation of 0.4. I have tried reducing the latter to 0.2 and increasing to 0.8 with little effect. The random effect of word corresponds to whatever factors influence the likelihood of reducing a word that are not captured by the word's frequency. The three numbers mentioned above are the free parameters of the model, but the qualitative predictions are unchanged across a range of possible values. The number of reduced and unreduced tokens for each word was then generated as a sample from the binomial distribution, as in (2), with probability of reduction (p_{red}) defined as the inverse logit of the log odds, (1), and number of trials defined as the frequency of the word.

$$(1)\ p_{red} = logit^{-1}(b_0 + b_{Freq} \times Freq + N(0, b_w))$$

$$(2)\ n_{red} \propto Binom(p_{red}, Freq).$$

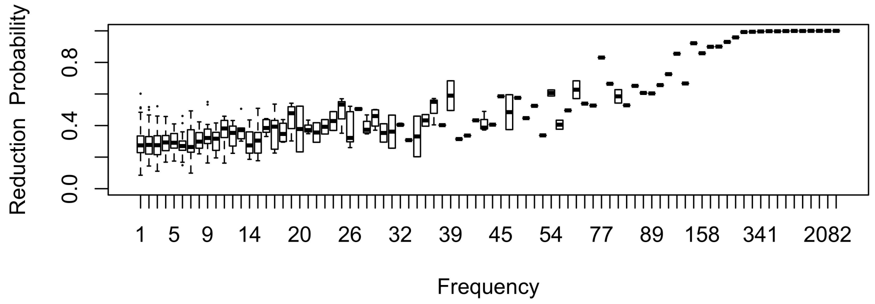


FIGURE 1 | The effect of frequency in the first generation, prior to passing the language through inference. Note that the frequency axis is rank-transformed (with the highest frequency on the right). Boxes consisting only of the median line contain a single word.

The effect of word frequency in this first generation is illustrated in **Figure 1** for $b_0 = -1$ and $b_{\text{Freq}} = 0.02$. The shape of the effect in Generation 1 represents what one would expect the shape of the frequency effect to be if inference played no role in articulatorily-motivated sound change. As one might expect, the effect of frequency is monotonic, with greater reduction in frequent words. Because reduction in (1) is proportional to raw frequency, and the frequency distribution is Zipfian, reduction probability is much higher in the highest-frequency words than in the bulk of the lexicon: reduction is nearly categorical in the most frequent words, while the mean reduction probability is 32%, close to the expected probability for a word of zero frequency, $b_0 = 27\%$. Lowering b_0 lowers the curve, lowering b_{Freq} reduces its slope, and lowering b_w (standard deviation) reduces the degree to which individual words deviate from the mean reduction probability at each point along the frequency axis.

Notice that the generative model in (1–2) is exactly that assumed by mixed-effects logistic regression with a by-word random intercept. Each generation was therefore assumed to use logistic regression to infer b_0, b_w and b_{Freq} or some subset thereof (**Table 1**). The regression was implemented using the lme4 package for R (Bates et al., 2015).[1]

Each generation then regenerated the corpus. In the model version that did not infer an effect of frequency (top two rows in **Table 1**), the inferred random effects of words replaced $N\left(0, b_w\right)$ in (1), and the inferred fixed-effects intercept replaced b_0 while the original b_{Freq} was retained. This represents the assumption that the effect of word frequency is due entirely to articulatory automatization. In the model that did infer the frequency effect (bottom row in **Table 1**), b_{Freq} was the sum of the inferred b_{Freq} and the original b_{Freq}. This corresponds to the possibility that words can be reduced either because reduction is inferred to be appropriate in this context, or because of articulatory automatization.

The language passed through up to 20 (or 100 or 300, where noted) generations. Iteration was stopped early if average probability of reduction across the tokens of the regenerated corpus exceeded 99% or fell under 1%, which defined the change running to completion or sputtering out, respectively.[2] 100 replications of the iterated learning process were performed for each parameter setting.

As mentioned above, the hierarchical structure assumed here is intended to be the simplest possible structure that can illustrate the effects of hierarchical inference on sound change. Additional influences on pronunciation can be easily incorporated into the model as additional fixed or random effects in Equation (1) above. For example, words can be nested within phonological contexts or morphemes to capture the fact that some morphemes can favor reduction across words, e.g., *-ado* favors Spanish intervocalic stop lenition (Bybee, 2002). Utterances or word senses can be nested in words to capture the fact that some uses of a word are more likely to be reduced. For example, *don't* is more likely to be reduced in *I don't know* than in *I don't think* and especially if *I don't know* is used to indicate uncertainty (Bybee and Scheibman, 1999). English auxiliaries are more likely to be reduced in some syntactic constructions than in others (Barth and Kapatsinski, 2017). Speakers (nested in social groups) can also be added as an additional random effect crossed with words, to implement inference of who flaps and who doesn't. Interactions between random effects can also be added, e.g., to capture knowledge of differences in the effect of phonological context effects on t/d deletion across English dialects (Coetzee and Pater, 2011).

SIMULATION RESULTS

Inference of a Random Effect of Lexical Identity: Lexicalization, Polarization, Stable Variation and a U-Shaped Frequency Effect

By treating lexical identity as a random effect, the model sidesteps the problem of estimating the effects of individual rare words, assuming that they will behave approximately like the average word, i.e., their reduction probabilities are drawn toward the mean reduction probability across all words. Partial pooling is

[1] $glmer(variant \sim (1 \mid w)\,, family\ =\ "binomial")$. or, if the effect of frequency is estimated, $glmer(variant \sim (1 \mid w)\ +\ Freq, family\ =\ "binomial")$. lme4 is used here because it converges relatively quickly, but informative prior beliefs can be added by replacing glmer() with brm() from the brms package (Bürkner, 2017) and specifying the desired prior.

[2] The early stoppage rule was introduced because the logistic regression command produces an error, which stops the iteration over replications when the response is constant, i.e., sound change has run to completion.

TABLE 1 | The model versions explored in the present paper.

Learner estimates	Reduction is influenced by	Section	Figures
b_0, b_w	Raw frequency	3.1	2, 3, 5, 7
	Log frequency	3.2	9, 10
b_w	Raw frequency	3.1	4, 6
b_0, b_w, b_{Freq}	Raw frequency	3.3	11

of course necessary for the rarest of the rare, the words that the speaker has never before encountered, because the model has no information about whether a novel word favors or disfavors reduction. However, it is also rational for more frequent words: the speaker would have considerable uncertainty about the acceptability of a flap in a word s/he observed two or three times if s/he could not use information about the acceptability of the flap in other (similar) contexts to make this determination (Gelman and Hill, 2007: 252–259).

Treating lexical identity as a random effect means that the regression model performs partial pooling of the information about variant probability across words, optimally weighting information from tokens of the word against information from tokens of the same sublexical unit occurring in other words (Gelman and Hill, 2007: 252–259). In partial pooling, the extent to which a word is drawn to the mean is inversely proportional to its frequency. The less frequent a word, the less information we have about the effect of that word on pronunciation (or on anything else). Thus, to know how an infrequent word behaves, a rational learner will partially rely on information about the behavior of other (similar) words. In contrast, to know how a frequent word behaves it is not necessary to rely on information about the behavior of other words: tokens of a word are more relevant for inferring its behavior than tokens of other words, and so should be relied on to the extent that they are available in sufficient quantities to draw a reliable inference.

The influence of inference on the word frequency effect is shown in **Figures 2–4**. The top panel shows the effect of frequency after the first pass through the inference process (Generation 1). At this point, the reduced variant is in the minority, and therefore the frequency effect is always monotonic, greater frequency favoring reduction. The middle panel shows a generation for which the reduced variant has become the majority variant, but has not yet achieved dominance: for this generation, the reduced variant accounts for 60–70% of tokens. The bottom panel shows a generation for which the reduced variant is statistically dominant, accounting for 90% of tokens.

In **Figure 2**, the learner estimates b_0, an overall probability of the reduced/innovative variant and the random effect of word on the choice, but does not estimate b_{Freq}, the effect of frequency. Reduction results only from use / automatization of production, increasing with raw frequency as in (1–2). A pronounced U develops in the shape of the frequency effect (as shown in the middle and bottom panels of **Figure 2**). By Generation 9 (middle panel), the median reduction probability for hapax legomena (frequency = 1) is much higher than for words that are more frequent. By Generation 18 (bottom panel), the words with frequencies below 8 or above 20 are almost always reduced, but the median reduction probability is at 90–95% for words of intermediate frequencies. As shown in **Figure 3**, this U is caused by the random effect of word, which maintains a set of exceptionally conservative words. These words must be frequent enough for their effect on reduction probability to be reliably estimable, but not so frequent as to become reduced through automatization of articulation.

The change in this model tends to stall at around 91% reduction (bottom panels of **Figures 2**, **3**). That is, the model gradually converges on nearly categorical use of the reduced variant, but the rate of change slows down dramatically once the probability of the reduced variant exceeds 90%. An individual chain can persist in the state shown in the bottom panels of **Figures 2**, **3** for a hundred generations. Furthermore, increasing or decreasing the size of the frequency effect by an order of magnitude changes how fast the model converges on ∼91% reduction but does not appear to help the model achieve greater reduction probability.

Table 2 shows the overall distribution of reduction probabilities across word types at Generation 100. The distribution shows what Zuraw (2016) has called *polarized variation*, which is characteristic of changes that have become lexicalized: the distribution of choice probabilities across words is highly bimodal, with clear peaks at 0 and 100%. A small number of words show intermediate behavior, with the vast majority of words (678.55 on average) always occurring with the reduced variant.

About 10% of the words (63.11 on average) become exceptionally conservative, reducing 0% of the time, with 4.16 more words reducing with a 1% probability. These are the outlier points at the bottom of the probability scale in the bottom panel of **Figure 2**. These rare reductions occurs because reduction can result from either inference that the word should be produced with the reduced variant, or from automatization of production. The automatization of production is blind to lexical idiosyncracies, and is always pushing words to reduce. However, inference resists this push for words that are inferred to be conservative.

As shown in **Figure 3**, emergence and persistence of polarized variation happens because the model learns of a set of exceptionally conservative medium frequency words (bottom panel). When most words are reduced 100% of the time, their random effects are essentially zero. However, exceptionally conservative words are maintained because their random effects are strongly negative. As long as these exception words are frequent enough, it appears that they can be maintained indefinitely.

Even though change in this model is driven entirely by frequency of use, the correlation between frequency and reduction probability weakens dramatically over time. Thus, log frequency accounts for 27% of the variance in reduction probability at Generation 2, but only 8% by Generation 9, and essentially 0 variance by Generation 18 (0.02%). Thus, the effect of word frequency in this model is expected to weaken dramatically as an articulatorily-motivated change progresses. Some support for this prediction can be found in Cohen-Goldberg (2015), who found an effect of lexical frequency on /r/

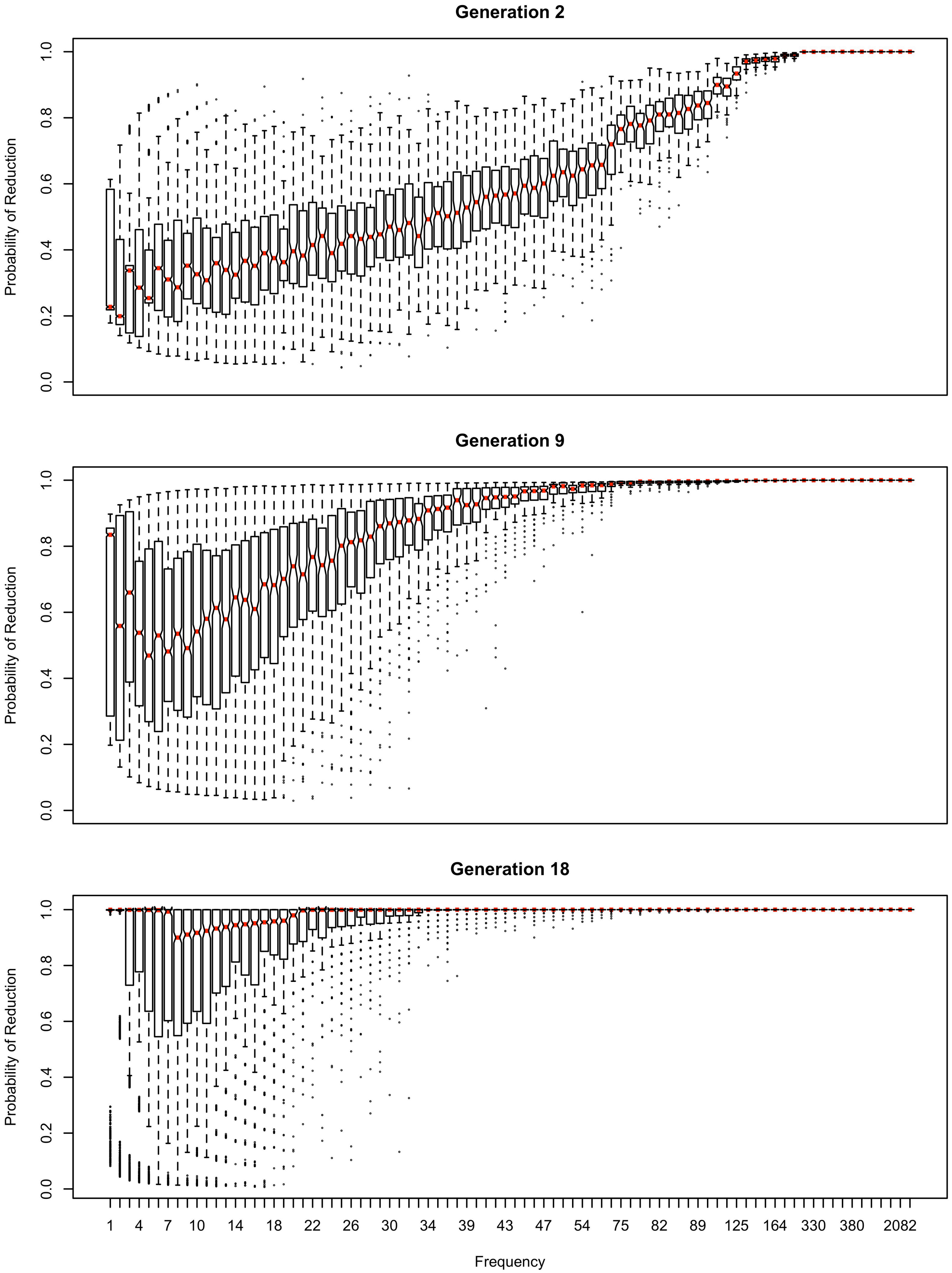


FIGURE 2 | The effect of frequency if the learner estimates overall probability of reduction (b_0) and the random effect of word, but not the effect of frequency across generations. Thick red lines show median probability of reduction at each frequency. Notches show the 95% confidence interval for the median.

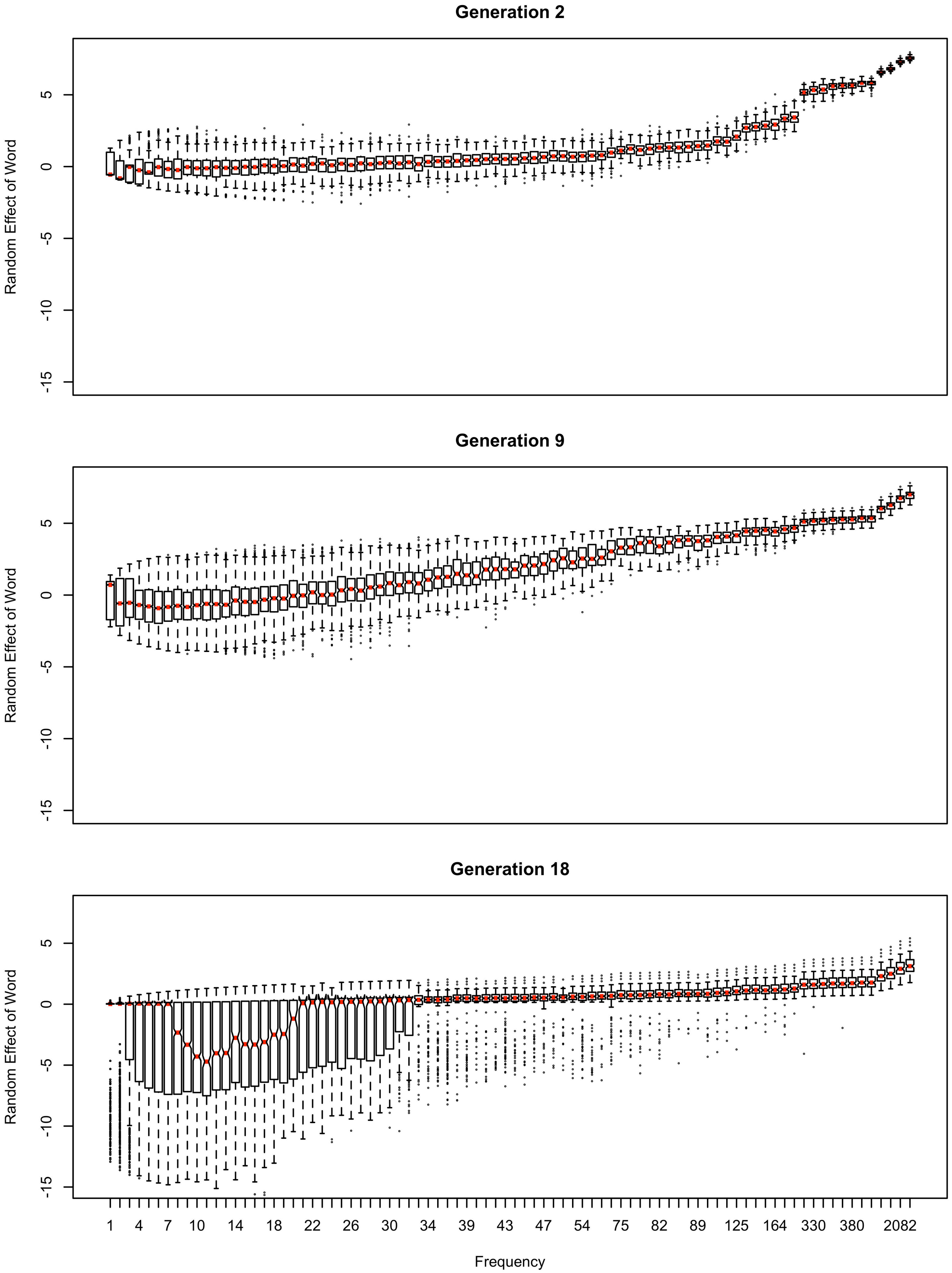


FIGURE 3 | The random effect of lexical identity across generations. A negative random effect for a word indicates that the word is associated with the conservative variant. A positive one indicates that the word is associated with the innovative variant.

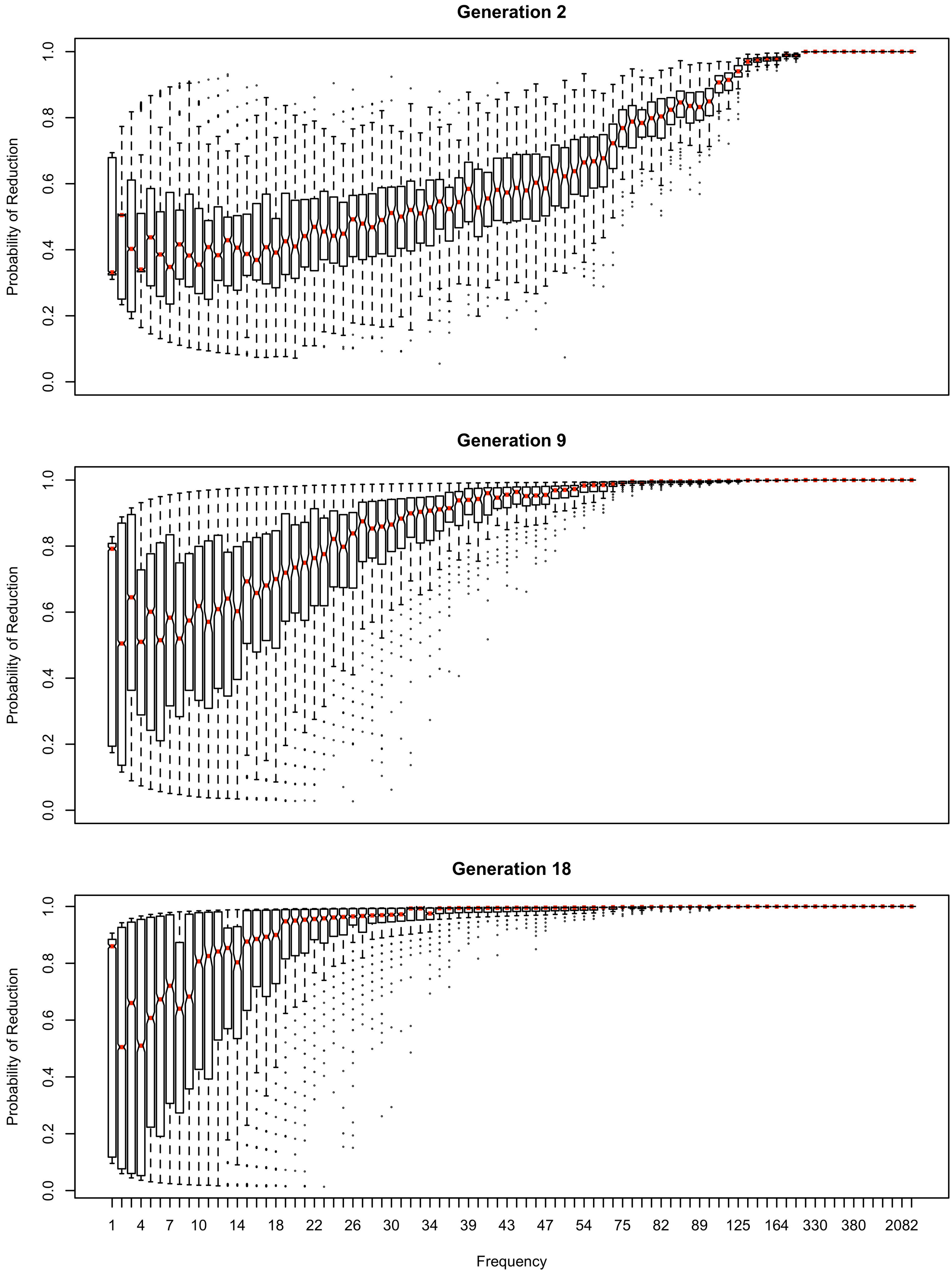


FIGURE 4 | The effect of frequency if the learner estimates only the random effect of word across generations.

TABLE 2 | The distribution of reduction probabilities in word types across 100 chains at generation 100.

0	0.01	0.1–0.2	0.21–0.3	0.31–0.4	0.41–0.5	0.51–0.6	0.61–0.7	0.71–0.8	0.81–0.9	1
6311	416	8	16	20	4	27	8	8	2	67855

There are no words with reduction probabilities between 0.01 and 0.13, or between 0.89 and 1.

deletion in a largely rhotic variety but not in a largely non-rhotic one. Furthermore, findings of weak or non-significant frequency effects in advanced changes (e.g., American English flapping in Warner and Tucker, 2011) are to be expected under this model, and do not provide evidence against articulatorily-motivated sound change being led by high-frequency words.

In **Figures 4**, **5**, the learner estimates only a random effect of word, and does not estimate either the effect of frequency on choice (b_{Freq}), or the overall probability of reduction (b_0). This version of the model behaves like the model in **Figures 2**, **3** in developing a U-shaped frequency effect because the words are still implicitly grouped together through partial pooling, resulting in the rare words being pulled toward the mean for the lexicon. However, the pace of change is slower, and the model does not converge to strongly favor the reduced variant. Instead, the model oscillates around a 61–64% reduction probability with the frequency effect illustrated in the bottom panel of **Figure 4** for at least 100 generations.

Thus, this model predicts that stable polarized variation will eventually develop, and that an initially phonetic change will become lexicalized, a trajectory that Bybee (2001) has argued to be a diachronic universal (see also Janda, 2003). The level at which the change stalls depends on whether the learner estimates an overall probability for a variant that is independent of individual words (b_0); in other words, estimating which variant is more likely overall, or if s/he only estimates how variant choice is conditioned by context. This seems intuitively satisfying for known cases of stable variation, such as the choice between *-ing* and *-in'* in English progressives, where the choice is invariably strongly conditioned by contextual factors such as register (see Gardiner and Nagy, 2017, for a review). In the present simulations, the conditioning contexts are lexical, thus the change becomes lexicalized, but other conditioning variables can be easily added to the model to investigate how a variant can become restricted to other environments, like speech styles or social personae.

The initial random intercepts with which words are seeded are not strong enough to resist reduction after the innovative variant becomes the default. How then do exceptionally conservative words become exceptionally conservative, enabling the conservative pronunciation to survive indefinitely? The bottom panel of **Figure 5** shows the random effect of word across the generations in one run of the model from **Figure 2**. In this model, when the innovative variant becomes dominant, most words succumb to the analogical pressure to reduce, regressing to the mean of the lexicon. However, a minority of words have random intercepts that are low enough for them not to regress to the mean of the lexicon at this point. These words then become "radicalized": the random intercepts of these words become ever more extremely negative in order to account for their lack of reduction. This makes these words increasingly resistant to reduction, stabilizing the system. Essentially, as the likelihood of reduction increases, the learner becomes increasingly confident that there is something special about the words not affected by reduction that prevents it from affecting them, resulting in lexicalization of the sound change.

Radicalization also happens in the model without an overall intercept shown in **Figure 4**, although here it is less extreme and affects both innovative and conservative words. Because all variability must be attributed to lexical identity, reduction caused by automatization of production leads to an increase in the corresponding random intercepts. The random intercepts of conservative words then must decrease to account for them now being farther from the lexicon mean. Because there is no overall intercept favoring the reduced variant across words, analogical pressure to regress to the mean is weaker, and variation stabilizes at a less skewed distribution. Interestingly, this distribution is also somewhat less polarized, with modes at 0.05 instead of zero and both at 0.95 and 1. Nonetheless, the variation remains stable after Generation 20.

The results in **Figures 1**–**6** replicate on a different lexicon, the set of English words that start with /ð/. As mentioned above, this sublexicon is representative of a change that affects or is triggered by an unproductive sublexical unit, and therefore can be seen to lie on the opposite end of the continuum of productivity from the set of words examined in **Figures 1**–**6** (Baayen, 1993; Bybee, 1995). Nonetheless, the results in this dataset are qualitatively very similar to those above: a U-shaped frequency effect develops as the reduced variant becomes dominant (**Figure 7**), and the change stalls as it becomes lexicalized, because exceptionally conservative words become radicalized when the reduced variant comes to dominate the lexicon. Thus, I expect these predictions to hold across a wide range of naturalistic sublexica eligible to undergo a particular sound change.

As with the flap sublexicon, not estimating an overall probability of reduction results in settling on a lower reduction probability. Interestingly, the overall reduction probability stabilizes much longer than individual words do. Thus, although probability of the reduced variant fluctuates around 0.67 for a long time, this stability initially masks large changes in the behavior of individual words from generation to generation as automatization-driven reduction battles entrenchment in conservatism for frequent words. Specifically, the mean reduction probability is about the same in both panels of **Figure 7** (0.65 on the left, 0.68 on the right) but the state represented in the left panel of **Figure 7** is unstable, and the model eventually converges to the state resembling the right panel, with all frequent words being categorically reduced.

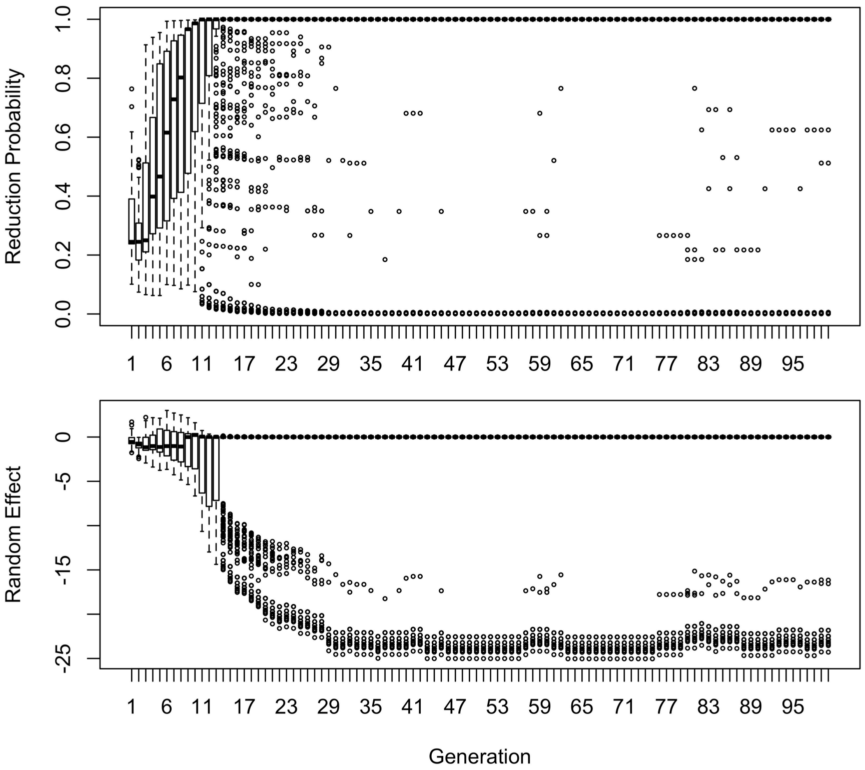


FIGURE 5 | Reduction probabilities and random effects of words that had below-average reduction probabilities at Generation 2. One run of the model shown. In this model, "the middle doesn't hold", and variation becomes polarized, with individual words reducing or not reducing close to 100% of the time (top panel). The bottom panel shows that words favoring reduction favor it because reduction is probable in the lexicon as a whole. The words that disfavor it instead become "radicalized", developing very strong negative random effects in favor of the conservative variant.

Stable Variation Depends on the Frequency Distribution and Its Effect on Reduction

The behavior of the model is dependent on the assumption that Equation (1) uses raw frequency and not log frequency. One might object to this assumption because log frequency is observed to be a better linear predictor of many behavioral dependent variables (e.g., Kapatsinski, 2010a, for error rate; Oldfield and Wingfield, 1965, for production latency). However, interestingly, this superior fit of log frequency turns out to also be true in the data generated by (1–2), even though they were generated using raw frequency: log frequency captures 27% of the variance in the generated reduction probabilities at Generation 1, compared to 18% captured by raw frequency. Thus, log frequency can fit the data better than raw frequency even if the data are generated by a model that is sensitive to raw frequency, i.e., a system in which every token of experience matters equally (as argued by Heathcote et al., 2000, for the effects of practice in general). This happens because there is an upper limit on reduction probability, so it always looks like the effect of frequency on reduction decelerates as reduction probability approaches the upper limit.

If log frequency is used in Equation (1), as illustrated in **Figure 8** (cf., **Figure 1**), the sound change progresses more quickly (**Figure 9**), even if mean $b_{\text{Freq}} \times \mathit{Freq}$ is equal to mean $b_{\text{Freq}} \times \log(\mathit{Freq})$. As mentioned earlier, mean reduction probability is in the 81-91% range across replications by Generation 20 with raw frequency, and can persist in that range for a hundred generations. In contrast, sound change completes at Generation 13–14 when reduction is driven by log frequency, even though it looks less advanced prior to training (**Figure 8** vs. **Figure 1**). This is due to the Zipfian distribution of word frequencies. With raw frequency driving reduction, the reduced words form a small minority for a long time: for a randomly chosen word type, the reduction probability is almost as low as that of a novel word. Therefore, the overall probability of reduction grows slowly. This allows some words time to become entrenched in their conservatism: they coexist with highly reduced frequent words. Furthermore, the frequent words in **Figure 1** are clear outliers relative to the mean of the lexicon. Their behavior is due to frequency but the learner does not know this, and therefore attributes it to a random effect of word. As result, the learner comes to believe that words have substantial idiosyncrasy: it is possible for a word to be really far from the lexicon mean in their reduction probability (as far as 5 standard deviations above in the top panel in **Figure 3**).

Because word is treated as a random effect, the learner estimates how variable the population of words is. Because of partial pooling, outlier words regress to the average reduction probability across words to the extent that

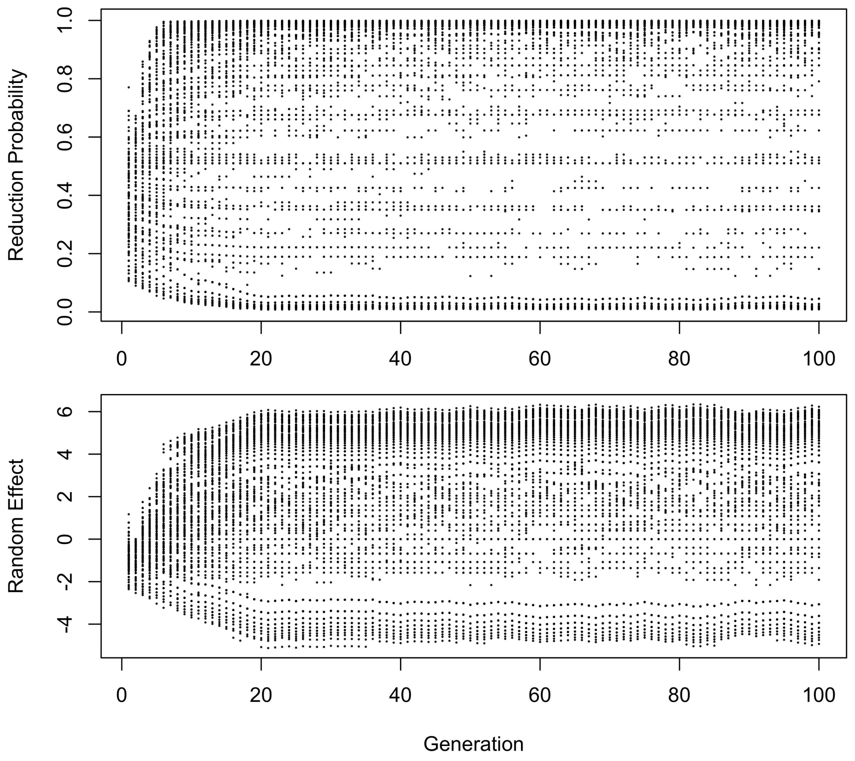


FIGURE 6 | Changes in the distribution of reduction probabilities and random effects for words with below-average reduction probability at Generation 2 in the model without an overall intercept. Negative random effects are in favor of the conservative variant.

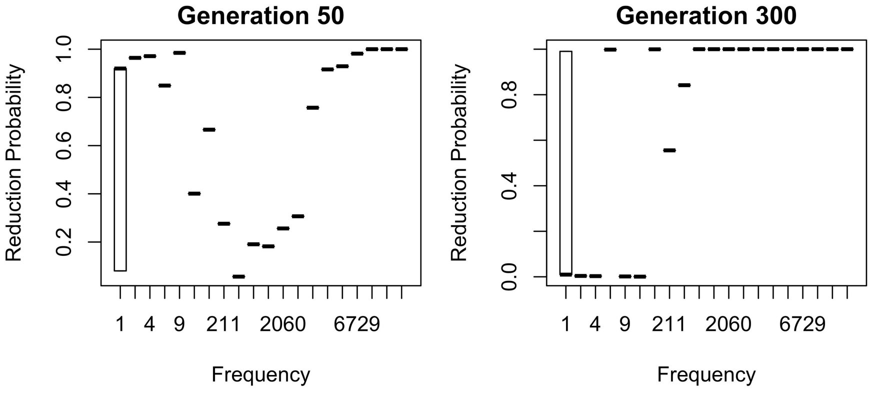


FIGURE 7 | Two generations with a similar mean probability of reduction of the /ð/ sublexicon. Note. b_{Freq} was reduced to 0.0002 for this simulation from 0.02 in **Figures 1–6** and b_0 was reduced to –3 from –1. This causes the model to converge more slowly, but the results are qualitatively similar if these parameters are higher.

words in general are tightly clustered around the average reduction probability. Therefore, estimating that words are highly variable in reduction probability allows exceptionally conservative words to not converge to the lexicon mean (**Figures 3**, **5**), which is what allows the conservative pronunciations to then be replicated across generations indefinitely. If reduction is proportional to log frequency, random effects are not so extreme: words look much more alike to the learner (**Figure 10**, which is on the same scale as **Figure 3**), hence sound change can run to completion relatively easily.

In addition, the U-shape predicted to occur in the later stages of a sound change is reduced, though not eliminated, and occurs at a higher average reduction probability than if raw frequency is used in (1). Nonetheless, the qualitative prediction is the same: once the reduced variant comes to dominate the lexicon, novel words should be reduced more than conservative medium-frequency words.

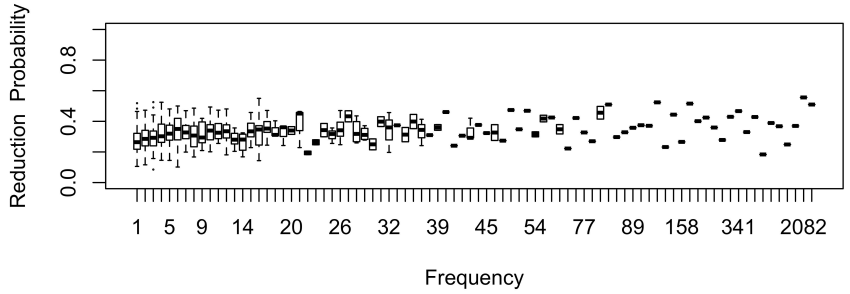


FIGURE 8 | Effect of frequency on reduction probability for Generation 1 (before the language is subjected to inference) with reduction driven by log frequency in Equation (1). b_{Freq} = 0.11 rather than 0.02, so that mean b_{Freq} × *Freq* in **Figure 1** is the same as mean b_{Freq} × log(*Freq*) here.

The difference between the initial distributions in **Figures 1**, **8** is that, in **Figure 1**, there are words whose behavior strongly deviates from that of the majority. This deviation is due to their higher frequency, coupled with the shape of the effect of frequency on reduction and the Zipfian distribution of word frequencies. Because their raw frequency is dramatically higher than that of the average word, and reduction rate tracks raw frequency, the frequent words look exceptional to a learner that cannot conceive of frequency as an explanation for these words' high degree of reduction.

That is, polarized variation requires the sublexicon of words affected by a change to contain apparently exceptional words. Although the first apparently exceptional words are exceptionally innovative, and these words become unexceptional as the lexicon approaches their reduction rates over generations, their existence is what allows for exceptionally conservative words to emerge and persist. This leads to a rather counterintuitive prediction: removing the highest-frequency, most reduced words from the sublexicon affected by a sound change should allow the sound change to run to completion even if the effect of frequency on reduction tracks raw frequency. This prediction is counterintuitive because this change makes the initial average probability of reduction lower. I have confirmed this prediction by creating an artificial version of the /ð/ sublexicon by removing words from the head of the frequency distribution (creating a 'headless' distribution; Harmon et al., 2021). Specifically, I removed all words with a frequency above 900 tokens in Switchboard, leaving only the 3 hapax legomena and 7 more frequent words (with frequencies 3, 4, 7, 9, 20, 30, and 211 tokens). This type of distribution might characterize a rare sound that occurs only in a small set of borrowed words (which are likely to be infrequent), such as the /ʒ/ word onset exemplified by *genre*. Even though removing the head of the distribution reduces the initial probability of the innovative variant, it allows the change to run to completion, with the innovative variant eventually dominating the production of all words. In other words, a change that affects a sublexicon of words of similar frequency is more likely to run to completion than a change that affects a sublexicon of words whose frequencies are very diverse. On the other hand, the change is also more likely to sputter out, with all words converging to the conservative variant. What does not frequently happen is a state of stable polarized variation (**Table 3**, left column), although two chains did converge on reduction in the most frequent word and lack of reduction elsewhere.

Similarly, a change is more likely to run to completion if the size of the effect of practice on reduction is small, because the small effect size ensures that no words are inferred to be exceptional. That is, articulatorily-motivated changes to segments that are less likely to change as a result of practice paradoxically have a greater chance of running to completion (although they also have a greater chance of sputtering out). With b_{Freq} = 0.02, the headless /ð/ sublexicon tends to quickly become lexicalized because the more frequent words are reduced much more than the less frequent words (**Table 3**, right column), even though the change frequently runs to completion with b_{Freq} = 0.0002 (**Table 3**, left column); a significant difference, $p < 0.0001$ (Fisher exact test). Because the initial probability of reduction is low (0.05), the final stable state tends to involve either 2 or 3 most frequent words categorically adopting the innovative variant, with the rest being categorically conservative (24 and 44 chains, respectively). However, occasionally the innovative variant spreads to most words, with a couple medium-frequency holdouts (4 chains), and sometimes even runs to completion (2 chains).

The results are similar with the larger flap lexicon, but differences in outcome between chains are smaller because the lexicon is larger, thus estimates of reduction probability are more stable and less affected by the exclusion of the few high-frequency words. In particular, strong reduction in the headless flap lexicon restricted to have the same maximum token frequency as the headless /ð/ lexicon always converges on stable polarized variation, but the final probability of reduction is much less variable, falling within 0.04 of 0.22. Weak reduction can still both sputter out or run to completion but the pace of change is much slower than in the /ð/ lexicon.

If Novel Words Are Thought to Be Like Rare Words, Frequency Effect Will Stay Monotonic

In all simulations reported above, a U is predicted to emerge in the shape of the frequency effect when the innovative, reduced variant becomes the default for the sublexicon. At that point, novel words would enter the lexicon with the reduced variant,

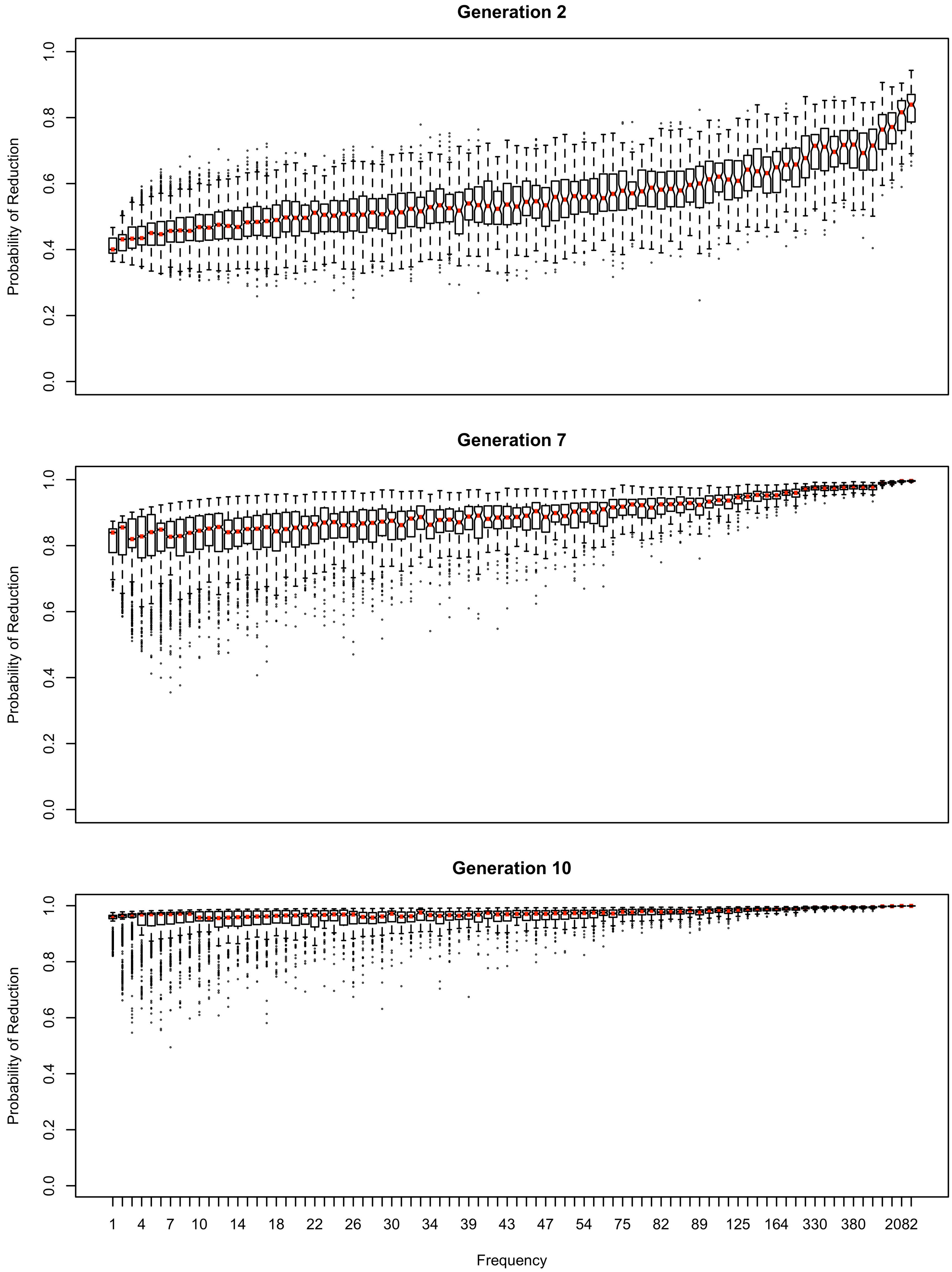


FIGURE 9 | The effect of word frequency if reduction is driven by log frequency. The learner does not estimate the effect of word frequency in this simulation.

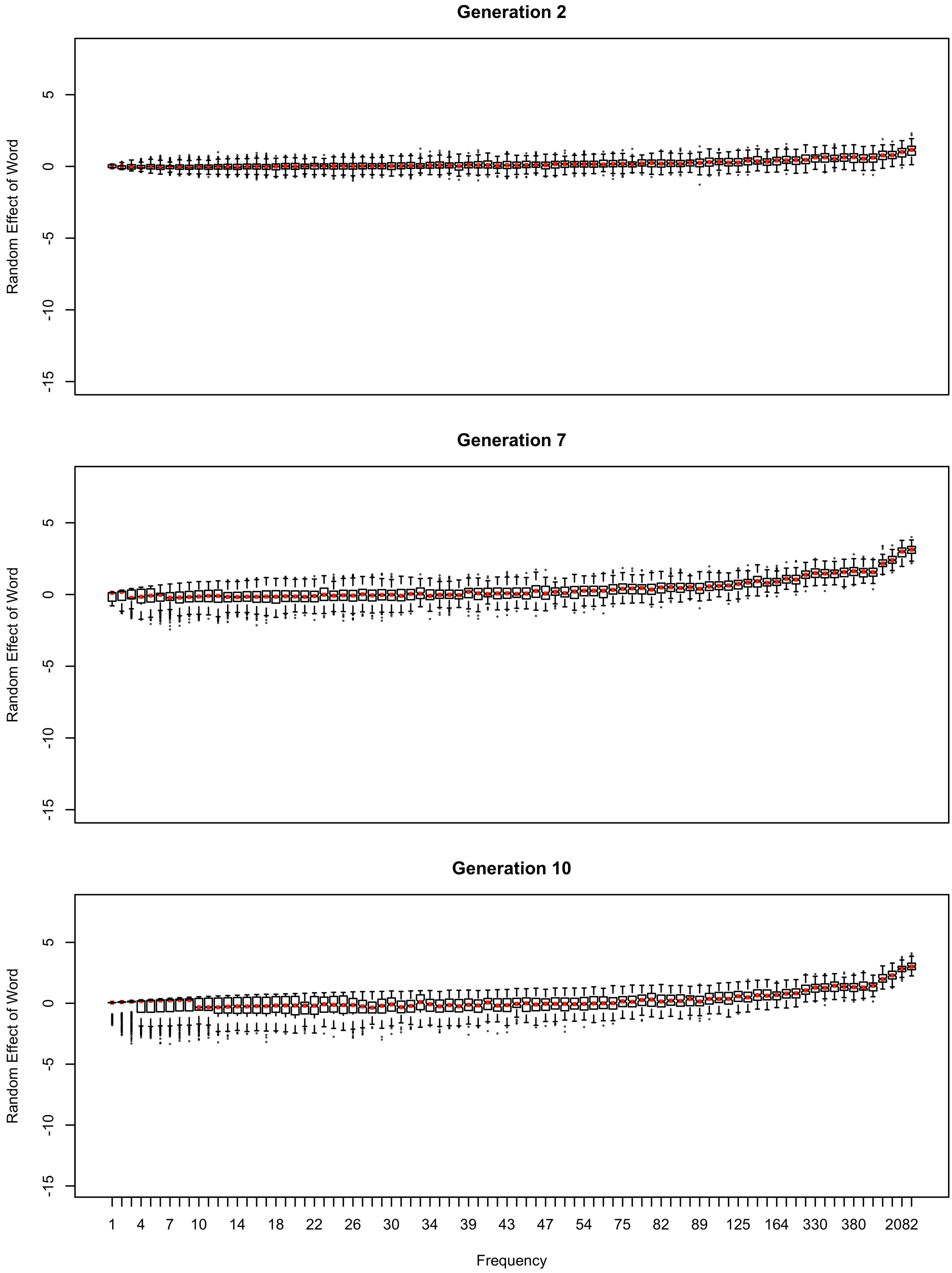


FIGURE 10 | Random effects in the model if reduction is driven by log frequency in Equation (1).

TABLE 3 | The effect of frequency on reduction, and the outcome of change.

Outcome of change	Weak reduction (b_{Freq} = 0.0002)	Strong reduction (b_{Freq} = 0.02)
Sputtered out	65	0
Run to completion	33	2
Stable polarized variation	2	98

b_0 = –3, headless /ð/ lexicon, learners estimate b_0 and a random effect of word.

while existing exceptionally conservative words would still be produced with the conservative variant.

As shown in **Figure 11** for the flap sublexicon, this prediction does not arise if the learner estimate the effect of frequency on variant choice, thus estimating all three b parameters in (1). In this model, the choice of the reduced variant can result either from the speaker's belief that this variant is more appropriate / likely, or from use / automatization of production: for every generation after the first one, the inferred b_{Freq} is added to the original b_{Freq}. As can be seen in **Figure 11**, no U-shape develops: the effect of frequency remains monotonic through the generations. The results in **Figure 9** do not change substantially if log frequency is used instead of raw frequency in Equation (1).

If the effect of frequency is estimated, the likelihood of the change running to completion is strongly dependent on the size of the frequency effect (b_{Freq}): with a strong reduction pressure (e.g., 0.1), the change runs to completion regardless of other parameters. However, with a weaker effect (e.g., 0.02), change does not run to completion. The change settles into stable variation at a reduction probability that depends on whether the learner estimates an overall intercept (b_0, the probability of variant choice). If they don't, the final reduction probability is quite high (above 90% in the flap sublexicon). If they do, then individual chains of learners estimating both b_{Freq} and b_0 settle on oscillating around ∼55% of innovative variant choice with the same parameter setting (b_{Freq} = 0.02). Indeed, average probability of reduction is able to progress beyond the initial state in **Figure 1** at all in this model only because of the additional reduction that comes from the incrementation of reduction probability by automatization of articulation: if only the inferred b_{Freq} is used, or the inferred and original b_{Freq} are averaged, the overall probability of the innovative variant does not increase across generations.

Variation in this model is not polarized: there is little variation in reduction probability between words of the same frequency; indeed, the random lexical variation the model is seeded with (**Figure 1**) reduces over time (**Figure 11**). Instead, stability comes from the model settling into a state in which only the lowest-frequency words (hapax legomena) are relatively unlikely to be reduced. The state to which this model converges if it does not estimate b_0 is similar to that shown by flapping in American English: there are no known words in which it is categorically impermissible, it occurs > 90% of the time, existing words reduce at similar rates across most of the frequency range, but novel words or words are produced with a full stop more often than known words (Herd et al., 2010; Warner and Tucker, 2011). The present model suggests that variation does not become polarized if differences in reduction rates across words are attributed to something other than their lexical identity. A rational learner that attributes the differences in reduction probabilities between frequent and infrequent words to frequency does not attribute this difference to lexical identity: frequency explains away apparent lexical idiosyncrasy. The model in **Figure 9** attributes them to frequency, but this is of course not the only possible factor conditioning variant choice. More generally, inference predicts that lexicalization should not happen when there are clear conditioning factors that account for between-word differences, whether these factors are social, stylistic, language-internal, or (like the effect of frequency) experiential.

DISCUSSION

This paper has examined the consequences of assuming that rational probabilistic inference is involved even in sound changes that are driven by automatization of production. Unlike analogical changes, these are sound changes that affect frequent words first (Schuchardt, 1885; Fidelholtz, 1975; Hooper, 1976; Phillips, 1984, 2001; Mowrey and Pagliuca, 1995; Bybee, 2001). In usage-based work, such changes have been discussed as resulting from automatization of holistic production plans associated with frequent words and collocations (Mowrey and Pagliuca, 1995). However, this hypothesis did not account for the fact that certain articulations are more likely to be affected by reduction than others, in a way that is specific to a particular language variety (e.g., Cohen Priva, 2017). To account for this property of change, Pierrehumbert (2002) proposed that articulatorily-motivated sound change affects sublexical articulatory units tagged with the larger lexical contexts in which they occur. The present model builds on this insight by allowing the learner to optimally allocate credit for an observed pronunciation between a segment and the larger context using hierarchical inference. In this paper, I examined how the predicted trajectories and outcomes of articulatorily-motivated sound change are affected by the assumption that the first language learner engages in this type of inference.

Sound change is commonly seen to result in a pattern of stable, lexicalized variation in which some words remain exceptionally conservative (e.g., Bybee, 2001). Zuraw (2016) points out that lexicalization results in a pattern of polarized variation, where some words occur with one pronunciation variant 100% of the time or nearly so, and others (almost) never occur with the variant. A model of articulatory optimization that does not provide a role for inference predicts that an articulatorily-motivated sound change will ultimately affect all words as their productions are optimized over generations. Hierarchical inference explains why changes might stall, and how polarized variation arises. Specifically, polarized variation occurs if articulatory reduction affects different words at very different rates, and the learner attributes these differences to lexical identity rather than their true cause. Here, that true cause is simple frequency of use, but it could also be occurrence in reduction-favoring linguistic or social contexts (as in Bybee, 2002, 2017; Brown, 2004; Raymond and Brown, 2012). An important direction for future work is to differentiate between frequency

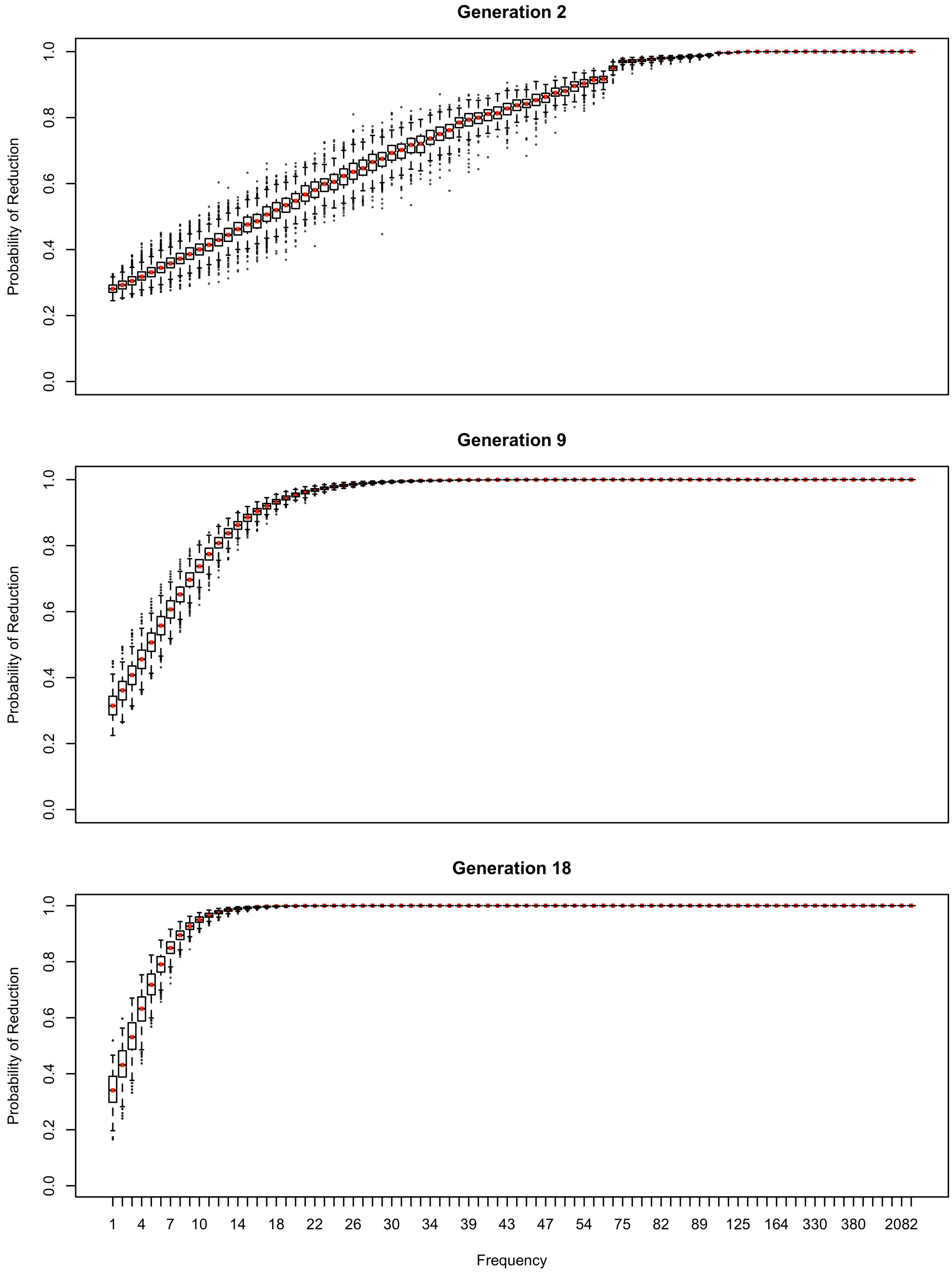


FIGURE 11 | The effect of frequency over time if the learner estimates the influence of frequency on variant choice as well as an overall probability of variant and the random effect of word.

of occurrence in reduction-favoring vs. disfavoring contexts. The literature is ambiguous regarding whether occurrence in reduction-disfavoring (e.g., formal) contexts merely delays change, or can actually lead the change to reverse direction. That is, it is not yet clear whether an additional token of occurrence in a reduction-disfavoring context, should decrement the probability of using the reduced variant in other contexts. It would be interesting to examine the consequences of this assumption.

Polarized variation arises through radicalization of exceptionally conservative words. Radicalization occurs because of the co-existence of conservative words with exceptionally innovative words in earlier generations, which leads the learner to estimate a large random effect of word. As the innovative pronunciation spreads through the lexicon, previously innovative words become the new mainstream, but their prior exceptionality allows exceptionally conservative words to retain their conservative pronunciations. That is, exceptions beget exceptions, even though the composition of the set of exceptions changes radically over time.

Hierarchical inference predicts that an articulatorily-motivated change can sputter out. Without this mechanism, articulatorily-motivated change inexorably marches on through the lexicon, converging to the reduced variant. However, in real life, the same change can sometimes take off, and sometimes not. In their foundational monograph on language variation and change, Weinreich et al. (1968) called this the actuation problem, and suggested that the answer to it is to be found in social dynamics – how an incipient change diffuses through society. The present simulations suggest that actuation also depends on lexical diffusion of the change: depending on the frequency distribution in the sublexicon of words that contain the structure affected by the change, and how the words that tend to occur in reduction-favoring contexts are distributed over the frequency spectrum, a change may not take off. In particular, if the effect of practice on the articulation is relatively weak for the sound in question, the sublexicon affected by the change happens to contain few high-frequency words (which are the words strongly affected by the reductive effect of practice), and the innovative pronunciation variant is initially rare, the change often sputters out. I submit that sputtering out is how changes ‘do not happen:’ variants that spread and take over in other languages arise and then disappear because they are inferred to have a low production probability. In essence, the speaker guards against reductions that they consider to be errors, suppressing their production. Covert error monitoring and suppression is of course well known to occur in language production (Motley et al., 1982). The present model shows one diachronic trajectory by which errors come to be seen as errors. Of course, there is always a chance for one of these variants to arise again because automatization of production continues to favor it over the conventionalized conservative alternative.

What can influence the strength of the influence of practice on articulation (b_{Freq})? The most obvious influence on this parameter is the fact that certain articulations are easier to produce in the context in which they occur than others. Articulations would not be particularly subject to the effect of practice. However, some articulations may also not change much as a result of practice even though they are not easy to articulate in context. For example, the tongue blade is a relatively fast, light and (at least for an adult) easily controllable articulator. It therefore appears relatively easy to speed up the production of a blade-raising gesture during the production of an alveolar stop with practice, turning it into a flap. In contrast, the tongue body is slow and heavy, making it much harder to speed up the production of a velar stop. Quantal effects, where certain articulatory changes lead to large changes in acoustics and other articulatory changes of the same magnitude do not (Stevens, 1989), can also make certain articulations more malleable due to absence of corrective feedback from interlocutors or the speaker’s own perceptual system.

What can influence the initial probability of reduction (b_0)? It seems likely that some changes originate from selection of variants that fall within a range of acceptable articulations before the change happens (Blevins, 2004). For example, there is a wide range of acceptable palatal constriction magnitudes for a Spanish [j]∼[ʒ], and selection of variants from within this range can drive divergence between dialects (Harris and Kaisse, 1999). Tongue positioning during a vowel is also quite variable, as is constriction magnitude in the production of an English flap (De Jong, 1998). In contrast, other changes might originate in speech errors, which may initially be very rare. A possible example is [θ] > [f] (Honeybone, 2016), because [f] and [θ] are not part of a continuous articulatory range of variants. In addition to changes that are not within an articulatory range associated with a production target, low initial production probability may hold for variants that are saliently perceptually different from the conservative variant, and therefore likely to be noticed by the listener (and perceived as a mismatch with intended acoustics by the speaker). Thus, changes that cross a quantal boundary might start out from a lower production probability. The simulations in the present paper show that such changes are likely to die out, but can also gain strength over time and even run to completion.

A take-home point of the present paper is that inference makes the dynamics of sound change rather chaotic; particularly so when the sublexical structure affected by the change has a low type frequency (like an initial /ð/ in English). Depending on small differences in initial conditions, and noise inherent to probabilistic selection of variants to produce, the same change affecting the same lexicon will sometimes go to completion, sometimes lexicalize, and sometimes sputter out. This is true in the present simulations even though there is no social environment to provide an additional source of variation. This means that the actuation problem is likely unsolvable. We should not expect to be able to predict whether a change will or will not happen. However, a theory of sound change can predict the directions in which change will proceed if it does, and a model that incorporates inference can help identify the factors that make a change more or less likely to be actuated, and to be lexicalized.

An intriguing prediction of hierarchical inference is that exceptionally conservative words should emerge in a ‘sweet spot’ in the frequency range when an articulatorily-motivated sound change enters a late stage in its development. When the reduced pronunciation becomes more likely, across the lexicon, than the original one, new words entering the lexicon should adopt the reduced pronunciation. Therefore, these new words should be more reduced than exceptionally conservative words. An important direction for future research is to model the impact of new words entering the lexicon on change. A limitation of the present implementation is that the lexicon is constant throughout. However, new words actually enter the lexicon all the time, and not at a constant rate (Gershkoff-Stowe and Smith, 1997). It would be interesting to see how the trajectory of change is influenced by state of the sublexicon when a large number of new words is encountered. An additional complication arises from the finding that words that have difficult articulations are especially likely to be replaced with other words because their articulation difficulty makes it less likely that they are selected for production (Berg, 1998; Martin, 2007).

Hierarchical inference predicts the effect of word frequency to be non-monotonic in the later stages of a reductive sound change. The most frequent words will be reduced because of two reasons: (1) the articulatory pressure toward reduction, as well as (2) because they were reduced in the input to the current generation of learners and thus will be associated with the reduced variant of the phone. The least frequent words will be reduced because they are not associated with any variant of the phone, and the reduced variant is more frequent. At intermediate frequency levels, some words, which happened to be often used with the unreduced variant of the phone by previous generations, can become associated with the unreduced pronunciation variant. As mentioned earlier, this prediction presupposes that a particular way of pronouncing a sublexical unit can spread from word to word, as suggested by Pierrehumbert (2002). This assumption is supported by the empirical results on new dialect acquisition in German et al. (2013), where speakers of American English were shown to rapidly learn new pronunciations for particular phones, e.g., a glottal stop in place of a flap, with no evidence of learning being restricted to individual words experienced during training (see also McQueen et al., 2006; Peperkamp and Dupoux, 2007; Maye et al., 2008; Nielsen, 2011).

An important contribution of the present simulations is to show the conditions under which exceptionally conservative words should emerge. This prediction of a U-shaped frequency effect in the later stages of an articulatorily-driven sound change is inevitable as long as (1) the sublexicon affected by the change includes frequent words that reduce at much higher rates than the rest of the sublexicon, and (2) the relationship between word frequency and variant choice is due solely to automatization of production, rather than to inference. That is, the learner should assume that novel words are likely to behave like the typical word, rather than like the typical *rare* word. This assumption is often made in research on productivity, because speakers tend to apply grammatical patterns to novel words based on the proportion of known words that obey them (see Kapatsinski, 2018b, for a review). However, Pierrehumbert and Granell (2018) found that the morphological behavior of hapax legomena is predicted by the behavior of rare words better than by the behavior of frequent words (see also Baayen, 1993; Zeldes, 2012; but cf. Albright and Hayes, 2003). Because productivity of a pattern is defined as its applicability to novel words, the particular importance of rare words in increasing productivity of a pattern suggests that learners infer the behavior of novel words from the behavior of (other) rare words, rather than from the entire lexicon. The question is whether speakers also implicitly know that the same phone (or letter) is likely to be pronounced differently in rare and frequent words, and make use of this knowledge in production.

It is also possible that speakers infer the likely pronunciations of words that they encounter more indirectly, by inferring the word's provenance. For example, speakers often need to infer the linguistic origin of a word to know how to pronounce it 'properly'. Relatedly, Vaughn and Kendall (2019) show that American English readers use the orthographic cue of an apostrophe at the end of a verb like *walkin'* to change their pronunciation of the rest of the utterance in a way that sounds more casual and Southern. For an adult native speaker's extensive experience with the language, the fact that the word is novel suggests that it is the kind of word that occurs in contexts with which the speaker has had little experience. For the typical experimental participant, a native-speaker university student, most newly encountered likely come from formal, academic contexts. They may therefore infer a novel word to likely be of similar provenance and thus pronounce it in a more formal fashion.

An important direction for future research is to extend the model to continuous articulatory variability. In principle, nothing in the proposed model depends on categoricity of the choice. For example, although we model reduction as the choice of a discrete variant here, a U-shape should also emerge if it were treated as a continuous acoustic or articulatory parameter (such as duration or degree of closure for a stop/flap/approximant continuum). The U shape depends on treating word as a random effect, and would emerge whether the learner estimates a logistic regression model (as here) or a linear regression model, as would be appropriate for a continuous variable. Nonetheless, a categorical choice produces certain discretization of the probability space because a difference in choice probabilities is only observable when it corresponds to a difference in token counts. This makes extreme probabilities more likely to converge to zero and 1, especially in rare words (e.g., Kapatsinski, 2010a). Variation could therefore, perhaps, be less polarized if the speakers were estimating a continuous parameter that is faithfully represented in the signal.

CONCLUSION

In this paper, I have explored the role of hierarchical probabilistic inference in articulatorily-motivated sound change, motivated by the findings that units at many levels of the linguistic hierarchy simultaneously influence pronunciation of a sound embedded in a particular context (Pierrehumbert, 2002). For example, pronouncing a /t/ as a flap in a particular phonological context could be due to the high probability of flapping in a favorable phonological context of a following unstressed vowel, or a high-frequency or informal word like *whatever*, which can lead to reduction outside of favorable phonological contexts (Shport et al., 2018). Because units at multiple levels (sublexical, lexical, and collocational) are jointly responsible for a particular pronunciation, a rational learner should allocate credit for a particular pronunciation across the levels via hierarchical inference. The proposed model provides a way to resolve the long-standing debate between proponents of regular sound change and proponents of lexical diffusion: it is not that "sounds change" or "words change". It is both. Hierarchical inference provides a way to estimate the contribution of both sounds and words to particular pronunciations. The present model suggests that speakers make use of this power.

The proposed model therefore incorporates the following assumptions: (1) there are both words and sounds, (2) a word's use causes reduction of the sounds in that word, and (3) both words and sounds (modeled as groups of words) are associated with reduction probabilities, with rational hierarchical inference adjudicating how much credit for a particular pronunciation of a sound in a word is assigned to the word vs. the sound.

The model explains how an articulatorily-motivated change can become lexicalized, even though there is a consistent pressure pushing all words to reduce. It also demonstrates the emergence of polarized variation (Zuraw, 2016). The model makes specific predictions about the circumstances under which a sound change can become lexicalized, the conditions under which it can sputter out, and the conditions under which it is likely to run to completion. Because chance plays an important role in determining the outcome of change, even in the absence of social influences, these predictions require a large-scale study of the characteristics of sublexica affected by changes that do and do not become lexicalized.

The hypothesis that speakers infer how to pronounce novel words based on generalization from a population of known words begs the question of what the relevant population is. Because rare words are often systematically different from frequent words (Bybee, 2001; Pierrehumbert and Granell, 2018), it can be considered rational for the learner to infer that a novel word will behave like other *rare* words, rather than being a typical representative of the whole sublexicon containing a sound eligible to undergo a change. When the innovative, reduced pronunciation becomes the majority variant, a learner who does not estimate the effect of frequency on pronunciation should favor the reduced pronunciation in novel words compared to known exceptionally conservative words. In contrast, a learner who does estimate the effect of frequency on variant choice should always show a monotonic frequency effect, with novel words being the least reduced. This provides another interesting direction for future empirical work.

DATA AVAILABILITY STATEMENT

The code for the model can be found in the **Supplementary Material**.

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was partially supported by the Faculty Excellence Fund Award and by the Article Processing Charge Fund Award from the University of Oregon.

ACKNOWLEDGMENTS

I would like to thank Danielle Barth, Joan Bybee, and Zara Harmon, for helpful discussion of the issues as well as Benjamin Tucker, Rory Tunbull and two reviewers for helpful comments on an earlier version. All errors are of course mine.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.652664/full#supplementary-material

REFERENCES

Abramowicz, Ł (2007). Sociolinguistics meets exemplar theory: frequency and recency effects in (ing). *Univ. Pennsyl. Working Papers Ling.* 13:3.

Albright, A., and Hayes, B. (2003). Rules vs. analogy in english past tenses: a computational / experimental study. *Cognition* 90, 119–161. doi: 10.1016/s0010-0277(03)00146-x

Baayen, H. (1993). *On Frequency, Transparency and Productivity. In Yearbook of Morphology 1992*. Dordrecht: Springer, 181–208.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *J. Phonetics* 40, 177–189. doi: 10.1016/j.wocn.2011.09.001

Bannard, C., Klinger, J., and Tomasello, M. (2013). How selective are 3-year-olds in imitating novel linguistic material? *Dev. Psychol.* 49, 2344–2356. doi: 10.1037/a0032062

Barth, D., and Kapatsinski, V. (2017). 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 Ling. Linguistic Theory* 13, 203–260. doi: 10.1515/cllt-2014-0022

Barth, D., and Kapatsinski, V. (2018). "Evaluating logistic mixed-effects models of corpus-linguistic data in light of lexical diffusion," in *Mixed Effects Models in Linguistics*, eds D. Speelman, K. Heylen, and D. Geeraerts (Cham: Springer), 99–116. doi: 10.1007/978-3-319-69830-4_6

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). *lme4: Linear Mixed-Effects Models Using Eigen and S4. R Package Version 1.0-4.*

Berg, T. (1998). *Linguistic Structure and Change: An Explanation From Language Processing*. Oxford: Oxford University Press.

Browman, C. P., and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology* 6, 201–251. doi: 10.1017/s0952675700001019

Brown, E. L. (2004). *The Reduction of Syllable -Initial /s/ in the Spanish of New Mexico and Southern Colorado: A Usage-Based Approach. Ph. D,. Thesis.* University of New Mexico.

Bürkner, P. C. (2017). brms: an R package for Bayesian multilevel models using Stan. *J. Statist. Soft.* 80, 1–28.

Buz, E., Tanenhaus, M. K., and Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: feedback from interlocutors affects speakers' subsequent pronunciations. *J. Memory Lang.* 89, 68–86. doi: 10.1016/j.jml.2015.12.009

Bybee, J. (1985). *Morphology: A Study of the Relation Between Meaning and Form*. Amsterdam: John Benjamins.

Bybee, J. (1995). Regular morphology and the lexicon. *Lang. Cogn. Proc.* 10, 425–455. doi: 10.1080/01690969508407111

Bybee, J. (2001). *Phonology and Language Use*. Cambridge, UK: Cambridge University Press.

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

Bybee, J. (2017). Grammatical and lexical factors in sound change: a usage-based approach. *Lang. Variat. Change* 29, 273–300. doi: 10.1017/s0954394517000199

Bybee, J., and Scheibman, J. (1999). The effect of usage on degrees of constituency: the reduction of don't in English. *Linguistics* 37, 575–596.

Blevins, J. (2004). *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge: Cambridge University Press.

Chomsky, N., and Halle, M. (1965). Some controversial questions in phonological theory. *J. Linguist.* 1, 97–138.

Coetzee, A. W., and Pater, J. (2011). "13 the place of variation in phonological theory," in *The Handbook of Phonological Theory*, 2nd Edn, eds J. Goldsmith, J. Riggle, and A. C. L. Yu (Wiley), 401–434. doi: 10.1002/9781444343069.ch13

Cohen Priva, U. (2017). Informativity and the actuation of lenition. *Language* 93, 569–597. doi: 10.1353/lan.2017.0037

Cohen-Goldberg, A. M. (2015). Abstract and lexically specific information in sound patterns: evidence from/r/-sandhi in rhotic and non-rhotic varieties of English. *Lang. Speech* 58, 522–548. doi: 10.1177/0023830914567168

De Jong, K. (1998). Stress-related variation in the articulation of coda alveolar stops: flapping revisited. *J. Phonetics* 26, 283–310. doi: 10.1006/jpho.1998.0077

Drummond, R. (2018). Maybe it's a grime [t]ing: th-stopping among urban British youth. *Lang. Soc.* 47, 171–196. doi: 10.1017/s0047404517000999

Edwards, J., Beckman, M. E., and Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *J. Speech Lang. Hear. Res.* 47, 421–436. doi: 10.1044/1092-4388(2004/034)

Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychol. Rev.* 116, 752–782. doi: 10.1037/a0017196

Fidelholtz, J. L. (1975). Word frequency and vowel reduction in English. *Chicago Ling. Soc.* 11, 200–213.

Gardiner, S., and Nagy, N. (2017). Stable variation vs. language change and the factors that constrain them. *Univ.Pennsyl. Working Papers Ling.* 23:10.

Gahl, S. (2008). Time and thyme are not homophones: the effect of lemma frequency on word durations in spontaneous speech. *Language* 84, 474–496.

Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

German, J. S., Carlson, K., and Pierrehumbert, J. B. (2013). Reassignment of consonant allophones in rapid dialect acquisition. *J. Phonetics* 41, 228–248. doi: 10.1016/j.wocn.2013.03.001

Gershkoff-Stowe, L., and Smith, L. B. (1997). A curvilinear trend in naming errors as a function of early vocabulary growth. *Cogn. Psychol.* 34, 37–71. doi: 10.1006/cogp.1997.0664

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). "SWITCHBOARD: telephone speech corpus for research and development," in *Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 517–520.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychol. Rev.* 105, 251–279. doi: 10.1037/0033-295x.105.2.251

Harmon, Z., Barak, L., Shafto, P., Edwards, J., and Feldman, N. H. (2021). "Making heads or tails of it: a competition–compensation account of morphological deficits in language impairment," in *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*.

Harris, J. W., and Kaisse, E. M. (1999). Palatal vowels, glides and obstruents in Argentinian Spanish. *Phonology* 16, 117–190. doi: 10.1017/s0952675799003735

Heathcote, A., Brown, S., and Mewhort, D. J. (2000). The power law repealed: the case for an exponential law of practice. *Psychonomic Bull. Rev.* 7, 185–207. doi: 10.3758/bf03212979

Herd, W., Jongman, A., and Sereno, J. (2010). An acoustic and perceptual analysis of/t/and/d/flaps in American English. *J. Phonetics* 38, 504–516. doi: 10.1016/j.wocn.2010.06.003

Honeybone, P. (2016). Are there impossible changes? θ> f but f≯ θ. *Papers Histor. Phonol.* 1, 316–358.

Hooper, J. B. (1976). "Word frequency in lexical diffusion and the source of morphophonological change," in *Current Progress in Historical Linguistics*, ed. W. Christie (Amsterdam: North-Holland), 96–105. doi: 10.1057/9780230286610_4

Janda, R. D. (2003). "Phonologization" as the start of dephoneticization–or, on sound change and its aftermath: of extension, generalization, lexicalization, and morphologization," in *The Handbook of Historical Linguistics*, eds R. D. Janda and B. Joseph (Wiley), 401–422. doi: 10.1002/9781405166201.ch9

Kapatsinski, V. (2010b). What is it i am writing? Lexical frequency effects in spelling russian prefixes: uncertainty and competition in an apparently regular system. *Corpus Ling. Linguistic Theory* 6, 157–215.

Kapatsinski, V. (2010a). Velar palatalization in russian and artificial grammar: constraints on models of morphophonology. *Laboratory Phonol.* 1, 361–393.

Kapatsinski, V. (2018a). *Changing Tools Changing Minds: From Learning Theory To Language Acquisition To Language Change*. Cambridge, MA: MIT Press.

Kapatsinski, V. (2018b). "Words versus rules (storage versus online production/processing) in morphology," in *Oxford Research Encyclopedia of Linguistics*, ed. M. Aronoff (Oxford: Oxford University Press).

Kapatsinski, V., Easterday, S., and Bybee, J. (2020). Vowel reduction: a usage-based perspective. *Rivista di Linguistica* 32, 19–44.

Kleinschmidt, D. F., and Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychol. Rev.* 122, 148–203. doi: 10.1037/a0038695

Kraljic, T., Samuel, A. G., and Brennan, S. E. (2008). First impressions and last resorts: how listeners adjust to speaker variability. *Psychol. Sci.* 19, 332–338. doi: 10.1111/j.1467-9280.2008.02090.x

Labov, W. (1981). Resolving the neogrammarian controversy. *Language* 57, 267–308. doi: 10.2307/413692

Labov, W. (1989). The child as linguistic historian. *Lang. Variat. Change* 1, 85–97. doi: 10.1017/s0954394500000120

Levelt, W. J. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Levelt, W. J., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behav. Brain Sci.* 22, 1–38.

Lieberman, E., Michel, J. B., Jackson, J., Tang, T., and Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature* 449, 713–716. doi: 10.1038/nature06137

Marslen-Wilson, W., Nix, A., and Gaskell, G. (1995). Phonological variation in lexical access: abstractness, inference and English place assimilation. *Lang. Cogn. Proc.* 10, 285–308. doi: 10.1080/01690969508407097

Martin, A. T. (2007). *The Evolving Lexicon PH. D, Thesis*. UCLA.

Maye, J., Aslin, R. N., and Tanenhaus, M. K. (2008). The weckud wetch of the wast: lexical adaptation to a novel accent. *Cogn. Sci.* 32, 543–562. doi: 10.1080/03640210802035357

McQueen, J. M., Cutler, A., and Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cogn. Sci.* 30, 1113–1126. doi: 10.1207/s15516709cog0000_79

Motley, M. T., Camden, C. T., and Baars, B. J. (1982). Covert formulation and editing of anomalies in speech production: evidence from experimentally elicited slips of the tongue. *J. Verbal Learn. Verbal Behav.* 21, 578–594. doi: 10.1016/s0022-5371(82)90791-5

Mowrey, R., and Pagliuca, W. (1987). "Articulatory evolution," in *Papers From the 7th International Conference on Historical Linguistics*, eds A. G. Ramat, O. Carruba, and G. Bernini (Amsterdam: John Benjamins), 459–472. doi: 10.1075/cilt.48.34pag

Mowrey, R., and Pagliuca, W. (1995). The reductive character of articulatory evolution. *Rivista di Linguistica* 7, 37–124.

Navarro, D. J., Perfors, A., and Vong, W. K. (2013). Learning time-varying categories. *Memory Cogn.* 41, 917–927. doi: 10.3758/s13421-013-0309-6

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *J. Phonetics* 39, 132–142. doi: 10.1016/j.wocn.2010.12.007

O'Donnell, T. J. (2015). *Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage*. Cambridge, MA: MIT Press.

Oláh, K., and Király, I. (2019). Young children selectively imitate models conforming to social norms. *Front. Psychol.* 10:1399.

Oldfield, R. C., and Wingfield, A. (1965). Response latencies in naming objects. *Quart. J. Exp. Psychol.* 17, 273–281. doi: 10.1080/17470216508416445

Osthoff, H., and Brugmann, K. (1878). *Morphologische Untersuchungen Auf Den Gebiete Der Indogermanischen Sprachen*. S. Hirzel, Germany.

Peperkamp, S., and Dupoux, E. (2007). "Learning the mapping from surface to underlying representations in an artificial language," in *Laboratory Phonology 9*, eds J. Cole and J. Hualde (Berlin: Mouton de Gruyter), 315–338.

Perfors, A., Tenenbaum, J. B., and Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition* 118, 306–338. doi: 10.1016/j.cognition.2010.11.001

Phillips, B. S. (1984). Word frequency and the actuation of sound change. *Language* 60, 320–342. doi: 10.2307/413643

Phillips, B. S. (2001). "Lexical diffusion, lexical frequency, and lexical analysis," in *Frequency and the Emergence of Linguistic Structure*, eds J. L. Bybee and P. J. Hopper (Amsterdam: John Benjamins), 123–136. doi: 10.1075/tsl.45.07phi

Pierrehumbert, J., and Granell, R. (2018). "On hapax legomena and morphological productivity," in *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 125–130.

Pierrehumbert, J. B. (2001). "Exemplar dynamics: word frequency, lenition and contrast," in *Frequency and the Emergence of Linguistic Structure*, eds J. L. Bybee and P. J. Hopper (Amsterdam: John Benjamins), 137–158. doi: 10.1075/tsl.45.08pie

Pierrehumbert, J. B. (2002). "Word-specific phonetics," in *Laboratory Phonology 7*, eds C. Gussenhoven and N. Warner (Berlin: Mouton de Gruyter), 101–139. doi: 10.1515/9783110197105.1.101

Raymond, W. D., and Brown, E. L. (2012). "Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish," in *Frequency Effects in Language Learning and Processing*, eds S. T. Gries and D. Divjak (Berlin: Mouton de Gruyter), 35–52.

R Core Team (2020). *R: A Language and Environment for Statistical Computing. Software*. R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Schuchardt, H. (1885). *Ueber Die Lautgesetze: Gegen Die Junggrammatiker*. Berlin: R. Oppenheim.

Shport, I. A., Johnson, G., and Herd, W. (2018). "Flapping before a stressed vowel – whatever!" in *Proceedings of the Meetings in Acoustics, 31, Paper 060004*.

Stevens, K. N. (1989). On the quantal nature of speech. *J. Phonetics* 17, 3–45. doi: 10.1016/s0095-4470(19)31520-7

Todd, S., Pierrehumbert, J. B., and Hay, J. (2019). Word frequency effects in sound change as a consequence of perceptual asymmetries: an exemplar-based model. *Cognition* 185, 1–20. doi: 10.1016/j.cognition.2019.01.004

Tomaschek, F., Tucker, B. V., Fasiolo, M., and Baayen, R. H. (2018). Practice makes perfect: The consequences of lexical proficiency for articulation. *Linguist. Vanguard* 4.

Vaughn, C., and Kendall, T. (2019). Stylistically coherent variants: cognitive representation of social meaning. *Revista de Estudos da Linguagem* 27, 1787–1830. doi: 10.17851/2237-2083.0.0.1787-1830

Warner, N., and Tucker, B. V. (2011). Phonetic variability of stops and flaps in spontaneous and careful speech. *J. Acoust. Soc. Am.* 130, 1606–1617. doi: 10.1121/1.3621306

Weide, B. (1995). *The CMU Pronouncing Dictionary. Version 0.7*.

Weinreich, U., Labov, W., and Herzog, M. I. (1968). "Empirical foundations for a theory of language change," in *Directions for Historical Linguistics: A Symposium*, ed. W. P. Lehmann (Austin: University of Texas Press), 95–195.

Wolf, M. (2011). "Exceptionality," in *The Blackwell Companion to Phonology: Phonological Interfaces*, Vol. 4, eds M. van Oosetendorp, C. J. Ewen, E. Hume, and K. Rice (Hoboken, NJ: Wiley-Blackwell), 106.

Xu, F., and Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychol. Rev.* 114, 245–272.

Zeldes, A. (2012). *Productivity in Argument Selection*. Berlin: Mouton De Gruyter.

Zipf, G. K. (1935). *The Psycho-Biology of Language*. Oxford: Mifflin Houghton.

Zuraw, K. (2016). Polarized variation. *Catalan J. Ling.* 15, 145–171. doi: 10.5565/rev/catjl.185

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 in Psychology

ORIGINAL RESEARCH
published: 19 August 2021
doi: 10.3389/fpsyg.2021.661479

The Acquisition of Noun and Verb Categories by Bootstrapping From a Few Known Words: A Computational Model

Perrine Brusini[1,2*†], *Olga Seminck*[3*†], *Pascal Amsili*[3] *and Anne Christophe*[2]

[1] *Department of Psychological Sciences, University of Liverpool, Liverpool, United Kingdom,* [2] *Laboratoire de Sciences Cognitives et Psycholinguistique, Centre National de la Recherche Scientifique, École Normale Supérieure/PSL University, Paris, France,* [3] *Laboratoire Langues, Textes, Traitements Informatiques, Cognition (Lattice), Centre National de la Recherche Scientifique, École Normale Supérieure/PSL University, Université Sorbonne Nouvelle, Paris, France*

OPEN ACCESS

Edited by:
Gerhard Jäger,
University of Tübingen, Germany

Reviewed by:
Sandra R. Waxman,
Northwestern University, United States
Tao Gong,
Educational Testing Service,
United States

***Correspondence:**
Perrine Brusini
p.brusini@liverpool.ac.uk
Olga Seminck
olga.seminck@cri-paris.org

† These authors have contributed equally to this work and share first authorship

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

Received: 30 January 2021
Accepted: 14 July 2021
Published: 19 August 2021

Citation:
Brusini P, Seminck O, Amsili P and Christophe A (2021) The Acquisition of Noun and Verb Categories by Bootstrapping From a Few Known Words: A Computational Model. Front. Psychol. 12:661479. doi: 10.3389/fpsyg.2021.661479

While many studies have shown that toddlers are able to detect syntactic regularities in speech, the learning mechanism allowing them to do this is still largely unclear. In this article, we use computational modeling to assess the plausibility of a context-based learning mechanism for the acquisition of nouns and verbs. We hypothesize that infants can assign basic semantic features, such as "is-an-object" and/or "is-an-action," to the very first words they learn, then use these words, the *semantic seed*, to ground proto-categories of nouns and verbs. The contexts in which these words occur, would then be exploited to bootstrap the noun and verb categories: unknown words are attributed to the class that has been observed most frequently in the corresponding context. To test our hypothesis, we designed a series of computational experiments which used French corpora of child-directed speech and different sizes of semantic seed. We partitioned these corpora in training and test sets: the model extracted the two-word contexts of the seed from the training sets, then used them to predict the syntactic category of content words from the test sets. This very simple algorithm demonstrated to be highly efficient in a categorization task: even the smallest semantic seed (only 8 nouns and 1 verb known) yields a very high precision (∼90% of new nouns; ∼80% of new verbs). Recall, in contrast, was low for small seeds, and increased with the seed size. Interestingly, we observed that the contexts used most often by the model featured function words, which is in line with what we know about infants' language development. Crucially, for the learning method we evaluated here, all initialization hypotheses are plausible and fit the developmental literature (semantic seed and ability to analyse contexts). While this experiment cannot prove that this learning mechanism is indeed used by infants, it demonstrates the feasibility of a realistic learning hypothesis, by using an algorithm that relies on very little computational and memory resources. Altogether, this supports the idea that a probabilistic, context-based mechanism can be very efficient for the acquisition of syntactic categories in infants.

Keywords: language development, acquisition of syntax, computational modeling, semantic seed, noun, verb, French

INTRODUCTION

In the past decades, many experimental studies have shown that young children start gathering knowledge about the syntactic structure of their native language much earlier than was initially thought. For instance, infants are sensitive to the function words of their language before their first birthday (e.g., Shafer et al., 1998; Shi et al., 2006a; Halle et al., 2008), and they start exploiting them to speed up their lexical access to already acquired content words between 12 and 18 months (e.g., in English or French, determiners are followed by nouns, personal pronouns by verbs, Kedar et al., 2006, 2017; Zangl and Fernald, 2007; van Heugten and Johnson, 2011; Cauvet et al., 2014). In addition, when presented with novel content words in several contexts, infants are able to infer which other contexts are expected for these novel words: for instance, after hearing *the blick*, then *a blick* would be expected but not *I blick* (for German: Höhle et al., 2004; for French: Shi and Melançon, 2010). Starting at 12–14 months of age, toddlers can exploit the syntactic contexts of novel content words to infer their plausible meaning—for instance, a novel word presented in a noun context, such as *it is a blick*, is assumed to refer to an object (e.g., Waxman, 1999; Waxman and Booth, 2001), while it is assumed to refer to an action if it is heard in a verb context, such as *he's blicking* (from 18 months on, Bernal et al., 2007; Waxman et al., 2009; Oshima-Takane et al., 2011; He and Lidz, 2017; de Carvalho et al., 2019). Around 20 months, toddlers also start to exploit the syntactic structure in which novel verbs appear to constrain their possible meaning—specifically mapping verbs appearing in transitive structures to causal actions (Yuan and Fisher, 2009; Arunachalam and Waxman, 2010; Fisher et al., 2010; Dautriche et al., 2014; de Carvalho et al., 2021).

These studies have established that the syntactic structure in which a word appears is exploited by toddlers to guess some of the probable characteristics of its referent. Depending on their syntactic contexts, words are attributed plausible semantic features, such that for instance, nouns are considered likely to refer to objects, and verbs likely to refer to actions (and similarly for different kinds of actions, such as 1-participant vs. 2-participants actions, and properties for adjectives). This wealth of experimental research was triggered by the *syntactic bootstrapping* hypothesis proposed by Lila Gleitman in the 80s (Landau and Gleitman, 1985; Gleitman, 1990), stating that very young children could exploit syntactic structure to constrain their learning of word meanings, by relying on the link between grammatical form and semantic characteristics (see also Waxman and Hall, 1993; Fisher et al., 1994; Fisher, 1996 and the excellent discussion in Waxman and Lidz, 2006). Since then, many studies have successfully demonstrated that some syntactic knowledge is available to children early in development, when they still have a fairly limited lexical knowledge. However, all these experimental results raise the question of *how* toddlers manage to figure out which contexts correspond to specific syntactic categories.

One possibility is that infants are able to analyze the distributional information of their input to identify words which occur in the same contexts as words from specific categories (Redington et al., 1998; Seidenberg and MacDonald, 1999). Several unsupervised computational models used the local context of words to assign them a category (Redington et al., 1998; Mintz, 2003; Parisien et al., 2008; Chemla et al., 2009; Chrupała and Alishahi, 2010; Weisleder and Waxman, 2010; Wang et al., 2011). They all presented better-than-chance performance in a categorization task, showing that local contexts do indeed contain relevant information. Because these models are unsupervised, they present the advantage that they presuppose no specific linguistic knowledge from infants. However, they run into several difficulties, that vary depending on the implementation choices that were made. For instance, Redington et al.'s model attempts categorization only for words which have been observed very often (the 1,000 most frequent words of the corpus), and groups words together based on the similarity of the contexts they occur in. Because it possesses very rich information regarding all the contexts that each to-be-categorized word may enter, it outputs a rich and accurate set of categories, for both content and function words (which are much represented in the 1,000 most frequent words). However, because this model does not even attempt categorization for new words or the ones that are seen only a few times, it is not particularly useful to describe how toddlers constrain word meaning acquisition, since these are precisely the words where additional information would come in handy to guess their meaning.

Other models have focused on frequent contexts rather than frequent to-be-categorized words, with the advantage that these models can categorize even words that are seen for the first time. In these models, the clustering mechanisms typically yield many different classes, with several classes for each target linguistic category (Mintz, 2003; Chemla et al., 2009; Gutman et al., 2015). For instance, in the "frequent frames" framework developed by Mintz (2003), the model starts by identifying the pairs of words that co-occur most frequently, with a gap of 1 word in-between. It turns out that words that are sandwiched within these contexts of frequently co-occurring words tend to share their category: for instance, *you _ it* selects verbs, while *the _ is* selects nouns. The end result of this procedure returns several groups of word for each syntactic category; for instance, there are several noun classes, corresponding to the frames *the _ is*, and *a _ is*, among others. Attempts to group classes together on the basis of shared words are not trivial, because many words belong to more than one category (e.g., noun/verb, "I bear," "the bear"). In an attempt to escape the tension between categorizing only a restricted number of frequent words and building many classes for the same categories, we present a model that is trained on a corpus in which a few words are initially categorized: the *semantic seed*.

The semantic seed refers to a plausible assumption: by the time children start addressing the categorization problem, they already have managed to learn the meaning of a few highly frequent content words. In addition, we hypothesized that infants are able to group those known words according to some semantic feature (e.g., words referring to objects, words referring to actions). Findings from the literature make both parts of this hypothesis highly plausible. First, several studies have shown that infants have already built a small lexicon before their first birthday (Bergelson and Swingley, 2012, 2013, 2015; Parise and Csibra, 2012; Syrnyk and Meints, 2017). For instance, Bergelson and Swingley (2012, 2013, 2015) have shown that 6-

and 9-month-old babies already know some nouns and some verbs. This demonstrates that word learning can occur very early, even when infants have very little linguistic knowledge yet. In some situations, the non-linguistic context is sufficiently supportive to promote word learning: namely when words have clear, concrete referents (objects and actions in the here and now, Medina et al., 2011; Taxitari et al., 2020), and when the context of the conversation contains rich socio-pragmatic cues (Tomasello and Akhtar, 1995; Akhtar et al., 1996). Second, it has been proposed that infants are able to detect specific semantic features in their environment and group them to form semantic categories such as agents, artifacts, or actions (Saxe et al., 2006; Carey, 2009). In addition, infants' ability to form categories is enhanced by speech, such that speech sounds seem to promote the formation of an object category in infants (Ferry et al., 2010, 2013), and labeling two objects with different words allows 9-month-old infants to consider them as different kinds (Xu, 2002, see Ferguson and Waxman, 2017 for a review). Other studies focusing on how language encodes some semantic features, such as gender, animacy and number, demonstrated that when semantic attributes are encoded in language, this is learned by infants (Berko, 1958; van Heugten and Shi, 2009; Shi, 2014; Lukyanenko and Fisher, 2016; Ferry et al., 2020). In fact, the range of semantic attributes that are morphosyntactically encoded in languages has been hypothesized to be part of what has been called the *core knowledge* system (Spelke, 2000; Strickland, 2017).

In the present work, we marked the different words known by the model, the semantic seed, as either action-referring words to form the seed of the "verb" category, or object-referring words to form the seed of the "noun" category. This is supported by a body of work showing that toddlers differentiate actions and objects and tend to map the first on verb items and the latter on noun items (Bernal et al., 2007; Waxman et al., 2009; Oshima-Takane et al., 2011; He and Lidz, 2017; de Carvalho et al., 2019). For instance, let's assume that a given infant managed to learn the meaning of "book," "teddy," "eat," "banana," "go," and "drink," (because they are highly frequent and refer to concrete objects and actions), they may be able to group them into [book, banana, teddy]$_{\text{object referents}}$ and [go, eat, drink]$_{\text{action referents}}$. Starting from this seed, infants would then need a learning mechanism that extends those proto-categories, relying for example on information from their context. By noticing in which contexts the object referents often appear (e.g., after "*and the*," or "*like a*"), children might be able to decide that an unknown word, such as "*bunny*" in "*and the bunny jumped*," also belongs to the object-referents category. The model we present here precisely attempts to test the efficacy of such a process.

The model stores two-word contexts for each word from a training corpus, in which a few words are categorized (the *semantic seed*). It then uses these contexts to categorize words in an unseen test corpus. We report here a series of experiments, in which we present the performance of this learning mechanism. We consider different sets of parameters, namely different sizes of the semantic seed and three different types of two-word-sized contexts: left, right and framing contexts. Evaluation of the model was obtained by carrying out a categorization task targeting unknown words. To study the impact of the size of the vocabulary known initially, we varied parametrically the size of the semantic seed (starting with only a handful of known words, up to a much more sizeable vocabulary).

To sum up, the aim of this study is to conduct a feasibility experiment and check how much knowledge infants could gather about the noun and verb categories, if they had access to the kind of computation hypothesized by the model. The model rests on two main assumptions which are both plausible and grounded in the infant literature. First, the *semantic seed* assumption proposes that when they approach the categorization task, infants have already succeeded in learning the meaning of a few words (frequent, referring to concrete objects and actions, presented in pragmatically helpful situations), and are able to group them into semantic classes: object referents and action referents (both parts of the assumption well supported by the infant literature, as seen above). Second, the model supposes that infants are able to keep track of bi- and trigram frequencies: a number of experiments support this assumption, showing that infants as young as 12 months pay attention to this type of distributional information, both when exposed to artificial languages (e.g., Gomez and Gerken, 1999; Marchetto and Bonatti, 2013), or when listening to sentences in their mother tongue (e.g., Santelmann and Jusczyk, 1998; Höhle et al., 2006; van Heugten and Johnson, 2010; van Heugten and Christophe, 2015). Note that the model is mimicking comprehension, since it attempts to categorize words from its input (on the basis of their linguistic context), in the hope of guessing their potential meanings, just as an infant would do when attempting to decode language.

In addition to these assumptions, the model has another important property: It categorizes words only in context. In other words, the model's main aim is not to produce a lexicon in which each word is listed together with its category—or, in the (rather frequent) case of words with more than one category, with its possible categories. Instead, each to-be-categorized word is classified as a function of its immediate context, irrespective of the nature of the word itself. Because of this characteristic, the model can classify words that are encountered for the first time (a useful feature if categorization is going to help word meaning acquisition) and should not suffer when it encounters an ambiguous word.

MATERIALS AND METHODS

Our model is based on a corpus of child-directed speech and keeps track of the frequency of triplets of adjacent words. It starts out knowing the categories of a few content words, that are grouped into semantic classes: object-referring and action-referring. At test, the model attempts to categorize some target words by looking at their two words of context. The model targets words that are not too frequent (namely, below a given frequency threshold), since frequent words are less likely to be unknown. As a consequence, the model will mostly target content words, since highly frequent words tend to be function words (for instance, upon hearing the string of words *the door*, one may expect to next find a verb, as in *the door creaks*; if, however the next word is *of*, as in *the door of the house*, the model will not attempt to

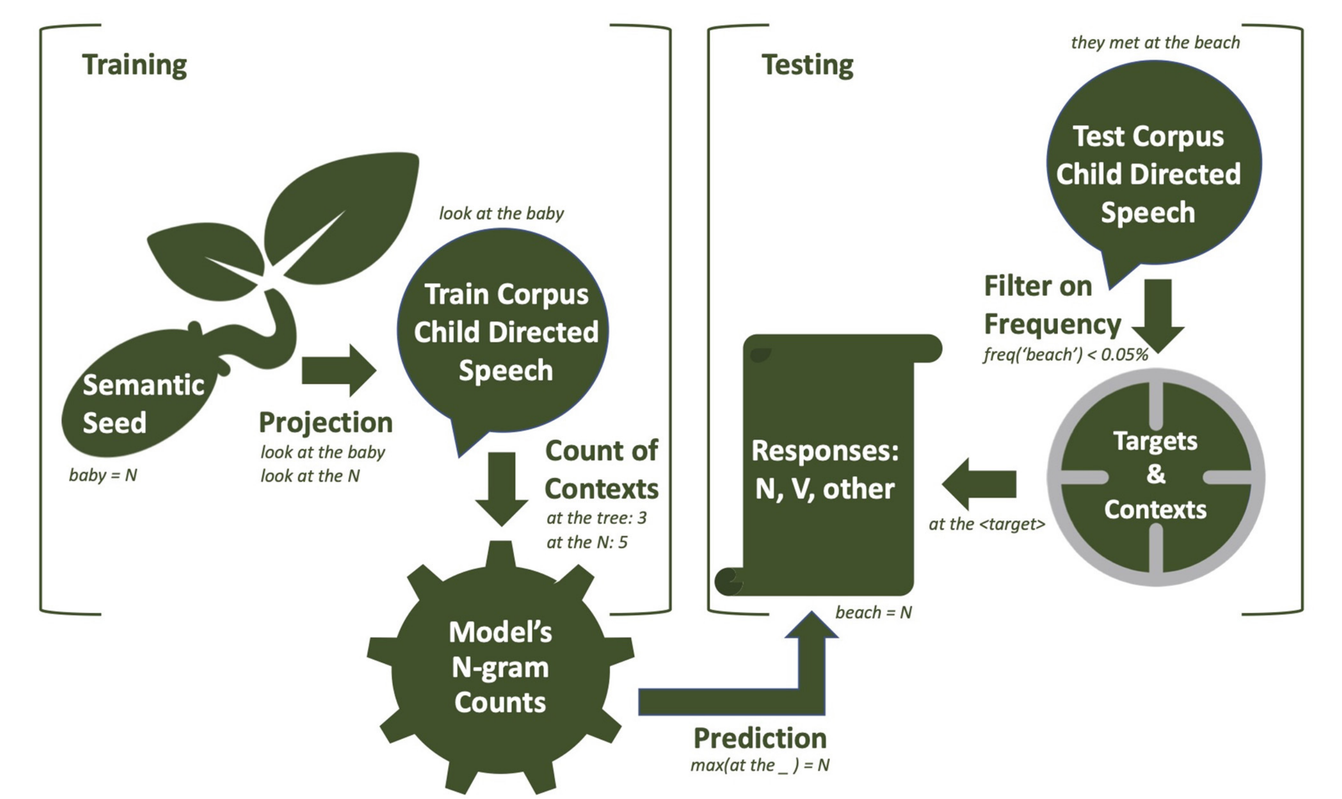


FIGURE 1 | A representation of the different steps of training and testing from our model. Details about the mechanisms can be found in the section Training and Testing.

categorize *of* because it is so frequent). The highly frequent words are, however, used as contexts.

To investigate the impact of the position of the words of context relative to the to-be-categorized word, three different contexts are implemented in three different models: two words immediately preceding the target word—left context; two words immediately following—right context; or one word before and one after—framing context. If these two words belong to trigrams that were observed during training, the model picks as its response the most frequent item occurring with these two words of context. We compare these three contexts to a baseline model: a model that does not rely on context to predict the syntactic category of low frequency words but that randomly predicts "noun," "verb," or "other" pondered by the percentage of known nouns and verbs from the corpus.

In this section, we present the details about the model's implementation. In **Figure 1**, the whole pipeline of our experiment is illustrated by a flow chart. The corpora and scripts are available in a GitHub repository, with the following link: https://github.com/oseminck/bootstrapping_model.

Corpus

The corpus is a transcription of spontaneous speech produced by French mothers during several play sessions with their child, and available in the CHILDES database (MacWhinney, 2000). The model used transcriptions of two mother/child pairs from the Lyon corpus (available at http://childes.psy.cmu.edu/data/Romance/French/Lyon.zip), Marie and Theotime, aged between 17 and 30 months during the recordings (Demuth and Tremblay, 2008).

The speech produced by the mothers of Marie and Theotime was extracted from the corpus, for a total of 58,241 utterances (265 K tokens). Each word of the corpus was then assigned a category (Part-of-Speech, or POS-tag) to evaluate the model's responses (by comparing the category predicted by the model with the actual category of the word). For the POS-tagging, we used the disambiguation grammar POST developed by Christophe Parisse that is integrated in the CLAN software (the program developed to exploit the CHILDES corpus; MacWhinney, 2000). We merged different types of noun categories and verb categories together (for example, we included modal verbs into the broader category of "verbs"). We performed a manual evaluation of the 640 first tokens of the corpus and found that 9% of the tokens were tagged with the wrong POS-tag. Because we are particularly interested in nouns and verbs, we also evaluated the error rate for these categories. The error rate of tokens tagged as verbs was 0%, but for nouns it was very high: 19%, meaning that 19% of the words that were tagged as nouns did not belong to that category. We therefore applied a correction to the tokens tagged as nouns in the following manner: we extracted all the noun lexemes from the corpus and sorted them by frequency. We then manually judged the 834 most

TABLE 1 | Words of the semantic seeds of various sizes.

	Nb noun lexemes in semantic seed	Percentage of projected nouns	Noun lexemes	Nb verb lexemes in semantic seed	Percentage of projected verbs	Verb lexemes
V0	8	7.3%	bébé, livre, doudou, main, tête, eau, voiture, pied	1	10.9%	aller
V1	16	11.8%	V0 + micro, nez, maison, lapin, train, lait, fleur, poisson	2	21.5%	V0 + faire
V2	32	18.7%	V1 + trou, oiseau, lit, cheval, gâteau, oreille, chat, éléphant, jeu, place, bouche, chien, morceau, chambre, pomme, doigt	3	26.6%	V1 + garder
V3	64	28.1%	V2 + poussin, canard, poule, carte, verre, montre, matin, monsieur, yeux, vache, boîte, camion, porte, oeuf, biberon, sac, rose, caméra, page, chausson, image, ballon, animal, assiette, mouchoir, cuillère, chanson, bras, fille, table, feuille, banane	6	34.6%	V2 + mettre, dire, tenir
V4	128	39.5%	V3 + mouton, balle, chaussure, bout, souris, bouton, bateau, téléphone, musique, carotte, ferme, nounours, puzzle, enfant, arbre, ours, chaise, mamie, soleil, cheveu, papillon, tour, souffle, tasse, fil, panier, café, bonhomme, chapeau, lettre, lumière, soeur, terre, pelle, dent, cochon, pantalon, vélo, sapin, jouet, fenêtre, école, forme, fruit, avion, garçon, crocodile, miette, argent, crèche, chaussette, château, photo, dessin, ventre, colle, clown, renard, pot, cuisine, lune, tétine, neige, tapis	12	41.9%	V3 + prendre, venir, manger, jouer, appeler, trouver
Vm	2159	100%	All nouns in the corpus	860	100%	All verbs in the corpus

frequent nouns (7 occurrences or more). We selected the lexemes that we suspected not to be nouns. For example, we found the word "*pour*" (the preposition *"for*" in French) in this list. This resulted in a list of 112 suspected lexemes. We then checked in the corpus whether the use was indeed non-nominal and not ambiguous between nouns and another syntactic category. For example, "*pour*" was never a noun, but "*touche*" (to touch/a button) was ambiguous between noun and verb. 100 lexemes were unambiguously non-nominal. We then corrected all the unambiguous lexemes in the corpus, which resulted in 6911 tokens being retagged. The list of the suspected lexemes and the corrected lexemes can be found in the additional materials of this article as well as in the GitHub repository.

Projection

To implement the idea that a small number of words are already correctly categorized by the learner, we placed an incomplete tier of categories on top of the tier of tokens in the training corpus. We call this tier of POS-tags the projection of the corpus. The category of all the words that belong to the semantic seed are identified in this tier.

Selection of the Semantic Seed

The semantic seed is composed of the most frequent nouns and verbs from the corpus that respectively refer to objects (including animate entities) and actions. The list of these words is given in **Table 1**. We varied parametrically the size of the semantic seed, so as to study the impact of the number of tokens initially categorized. As a starting point, we selected a situation in which the learner knows initially only very few of the verb and noun tokens: this corresponds to 8 nouns (7.1% of the noun tokens) and 1 verb (10% of the verb tokens). We then constructed 4 larger vocabulary sets, doubling the number of known nouns at each step, and adjusting the number of verbs such that the percentages of projected noun and verb tokens were relatively similar (increasing the percentage of the projection with about 5–10% for each new semantic seed, see **Table 1**). The reason why the number of verbs in the smaller semantic seeds is so low, is that these verbs are highly frequent, much more so than the most frequent nouns (see **Figure 2**)[1]. As a comparison point, one last set of vocabulary was created, containing all the nouns and verbs present in the training corpus (2,159 nouns and 860 verbs). This last vocabulary is obviously not a plausible representation of the lexical knowledge of a toddler, but it gives us an estimate of the best possible performance of the models we are implementing.

It might be important to note that for our model, we used the classical notation of nouns and verbs, but that we could as well have referred to object-referring-words and action-referring-words, if it weren't for the fact that we used a syntactic POS-tagger to evaluate the model's outcome. In principle, the model could work with other categories, such as finer-grained noun categories (e.g., animate/inanimate, human/non-human, edible/non-edible), or finer-grained verb categories (e.g., causative verbs, etc).

[1] In pilot experiments, we tested other configurations for the size of the semantic seed, for instance relying solely on frequency for the choice of the semantic seed or implementing a stronger filter to retain only concrete and observable words. The results are highly comparable, the model seems to be very robust with respect to these parameters.

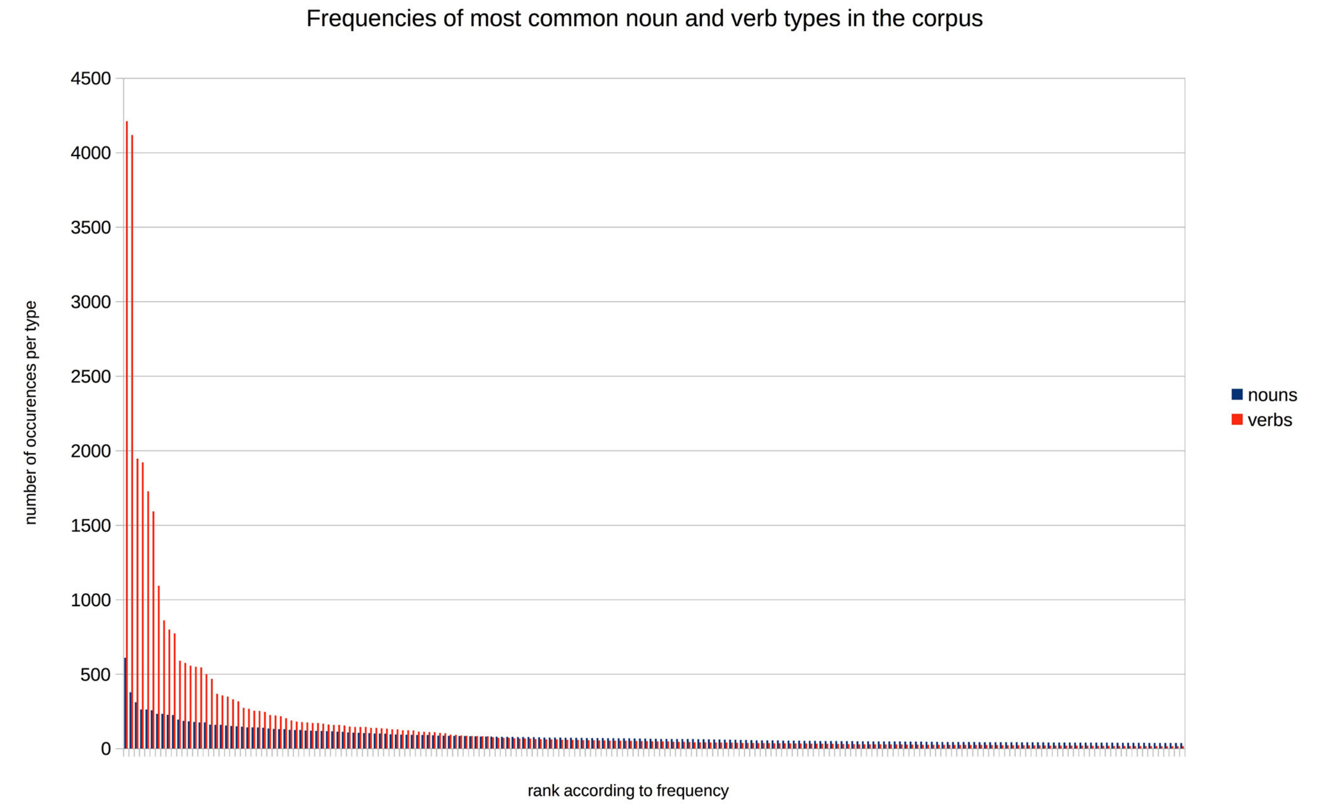


FIGURE 2 | The number of occurrences of the 200 most frequent noun and verb types in the corpus, ranked by frequency.

Training and Testing

We divided the corpus into training and test sets. To evaluate the robustness of the model, we first split the corpus into ten mini-corpora (each of them containing a tenth of the total corpus), then split each of them into a training (two thirds of the mini-corpus) and a test corpus (one third of the mini-corpus). This manipulation that leads to small non-overlapping corpora allows us to compute the variability of the model's performance, over each of the 10 runs.

To train the model, we collected the frequencies of each sequence of bigrams and trigrams of words encountered in the training corpus. In principle, our model relies on trigram frequencies, but in the test phase, when it makes predictions about unknown words, it relies on bigrams if the trigram that forms the context of this word has not been encountered during the training phrase. An example of how the model counts trigrams in an utterance is given in **Table 2**. Utterance boundaries (transcribed as strong punctuation in the corpus, coded as "*{*" and "*}*") were used as elements of context, but no n-gram could span over such boundaries (for example in "*Take that. Yes, that*," the 3-gram "*that } {*" is not counted).

Testing

During the test phase, the n-gram frequencies learnt during training, together with the local context of target words, were used to predict their syntactic category. To make a prediction, the context of the target word was compared with the set of n-grams collected during training. If this specific two-word context had been encountered during training as part of at least one trigram, the model selected as its prediction the most frequent item completing the trigram. If no trigram featured this two-word context, the process was reiterated with only one word of context (the left one for framing contexts). In a case where the one-word context was never encountered as part of a bigram, the model did not attempt to make a prediction.

One may note that our choice of model is extremely simple, since it consists of a table of trigrams, and does not attempt to assign probabilities to unseen events, as do more sophisticated models typically used in Natural Language Processing (e.g., deep-learning models, Markov chains, or regression models). The main reason for this choice is the interpretability of the model's parameters. The chosen framework allows us to easily analyze which contexts do most of the job (to glimpse ahead: those with pronouns for verbs and those with determiners for nouns). This would not have been the case using other models, for instance, neural networks (besides, the corpora we used are probably too small to train a neural-network). The simplicity of the model also makes the comparison between left, right and framing contexts extremely easy. A final argument in favor of our algorithm is that despite its simplicity, it is very effective. This suggests that infants do not need highly complex calculations to use statistical information from contexts.

TABLE 2 | The trigrams that are counted for the sentence "*Mais regarde, le bébé éléphant il va manger.*" (*But look, the baby elephant is going to eat.*).

Framing context		**Left context**		**Right context**	
Context	**Target Word**	**Context**	**Target Word**	**Context**	**Target Word**
{ _ regarde	mais	{{ _	mais	_ regarde }	mais
mais _ }	regarde	{ mais _	regarde	_ }}	regarde
{ _ bébé	le	{{ _	le	_ bébé éléphant	le
le _ éléphant	N	{ le _	N	_ éléphant il	N
bébé _ il	éléphant	le bébé _	éléphant	_ il va	éléphant
éléphant _ va	il	bébé éléphant _	il	_ va manger	il
il _ manger	V	éléphant il _	V	_ manger }	V
va _ }	manger	il va _	manger	_ }}	manger

The words 'bébé' and 'aller' (meaning respectively 'baby' and 'to go') are in the semantic seed.
Projection Tier: { mais regarde le N éléphant il V manger }.
Tokens: { mais regarde le bébé éléphant il va manger }.

Targets

To test the model, we took an unseen part of the corpus. As was said earlier, the model did not attempt to make a prediction for each word in the corpus. Rather, target words for which the model attempted a prediction had to fulfill the following two conditions: first, the context word closest to the target must have been seen by the model during training. In other words, the model did not attempt a prediction when it had no information on which to base its prediction. Second, target words should not be too frequent. In practice, words that had a frequency of 0.05% or more during training were excluded from categorization (corresponding to having been encountered 17 times or more during training). At this threshold, most function words were excluded, while most content words remained suitable candidates for categorization (more precisely 97.53% of the noun types and 94.63% of the verb types were selected, and among the few excluded nouns and verbs, most belonged to the smallest semantic seeds and were consequently known by the model).

Evaluation

To evaluate the model's performance, we calculated precision and recall for the noun and verb targets (see below) and compared the performance of the context-aware models (left, framing and right) to a chance model that constitutes a baseline for our experiments.

Precision and Recall

The use of the semantic seed entails that the training corpora contain some categorized words (N or V, the known words from the semantic seed), and a lot of tokens for which the category remains unknown (articles, adjectives, adverbs and the vast majority of the nouns and verbs that are not in the semantic seed). This fact has a consequence on the set of possible responses the model can produce in the categorization task. Because the model chooses as its response the most frequent item that was encountered in a given context, it may respond either with a category (N, V), or with a specific word-form (see **Table 3** for an example).

In this way, the model's responses were coded into three categories: noun, verb, and other. They were compared to the actual category present in the test corpus and used to compute hit, miss, and false alarm rates, separately for nouns and verbs. A hit was recorded whenever the model's response was either *"N"* or *"V"* and matched the actual category of the target word. A miss was recorded when the model should have responded *"N"* or *"V"* but instead replied something else, for example *"giraffe"* or *"V"* when the correct answer was *"N."* A false alarm (FA) was counted when the model responded *"N"* or *"V*,*"* whereas the target did not belong to that category. We should note that wrongly responding *"giraffe"* leads only to a miss (for nouns) but answering *"N"* when the correct answer is *"V"* leads to a miss for verbs and a false alarm for nouns.

These measures enable us to compute the precision and recall of the model. Precision is the hit rate divided by the total number of responses of a given category: hit/(hit + FA). If the precision is high, this means that when the model responds noun (or verb), it is usually correct. Recall is the hit rate divided by the total number of target words from a given category in the corpus: hit/(hit + miss). A high recall means that most of the nouns (resp. verbs) present among the target words have been categorized as such by the model.

Baseline: Chance Model

To evaluate objectively the performance of the learning mechanism, we created a different model that plays the role of a baseline. This model randomly categorized nouns and verbs without taking into account the context of the target words. The only information available to this model was the number of projection of nouns and verbs in the training corpus, which varies according to the size of the semantic seed. For example, if the training corpus contains 10% of known verbs, 10% of known nouns and 80% of words belonging to other categories, the baseline model randomly attributes a verb category 10% of the time, a noun category 10%, and neither noun nor verb for the remaining 80% of the words. For this model—as for the others—we computed the precision and the recall for the noun and verb categories, and we did this 10 times, using the 10 mini-corpora. Note that contrary to the other three models

TABLE 3 | Example of how the left context model would decide how to categorize a target word in two different scenarios.

Semantic Seed

N: baby, blankie, bottle

V: go, do

Context: *{ the _*

	Trigram counts from training	
	Scenario 1	**Scenario 2**
'{ the giraffe'	2	4
'{ the baby'	4	2
Model's Prediction	N	giraffe

If the model encountered the following trigrams during training: "{ the giraffe" twice and "{ the baby" 4 times, with 'baby' in the semantic seed, then the left context "{ the _" will trigger the prediction "N", since it is the item encountered most frequently within this context. If, in contrast, "{ the giraffe" had been encountered more frequently than "{ the baby", the model would have predicted "giraffe" to occur in the context of "{ the _".

which are deterministic, the baseline model contains a chance component, which means that running the model twice over the same corpus will yield slightly different results. It turns out that the performance of the chance model is stable over the 10 mini-corpora (see **Figure 3**), so that we estimated that running the baseline model several times over each mini-corpus was not necessary. If the two-word local contexts contain useful information for noun/verb categorization, then the context-aware models should exhibit a better performance than the chance model.

RESULTS

We first present here the results for the main categorization task, the precision and recall for nouns and verbs, for various semantic seed sizes and the four models we implemented (left, framing, right and chance). Then, we present some *post-hoc* analyses conducted to better understand the behavior of the models: an analysis of the misses for the smallest semantic seed, and a table presenting the most frequently used contexts.

Precision and Recall

The precision (top) and recall (bottom) of the left context (red), right context (yellow), framing context (blue) and chance (black) model are presented in **Figure 3**, with nouns on the left side and verbs on the right side. The x-axis in all graphs represents the different semantics seeds.

We ran mixed effects models in R (R Core Team, 2013) with the package lme4 (Bates and Sarkar, 2007; Bates et al., 2015). The statistical models we created aim to analyze the relation between our measures, precision and recall (*precision* [0–1], *recall* [0–1]*)* and the predictor variables: model type (*model:* baseline, right, left, framing), semantic seed size, (*voc*: V0, V1, V2, V3, V4, Vm), and the targets: nouns or verbs (*n_v*). Random intercepts and slopes for the 10 mini-corpora (the *fold)* were modeled for the predictor variables semantic seed size (*voc*) and noun or verb targets (*n_v*). This resulted in the following model:

$$\text{precision} \sim \text{model} * \text{n\_v} * \text{voc} + (\text{n\_v} * \text{voc}|\text{fold})$$

We built a similar model for recall (*recall* [0-1]):

$$\text{recall} \sim \text{model} * \text{n\_v} * \text{voc} + (\text{n\_v} * \text{voc} \mid \text{fold})$$

In order to be able to compare all types of models against each other, we repeated our analyses three times, changing every time the base value of the *model* variable (either right, left or framing). This resulted in a total of six mixed models, accounting for the 2 measures, and therefore we adapted our level of significance to 0.05/6 = 0.0083 instead of 0.05, according to a Bonferroni correction. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. The full output of all models can be found in the **Supplementary Material** of this article **(Data Sheet 1)**.

Overall, the left context and framing context models typically yield better precision than the baseline (precision: left model: $\beta = -3.675\text{e}{-01}$, $t = -9.804$, $p < 0.001$, framing model: $\beta = -3.233\text{e}{-01}$, $t = -8.626$, $p < 0.001$). The right context model performs more poorly, with no significant overall difference in precision relative to baseline ($\beta = -0.012865$, $t = -0.343$, $p = 0.73$).

The first striking result is the excellent precision that is obtained by the left and framing models, independently of the size of the semantic seed, which was not a significant predictor variable when modeling the precision of the left and the framing context models ($\beta = 2.899\text{e}{-03}$, $t = 0.368$, $p = 0.71$, and $\beta = 2.455\text{e}{-03}$, $t = 0.312$, $p = 0.76$, respectively). Precision is above 80%, for nouns and verbs, for both models. This means that even when the semantic seed is very small, and only a small number of contexts can be learned, these contexts are good contexts, that provide error-free categorization. In contrast, recall depends highly on the number of nouns and verbs categorized in the training corpus, with a low recall when the semantic seed is small, and a clear improvement as it increases ($\beta = 0.118103$, $t = 24.435$, $p < 0.001$)[2]. This reflects the fact that with a small semantic seed, the model can learn only a limited number of noun and verb contexts, and consequently, that it can categorize only a limited number of new nouns and verbs (albeit with a good precision).

The kind of contexts used by the model impacts the results. The right-context model is clearly the least efficient at correctly predicting nouns and verbs, with both precision and recall significantly lower than the other two models ($\beta = -3.546\text{e}{-01}$, $t = -9.461$, $p < 0.001$; $\beta = -3.104\text{e}{-01}$, $t = -8.283$, $p < 0.001$, for the right model compared to the left model and the framing model respectively). The others two models, relying on left and framing contexts, exhibit consistently good results, with a precision far above the baseline at all semantic seed sizes, as indicated above (~0.9 for nouns and ~0.8 for verbs), and a recall that rapidly rises above baseline as the semantic seed grows (results for the interaction of semantic seed size and model type when comparing the baseline model and the left model: $\beta =$

[2] We used the statistical model that uses the left model as the base level for the variable "model," but the two other statistical models for recall yield similar results at the same level of significance. Please see the **Supplementary Material** for more details.

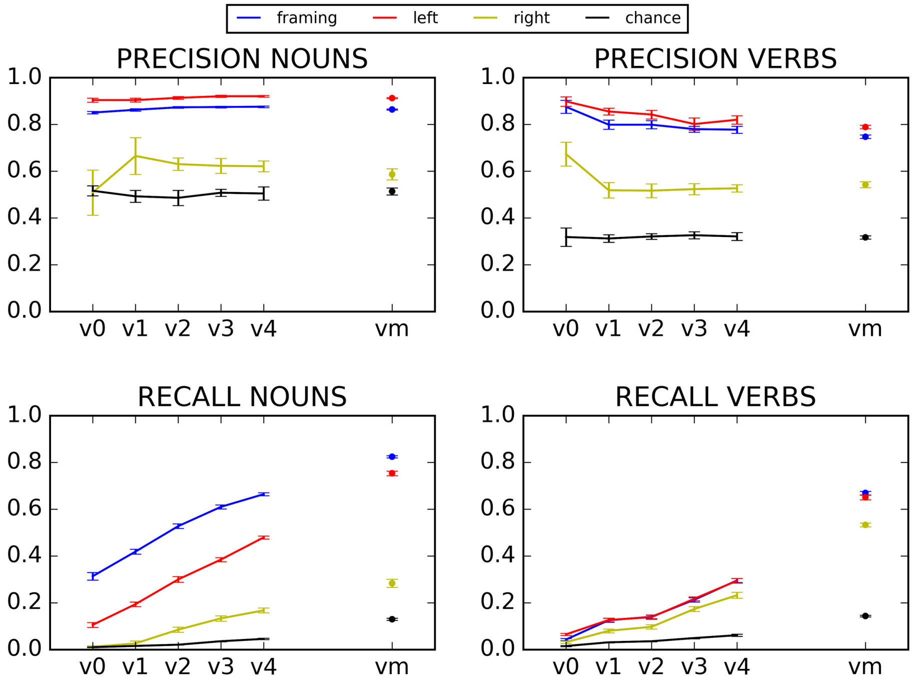


FIGURE 3 | Precision and recall for N and V, for various semantic seed sizes. Error bars represent the standard error of the mean of the ten different mini-corpora.

-0.097120, $t = -14.208$, $p < 0.001$; results for the interaction of semantic seed size and model type when comparing the baseline model and the framing model: $\beta = -0.074068$, $t = -10.836$, $p < 0.001$). The performance of these two models is very similar, with a small, nonsignificant advantage for the left model for noun and verb precision ($\beta = 4.416\text{e}{-02}$, $t = 1.178$, $p = 0.24$), and a rather large significant advantage of the framing model for the recall of nouns ($\beta = 0.028534$, $t = 2.952$, $p < 0.004$).

Finally, the framing and left context models exhibit a better precision for nouns than for verbs (although this does not reach significance, $\beta = -0.098109$, $t = -1.851$, $p = 0.05$ when we look at the interaction between the left model and the verb category and $\beta = -0.096205$, $t = -1.815$, $p = 0.07$ when we look at the interaction between the framing model and the verb category), and recall is also higher for nouns (significant difference when taking the framing model as a base level: $\beta = 0.356097$, $t = 13.377$, $p < 0.001$). This difference between nouns and verbs might come from the fact that the syntactic dependents of a noun are generally closer to their head than is the case for verbs [a similar advantage for nouns over verbs was observed in Bannard et al. (2009), in a model of young children's productions]. This is also consistent with the developmental literature, since nouns are typically understood and produced earlier than verbs (Gleitman, 1990; Waxman and Markov, 1995; Gentner, 2006; Bergelson and Swingley, 2012, 2013). It should also be noted that precision varies slightly more for verbs than for nouns (larger error bars for verbs for the framing and left context models), this is probably due to the lower recall for verbs (lower recall is caused by less hits and variance increases for lower numbers). Furthermore, the category of verbs is more heterogenous than the one of nouns: typically, we can describe a verb as intransitive, transitive, ditransitive, modal, stative, dynamic, etc. The syntactic selection of these different types of verbs influences the context they appear in. The variety inside the class of verbs and the low number of verbs in the smallest sizes of the semantic seed can also explain why the precision of verbs decreases a bit with the growth of the semantic seed throughout our experiences (although not significantly, as stated above). Because the smaller semantic seeds are only composed of 1, 2 or 3 verbs, these verbs might lead to more homogenous contexts than when more verbs are added.

Error Analysis of Misses

Since the recall was low for the smallest semantic seed, there were many misses: this is the reason why we focused our analysis of the model errors on the misses. The very high precision, on the other hand, means that false alarms were very rare. Our study of misses allows us to investigate what our model predicts when it should predict "*N*" or "*V*" and fails to do so.

Figure 4 presents the misses of the left model with the smallest vocabulary size (V0)[3]. The graphic on the left represents the noun misses (cases where the test corpus contained a noun, and something else than "*N*" was predicted). In **Figure 4**, we group together the different responses given instead of "*N*." Since the model could give as response either "*N*," "*V*," or a specific wordform (e.g., *giraffe, slowly, carry, not*...), we classified the errors that involved specific wordforms using classical categories:

[3] We chose the left model because it shows the best performance in terms of precision. We chose the V0 vocabulary because the number of misses is the highest as small semantic seeds lead to the lowest recall.

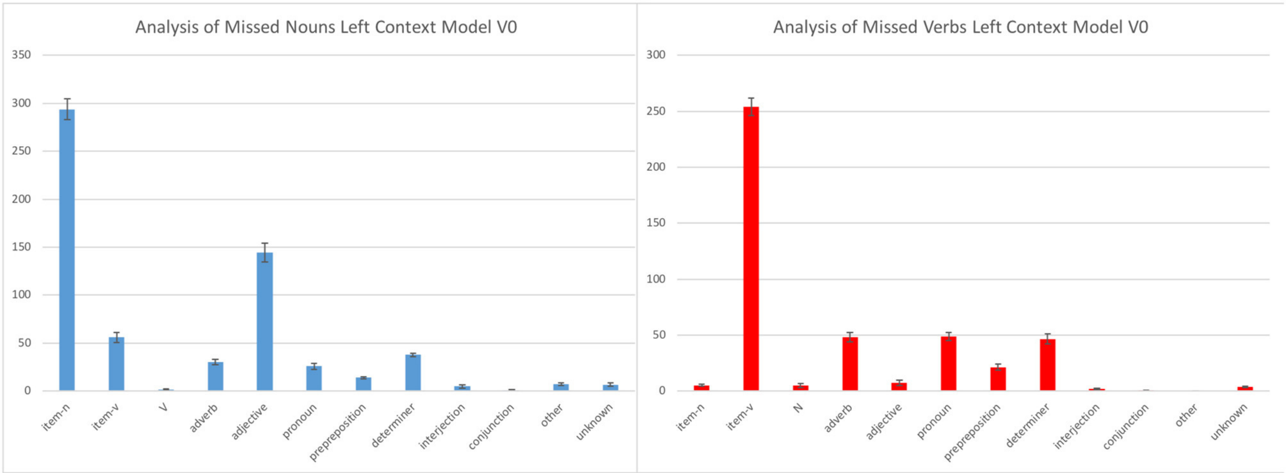


FIGURE 4 | Misses of the left context model with the smallest vocabulary size.

item-N and *item*-V for specific nouns and verbs (to distinguish them from the N and V categories built around the semantic seed), and adjective, adverb, pronoun, preposition, etc. for all other specific wordforms. The graphic on the right gives the corresponding results for verbs misses.

The most common type of miss is the prediction of a specific item of the correct category ("item-N" for nouns and "item-V" for verbs), which means that the model confuses specific items with their actual category. Developmentally, this type of error has the least negative impact for an infant. As can be expected, the number of such errors decreases with the number of verbs and nouns in the semantic seed[4] (congruent with the fact that the recall increases with the size of the vocabulary).

The other types of misses are much less frequent. When the model misses a noun and does not predict a specific noun item, its answer is most of the time an item of the adjective category. This is perfectly plausible, as a lot of frequently used adjectives in French are placed in between determiners and nouns. For example, when we have a context such as *"voit le _"* (*"sees the _"*), the word in the gap could perfectly be an adjective as well as a noun: *"voit le petit lapin"* (*"sees the little rabbit"*). The misses that are caused by *"item-V"* can also be explained by some specific contexts, such as the *"veut le _"* context (*"wants the"* or *"wants to _ him"*): it can be followed by a verb, as for instance in *"Marie veut le caresser* (*'Mary wants to pet him'*), or by a noun, as in *'Marie veut le poney'* (*'Mary wants the poney"*).

When a miss is recorded for a verb and the model does not predict a specific verb its answer is most of the time an adverb, a pronoun or a determiner. As for the misses for nouns, these guesses can be explained by contexts that can also receive these categories, such as *"Marie veut _"* (*"Mary wants _"*). It can be completed by either a verb, an adverb, a determiner, or a pronoun: *"Marie veut* ***danser****"* (*"Mary wants* ***to dance****"*), *"Marie veut* ***bien*** *danser"* (*"Mary would* ***gladly*** *dance"*), *"Marie veut* ***un*** *poney"* (*"Mary wants* ***a*** *poney"*), *"Marie veut* ***le*** *caresser"* (*"Mary wants to pet* ***him"***).

Frequently Used Contexts

In this subsection, we examined the contexts most frequently used by the left-context model to classify noun and verb targets. The qualitative study of these contexts helped us understand why the model performs well and what its pitfalls are.

The contexts are represented in **Table 4**. In each subtable, the first column gives the most frequently used contexts (ordered by decreasing frequency), the second one the translation, the third and fourth ones the number of times the model used this specific context during the test (2 columns giving the number of times this context was followed by a noun or by a verb) and finally the answer chosen by the model whenever it encountered this context. Thus, an *"N"* in the last column of the first table, along with a large number in the fourth column is evidence that the model gives a correct answer most of the time. For example, for the *"{ un _"* (*"{ a _"*) context, which is the most frequent context used by the model when categorizing nouns, out of the 179 encounters of this context, it was followed by a noun 170 times in the test corpus, and only once by a verb (the remaining times it was followed by something else, adjectives or adverbs). Since the model predicted *"N"* whenever it encountered this context, this means that it gave a correct answer 170 times, and a false alarm for the noun category 9 times. The same reasoning applies to the verb contexts.

We can note that the 20 most frequently used contexts for *"N"* all include at least one function word; more specifically the 19 most frequently used contexts contain a determiner. This is potentially not surprising given the crucial role played by function words in grammatical structure; yet no concept of function word was built in our model, let alone a concept of determiner. This means that the sheer frequency of function words, together with their distributional properties, were sufficient to make function words a key ingredient for the efficient

[4] The data for the other semantic seeds can be found in the GitHub repository.

TABLE 4 | Most frequent contexts used by the left context model during categorization, with a maximal projection (Vm).

Context	Translation	Number of Uses	Target = N	Target = V	Answer from Model
Most Frequently Used Contexts Used to Predict Noun Targets					
{ un	{ a	179	170	1	N
est un	is a	144	135	1	N
{ le	{ the	133	124	3	N
{ une	{ a	121	107	3	N
dans la	in the	105	101	3	N
de la	from the	109	101	5	N
{ les	{ the	103	97	2	N
{ la	{ the	93	88	2	N
est le	is the	92	83	0	N
dans le	in the	79	79	0	N
est une	is a	89	79	4	N
un petit	a little	79	78	0	N
à la	to the	78	75	0	N
sur le	on the	67	64	0	N
{ des	{ some	62	59	0	N
sur la	on the	56	54	0	N
est la	is the	66	49	2	N
à l'	to the	49	46	3	N
le petit	the little	44	44	0	N
c' est	it is	333	43	42	pas (*not*)
Most Frequently Used Contexts Used to Predict Verb Targets					
{ tu	{ you	323	41	258	V
{ on	{ we	133	12	110	V
tu as	you have	143	36	93	V
on va	we are going to	87	0	75	V
{ il	{ he	83	7	67	V
il est	he is	121	7	67	pas (*not*)
{ ça	{ it	86	8	62	V
que tu	that you	75	11	52	V
tu veux	you want	55	1	50	V
{ je	{ I	50	0	44	V
tu me	you me$_{\text{(direct object)}}$	51	6	43	V
c' est	it is	333	43	42	pas (*not*)
qu' on	that we	59	2	42	V
tu le	you it$_{\text{(direct object)}}$	52	12	40	V
tu te	you yourself$_{\text{(direct object)}}$	47	6	40	V
tu vas	you are going to	49	0	40	V
je te	I you$_{\text{(direct object)}}$	41	2	36	V
tu t'	you yourself$_{\text{(direct object)}}$	33	0	33	V
on le	we it$_{\text{(direct object)}}$	40	9	29	V
{ elle	{ she	36	1	29	est (*is*)

discovery of the noun category. We find a similar situation for verbs, where this time the most useful cues are pronouns, which occur in 20 contexts out of 20.

It is interesting to note that the most frequent contexts for verb targets also feature some contexts predicting the negation particle *"pas."* Indeed, in French, this small word is often considered as belonging to the category of adverbs, but is placed in the same position as a verb when we only consider the two-word context to the left, especially since in natural speech the pre-verb particle "ne" is often dropped ("Je veux pas" *I don't want* vs. "Je veux manger" *I want to eat*). The fact that the left context model predicts *"pas"* for some very frequent contexts during the maximal vocabulary (Vm) experiment, explains (partly) why a hundred percent recall is not reached even in this condition.

Furthermore, these contexts show why the precision for nouns is higher than for verbs. When we look at the number of verb targets among the contexts that are used most frequently to predict nouns, we globally observe a lower number than when we look at the noun targets among the contexts that are used most frequently for the prediction of verbs. Indeed, most of the time, a context such as *"tu as _"* is followed by a verb. However, about a third of the *"tu as _"* contexts are followed by a noun (for example in *"tu as faim";* literally, "*you have hunger*," meaning *"you are hungry"*). Nevertheless, the model classifies all targets in this context as a verb, leading to 36 false alarms in this case and thus to a lower precision for verbs than for nouns.

DISCUSSION

We presented a learning mechanism aiming to explain the formidable ability of infants to guess the probable meaning of unknown words by using their syntactic contexts. To do this we implemented a computational model that aims at categorizing nouns and verbs on the basis of their local contexts. Our algorithm is driven by frequency and expectation. We compared three different types of contexts and showed that both left and framing contexts were effective, whereas the right context gave poor information to predict categories. Overall, this model demonstrates that relying on local contexts and on a semantic seed is an efficient and simple method that may allow children to learn which contexts correspond to nouns, and which to verbs, as demonstrated with infants in several psycholinguistic experiments (Cauvet et al., 2014; Shi, 2014; Brusini et al., 2017; Babineau et al., 2020).

This model rests on two assumptions, that we argue are highly plausible. First, infants are supposed to be able to build a semantic seed. The semantic seed is a handful of words for which infants have succeeded in learning a meaning (frequent words, referring to concrete objects and actions, presented in pragmatically helpful situations), and that they are able to group together: a small number of known object-referents to form a proto-category of nouns and a few known action-referents to form a proto-category of verbs (Carey, 2009). Second, the model rests on the assumption that infants keep track of bi- and tri-gram frequencies, a hypothesis supported by many experiments (e.g., Santelmann and Jusczyk, 1998; Gomez and Gerken, 1999; Höhle et al., 2006; van Heugten and Johnson, 2010; Marchetto and Bonatti, 2013). The number of nouns and verbs supposedly known is very low: only 8 nouns and 1 verb at the smallest size of the semantic seed, a vocabulary which might plausibly be known by infants around the age of 10–12 months. Bergelson and Swingley (2012, 2013) present data suggesting that 10–13-month-olds already know 2 verbs, while 9-month-olds already know 10 nouns, rendering our initialization hypothesis highly plausible. We showed here that as soon as infants are able to group known words on semantic grounds, the use of local contexts is highly efficient to spread these proto-categories to many unknown words. We suspect that such a mechanism would be just as efficient on the learning of syntactic categories other than nouns or verbs: whenever there is a link between a semantic feature and a local morphosyntactic context, young children could rely on the local contexts to spread this semantic feature to other, yet unknown, words. Consistent with this hypothesis, a large-scale cross-linguistic study of the kind of semantic features that are commonly encoded in morphosyntax revealed that these correspond to core knowledge distinctions, that are perceived very early by infants (e.g., the mass/count distinction, or animate/inanimate, Strickland, 2017). Our experiments demonstrate the interest of computational approaches in developmental and cognitive science, as the models we built allowed us to evaluate different cognitive mechanisms in an efficient manner and confront their outcomes with results from experimental work. The model possesses two important characteristics that make it particularly attractive as a model of early lexical acquisition: the efficiency of the semantic seed, and the fact that it categorizes words in context. As we saw above, the semantic seed is highly plausible, and it is also highly efficient: even at the smallest size of the semantic seed, the model already achieves an excellent precision, both for nouns and for verbs. Unsupervised learning algorithms seeded with semantic information have been presented before in the computational linguistic literature (to solve other problems), with excellent results (Yarowsky, 1995). Arguably, we can oppose that the method presented here is not a complete mechanism for bootstrapping the nouns and verbs categories. Indeed, the models we presented here do not use the words they managed to categorize in order to expand their semantic seed to learn even more categorizing contexts, something we would expect real learners to be able to achieve.

The second important characteristic of the model is that it categorizes words in context. It does not attempt to build a "mental dictionary," a list of word-forms, where each word-form would be assigned a syntactic category—or several possible ones for each possible meaning. Instead, the model categorizes words solely on the basis of their immediate context (whenever it is sufficiently informative). This feature buys the model two important advantages: first, novel words, that are encountered for the first time, can be categorized (provided they occur in a known context). This is important as it means that a child could deduce the category of a word she/he heard for the first time and use it to guess the meaning of the novel word, as has been observed in many infant experiments (Bernal et al., 2007; Waxman et al., 2009; Oshima-Takane et al., 2011; He and Lidz, 2017; de Carvalho et al., 2019). Second, the model does not suffer from the fact that many words possess more than one syntactic category, in fact, it does not even notice such cases. This particular aspect of the model's behavior is also consistent with recent experimental work testing how toddlers handle homophones: not only do 20-month-olds understand noun-verb homophones in their native language (Veneziano and Parisse, 2011; de Carvalho et al., 2017), they are also willing to learn a novel meaning for a word-form they already know (e.g., "to give"), provided that the novel word appears in a context that would be inappropriate for the known meaning, e.g., it belongs to a different syntactic category (e.g., they can taught that *a give* is the name of a novel animal; Dautriche et al., 2015, 2018).

These two characteristics, that mesh well with the developmental literature on word learning in infants, gives a real plausibility boost in favor of the present model, compared to previous work relying on local contexts for categorization, at least at the earliest stages of learning. For example, the Redington et al.'s model yielded fine-grained syntactic categories (much more precise than simply noun vs. verb), but attempted categorization only on the most frequent words of the corpus, the words that a child would have heard many times in her input. As a result, this model would not even have attempted to categorize a word on first encounter. Since it builds a diagram of similarities between word-forms, it also ignores word homophony and falls back on assigning to each word-form the syntactic category that is most frequent, at the risk of confusion (e.g., a ring, to ring). One might think that the two approaches could be usefully combined by children: on one hand, an on-line categorization approach based on immediate context, as in the present model, could provide infants with a first hint as to the possible meaning of a word (even on first encounter); on the other hand, the fine-grained categorization provided by the analysis of a large number of contexts (as implemented in Redington et al., 1998) could give slightly older children more precise information about a word's meaning, which could be especially helpful for acquiring the meaning of verbs (Gleitman, 1990; Naigles, 1990; Yuan and Fisher, 2009; Arunachalam and Waxman, 2010), or of some other more abstract words (e.g., quantifiers, preposition, etc., see Waxman and Lidz, 2006).

The present model also improves over the Frequent Frames model proposed by Mintz (2003), from which it was partly inspired. The Frequent Frames model also aligns with developmental data and has the capacity to categorize a word on first encounter, provided the context is known (indeed, this characteristic was borrowed from the Frequent Frames model). Its main drawback is the fact that it builds several classes for each syntactic category: for instance, the frames *"the _ is"* and *"a _ is"* both select nouns. The present model escapes this difficulty through seeding the categorization process with a few known words, which are categorized precisely because we supposed their meaning known (objects and actions). Not surprisingly, adding more information in the input yields a better performance in the end.

The *post-hoc* analysis of the most frequently used contexts demonstrated that the efficiency of the model is in a great part due to function words. These words play an important linguistic role in the structure of sentences. Many experiments have demonstrated that infants notice these words early in development, thanks to their acoustic and distributional characteristics (Shady, 1996; Shafer et al., 1998; Shi et al., 1998, 2006a,b; Shi and Lepage, 2008). Then, from around 14–18 months of age, infants can use them to build expectations about novel words (Bernal et al., 2007; Shi and Melançon, 2010; Brusini et al., 2016; Babineau et al., 2020). Here, the algorithm used by the model did not attribute any specific role to these words, but their frequency and their natural pertinence regarding the categorization task enhanced their role naturally. This alignment between what we know of toddlers processing of function words, and the way they are used by our model, confirms its developmental plausibility regarding the acquisition of the noun and verb categories. Additionally, the results presented here also show that it is not necessary to form categories of function words, such as determiner or pronoun, to be able to use them to predict nouns and verbs. The idea that children group function words together into categories is rather intuitive (Shi and Melançon, 2010) but remains disputed (Pine and Martindale, 1996; Valian et al., 2009; Pine et al., 2013; Yang, 2013). Here, we demonstrated that this step is in fact unnecessary. The simple knowledge of the phonological form of the function words could be enough to bootstrap the growth of content word categories. Here, we see how the use of modeling work enlightens current developmental hypotheses.

For our research, we compared three types of context: left, right and framing. We found that the left context leads to the best precision. Two hypotheses might be proposed to explain why. The first is that many of the most frequently-used contexts (see **Table 4**) include a marker of the beginning of the sentence. Indeed, a determiner such as *"le"* or *"la"* ("the") is homophonous with clitic object pronouns in French (*"him/her"*). Knowing that *"le"* or *"la"* ("the") is placed at the beginning of the sentence gives crucial information that the function word is a determiner and consequently likely to be followed by a noun (or an adjective). Another explanation for the better performance of left contexts would be that French, like English, is mostly right-branching: there is a large number of syntactic phrases in which the head is at the beginning (right-branching phrases are also called head-initial phrases). Since heads are by definition words that constrain the category of the phrase and the nature of their dependents, it can be expected that finding the head at the left edge of the phrase is very informative, and, accordingly, that in general words located on the right of the target will be much less informative. Since French comprises both left-branching and right-branching structures (albeit skewed in favor of right-branching ones) it might favor both left and framing contexts. If this analysis is correct, we expect that we would get different results for languages in which the distribution of left-branching and right branching structures is different. In this respect, it would be interesting to do the same study with a language such as Japanese, which is well-known to be almost fully left-branching.

Despite all the qualities of the *semantic seed* model, the way it is currently implemented, it possesses several characteristics that lack psychological plausibility: (1) it has a perfect memory; (2) it has no way of increasing its vocabulary of known words; and (3) it works from an input segmented into words. We think that none of these aspects are crucial for the good performance of the model, and that each could be modified to make it more plausible (and perhaps even further improve its performance). We will discuss each of these in turn. First, as currently implemented, the model never forgets any of the word triplets presented during training, thus assuming perfect memory on the part of the infant (which is clearly undesirable). However, since the model's performance relied on those word triplets which had been encountered most frequently, it should be possible to incorporate a forgetting mechanism through which triplets which have been encountered only a few times (in a to-be-defined number of utterances) are forgotten. This would probably not impact the performance too dramatically (as an aside, most models suppose perfect memory to test the feasibility of a method; e.g., Redington et al., 1998).

Second, the model currently has no way to increase its vocabulary. It starts out with a small initially known vocabulary (the semantic seed), memorizes word triplets from the training corpus, then uses these to categorize content words. Ideally, the model should be able to rely on its high precision to learn from its own predictions a new set of newly-learned words, perhaps with a simple threshold of confidence (although we should note that real learners would presumably exploit the categorizing that they performed in order to learn something about the semantics of the words they categorized, before adding them to their semantic seed). In that way, the model could perhaps start out with the smallest semantic seed (which already demonstrates a high precision), and increase the number of words it categorizes, namely the recall, by accumulating new contexts, precisely the ones it can extract thanks to the newly-learnt words. Thus, the model could start with as little as 8 nouns and 1 verb, and categorize many more words in an iterative fashion.

Third, the model takes as input a transcribed corpus (like all other computational models attempting to categorize lexical items so far), and it therefore assumes that the continuous speech stream is segmented into words. This is a reasonably plausible assumption in light of the many experiments showing that infants already possess rather refined word-segmentation abilities within their first 18 months of life (Jusczyk and Aslin, 1995; Gout et al., 2004; Nazzi et al., 2005, 2006; Fló et al., 2019), although we do not know when exactly children might have access to an adult-like segmentation of speech (Ngon et al., 2013). Future work should ideally attempt to start from an unsegmented input and adopt a plausible word-segmentation strategy as a first step (Johnson et al., 2015). Last, a final improvement of the model could be to use grammatical categories with maximal cognitive plausibility. In the present experiments, we chose to work with the noun and verb categories for three reasons. First, the experimental literature reviewed in the introduction shows that 18-month-olds are able to exploit local contexts to map nouns to objects and verbs to actions (e.g., He and Lidz, 2017). Second, and this is a practical reason, nouns and verbs can be identified by off-the-shelf part-of-speech taggers. Third, these categories seem to be generally present cross-linguistically. However, we are well aware that these categories are not necessarily universal (Feng et al., 2020), and definitely not homogeneous. The verb category is an ideal example of that: verbs can be divided in numerous subcategories for which children have some sensitivity, for example 1-participant action verbs vs. 2-participants action verbs (Yuan and Fisher, 2009).

More generally, we think that the mechanism tested in our model would be relevant for any categories, not just nouns and verbs: namely, using known content words to learn about the contexts they appear in, then, whenever a novel content word is encountered, using these contexts to project some of the properties of the known content words on the novel content word. For instance, some languages implement specific morphology for the animate/inanimate distinction, mass/count, human/non-human, and so on. Infants learning these languages could exploit these markers to narrow down their hypotheses about the meaning of words occurring in these contexts. Consistent with this hypothesis, a large-scale cross-linguistic study of the kind of semantic features that are commonly encoded in morphosyntax revealed that these correspond to *core knowledge* distinctions (Spelke, 2000), that are perceived very early by infants (e.g., the mass/count distinction, or animate/inanimate, Strickland, 2017). One possible interpretation for this fact is the idea that languages are shaped by the generations of children who acquire them (e.g., Christiansen and Chater, 2008): indeed, morphosyntactic markers that encode semantic distinctions that are relevant and salient for infants (*core knowledge* distinctions), will both be learned more easily, and make language learning easier for infants, since they will be able to exploit these markers to rapidly guess the possible meaning of novel words. This is consistent with many modeling studies showing that natural languages are shaped by acquisition and processing constraints (e.g., Piantadosi et al., 2011; Dautriche et al., 2017), as well as with models of language emergence (e.g., Kirby et al., 2008; Gong, 2011).

Notwithstanding the implementation limitations that we raised above, the model can already be used to make predictions regarding the acquisition of novel words, and these predictions can be experimentally tested in children: For instance, by using well-known words to teach them novel syntactic contexts in their native language, and seeing whether they would be ready to rely on those newly-learnt contexts to categorize novel content words (into object-referents vs. action-referents, for instance). This is precisely what Babineau et al. (2021) did in a recent experiment, teaching two groups of 3- to 4-year-olds a novel function word "ko," in French; in half the children, "ko" replaced all determiners, and preceded well-known nouns and adjectives (e.g., *ko rabbit*, *ko little chicken*), in a video where a speaker was playing with toys and telling a story; the other half of the children watched the same video, in which "ko" replaced all personal pronouns, and preceded verbs and auxiliaries (e.g., *ko plays*, *ko will jump*). At test, all children were presented with a choice of 2 videos, one exhibiting a novel object, and the other one a novel action, while they heard "Regarde! Ko bamoule!" (*look! Ko bamoule*). The results showed that children who had heard "ko" in the position of personal pronouns looked more at the novel action than children who had heard "ko" in the position of determiners, who looked more at the novel object. These results thus suggest that young children, just like the model, are able to exploit content words they already know, in order to learn some of the properties of novel function words, then use these novel function words to guess the probable meaning of an unknown content word (*bamoule*). Although this experiment was performed with rather "old" children (3–4-year-olds) and should be replicated with younger children, it already is a very encouraging confirmation of the main hypothesis behind the model.

CONCLUSION

The computational model presented here clearly shows the relevance of local contexts to categorize nouns and verbs in sentences. Two crucial characteristics of the current model make it particularly relevant to describe lexical acquisition during infancy. The *semantic seed*—minimal information regarding a handful of known words, grouped into object-referents and action-referents—allows it to group words together with very high precision, even for words that are encountered for the first time (provided they occur in known contexts). And the fact that the model categorizes words in context neatly bypasses the potential difficulties posed by homophones—in this case, noun/verb homophones, which are frequent in many languages. It is noteworthy that, just like adult speakers, toddlers seem to be completely impervious to homophones, not even noticing them: our model behaves in just the same way. Importantly, *any* semantic feature that has a realization in language, can be identified by infants and has the potential to be generalized in that way. The present model thus exhibits a plausible mechanism through which toddlers could succeed in learning about the contexts of nouns and verbs in their native language—knowledge which we know they possess from 18 months on—and perhaps, more generally, could be extended to learning the contexts of more fine-grained categories (such as different subclasses of verbs, adjectives, animates etc.).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://phonbank.talkbank.org/access/French/Lyon.html.

ETHICS STATEMENT

Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

PB elaborated the idea for this model and conducted a first series of experiments with it for her PhD-thesis at the Laboratoire de Sciences Cognitives et Psycholinguistique under the supervision of AC and PA. OS conducted a second series of experiments for her Master thesis at Université Paris Diderot under the supervision of AC and PA. She recoded the experiments and enhanced Brusini's model. All four authors contributed to the redaction of the manuscript. The figures, tables, statistic analyses, and computer code were elaborated by OS. All authors contributed to the article and approved the submitted version.

FUNDING

This researched was funded by the Fondation de France, the Ecole Normale Supérieure, the Laboratoire de Sciences Cognitives, the Laboratoire de Linguistique Formelle (UMR 7110) and the following grants from l'Agence Nationale de la Recherche: ANR-13-APPR-0012 LangLearn, ANR-17-CE28-0007-01 LangAge, and ANR-17-EURE-0017 FrontCog. PB was supported by the ESRC International Centre for Language and Communicative Development (LuCiD) while writing the last version of this manuscript [ES/L008955/1].

ACKNOWLEDGMENTS

We thank Emmanuel Chemla for his help and insights with the first versions of our computational model.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.661479/full#supplementary-material

REFERENCES

Akhtar, N., Carpenter, M., and Tomasello, M. (1996). The role of discourse novelty in early word learning. *Child Dev.* 67, 635–645. doi: 10.1111/j.1467-8624.1996.tb01756.x

Arunachalam, S., and Waxman, S. R. (2010). Meaning from syntax: evidence from 2-year-olds. *Cognition* 114, 442–446. doi: 10.1016/j.cognition.2009.10.015

Babineau, M., Carvalho, A., Trueswell, J., and Christophe, A. (2021). Familiar words can serve as a semantic seed for syntactic bootstrapping. *Dev. Sci.* 24:e13010. doi: 10.1111/desc.13010

Babineau, M., Shi, R., and Christophe, A. (2020). 14-month-olds exploit verbs' syntactic contexts to build expectations about novel words. *Infancy* 25, 719–733. doi: 10.1111/infa.12354

Bannard, C., Lieven, E., and Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proc. Natl. Acad. Sci. U.S.A.* 106, 17284–17289. doi: 10.1073/pnas.0905638106

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

Bates, D. M., and Sarkar, D. (2007). *lme4: Linear Mixed-effects Models Using S4 Classes.* Available online at: http://cran.r-project.org/ (accessed July 25, 2021).

Bergelson, E., and Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proc. Natl. Acad. Sci. U.S.A.* 109, 3253–3258. doi: 10.1073/pnas.1113380109

Bergelson, E., and Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition* 127, 391–397. doi: 10.1016/j.cognition.2013.02.011

Bergelson, E., and Swingley, D. (2015). Early word comprehension in infants: replication and extension. *Lang. Learn. Dev.* 11, 369–380. doi: 10.1080/15475441.2014.979387

Berko, J. (1958). The child's learning of english morphology. *Psycholinguistics: A Book of Readings,* Holt: Rinehart & Winston, 150–177.

Bernal, S., Lidz, J., Millotte, S., and Christophe, A. (2007). Syntax constrains the acquisition of verb meaning. *Lang. Learn. Dev.* 3, 325–341. doi: 10.1080/15475440701542609

Brusini, P., Dehaene-Lambertz, G., Dutat, M., Goffinet, F., and Christophe, A. (2016). ERP evidence for on-line syntactic computations in 2-year-olds. *Dev. Cogn. Neurosci.* 19, 164–173. doi: 10.1016/j.dcn.2016.02.009

Brusini, P., Dehaene-Lambertz, G., van Heugten, M., de Carvalho, A., Goffinet, F., Fiévet, A., et al. (2017). Ambiguous function words do not prevent 18-month-olds from building accurate syntactic category expectations: an

ERP study. *Neuropsychologia* 98, 4–12. doi: 10.1016/j.neuropsychologia.2016.08.015

Carey, S. (2009). The origin of concepts. *The Origin of Concepts*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195367638.001.0001

Cauvet, E., Limissuri, R., Millotte, S., Skoruppa, K., Cabrol, D., and Christophe, A. (2014). Function words constrain on-line recognition of verbs and nouns in French 18-month-olds. *Lang. Learn. Dev.* 10, 1–18. doi: 10.1080/15475441.2012.757970

Chemla, E., Mintz, T. H., Bernal, S., and Christophe, A. (2009). Categorizing words using 'frequent frames': what cross-linguistic analyses reveal about distributional acquisition strategies. *Dev. Sci.* 12, 396–406. doi: 10.1111/j.1467-7687.2009.00825.x

Christiansen, M. H., and Chater, N. (2008). Language as shaped by the brain. *Behav. Brain Sci.* 31, 489–509. doi: 10.1017/S0140525X08004998

Chrupała, G., and Alishahi, A. (2010). "Online entropy-based model of lexical category acquisition," in *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, 182–191.

Dautriche, I., Cristia, A., Brusini, P., Yuan, S., Fisher, C., and Christophe, A. (2014). Toddlers default to canonical surface-to-meaning mapping when learning verbs. *Child Dev.* 85, 1168–1180. doi: 10.1111/cdev.12164

Dautriche, I., Fibla, L., Fievet, A., and Christophe, A. (2018). Learning homophones in context: easy cases are favored in the lexicon of natural languages. *Cogn. Psychol.* 104, 83–105. doi: 10.1016/j.cogpsych.2018.04.001

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., and Piantadosi, S. T. (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition* 163, 128–145. doi: 10.1016/j.cognition.2017.02.001

Dautriche, I., Swingley, D., and Christophe, A. (2015). Learning novel phonological neighbors: syntactic category matters. *Cognition* 143, 77–86. doi: 10.1016/j.cognition.2015.06.003

de Carvalho, A., Dautriche, I., Fiévet, A., and Christophe, A. (2021). Toddlers exploit referential and syntactic cues to flexibly adapt their interpretation of novel verb meanings. *J. Exp. Child Psychol.* 203, 105017. doi: 10.1016/j.jecp.2020.105017

de Carvalho, A., Dautriche, I., Lin, I., and Christophe, A. (2017). Phrasal prosody constrains syntactic analysis in toddlers. *Cognition* 163, 67–79. doi: 10.1016/j.cognition.2017.02.018

de Carvalho, A., He, A. X., Lidz, J., and Christophe, A. (2019). Prosody and function words cue the acquisition of word meanings in 18-month-old infants. *Psychol. Sci.* 30, 319–332. doi: 10.1177/0956797618814131

Demuth, K., and Tremblay, A. (2008). Prosodically-conditioned variability in children's production of French determiners. *J. Child Lang.* 35, 99–127. doi: 10.1017/S0305000907008276

Feng, S., Qi, R., Yang, J., Yu, A., and Yang, Y. (2020). Neural correlates for nouns and verbs in phrases during syntactic and semantic processing: an fMRI study. *J. Neurolinguistics* 53, 100860. doi: 10.1016/j.jneuroling.2019.100860

Ferguson, B., and Waxman, S. R. (2017). Linking language and categorization in infancy. *J. Child Lang.* 44, 527–552. doi: 10.1017/S0305000916000568

Ferry, A. L., Hespos, S. J., and Waxman, S. R. (2010). Categorization in 3- and 4-month-old infants: an advantage of words over tones. *Child Dev.* 81, 472–479. doi: 10.1111/j.1467-8624.2009.01408.x

Ferry, A. L., Hespos, S. J., and Waxman, S. R. (2013). Nonhuman primate vocalizations support categorization in very young human infants. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15231–15235. doi: 10.1073/pnas.1221166110

Ferry, A. L., Nespor, M., and Mehler, J. (2020). Twelve to 24-month-olds can understand the meaning of morphological regularities in their language. *Dev. Psychol.* 56, 40–52. doi: 10.1037/dev0000845

Fisher, C. (1996). Structural limits on verb mapping: the role of analogy in children's interpretations of sentences. *Cogn. Psychol.* 31, 41–81. doi: 10.1006/cogp.1996.0012

Fisher, C., Gertner, Y., Scott, R. M., and Yuan, S. (2010). Syntactic bootstrapping. *Wiley Interdiscip. Rev. Cogn. Sci.* 1, 143–149. doi: 10.1002/wcs.17

Fisher, C., Hall, D. G., Rakowitz, S., and Gleitman, L. (1994). When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *Lingua* 92, 333–375. doi: 10.1016/0024-3841(94)90346-8

Fló, A., Brusini, P., Macagno, F., Nespor, M., Mehler, J., and Ferry, A. L. (2019). Newborns are sensitive to multiple cues for word segmentation in continuous speech. *Dev. Sci.* 22:e12802. doi: 10.1111/desc.12802

Gentner, D. (2006). "Why verbs are hard to learn," in *Action Meets Word: How Children Learn Verbs*, eds K. Hirsh-Pasek and R. M. Golinkoff (Oxford: Oxford University Press), 544–564. doi: 10.1093/acprof:oso/9780195170009.003.0022

Gleitman, L. (1990). The structural sources of verb meanings. *Lang. Acquis.* 1, 3–55. doi: 10.1207/s15327817la0101_2

Gomez, R. L., and Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition* 70, 109–135. doi: 10.1016/S0010-0277(99)00003-7

Gong, T. (2011). Simulating the coevolution of compositionality and word order regularity. *Interac. Stud.* 12, 63–106. doi: 10.1075/is.12.1.03gon

Gout, A., Christophe, A., and Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access II. Infant data. *J. Memory Lang.* 51, 548–567. doi: 10.1016/j.jml.2004.07.002

Gutman, A., Dautriche, I., Crabb,é, B., and Christophe, A. (2015). Bootstrapping the syntactic bootstrapper: probabilistic labeling of prosodic phrases. *Lang. Acquis.* 22, 285–309. doi: 10.1080/10489223.2014.971956

Halle, P. A., Durand, C., and de Boysson-Bardies, B. (2008). Do 11-month-old French infants process articles? *Lang. Speech* 51, 23–44. doi: 10.1177/00238309080510010301

He, A. X., and Lidz, J. (2017). Verb learning in 14-and 18-month-old English-learning infants. *Lang. Learn. Dev.* 13, 335–356. doi: 10.1080/15475441.2017.1285238

Höhle, B., Schmitz, M., Santelmann, L. M., and Weissenborn, J. (2006). The recognition of discontinuous verbal dependencies by german 19-month-olds: evidence for lexical and structural influences on children's early processing capacities. *Lang. Learn. Dev*. 2, 277–300. doi: 10.1207/s15473341lld0204_3

Höhle, B., Weissenborn, J., Kiefer, D., Schulz, A., and Schmitz, M. (2004). Functional elements in infants' speech processing: the role of determiners in the syntactic categorization of lexical elements. *Infancy* 5, 341–353. doi: 10.1207/s15327078in0503_5

Johnson, M., Pater, J., Staubs, R., and Dupoux, E. (2015). "Sign constraints on feature weights improve a joint model of word segmentation and phonology," in *NAACL HLT 2015 −2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*. doi: 10.3115/v1/n15-1034

Jusczyk, P. W., and Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cogn. Psychol.* 29, 1–23. doi: 10.1006/cogp.1995.1010

Kedar, Y., Casasola, M., and Lust, B. (2006). Getting there faster: 18- and 24-month-old infants' use of function words to determine reference. *Child Dev.* 77, 325–338. doi: 10.1111/j.1467-8624.2006.00873.x

Kedar, Y., Casasola, M., Lust, B., and Parmet, Y. (2017). Little words, big impact: determiners begin to bootstrap reference by 12 months. *Lang. Learn. Dev*. 77, 325–328. doi: 10.1080/15475441.2017.1283229

Kirby, S., Cornish, H., and Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10681–10686. doi: 10.1073/pnas.0707835105

Landau, B., and Gleitman, L. R. (1985). *Language and Experience: Evidence from the Blind Child. Vol. 8*. Harvard: Harvard University Press. Available online at: https://psycnet.apa.org/record/1985-97756-000 (accessed July 25, 2021).

Lukyanenko, C., and Fisher, C. (2016). Where are the cookies? Two- and three-year-olds use number-marked verbs to anticipate upcoming nouns. *Cognition* 146, 349–370. doi: 10.1016/j.cognition.2015.10.012

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk: Volume I: Transcription Format and Programs*, *Volume II: The Database*. Cambridge, MA: MIT Press.

Marchetto, E., and Bonatti, L. L. (2013). Words and possible words in early language acquisition. *Cogn. Psychol.* 67, 130–50. doi: 10.1016/j.cogpsych.2013.08.001

Medina, T. N., Snedeker, J., Trueswell, J. C., and Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proc. Nat. Acad. Sci.* 108, 9014–9019. doi: 10.1073/pnas.1105040108

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition* 90, 91–117. doi: 10.1016/S0010-0277(03)00140-9

Naigles, L. (1990). Children use syntax to learn verb meanings. *J. Child Lang.* 17, 357–374. doi: 10.1017/S0305000900013817

Nazzi, T., Dilley, L. C., Jusczyk, A. M., Shattuck-Hufnagel, S., and Jusczyk, P. W. (2005). English-learning infants' segmentation of verbs from fluent speech. *Lang. Speech* 48, 279–298. doi: 10.1177/00238309050480030201

Nazzi, T., Iakimova, G., Bertoncini, J., Frédonie, S., and Alcantara, C. (2006). Early segmentation of fluent speech by infants acquiring French: Emerging evidence for crosslinguistic differences. *J. Mem. Lang.* 54, 283–299. doi: 10.1016/j.jml.2005.10.004

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., and Peperkamp, S. (2013). (Non)words, (non)words, (non)words: Evidence for a protolexicon during the first year of life. *Dev. Sci.* 16, 24–34. doi: 10.1111/j.1467-7687.2012.01189.x

Oshima-Takane, Y., Ariyama, J., Kobayashi, T., Katerelos, M., and Poulin-Dubois, D. (2011). Early verb learning in 20-month-old Japanese-speaking children. *J. Child Lang.* 38, 455–484. doi: 10.1017/S0305000910000127

Parise, E., and Csibra, G. (2012). Electrophysiological evidence for the understanding of maternal speech by 9-month-old infants. *Psychol. Sci.* 23, 728–733. doi: 10.1177/0956797612438734

Parisien, C., Fazly, A., and Stevenson, S. (2008). "An incremental Bayesian model for learning syntactic categories," in *CoNLL 2008 - Proceedings of the Twelfth Conference on Computational Natural Language Learning*. doi: 10.3115/1596324.1596340

Piantadosi, S. T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proc. Natl. Acad. Sci. U.S.A.* 108, 3526–3529. doi: 10.1073/pnas.1012551108

Pine, J. M., Freudenthal, D., Krajewski, G., and Gobet, F. (2013). Do young children have adult-like syntactic categories? Zipf's law and the case of the determiner. *Cognition* 127, 345–360. doi: 10.1016/j.cognition.2013.02.006

Pine, J. M., and Martindale, H. (1996). Syntactic categories in the speech of young children: the case of the determiner. *J. Child Lang.* 23, 369–395. doi: 10.1017/s,0305000900008849

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Available online at: http://www.r-project.org/ (accessed July 25, 2021).

Redington, M., Chater, N., and Finch, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cogn. Sci.* 22, 425–469. doi: 10.1207/s15516709cog2204_2

Santelmann, L. M., and Jusczyk, P. W. (1998). Sensitivity to discontinuous dependencies in language learners: evidence for limitations in processing space. *Cognition* 69, 105–134. doi: 10.1016/S0010-0277(98)00060-2

Saxe, R., Tzelnic, T., and Carey, S. (2006). Five-month-old infants know humans are solid, like inanimate objects. *Cognition* 1, B1–B8. doi: 10.1016/j.cognition.2005.10.005

Seidenberg, M. S., and MacDonald, M. C. (1999). A Probabilistic Constraints Approach to Language Acquisition and Processing. *Cogn. Sci.* 23, 569–588. doi: 10.1207/s15516709cog2304_8

Shady, M. (1996). *Children' Sensitivity to Function Morphemes*. Buffalo, NY: State University of New York.

Shafer, V. L., Shucard, D. W., Shucard, J. L., and Gerken, L. A. (1998). An electrophysiological study of infants' sensitivity to the sound patterns of english speech. *J. Speech Lang. Hear. Res.* 41, 874–886. doi: 10.1044/jslhr.4104.874

Shi, R. (2014). Functional morphemes and early language acquisition. *Child Dev. Perspect.* 8, 36–41. doi: 10.1111/cdep.12052

Shi, R., Cutler, A., Werker, J., and Cruickshank, M. (2006a). Frequency and form as determinants of functor sensitivity in English-acquiring infants. *J. Acoust. Soc. Am.* 119, EL61–EL67. doi: 10.1121/1.2198947

Shi, R., and Lepage, M. (2008). The effect of functional morphemes on word segmentation in preverbal infants. *Dev. Sci.* 11, 407–413. doi: 10.1111/j.1467-7687.2008.00685.x

Shi, R., and Melançon, A. (2010). Syntactic categorization in French-learning infants. *Infancy* 15, 517–533. doi: 10.1111/j.1532-7078.2009.00022.x

Shi, R., Morgan, J. L., and Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: a cross-linguistic perspective. *J. Child Lang.* 25, 169–201. doi: 10.1017/S0305000997003395

Shi, R., Werker, J. F., and Cutler, A. (2006b). Recognition and representation of function words in English-learning infants. *Infancy* 10, 187–198. doi: 10.1207/s15327078in1002_5

Spelke, E. S. (2000). Core knowledge. *Am. Psychol.* 55, 1233–1243. https://psycnet.apa.org/buy/2000-14050-006

Strickland, B. (2017). Language reflects "core" cognition: a new theory about the origin of cross-linguistic regularities. *Cogn. Sci.* 41, 70–101. doi: 10.1111/cogs.12332

Syrnyk, C., and Meints, K. (2017). Bye-bye mummy—word comprehension in 9-month-old infants. *Br. J. Dev. Psychol.* 35, 202–217. doi: 10.1111/bjdp.12157

Taxitari, L., Twomey, K. E., Westermann, G., and Mani, N. (2020). The limits of infants' early word learning. *Lang. Learn. Dev.* 16, 1–21. doi: 10.1080/15475441.2019.1670184

Tomasello, M., and Akhtar, N. (1995). Two-year-olds use pragmatic cues to differentiate reference to objects and actions. *Cogn. Dev.* 10, 201–224. doi: 10.1016/0885-2014(95)90009-8

Valian, V., Solt, S., and Stewart, J. (2009). Abstract categories or limited-scope formulae? the case of children's determiners. *J. Child Lang.* 36, 743–778. doi: 10.1017/S0305000908009082

van Heugten, M., and Christophe, A. (2015). Infants' acquisition of grammatical gender dependencies. *Infancy* 20, 675–683. doi: 10.1111/infa.12094

van Heugten, M., and Johnson, E. K. (2010). Linking infants' distributional learning abilities to natural language acquisition. *J. Mem. Lang.* 63, 197–209. doi: 10.1016/j.jml.2010.04.001

van Heugten, M., and Johnson, E. K. (2011). Gender-marked determiners help Dutch learners' word recognition when gender information itself does not. *J. Child Lang.* 38, 87–100. doi: 10.1017/S0305000909990146

van Heugten, M., and Shi, R. (2009). French-learning toddlers use gender information on determiners during word recognition. *Dev. Sci.* 12, 419–425. doi: 10.1111/j.1467-7687.2008.00788.x

Veneziano, E., and Parisse, C. (2011). "Retrieving the meaning of words from syntactic cues: a comprehension study of 2 to 4 yrs old French-speaking children," in *IASCL 2011, International Conference on the Study of Child Language*.

Wang, H., Höhle, B., Ketrez, N. F., Küntay, A. C., Mintz, T. H., Danis, N., et al. (2011). "Cross-linguistic distributional analyses with frequent frames: the cases of german and turkish," in *Proceedings of 35th Annual Boston University Conference on Language Development*, 628–640.

Waxman, S. R. (1999). Specifying the scope of 13-month-olds' expectations for novel words. *Cognition* 70, B35–B50. doi: 10.1016/S0010-0277(99)00017-7

Waxman, S. R., and Booth, A. E. (2001). Seeing pink elephants: fourteen-month-olds' interpretations of novel nouns and adjectives. *Cogn. Psychol.* 43, 217–242. doi: 10.1006/cogp.2001.0764

Waxman, S. R., and Hall, D. G. (1993). The development of a linkage between count nouns and object categories: evidence from fifteen- to twenty-one-month-old infants. *Child Dev.* 64, 1224–1241. doi: 10.1111/j.1467-8624.1993.tb04197.x

Waxman, S. R., and Lidz, J. L. (2006). "Early Word Learning," in *Handbook Of Child Psychology: Cognition, Perception, And Language*, eds D. Kuhn, R. S. Siegler, W. Damon, and R. M. Lerner (Hoboken, NJ: Wiley), 299–335.

Waxman, S. R., Lidz, J. L., Braun, I. E., and Lavin, T. (2009). Twenty four-month-old infants' interpretations of novel verbs and nouns in dynamic scenes. *Cogn. Psychol.* 59, 67–95. doi: 10.1016/j.cogpsych.2009.02.001

Waxman, S. R., and Markov, D. B. (1995). Words as invitations to form categories: evidence from 12-to 13-month-old infants. *Cogn. Psychol.* 29, 257–302. doi: 10.1006/cogp.1995.1016

Weisleder, A., and Waxman, S. R. (2010). What's in the input? Frequent frames in child-directed speech offer distributional cues to grammatical categories in Spanish and English. *J. Child Lang.* 37, 1089–1108. doi: 10.1017/S0305000909990067

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition* 85, 223–250. doi: 10.1016/S0010-0277(02)00109-9

Yang, C. (2013). Who's afraid of George Kingsley Zipf? Or: Do children and chimps have language? *Significance* 10, 29–34. doi: 10.1111/j.1740-9713.2013.00708.x

Yarowsky, D. (1995). "Unsupervised word sense disambiguation rivaling supervised methods," in *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, 189–196. doi: 10.3115/981658.981684

Yuan, S., and Fisher, C. (2009). "Really? She blicked the baby?": Two-year-olds learn combinatorial facts about verbs by listening: research article. *Psychol. Sci.* 20, 619–626. doi: 10.1111/j.1467-9280.2009.02341.x

Zangl, R., and Fernald, A. (2007). Increasing flexibility in children's online processing of grammatical and nonce determiners in fluent speech. *Lang. Learn. Dev*. 3, 199–231. doi: 10.1080/15475440701360564

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 reviewer SW declared a past collaboration with one of the authors AC to the handling editor.

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 Psychology

ORIGINAL RESEARCH
published: 31 August 2021
doi: 10.3389/fpsyg.2021.677571

Multiple Sources of Surprisal Affect Illusory Vowel Epenthesis

*James Whang**

Department of Linguistics, Seoul National University, Seoul, South Korea

Illusory epenthesis is a phenomenon in which listeners report hearing a vowel between a phonotactically illegal consonant cluster, even in the complete absence of vocalic cues. The present study uses Japanese as a test case and investigates the respective roles of three mechanisms that have been claimed to drive the choice of epenthetic vowel—phonetic minimality, phonotactic predictability, and phonological alternations—and propose that they share the same rational goal of searching for the vowel that minimally alters the original speech signal. Additionally, crucial assumptions regarding phonological knowledge held by previous studies are tested in a series of corpus analyses using the Corpus of Spontaneous Japanese. Results show that all three mechanisms can only partially account for epenthesis patterns observed in language users, and the study concludes by discussing possible ways in which the mechanisms might be integrated.

Keywords: illusory vowel epenthesis, information theory, Japanese, phonology, phonotactic learning, alternation learning

OPEN ACCESS

Edited by:
Rory Turnbull,
Newcastle University, United Kingdom

Reviewed by:
Alexander Kilpatrick,
Nagoya University of Commerce and Business, Japan
Enes Avcu,
Massachusetts General Hospital and Harvard Medical School, United States

****Correspondence:***
James Whang
jamesw@snu.ac.kr

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

Received: *07 March 2021*
Accepted: *02 August 2021*
Published: *31 August 2021*

Citation:
Whang J (2021) Multiple Sources of Surprisal Affect Illusory Vowel Epenthesis.
Front. Psychol. 12:677571.
doi: 10.3389/fpsyg.2021.677571

1. INTRODUCTION

Illusory epenthesis, or perceptual epenthesis, is a phenomenon where listeners perceive C_1C_2 consonant clusters that are phonotactically illegal in their native language as C_1VC_2 sequences (Dupoux et al., 1999, 2011; Dehaene-Lambertz et al., 2000; Monahan et al., 2009; Durvasula and Kahng, 2015; Whang, 2019; Kilpatrick et al., 2020). The misperceived medial vowel is not present in the original speech signal but makes the resulting sequence phonotactically legal. Since C_1C_2 sequences are repaired to C_1VC_2 during perception, listeners have difficulties distinguishing such vowel-less vs. vowel-ful pairs accurately. For example, a series of studies by Dupoux et al. (1999, 2011) showed that Japanese listeners are unable to distinguish pairs such as [ebzo] and [ebuzo] reliably and exhibit a strong tendency toward perceiving both as [ebuzo]. Mainly three separate mechanisms have been proposed in the literature as driving the epenthetic process—phonetic minimality, phonotactic predictability, and phonological alternations. The current study investigates each in detail and shows that separately the mechanisms can only partially predict human epenthetic behavior and need to be integrated. In order to integrate the three mechanisms, the current study takes a rational approach, reframing illusory epenthesis as an optimization process (Anderson, 1990).

Rational analysis uses probabilistic approaches (e.g., Bayesian, information theoretic, and game theoretic frameworks) to explain the mechanisms that underlie human cognition. In linguistic research, the rational framework has been applied at various linguistic levels, such as pragmatic reasoning (Frank and Goodman, 2012; Lassiter and Goodman, 2013), word recognition (Norris, 2006), and speech perception (Feldman and Griffiths, 2007; Sonderegger and Yu, 2010). Of particular relevance to the current study are previous works that take an information theoretic approach to phonological processing, showing that speakers perform various manipulations at the loci of sudden surprisal peaks in the speech signal, making syllables longer or more prosodically prominent presumably to make processing easier for the listener (Aylett and Turk, 2004; Hume and Mailhot, 2013). Illusory epenthesis occurs when phonotactic violations are detected (i.e., between high surprisal sequences). The process, therefore, is simply another strategy for smoothing extreme surprisal peaks under a rational approach, and the three epenthesis mechanisms constitute different linguistic levels that a listener relies on to select the most probable output for a given input.

1.1. Phonetic Minimality

Phonetic minimality is the idea that the vowel that is physically shortest in a given language, and thus acoustically the closest to zero (∅; lack of segment), is used as the default epenthetic segment (Steriade, 2001; Dupoux et al., 2011). In the case of Japanese, this vowel is [u], which has an average duration of ∼50 ms but can be as short as 20 ms (Beckman, 1982; Han, 1994; Shaw and Kawahara, 2019). Dupoux et al. (2011) also found that in Brazilian Portuguese, where [i] is the shortest vowel, it is [i] that functions as the default epenthetic segment in the language instead of [u] as in Japanese, further bolstering the idea that phonetically minimal vowels are the default epenthetic segment in perceptual repair.

When framed rationally, the phonetic minimality account is arguing that listeners are selecting the most probable output based on acoustic similarity. As can be seen in (1), where A denotes the acoustic characteristics of the input [ebzo], the output that is most consistent with the input is naturally the faithful one, namely [ebzo]. However, [ebzo] is phonotactically illegal in Japanese. It has a near-zero probability in the language and is eliminated, denoted with a strikeout. The Japanese listener, therefore, assigns the highest probability to [ebuzo] instead because [u] is the shortest vowel in Japanese, and its epenthesis results in an output that conforms to the phonotactics of the language with the smallest possible acoustic change from the original signal.

Given [ebzo]: (1)
$$\cancel{P(\text{ebzo}|A)} > P(\text{ebuzo}|A) > P(\text{ebizo}|A)$$

The main weakness of the phonetic minimality account is that it incorrectly predicts the use of only one epenthetic segment in a given language, contrary to the fact that languages often employ more than one epenthetic vowel for phonotactic repair (e.g., Japanese—Mattingley et al., 2015; Korean—Durvasula and Kahng, 2015; Mandarin—Durvasula et al., 2018).

1.2. Phonotactic Predictability

Phonotactic predictability is the idea that the most frequent, and thus the most predictable, vowel in a given phonotactic context is the vowel that is epenthesized for perceptual repair. A recent study by Whang (2019) showed that while Japanese listeners do repair consonant clusters primarily through [u] epenthesis, there is also a consistent effect of the palatal consonants [ʃ, ç], after which [i] is epenthesized instead. The study calculated surprisal values (Shannon, 1948) for /u, i/ after the consonants that were used as stimuli ([b, ɡ, z, p, k, ʃ, ɸ, s, ç]) using the Corpus of Spontaneous Japanese. Crucially for the present study, Whang (2019) did not include non-high vowels and long vowels in the calculations under the assumption that Japanese listeners only consider short high vowels for epenthesis due to a lifelong experience of having to recover devoiced/deleted high vowels. The results showed that [i] had lower surprisal than [u] after the two palatal consonants [ʃ, ç] while [u] had lower surprisal after the rest. Based on these results, Whang argued that the choice between [u, i] must be driven at least in part by phonotactic predictability, that [u] is the default epenthetic segment in Japanese not only because it is the shortest vowel but also because it is the most common vowel in most contexts. When another vowel has lower surprisal in a given context (e.g., [i] after palatal consonants), the vowel with the lower surprisal is epenthesized instead, suggesting that phonotactic information can override the use of a phonetically minimal segment. Kilpatrick et al. (2020) also found similar results with [ɡ, ʧ, ʃ], where [u] was epenthesized more often after [ɡ] but [i] was epenthesized after the palatalized consonants [ʧ, ʃ].

When framed rationally, the phonotactic predictability account appeals to an idealized optimal listener's knowledge of context-specific segmental frequency to select the most probable output for a given input. This can be summarized as in (2) and (3), where relative probabilities are assigned according to the listener's knowledge of native phonotactics K_p rather than the physical signal A as was the case for phonetic minimality in (1). Since heterorganic clusters such as [bk, ʃp] are prohibited, the Japanese listener assigns near-zero probabilities to the faithful outputs, namely [ebko] and [eʃpo]. The listener then selects alternative candidates that contain the most frequent vowel in a given context. As shown in (2), [u] is the most frequent vowel after [b], whereas (3) shows that [i] is the most frequent after [ʃ]. This results in the outputs [ebuko] and [eʃipo], respectively. Note that in the case of (3), the phonetic minimality account would incorrectly predict [eʃupo] as the perceived output.

Given [ebko]: (2)
$$P(\text{ebuko}|K_p) > P(\text{ebiko}|K_p) > P(\text{ebko}|K_p)$$

Given [eʃpo]: (3)
$$P(\text{eʃipo}|K_p) > P(\text{eʃupo}|K_p) > P(\text{eʃpo}|K_p)$$

The main weakness of the phonotactic predictability account is not necessarily inherent to the approach itself but lies instead in the specific assumptions that previous studies have made. First, both Whang (2019) and Kilpatrick et al. (2020) assume *a priori* that only high vowels are considered for perceptual epenthesis due to knowledge of high vowel devoicing, and fail to show empirically that non-high vowels do not participate in illusory epenthesis. Second, and more importantly, the two studies do not distinguish voiced and devoiced vowels before calculating predictability, assuming that they belong to the same underlying vowel. In other words, voiced and devoiced vowels are assumed to be allophones of each other that alternate depending on the context. This means that as it currently stands the phonotactic predictability account subsumes phonological alternations, making it difficult to tease apart the independent effects of the two types of linguistic knowledge.

1.3. Phonological Alternations

Phonological alternations are different from both phonetic minimality and phonotactic predictability in that it requires a lexicon that is detailed enough to keep track of the different ways in which words and morphemes show variation on the surface. For example, an English learner must know that the suffix *-s* after nouns means "more than one" before learning that the suffix has multiple surface forms [-s, -z, -əz] depending on the final segment of the noun stem it attaches to. The phonological alternation mechanism is such that a language user learns that certain units alternate in the lexicon as a result of various phonological processes and represents the alternations as equivalent in certain contexts. An example of this sort of context-specific phonological equivalence effect on speech perception can be found in Mandarin Chinese listeners. Mandarin Chinese has four lexical tones—high (55), rising (35), contour (214), and falling (51), where the numbers in parenthesis indicate the relative pitch of the tone on a five-level scale—and has a well-known tone sandhi process where the contour tone becomes a rising tone when another contour tone immediately follows. Huang (2001) tested whether Mandarin Chinese listeners' experience with the contour-rising tone alternation in their native grammar yields different perceptual patterns from American English listeners, who have no such experience. The results showed that Mandarin Chinese listeners had more difficulties distinguish the contour and rising tones than American English listeners, suggesting that the two tones are represented as being equivalent by Mandarin Chinese listeners in certain contexts. The need for phonological alternations in explaining perceptual epenthesis was perhaps most clearly shown in a series of experiments with Korean listeners by Durvasula and Kahng (2015). Korean phonotactic structure prohibits consonant clusters within a syllable, and Korean listeners typically repair illicit clusters by epenthesizing the high central unrounded vowel [ɨ] (e.g., [klin] → [kɨlin] "clean"). However, Durvasula and Kahng (2015) found that in contexts where another vowel other than [ɨ] more frequently alternates with zero in the lexicon, it is the more frequently alternating vowel that is perceived instead. To illustrate how a vowel alternates with zero in Korean, consider the phrase "although (it is) big." The phrase is a bimorphemic word in Korean /k^{h}ɨ + ədo/ "big + although," but /ɨə/ sequences that result from such adjectival morpheme concatenations undergo simplification. The first vowel /ɨ/ is deleted, deriving the output [k^{h}ədo], resulting in a regular alternation between [ɨ] and zero after [k]. [ɨ] is actually illegal in Korean after palatal fricatives such as [ʃ], and the vowel that most frequently alternates with zero in these contexts instead is [i]. What Durvasula and Kahng (2015) found was that it is this sort of phonological alternation that best predicts the identity of the perceptually epenthesized vowel—generally [ɨ] but [i] after palatal fricatives where [ɨ] is phonotactically illegal, and either [ɨ, i] after palatal stops where both vowels are allowed—and argue that sublexical mechanisms (i.e., phonetic minimality and phonotactic predictability) are employed hierachically, becoming active only when phonological alternations fail to provide an optimal candidate for epenthesis.

The basic rational framing for phonological alternations is similar to that of phonotactic predictability, where the listener relies on a particular kind of phonological knowledge to select the optimal output for a given input. However, instead of surface level phonotactics K_p, the listener relies on the knowledge of phonological alternations K_a to assign probabilities to possible candidates for perception.

Given [ebko]: (4)

$$P(\text{ebuko}|K_a) > P(\text{ebiko}|K_a) > P(\text{ebko}|K_a)$$

Given [eʃpo]: (5)

$$P(\text{eʃipo}|K_a) > P(\text{eʃupo}|K_a) > P(\text{eʃpo}|K_a)$$

1.4. Summary

Although discussed in the previous literature using various terminology, the three main ways that were argued to be the driving factors behind epenthetic vowel selection can be reframed as being motivated by a common goal of rational optimization: Select the output that is most probable given the original input. Epenthesizing the shortest vowel in a given language results in an output that is acoustically the most similar to the original signal, hence is most probable; epenthesizing the vowel with the lowest surprisal in a given context results in an output with total information that is most similar to the original signal, hence is most probable; epenthesizing the vowel that most frequently alternates with zero in a given context results in an output that is representationally equivalent to the original signal, hence is most probable. Note that all three mechanisms are triggered by phonotactic violations, which have extremely high surprisal due to their near-zero probabilities, and are repaired by inserting a segment that consequently removes the locus of high surprisal. Illusory epenthesis, therefore, can also be viewed in information theoretic terms (Shannon, 1948) as smoothing sudden peaks in surprisal. Numerous studies have shown that listeners take longer to process high surprisal (low frequency) words and segments than low surprisal (high frequency) ones (Jescheniak and Levelt, 1994; Vitevitch et al., 1997), suggesting processing difficulties for high surprisal elements. This would suggest that phonotactically illegal sequences that have near-zero probabilities (= near-infinite surprisal) in the listener's language are difficult to process as well. Language users seem to be aware of such processing bottle-necks and have been found to employ various methods to achieve a smoother probability distribution through various phonological manipulations such as syllable duration and prosodic prominence (Aylett and Turk, 2004; Shaw and Kawahara, 2019).

To summarize, there are three main factors involved in illusory vowel epenthesis, all of which are triggered by phonotactic violations in the input and share the goal of selecting the most probable, phonotactically legal alternative as the output. However, the respective contributions of each factor are difficult to tease apart due to a number of assumptions in the previous studies that often have not been tested explicitly. The present study, therefore, investigates the main assumptions behind each of the three epenthesis methods through a series of corpus analyses. The results show that no single method is able to fully account for the observed epenthetic patterns in language users. Section 2 first presents a simulation of how phonotactic restrictions might be learned by a Japanese learner and also describes the Corpus of Spontaneous Japanese, which is used for all simulations and calculations in this paper. Section 3 then discusses in information theoretic terms how the illusory vowel is chosen at a sublexical level according to the phonetic minimality and phonotactic predictability accounts. Section 4 simulates how a Japanese learner might build a lexicon based strictly on surface forms and consequently learn phonological alternations that contribute to illusory vowel epenthesis. Section 5 concludes the study, first by summarizing the overall results and discussing possible avenues for how the different factors involved in illusory vowel epenthesis might be unified into a single system based on convergent proposals from multiple lines of research, ranging from acquisition studies to psycholinguistics and theoretical phonology.

2. PHONOTACTIC LEARNING

Although previous studies generally agree that the process of perceptual epenthesis is the result of repairing phonotactically illegal consonant clusters, Japanese actually allows numerous consonant clusters on the surface. Japanese has a highly productive high vowel devoicing process, where high vowels [i, u] lose their phonation between two voiceless consonants (Fujimoto, 2015). Although devoiced vowels were traditionally analyzed as only losing their phonation while maintaining their oral gestures, recent studies show that there is often no detectable trace of devoiced vowels both acoustically (Ogasawara, 2013; Whang, 2018) and articulatorily (Shaw and Kawahara, 2018). This presents an interesting puzzle whereby Japanese listeners are frequently exposed to and produce consonant clusters, yet repair such sequences with epenthetic vowels during perception. Therefore, rather than assuming that consonant clusters are illegal in Japanese *a priori*, this section first establishes that phonotactic restrictions against heterorganic consonant clusters in Japanese can be learned from the data, using the Corpus of Spontaneous Japanese.

2.1. The Corpus of Spontaneous Japanese

All calculations in the present study are based on a subset of the Corpus of Spontaneous Japanese (CSJ; Maekawa and Kikuchi, 2005). The corpus in its entirety consists of ∼7.5 million words—660 h of speech—recorded primarily from academic conference talks. The subset used is the "core" portion of the corpus (CSJ-RDB), which contains data from over 200 speakers, comprising ∼500,000 words—45 h of recorded speech—that have been meticulously segmented and annotated with the aim to allow linguistic analyses from the phonetic level to the semantic level. The most relevant annotations for the present study are the "prosodic," "word," and "phonetic" level annotations. From the prosodic level, the present study primarily uses the intonational phrase for modeling phonotactic learning based on previous findings that infants as young as 6-months of age use prosodic boundary cues to segment clauses and words within speech streams (Jusczyk et al., 1993; Morgan and Saffran, 1995), which suggests that prosodic boundaries can be detected and used for linguistic processing even by the most naïve of listeners. The phonetic level is used for phonotactic learning as well as for calculating predictability in this section. The word and phonetic levels are used together in section 4 to build a lexicon, which is necessary for alternation learning. The word level provides Japanese orthographic representations of all words in the corpus as well as their syntactic categories (e.g., noun, verb, adjective, etc.). The phonetic level provides phonetically detailed transcriptions of the recorded speech, and crucially, indicates the voicing status of vowels.

Two modifications were made to the phonetic transcriptions provided by the CSJ-RDB before using the data as input for calculations. First, "phonetically palatalized" (e.g., /s̠i/ → [s̠ʲi]) vs. "phonologically palatalized" (e.g., /s̠ʲa/ → [s̠ʲa]) consonants, which the CSJ-RDB distinguishes for coronal and dorsal consonants, were collapsed as belonging to the same palatalized consonant. Phonetically palatalized consonants occurred exclusively before /i, iː/, which suggests that the purpose of phonetically palatalized annotations was to reflect coarticulation, where coronal consonants become backed while dorsal consonants become fronted toward a following high front vowel. However, this meant that phonetically palatalized consonants all had near-zero surprisal because only short [i] and long [iː] occurred after these consonants, and the short vowel is over 30 times more frequent than its long counterpart. Furthermore, although the phonetic/phonological distinction might be meaningful underlyingly, it is unclear how phonetically palatalized and phonologically palatalized consonants would differ meaningfully on the surface. Since all of the analyses of the present study assume that phonological learning begins without a lexicon and by extension without knowledge of underlying forms, the difference in palatalization was removed as unlikely to be salient to an uninformed listener.

Second, vowels transcribed as devoiced in the CSJ-RDB were deleted at a probability of 0.10. Recent experimental results show that there is often no detectable acoustic cue (Ogasawara, 2013; Whang, 2018) or articulatory gesture (Shaw and Kawahara, 2018) for vowels that have undergone devoicing, suggesting deletion. However, the CSJ-RDB never transcribes devoiced vowels as deleted. Instead, the CSJ consistently transcribes devoiced vowels as being part of the preceding consonant. For example, the final high vowel in the formal declarative copula *-desu* has a high devoicing rate, and the devoiced copula is segmented as [d], [e], [su̥], where the fricative and the devoiced vowel are segmented together. This shows that the annotators could not reliably separate the devoiced vowel from the preceding consonant but also that the vowel was assumed to be present. It is difficult to conclude with confidence that such unseparated segmentations indicate deletion, however, since there are multiple possible reasons for the annotators' reluctance to mark a segment boundary, such as extreme coarticulation between the segments, lack of obvious vowel spectra despite being audible, lack of vowel cue due to deletion, etc. The story is much the same in previous experimental studies. Despite there being evidence that devoiced vowels do delete, it is difficult to calculate the exact deletion rates due to limitations in the methodology (e.g., reliance on a single acoustic cue to determine deletion; Whang, 2018) or stimuli used (e.g., focusing on a single vowel in limited contexts; Shaw and Kawahara, 2018). The chosen deletion rate of 0.10 is admittedly arbitrary, but it was chosen to introduce some deletion in the data while limiting the number of changes to the original transcriptions that lack clear empirical support. Calculations were also run with deletion probabilities as high as 0.30, but the results were qualitatively similar.

2.2. Learning From Unsegmented Speech

The phonotactic learner is based on the Frequency-Driven Constraint Induction mechanism of StaGe (Adriaans and Kager, 2010). StaGe is a lexiconless model built for the purposes of word segmentation in continuous, unsegmented speech. The model is lexiconless based on infant language acquisition studies that showed that infants are sensitive to various aspects of the native language, such as phonetic categories (Werker and Tees, 1984; Werker and Lalonde, 1988; Maye et al., 2002) and phonotactics (Jusczyk et al., 1994; Mattys and Jusczyk, 2001) before the age of 1;0 (years;months) and as early as 0;6. Infants around this age have also been shown to be able to extract words from a continuous stream of speech (Jusczyk and Aslin, 1995; Saffran et al., 1996). In other words, infants already have sophisticated knowledge of their native phonology before acquiring a sufficiently detailed lexicon. The present study, therefore, also assumes that phonotactic learning in Japanese begins before a lexicon is formed and applies the learning mechanism to unsegmented intonational phrases rather than words, as annotated in the CSJ-RDB.

The Frequency-Driven Constraint Induction mechanism of StaGe calculates observed/expected ratios (O/E; Pierrehumbert, 1993; Frisch et al., 2004) of all biphones that occur in the input data and induces constraints by setting thresholds on the O/E ratios. O/E ratios compare how often a biphone actually occurs in the data (Observed) to how often each biphone should have occurred if all segments are assumed to have an equal likelihood of combining to form biphones (Expected) by dividing the probability of a biphone (xy) divided by the product of the summed probability of all biphones beginning with (x) and the summed probability of all biphones ending with (y). The resulting value indicates the magnitude of a given biphone's over-/underrepresentation. For example, O/E ratio of 1 indicates that a given biphone occurred exactly as often as expected, while O/E of 3.0 indicates that a biphone occurred thrice as often as expected.

$$\frac{O(xy)}{E(xy)} = \frac{Pr(xy)}{\sum Pr(xY) * \sum Pr(Xy)} \tag{6}$$

StaGe induces markedness constraints that flag a biphone as requiring repair for underrepresented biphones. StaGe also induces Contiguity constraints that keep biphones unchanged for overrepresented biphones. The strength of the induced constraints are the target biphones' expected probabilities $E(xy)$. The thresholds for under- and overrepresentation are arbitrary (perhaps language-specific), but in the original study, Adriaans and Kager (2010) set the O/E thresholds at 0.5 or lower for underrepresentation and 2.0 or higher for overrepresentation. For the present study, the thresholds are set at 0.75 for underrepresentation and 1.25 for overrepresentation so that the model induces constraints more aggressively.

TABLE 1 | Five over-/underrepresented biphones in Japanese with highest expected values.

Overrepresented			Underrepresented		
xy	*E(xy)*	*O/E*	*xy*	*E(xy)*	*O/E*
ta	7.61×10^{-3}	2.21	ti	4.44×10^{-3}	3.85×10^{-2}
ka	6.78×10^{-3}	2.91	tt	3.61×10^{-3}	2.20×10^{-3}
to	6.08×10^{-3}	3.55	kt	3.21×10^{-3}	4.19×10^{-2}
ko	5.42×10^{-3}	2.00	tk	3.21×10^{-3}	2.91×10^{-3}
na	4.96×10^{-3}	3.01	kk	2.86×10^{-3}	1.16×10^{-2}

To illustrate the phonotactic learning mechanism of StaGe, suppose that the model receives the following words as input: [ku̥toː, kta]. Focusing on the word-initial biphones, the model learns by calculating observed/expected (O/E) ratios that [ku̥, kt] are both likely to occur in the language. However, when the model receives [kubi, kumo, kugi, kuʤi] as additional input, the O/E of [ku] increases while the O/E for [ku̥, kt] decreases. In this way, the O/E ratios of biphones rise and fall based on the data, and when the O/E ratio of a particular biphone sequence falls below 0.75, the model induces a markedness constraint (e.g., **kt*: flag *kt* sequence as requiring repair). When the O/E ratio is 1.25 or higher, the model induces a Contiguity constraint (e.g., Contig-*ku*: keep *ku* sequence unchanged). Although it is possible to set constraint induction thresholds based on surprisal instead of O/E ratios, O/E ratios are used as in the original StaGe for the present study. Both surprisal and O/E ratios quantify unexpectedness based on frequency, but there is no obvious reference value for "exactly as expected" for surprisal, whereas this would simply be 1.0 for O/E ratios, making the latter more intuitive to interpret.

2.3. Phonotactic Learner Results

Out of 1,280 unique biphones total in the CSJ-RDB, there were 558 consonant-initial biphones. Of them, 127 were overrepresented with O/E ratios >1.25, and 370 were underrepresented with O/E ratios <0.75. The remaining 61 had O/E ratios between the 1.25 and 0.75 thresholds and did not induce constraints. All overrepresented biphones were consonant-vowel biphones, and more importantly all 213 consonant-consonant and 13 consonant-boundary (C#; i.e., word-final consonants) biphones observed in the CSJ-RDB had O/E ratios below 0.75. Shown in **Table 1** are five overrepresented consonant-initial biphones with the highest expected values (i.e., the strength of the induced Contiguity constraints that keep the biphone intact) and five underrepresented biphones with the highest expected values (i.e., the strength of the induced markedness constraints that mark the biphone as requiring repair) that illustrate the phonotactic structures that the model learned[1].

Table 1 shows that overrepresented consonant-initial biphones with the highest expected values in Japanese are all CV. Underrepresented consonant-initial biphones are all consonant clusters with the exception *[ti], which actually reflects another well-known phonotactic restriction in Japanese against high vowels after alveolar obstruents (Ito and Mester, 1995). It should also be noted that coda consonants were distinct from onset consonants in the CSJ-RDB, where [N] represented the placeless nasal coda of Japanese that assimilates in place with the following segment and [Q] represented the first half of a geminate consonant, and thus also placeless. Because the surface forms of [N, Q] are completely predictable based on the following segment, they were left unchanged before the analysis. Therefore, the biphones *[tt, kk] in **Table 1** are not geminates but clusters of consonants with independent place features.

The results show that the phonotactic learner learns both a strong preference for CV structure and a strong restriction against CC and C# sequences. However, learning that consonant clusters are prohibited in Japanese is not enough to explain how perceptual repair occurs. StaGe, which the current phonotactic learner is based on, detects phonotactically illicit sequences and inserts a word boundary. However, unlike in the case of the original Dutch data that StaGe was applied to, where many consonant clusters and codas are allowed, simply breaking up a cluster by inserting a word boundary in Japanese would result in C# sequences which are also prohibited. Phonotactic repair requires choosing a vowel to epenthesize when a consonant cluster is detected, which this paper now turns to in the following section.

3. SUBLEXICAL FACTORS IN PERCEPTUAL EPENTHESIS

In an experimental study, Dupoux et al. (1999) presented Japanese listeners with acoustic stimuli containing the high back rounded vowel [u] of varying durations ranging from 0 to 90 ms occurring between two consonants (e.g., [ebzo] ∼ [ebuːzo]). The results showed that Japanese speakers were unable to distinguish vowel-ful tokens from their vowel-less counterparts, erring heavily toward perceiving a vowel between consonant clusters (e.g., [ebzo] → [ebu̲zo]). The authors proposed that the results are due to the phonotactics of Japanese that disallows heterorganic consonant clusters. This is supported by the phonotactic learner results presented in the previous section, which showed that restrictions against consonant clusters can indeed be learned from the data. The authors further argue that there is a top-down phonotactic effect on perception, where phonotactically illegal sequences are automatically perceived as the *nearest* legal sequence rather than repaired at a higher, abstract phonological level. The nearest legal sequence is one that requires the most phonetically minimal repair, making [u] the best candidate due to its shortness (Beckman, 1982; Han, 1994).

Phonetic minimality captures an important generalization that it is high vowels that tend to be default epenthetic segments cross-linguistically (e.g., [i] in Brazilian Portuguese; [u] in Japanese; and [ɨ] in Korean) and also be targeted for deletion during production. However, reliance on phonetic minimality leads to the prediction that languages can only have one epenthetic segment, unless there are more than one vowel that are equally short. Languages often employ more than one epenthetic vowel for phonotactic repair, as discussed in the introduction. In the case of Japanese, [u] is the most frequent epenthetic vowel, but recent studies by Whang (2019) and Kilpatrick et al. (2020) found that Japanese listeners report hearing [i] instead in contexts where the high front vowel is the most phonotactically predictable.

3.1. Calculating Surprisal

Whang (2019) and Kilpatrick et al. (2020) identify the most phonotactically predictable vowel in a given context using surprisal, which is based on the conditional probabilities of vowels after a given consonant [i.e., $\Pr(v \mid C_1\_)$]. Surprisal is the negative $\log_2$ probability, which transforms the probability to bits that indicate the amount of information (effort) necessary to predict a vowel after a given C_1.

$$-\log_2 \Pr(v \mid C_1\_) \tag{7}$$

Although the choice of epenthetic vowel by Japanese listeners seems to be affected by phonotactic predictability, both Whang (2019) and Kilpatrick et al. (2020) make a number of assumptions in their calculations that confound surface level phonotactics with underlying representations, which are not subject to phonotactic restrictions. First, though not an assumption in and of itself, the contexts tested are limited depending on the study's focus. Second, as mentioned above, although Whang (2019) calculated the surprisal of vowels after a given consonant, only high vowels were considered after voiceless consonants under the assumption that Japanese listeners must have learned high vowel devoicing already. High vowel devoicing is essentially the reverse of high vowel epenthesis, where the former systematically removes high vowels while the latter recovers them, and thus the two processes most likely affect each other within the Japanese language. However, assuming knowledge of high vowel devoicing *a priori* to explain epenthesis begs the question of how then the devoicing process was learned. Lastly, both Whang (2019) and Kilpatrick et al. (2020) collapse voiced and devoiced vowels as belonging to the same vowel category before calculating surprisal. Indeed devoiced vowels are considered allophones of voiced vowels in Japanese (Fujimoto, 2015) and belong to the same underlying phonological category as their voiced counterparts (e.g., [u, u̥] → /u/), but underlying categories are also something that must be learned from alternations in the lexicon. Furthermore, underlying forms are not subject to phonotactic restrictions, which strictly apply to surface structures (Ito, 1986, *et seq.*). Previous infant studies suggest that phonotactic violations are learned at the surface level prior to detailed lexical acquisition (Jusczyk et al., 1994; Mattys and Jusczyk, 2001), and thus necessarily prior also to

[1] Full results of all analyses in the present paper can be found in the author's repository, the link to which can be found in the data availability statement.

alternation learning (Tesar and Prince, 2007). In other words, the studies on phonotactic predictability are conflating the effects of phonotactic predictability and phonological alternations, and it is necessary to recalculate phonotactic predictability without collapsing devoiced and voiced vowels in order to tease apart the effects of the two types of phonological knowledge.

3.2. Sublexical Surprisal Results

Using the same pre-processed data from the CSJ-RDB as with the phonotactic learner, surprisal was calculated for all vowels after all consonants in the data. The results, shown in **Table 2**, reveal that the phonotactic predictability account regarding the choice of epenthetic segment in Japanese is only partially upheld once assumptions of higher phonological knowledge is removed. As discussed above, devoiced and voiced vowels were kept distinct, and since Japanese has phonemic vowel length contrasts (e.g., /oba̲saN/) "aunt" vs. /oba̲ː̲saN/ "grandmother"), this resulted in a total of 20 possible vowels: five short voiced [i, e, a, o, u], five short devoiced [i̥, e̥, ḁ, o̥, u̥], five long voiced [iː, eː, aː, oː, uː], and five long devoiced [i̥ː, e̥ː, ḁː, o̥ː, u̥ː]. In the interests of space, below are the three vowels with lowest surprisal values after every obstruent consonant observed in the data.

The consonants in the "non-standard" rows are atypical, occurring only in loanwords, and are not (yet) regarded as phonemic in Japanese. These non-standard consonants [tʲ, dʲ, kʷ, ɸʲ, β] each occurred 360, 7, 1, 1, and 2 times, respectively, in the entire CSJ-RDB, and thus are excluded from discussion for the remainder of this paper.

Starting with the stop consonants, **Table 2** shows that based on phonotactic predictability the only context in which the "default" [u] would be epenthesized is after [b]. The epenthetic vowel after [p, k, g], [t], and [d] are predicted to be [a, o, e], respectively. In the case of [p, k, g], previous studies have shown repeatedly that it is in fact [u] that is epenthesized after these consonants (Dupoux et al., 1999, 2011; Whang, 2019; Kilpatrick et al., 2020). The results also show that neither [u] nor [i] are predicted to be epenthesized after the coronal stops [t, d]. Instead, [o] is predicted after [t] and [e] after [d]. In fact, high vowels are prohibited in the native and Sino-Japanese lexical strata of Japanese, and it is most often [o] that is epenthesized after coronal stops in loanwords (e.g., /faɪt/ → [ɸaito̲] "fight"; Ito and Mester, 1995). However, despite the expectation that the illusory vowel should then be [o] in these contexts, this is not borne out in experimental results. Monahan et al. (2009) tested precisely the issue of illusory epenthesis in coronal stop contexts and found that (i) Japanese listeners do not confuse tokens such as [e{t/d}ma] with [e{t/d}uma] but also that (ii) Japanese listeners do not confuse tokens such as [e{t/d}ma] with [e{t/d}oma] either, suggesting that unlike [u, i], the mid-back vowel [o] does not participate in illusory epenthesis. The authors propose that perhaps in coronal stop contexts, Japanese listeners represent the input as [etVma], which is distinct from both [etuma] and [etoma]. Additionally, older loans with coronal stop codas also do not show [o] epenthesis, opting instead for deletion (e.g., /pɑkɛt̲/ → [pokke_] "pocket") or [u] epenthesis, which also triggers spirantization (e.g., /waɪt ʃɜɹt̲/ → [waiʃaʦ̲u̲] "white shirt"; Smith, 2006). Although loanwords are not the focus of this paper, it seems worth pointing out that the phonotactic calculations and the available experimental evidence suggest that the prevalent use of [o] after [t] in loanwords is not due to illusory epenthesis but possibly due to surface phonotactics of Japanese[2]. This does mean, however, that the phonotactic account fails to predict what Japanese listeners actually perceive in this context as there is no option to posit a featureless vowel. Furthermore, unlike in the case of [to], there is little support for the predicted epenthesis of [e] after [d] in the literature except for the occasional substitution of high vowels with [e] after coronal stops in older loanwords (e.g., /k'akt'u̲ki/ → [kakute̲ki] "Korean radish kimchi"; /stɪ̲k/ → [sute̲kki] "(walking) stick"). Whether tokens such as [e{t/d}ema] are perceptually confused with [e{t/d}ma] remains to be tested rigorously.

Setting aside [h][3], a surprising result is found with the fricatives. The vowel with the lowest surprisal after [s] is the *devoiced* vowel [u̥]. Voiced [u], in fact, is the third most common vowel after [s], leading to the prediction that [u̥] would be epenthesized in this context. Aside from [pʲ, bʲ][4], **Table 2** additionally shows that the phonotactic predictability account would correctly predict the epenthesis of a short high front vowel after palatalized obstruents (Dupoux et al., 1999; Whang, 2019; Kilpatrick et al., 2020). However, as was the case with [s], it is a devoiced vowel that is predicted to be epenthesized after the palatal fricatives [ʃ, ç].

Although phonetic minimality correctly predicts the epenthesis of [u] after non-palatalized consonants [p, k, b, g, ʦ, ʥ, ɸ, s], it is completely unable to account for the consistent epenthesis of [i] after palatalized consonants. Phonotactic predictability, on the other hand, is able to account for the epenthesis of [i] after palatalized consonants (and perhaps also the non-illusory epenthesis of [o] in loanwords). However, it is a poor predictor for the epenthesis of [u] after non-palatalized consonants once assumptions regarding higher phonological knowledge of underlying forms and high vowel devoicing are removed. In short, both phonetic minimality and phonotactic predictability are unable to fully account for human perceptual epenthetic behavior.

[2] The full surprisal results show that the [t__] context occurred 83,399 times in the CSJ-RDB, of which more than a third (29,983) was the [to] sequence, perhaps due to the frequent use of the homophonous conjunctive and quotative particles /-to/. In other words, the use of [o] as the epenthetic vowel in loanwords in Japanese is grounded in the statistical tendencies of the native phonology, which is in line with other previous research on loan phonology that have argued that seemingly "novel" loanword patterns are actually instantiations of previously "covert" statistical generalizations in the native grammar (Zuraw, 2000; Kubozono, 2006; Rose and Demuth, 2006).

[3] To this author's knowledge, [h] has never been previously tested in the perceptual epenthesis literature because the consonant is susceptible to extreme coarticulation with surrounding segments due to its lack of oral gestures. It is often the allophones of the phoneme /h/, namely [ɸ, ç], which occur before [u, i], respectively, that are included in studies.

[4] For the palatalized consonants, recall that unlike coronal and dorsal consonants, there were no labial consonants that were transcribed as "phonetically palatalized" in the CSJ-RDB. This meant that after [pʲ, bʲ], there were zero instances of high front vowels [i, i̥, iː, i̥ː].

TABLE 2 | Vowels with the three lowest surprisal after all obstruents observed in the CSJ-RDB, in order of increasing surprisal.

	Context	Vowel	Surprisal	Vowel	Surprisal	Vowel	Surprisal	Target
Stops	p	a	1.628	u	2.273	aː	3.275	u
	t	o	1.476	e	1.800	a	1.835	o
	k	a	1.439	o	2.301	u	2.711	u
	b	u	1.502	a	1.800	e	3.127	u
	d	e	0.717	a	2.579	o	2.710	o
	ɡ	a	0.623	o	2.494	e	3.751	u
Affricates	ʦ	u	0.670	u̥	1.728	uː	4.832	u
	ʥ	u	1.441	e	2.166	a	2.589	u
Fricatives	ɸ	u	1.298	u̥	1.654	uː	2.541	u
	s	u̥	1.918	a	2.600	u	2.627	u
	h	a	1.162	o	1.936	oː	2.395	–
Palatalized stops	p^j	oː	0.277	uː	3.558	a	4.143	i
	k^j	i	0.801	i̥	2.351	oː	3.649	i
	b^j	oː	0.517	uː	2.687	a	2.821	i
	g^j	i	0.664	oː	2.006	a	3.776	i
Palatalized affricates	ʧ	i	1.078	i̥	2.810	oː	3.173	i
	ʤ	i	1.023	oː	2.529	uː	2.982	i
Palatalized fricatives	ʃ	i̥	1.117	i	2.003	o	4.175	i
	ç	i̥	1.214	i	1.844	oː	2.714	i
Non-standard	t^j	uː	0.561	u	1.710	u̥	5.907	–
	d^j	u	0.485	uː	1.807	–	–	–
	k^w	a	0.000	–	–	–	–	–
	$ɸ^j$	uː	0.00	–	–	–	–	–
	β	a	1.000	i	1.000	–	–	–

4. LEARNING ALTERNATIONS FROM THE LEXICON

As shown in Section 3 above, phonetic minimality and phonotactic predictability are both only partially successful in predicting the perceptual epenthesis patterns shown in language users. Here, the present study proposes that the limited success is due to reliance on sublexical, phrase-level phonology. This section shows that a lexicon is necessary to fully account for perceptual epenthesis in Japanese, and more specifically phonological alternations that can only be learned by comparing surface forms that map to the same meaning. Durvasula and Kahng (2015) showed the necessity of phonological alternations in explaining the perceptual epenthesis patterns of Korean listeners, and the parallel between the Korean account in Durvasula and Kahng (2015) and Japanese perceptual epenthesis is not difficult to see. Just as certain vowels regularly alternate with zero in Korean due to productive phonological processes, alternations between high vowels and zero should also be observed in the Japanese lexicon due to the productive process of high vowel devoicing. This section aims to first establish that vowel-zero alternations with a bias toward vowel-fulness can in fact be learned from a lexicon in Japanese, despite there being surface clusters that result from high vowel devoicing/deletion.

4.1. Building the Lexicon

A lexicon allows a language learner to keep track of what input forms correspond to what meaning (Apoussidou, 2007) and eventually acquire a paradigm over the lexicon. To learn alternations from a lexicon, one must first build a lexicon, and for a lexicon to be built with sufficient detail, it is necessary to differentiate meaning. To simulate meaning-based learning, the lexicon builder built for the present study relies on a combination of the orthographic representation and syntactic category of each word as provided by the CSJ-RDB. When the lexicon builder encounters a new word, it creates a new lexical entry with the orthographic form, the syntactic category, and phonetic form of the word. Note that the phonetic forms were the same, pre-processed transcriptions from the CSJ-RDB used for the phonotactic analysis, which included deleted vowels. Every time the same combination of orthographic form and syntactic category is encountered, it adds the phonetic form to the entry. If the same phonetic form was encountered before, the lexicon builder simply updates the count. If a new phonetic form is

TABLE 3 | Toy lexicon.

Word	Category	Gloss	Surface
した	Verb	“did”	[ʃi̥ta] (x7), [ʃta] (x2), [ʃita] (x1)
下	Noun	“down”	[ʃi̥ta] (x4), [ʃita] (x1)
舌	Noun	“tongue”	[ʃi̥ta] (x1), [ʃta] (x1)
ある	Verb	“exists”	[aru] (x10)
ある	Adjective	“a certain…”	[aru] (x5)

encountered, it starts a separate count for the new phonetic form. A toy example of a resulting lexicon is shown in **Table 3**.

The first three words in **Table 3** “did,” “down,” and “tongue” are homophonous, and at least one of each word’s phonetic forms overlaps with another word. However, these three words differ in meaning (orthographic forms) and thus are listed as separate entries. This allows the phonetic forms for the words to be counted separately. For example, despite the fact that the words “did” and “down” both occurred with a devoiced [i̥] and voiced [i], there are separate counts for the homophonous forms according to lexical entry. Additionally, the last two words “exists” and “a certain…” show why the use of syntactic category was also necessary in building the lexicon. Neither words show any alternation, and thus are completely homophonous; and although they differ in meaning, they have the same orthographic representation. What differentiates them for the lexicon builder in this case is their respective syntactic categories. The lexicon builder learned a total of 14,121 unique words, and of them 3,353 words had more than one phonetic form.

4.2. Alternation Learner

With the lexicon established, let us now turn to how phonological alternations might be learned. The lexicon does not yet have underlying phonological representations because it simply mapped one or more phonetic (surface) forms to the same meaning. This was by design as it is the job of the language learner to figure out what single form the alternating phonetic forms must map to. The learner used `SequenceMatcher` in the `difflib` package for Python to learn alternations in the lexicon. `SequenceMatcher` compares two strings (sequences of phones) by setting one as the baseline and identifying substrings in the other to *replace*, *delete*, *insert*, or keep *equal* in order to match the baseline. The baseline was always set to the most frequent phonetic form for a given lexical entry. For example, using the entry “did” in **Table 3** above, the baseline would be [ʃi̥ta] since it occurred seven times out of 10. `SequenceMatcher` then would compare [ʃta] and [ʃita] to the baseline and learn the following:

- With [ʃi̥ta] as baseline and [ʃta] as alternate…
  - Keep *equal* the initial segment [ʃ].
  - *Insert* [i̥] in the second position.
  - Keep *equal* the third segment [t].
  - Keep *equal* the final segment [a].
- With [ʃi̥ta] as baseline and [ʃita] as alternate…
  - Keep *equal* the initial segment [ʃ].
  - *Replace* the second segment [i] with [i̥].
  - Keep *equal* the third segment [t].
  - Keep *equal* the final segment [a].

Each *replace*, *delete*, *insert*, or *equal* operation was multiplied by the number of times the alternate form occurred. The baseline was also compared to itself and multiplied by its token frequency. In cases where the alternate form had the same frequency as the baseline, `SequenceMatcher` was run again with the baseline and alternate forms switched. This was to ensure that the model gives equal weight to lexical alternations with the same probability and does not learn an accidental bias introduced by the sorting method of a particular programming language. Additionally, words that did not alternate were also compared to themselves and multiplied by their respective token frequencies. Multiplying each operation by token frequencies meant that the alternation learner actually learns a bias toward keeping segments unchanged, (i) since only 3,353 words of the 14,121 total showed alternations and (ii) since it is rarely the case that the baseline and alternate forms are completely different. Lastly, since the purpose of this alternation learner was to investigate what alternations can be learned after a given consonant *à la* Durvasula and Kahng (2015), the operations were contextualized with the previous segment. For word-initial segments, the context was a word boundary. For example, the [i] replacement operation above was recoded as [ʃ]:[i] → [i̥] (when [i] occurs after [ʃ], replace with [i̥]). For every observed $x : y \rightarrow z$ alternation, surprisal was calculated for the $y \rightarrow z$ operation with x as the context, quantifying how unexpected it is for the phonological grammar to perform a particular *replace*, *delete*, *insert*, or keep *equal* operation after a given consonant.

$$-\log_2 \Pr(\varnothing \rightarrow v \mid C_1\_) \tag{8}$$

Shown in **Table 4** are the zero-to-vowel alternations that the model actually learned for every obstruent context. The results of the alternation learner shows that the model correctly learns the necessary alternations between zero and high vowels in almost all relevant contexts. Of equal importance is that because the alternation learner tends to learn a “keep *equal*” bias, the surprisal for vowel-to-zero (deletion) operations are often the highest among all learned operations in a given context. In short, alternation learning strengthens the phonotactic prohibition of consonant clusters in Japanese.

It should be pointed out that the alternation learning model predicts that a devoiced high vowel will be epenthesized after most voiceless consonants [k, ʦ, ɸ, s, kʲ, ʃ, ç]. The exceptions are [p, ʧ], after which voiced high vowels are predicted to be epenthesized, and [t, d], after which [o, e] are again predicted to be epenthesized, respectively, as with the phonotactic predictability results. The prediction for devoiced vowel epenthesis after [k, ʦ, ɸ, s, kʲ, ʃ, ç] is supported by Ogasawara and Warner (2009), who found in a lexical judgment task that Japanese listeners have shorter reaction

TABLE 4 | Nothing-to-vowel alternations with the three lowest surprisal after all obstruents observed in the CSJ-RDB, learned by the alternation learner.

	Context	Vowel	Surprisal	Vowel	Surprisal	Vowel	Surprisal	Target
Stops	p	u	1.000	u̥	1.737	i	3.322	u
	t	o	0.941	e	2.093	a	2.159	o
	k	u̥	0.900	u	1.567	a	4.042	u
	b	u	0.807	a	1.807	i	2.807	u
	d	e	0.142	a	4.000	o	6.000	o
	g	u	1.322	a	1.322	–	–	u
Affricates	ts	u̥	0.812	u	1.216	–	–	u
	dz	u	0.000	–	–	–	–	u
Fricatives	ɸ	u̥	0.283	u	2.605	–	–	u
	s	u̥	0.110	u	4.705	o	6.095	u
	h	a	0.678	o	1.415	–	–	–
Palatalized stops	p^j	–	–	–	–	–	–	i
	k^j	i̥	0.826	i	1.228	a	8.388	i
	b^j	–	–	–	–	–	–	i
	g^j	–	–	–	–	–	–	i
Palatalized affricates	tʃ	i	0.795	i̥	1.455	o	5.087	i
	dʒ	i	0.453	i̥	3.700	u	4.700	i
Palatalized fricatives	ʃ	i̥	0.177	i	3.471	u̥	5.910	i
	ç	i̥	0.131	i	3.617	a	7.524	i
Non-standard	t^j	–	–	–	–	–	–	–
	d^j	–	–	–	–	–	–	–
	k^w	–	–	–	–	–	–	–
	ɸ^j	–	–	–	–	–	–	–
	β	–	–	–	–	–	–	–

times when presented with devoiced forms of words where devoicing is typically expected compared to when presented with voiced forms. This suggests that Japanese listeners do not restore devoiced vowels as underlyingly voiced for lexical access, relying instead on the more common surface form (Cutler et al., 2009; Ogasawara, 2013). Additionally, the alternation learner predicts that [o, e] will be epenthesized after [t, d], respectively, repeating the shortcomings of phonotactic predictability account in explaining illusory epenthesis in these contexts but providing a possible source for the prevalent use of mid vowels where high vowels are prohibited.

There are two contexts in which the alternation learner is unable to predict the epenthetic vowel—after [g] where [a, u] are the only candidates for epenthesis but have the same surprisal, and after [g^j] where no alternation with zero was observed in the lexicon. The first problem is attributable to the fact that all devoiced vowels, both high and non-high, were assigned the same rate of deletion. However, it seems reasonable to speculate that devoiced high vowels would be deleted at higher rates than devoiced non-high vowels, since high vowels are the ones that are categorically targeted for devoicing (Fujimoto, 2015), whereas non-high vowel devoicing is a more phonetic process that results from the glottis failing to close sufficiently in time (Martin et al., 2014). Implementing higher deletion rates for devoiced high vowels would increase the overall number of alternations of high vowels with zero relative to non-high vowels in the data; but again, in the absence of accurate deletion rates, it is perhaps hasty to implement different deletion rates according to vowel height simply to increase the model's performance. The second problem can be resolved by implementing a generalization mechanism similar to the feature-based approach of StaGe, which would allow both the phonotactic and alternation models to learn that [i] is most frequent after palatalized consonants in general.

5. DISCUSSION AND CONCLUSION

The present study investigated the three main ways in which illusory vowel epenthesis has been argued to occur and showed that no single method is able to fully account for the epenthetic behavior of language users. Section 2 first established that phonotactic restrictions against consonant clusters can be learned from Japanese input, despite the language allowing surface clusters due to a productive high vowel devoicing process. Section 3 then discussed the roles of phonetic minimality and phonotactic predictability. Crucially, the section revealed that once the misassumption of access to underlying forms is corrected for, phonotactic predictability is only successful in predicting the epenthesis of [i] in palatal contexts but not the epenthesis of [u] in other contexts. The results also showed that the phonotactic predictability account predicts the epenthesis of mid vowels [o, e] after [t, d], which finds support in the loanword literature but not in the experimental literature. Lastly, Section 4 showed that phonological alternations are the most successful at predicting the identity of illusory vowels in given contexts even with a small number of deletion introduced to the data, but also that it too is unable to account for all contexts. Specifically, it predicts the epenthesis of devoiced high vowels after most voiceless consonants, which although somewhat surprising at first glance is supported by lexical access studies. Additionally, as with the phonotactic predictability account, the phonological alternations learned by the model predict the epenthesis of [o, e] after [t, d]. Lastly, it is unable to narrow down the choice of epenthetic vowel in certain contexts.

The main proposal of the present study was that all three methods for illusory epenthesis can be reframed as having the same rational goal of choosing the optimal epenthetic vowel that results in the smallest amount of change to the original speech signal, motivated by the need to smooth extreme surprisal peaks in the signal that make processing difficult (Jescheniak and Levelt, 1994; Vitevitch et al., 1997; Aylett and Turk, 2004). A phonotactic violation is detected when there is a spike in surprisal caused by a sequence of sounds that are rarely or never adjacent to each other in the listener's native language. In the case of Japanese, the language has a strong CV preference, and thus listeners epenthesize a vowel between illicit consonant-consonant sequences. For example, according to the phonotactic analysis discussed in section 3, when [k] is followed by [t], the surprisal is 8.635, but by epenthesizing [u] between the two consonants, the result is substantial smoothing of surprisal, where the surprisal after [k] is now lowered to 2.711 and the subsequent transition from [u] to [t] is 3.924. Even when the surprisal is summed, the transition from [k] to [t] is now lower than when there was no intervening vowel. A phonetically minimal vowel is one that least alters the acoustic characteristics of the original input, and thus is an optimal repair. Similarly, a vowel that has the highest phonotactic predictability in a given context is one that has the lowest information density, altering the least the total information content of the original input, and thus is an optimal repair. Lastly, epenthesis based on knowledge of phonological alternations moves the search for the optimal vowel for repair from the sublexical domain to the lexical domain. However, the rational motivation remains the same. A vowel that phonologically alternates with zero in a given context is equivalent to zero in that context, and thus epenthesizing the vowel that most frequently alternates with zero least alters the phonological representation of the original input.

The problem as the results in the current paper showed is that the "optimal" vowel can differ depending on the level of analysis. For example, after [kʲ], the optimal vowels are [u, i, i̥] according the phonetic minimality, phonotactic predictability, and phonological alternation accounts, respectively. The question, then, is what level takes precedence? There are multiple converging lines of research that suggest that lexically driven processes take precedence over sublexical processes. First, previous phonological literature that have long noted the importance of the lexicon in phonological processing within traditional generative approaches best exemplified by Lexical Phonology (Kiparsky, 1982; Mohanan, 1982, *et seq.*). Although the theoretical details differ slightly between Kiparsky (1982) and Mohanan (1982) and their respective related works, they share the intuition that there are lexical and postlexical levels of phonological processing, where phonological rules (and/or constraints) operate first on underlying, morphological units at the lexical level, the results of which are processed at the postlexical level as combined phrase-sized units. Second, more recent, functional approaches as exemplified by Message-Oriented Phonology (Hall et al., 2016) go a step further and argue that since the main purpose of language is communicating *meaning*, phonological grammars are shaped largely by lexical concerns, where processes that more directly aid lexical access take precedence. Although the motivations differ, generative and functional approaches agree that the lexicon plays a primary role in phonology. Phonological alternations rely on lexical knowledge, and thus Durvasula and Kahng (2015)'s proposal that epenthesis based on phonological alternations must take precedence, where phonetic/phonotactic factors only come into play when there are no strong expectations that arise from knowledge of phonological alternations is also in line with the theoretical literature. In other words, illusory epenthesis is a serial process, starting from phonological processes that require lexical knowledge followed by sublexical processes that rely on surface-level phonotactics and/or phonetic cues.

Another converging line of work on the primacy of the lexicon is exemplified by Mattys et al. (2005), who investigated the interaction of lexical, phonotactic, coarticulatory, and prosodic cues in speech segmentation, all of which have been shown in previous studies to have a significant effect. The results revealed a hierarchical relationship among the different cues, where listeners use sublexical cues only when noise or lack of relevant contexts make reliable use of lexical information difficult. The results further showed that at the sublexical level, segmental cues (phonotactics, coarticulation, etc.) take precedence over prosodic cues. Of particular relevance for the current discussion is that at the sublexical level, the relative "weights" of different segmental cues are proposed to be language-specific, and thus have no set hierarchy. Mattys et al. (2005) do not provide details on how the language specific weights might be calculated, but again there is a diverse body of works from multiple traditions that bear on this issue.

First, the integration of different segmental cues are rather straightforward under a rational framework. Turning back to illusory epenthesis, the optimization process can be formalized as in (10), where A denotes the acoustic characteristics of the input and K_p denotes the phonotactic knowledge of the listener.

$$\text{Given [ebko]:} \qquad (10)$$
$$P(\text{ebuko}|A, K_p) > P(\text{ebiko}|A, K_p) > P(\text{ebko}|A, K_p)$$

Indeed, when framed in this way, the question of an integrated sublexical evaluation becomes similar to that of Sonderegger and Yu (2010), who investigated how an optimal listener compensates for vowel-to-vowel coarticulatory effects[5]. The optimal listener does not rely on phonetic minimality (A) or phonotactic predictability (K_p) alone, which often give conflicting predictions. Rather the listener considers them together to arrive at an output that is optimal, interpreting acoustic-phonetic cues based on prior knowledge of how they vary in certain phonotactic contexts. It should be noted that because the rational account views the choice of output candidates as a consequence of the listener's optimization process, which in turn is based on the listener's prior linguistic experience, the account also predicts different probabilities depending on the listener. With listeners of sufficiently divergent linguistic experiences, the optimal outputs would differ, and thus the integration of multiple cues under a rational framework is also applicable to crosslinguistic perception.

Second Hume and Mailhot (2013) also propose a similar integration of contextual and phonetic information using an information theoretic framework, but additionally point out that it is non-trivial to precisely quantify the informativity of various phonetic cues. In the interests of space, the current paper simply suggests that an information theoretic framework might also be useful in quantifying the relative informativity of a given segment's acoustic/phonetic cues. For example, let us assume the following vowel system: [i, y, u], which can be distinguished along the height ([±high]; F1), backness ([±back]; F2), and roundedness ([±round]; F3) dimensions. Since all three vowels are high (low F1), height cues have very low surprisal and thus are not informative. This leads to the prediction that listeners of this language would be more sensitive to the backness and roundedness cues than to height cues.

Lastly, a more sophisticated view of phonetic minimality that looks inside segments in more detail to precisely quantify and model the informativity of transitional cues seems necessary. The models in this paper and the previous literature on which the models are based on assume that the basic unit of phonological processes is the segment, but various lines of theoretical research such as Aperture Theory (Steriade, 1993), Articulatory Phonology (Browman and Goldstein, 1992; Gafos, 2002, *inter alia*), and more recently Q Theory (Shih and Inkelas, 2014) all have shown that representing segments in more detail results in a substantial increase in a framework's capacity to capture gradient and autosegmental phonological phenomena. The advantages of a more detailed segmental representation is not just theoretical, although it does complement formal approaches to perceptibility effects on phonology (e.g., P-map, Steriade, 2001; Uffmann, 2006). It also affords a more precise way to quantify and model which transitional cues an optimal listener relies on to select an acoustically "minimal" epenthetic segment and also how low-level phonetic information interacts with phonotactic information[6].

Assuming an optimal perceptual structure, where lexical processes apply first, followed by a language-specific combination of sublexical processes only when the lexical processes fail to choose an optimal output, we now turn back to the issue of [ɡ] and [ɡ^j]. What is puzzling about the [ɡ] and [ɡ^j] cases is that the two contexts require different sublexical mechanisms to predict the correct epenthetic vowel. After [ɡ], phonological alternations regard [a, u] as equally likely options for epenthesis. If the vowel is chosen based on phonotactic predictability, the wrong vowel [a] would be chosen, since it is the most phonotactically probable vowel in the given context. So then the target vowel [u] must be chosen based on phonetic minimality in this case. In the case of [ɡ^j], the situation is reversed. There are no zero-vowel phonological alternations after [ɡ^j], so all vowels are possible candidates for epenthesis. If a vowel is chosen based on phonetic minimality as with [ɡ], however, the chosen vowel would be incorrect as Japanese listeners perceive [i] in this context (Whang, 2019). So then after [ɡ^j], the epenthetic vowel must be chosen based on phonotactic predictability. In an integrated system as described above, the decision may be made as follows by an optimal listener. First, the burst noise of *a*-coarticulated [ɡ] and *u*-coarticulated [ɡ] are not only acoustically different, there is also evidence suggesting that Japanese listeners are sensitive to such coarticulatory differences (Whang, 2019). In other words, representing and quantifying the coarticulatory information can help predict the perceived similarity between the [ɡ] burst in a [ɡ]-C sequence and in a [ɡ]-[u] sequence relative to the burst in a [ɡ]-[a] sequence that makes [u] the phonetically minimal epenthetic vowel. Additionally, if the [ɡ] burst in a [ɡ]-C sequence is judged to be similar to an *u*-coarticulated [ɡ], the most phonotactically predictable vowel in this context would naturally be [u], resolving the apparent conflict between the phonetic minimality and phonotactic predictability accounts. The same process applies to [ɡ^j]. The epenthetic vowel that would result in a C-V transition that is acoustically the most similar to the fronted velar burst of [ɡ^j] is [i], again corroborating the predictions based on phonotactic predictability. It is perhaps premature to speculate further on the predictions of such an implemented model based on just two contexts. The present study, therefore, simply presents it as an example of how a rational approach can be used to bring together insights from various lines of research to integrate the seemingly contradictory

[5] The author is grateful to the editor for suggesting this connection.

[6] Quantifying (combinations of) cues/features that decide "phonetically minimal" vowels may also provide additional insight into why it is [o] that became the default epenthetic segment after both [t, d] in loanwords despite [e] being the most frequent after [d]. It seems likely that the transition from [d] to [o] is acoustically more consistent with a [d] burst than the transition from [d] to [e]. In the same vain, the transition from [t, d] to [u] should be acoustically even more similar to the stop bursts in consonantal contexts, and it would be interesting to see if [u] eventually replaces [o] as the default epenthetic vowel in loanwords for these contexts as the restriction against [tu, du] continues to weaken in Japanese. More recent loans provide some evidence for this regularization of [u] epenthesis (e.g., [twaɪs] → [tu̲waisu] "twice"; [tɹu] → [tu̲ɾuː] "true"; [dwɛliŋ] → [du̲eɾiŋgu] "dwelling"), but what this means for illusory epenthesis in these contexts remains to be seen.

predictions from different levels of linguistic processing, leaving more rigorous investigations for future studies.

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/jdwhang/RAILS-data.

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 New Faculty Startup Fund from Seoul National University and by the Deutsche Forschungsgemeinschaft (DFG) as part of SFB 1102 Information Density and Linguistic Encoding, Project C1 at Saarland University.

ACKNOWLEDGMENTS

The author is grateful to the editor and the reviewers for their insightful comments.

REFERENCES

Adriaans, F., and Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *J. Mem. Lang.* 62, 311–331. doi: 10.1016/j.jml.2009.11.007

Anderson, J. R. (1990). *The Adaptive Character of Thought*. New York, NY: Lawrence Erlbaum Associates, Inc.

Apoussidou, D. (2007). *The learnability of metrical phonology* (Ph.D. thesis). University of Amsterdam, Amsterdam, Netherlands.

Aylett, M., and Turk, A. (2004). The smooth 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

Beckman, M. (1982). Segmental duration and the 'mora' in Japanese. *Phonetica* 39, 113–135. doi: 10.1159/000261655

Browman, C. P., and Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica* 49, 155–180. doi: 10.1159/000261913

Cutler, A., Otake, T., and McQueen, J. M. (2009). Vowel devoicing and the perception of spoken Japanese words. *Acoust. Soc. Am.* 125, 1693–1703. doi: 10.1121/1.3075556

Dehaene-Lambertz, G., Dupoux, E., and Gout, A. (2000). Electrophysiological correlates of phonological processing: a cross-linguistic study. *J. Cogn. Neurosci.* 12, 635–647. doi: 10.1162/089892900562390

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., and Mehler, J. (1999). Epenthetic vowels in Japanese: a perceptual illusion? *J. Exp. Psychol.* 25, 1568–1578. doi: 10.1037/0096-1523.25.6.1568

Dupoux, E., Parlato, E., Frota, S., Hirose, Y., and Peperkamp, S. (2011). Where do illusory vowels come from? *J. Mem. Lang.* 64, 199–210. doi: 10.1016/j.jml.2010.12.004

Durvasula, K., Huang, H. H., Uehara, S., Luo, Q., and Lin, Y. H. (2018). Phonology modulates the illusory vowels in perceptual illusions: evidence from Mandarin and English. *Lab. Phonol.* 9, 1–27. doi: 10.5334/labphon.57

Durvasula, K., and Kahng, J. (2015). Illusory vowels in perceptual epenthesis: the role of phonological alternations. *Phonology* 32, 385–416. doi: 10.1017/S0952675715000263

Feldman, N. H., and Griffiths, T. L. (2007). "A rational account of the perceptual magnet effect," in *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (Nashville, TN).

Frank, M. C., and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science* 336, 998–998. doi: 10.1126/science.1218633

Frisch, S., Pierrehumbert, J., and Broe, M. B. (2004). Similarity avoidance and the OCP. *Nat. Lang. Linguist. Theory* 22, 179–228. doi: 10.1023/B:NALA.0000005557.78535.3c

Fujimoto, M. (2015). "Chapter 4: Vowel devoicing," in *Handbook of Japanese Phonetics and Phonology*, ed H. Kubozono (Berlin; Hong Kong; Munich: Mouton de Gruyter), 167–214. doi: 10.1515/9781614511984.167

Gafos, A. (2002). A grammar of gestural coordination. *Nat. Lang. Linguist. Theory* 20, 269–337. doi: 10.1023/A:1014942312445

Hall, K. C., Hume, E., Jaeger, F., and Wedel, A. (2016). *The Message Shapes Phonology*. University of British Columbia; University of Canterbury; University of Rochester; Arizona University.

Han, M. S. (1994). Acoustic manifestations of mora timing in Japanese. *J. Acoust. Soc. Am.* 96, 73–82. doi: 10.1121/1.410376

Huang, T. (2001). *Tone Perception by Speakers of Mandarin Chinese and American English*. Ohio State University Working Papers in Linguistics, 55.

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*, ed A. C. L. Yu (Oxford: Oxford University Press), 29–47. doi: 10.1093/acprof:oso/9780199573745.003.0002

Ito, J. (1986). *Syllable theory in prosodic phonology* (Ph.D. thesis). University of Massachusetts, Amherst, MA, United States.

Ito, J., and Mester, A. (1995). "Japanese phonology," in *Handbook of Phonological Theory*, ed J. Goldsmith (Cambridge, MA: Blackwell), 817–838.

Jescheniak, J. D., and Levelt, W. J. M. (1994). Word frequency effects in speech production: retrieval of syntactic information and of phonological form. *J. Exp. Psychol.* 20, 822–843. doi: 10.1037/0278-7393.20.4.824

Jusczyk, P., and Aslin, R. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cogn. Psychol.* 29, 1–23. doi: 10.1006/cogp.1995.1010

Jusczyk, P. W., Frederici, A., Wessels, J. M., Svenkerud, V. Y., and Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *J. Mem. Lang.* 32, 402–420. doi: 10.1006/jmla.1993.1022

Jusczyk, P. W., Luce, P. A., and Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *J. Mem. Lang.* 33, 630–645. doi: 10.1006/jmla.1994.1030

Kilpatrick, A., Kawahara, S., Bundgaard-Nielsen, R., Baker, B., and Fletcher, J. (2020). Japanese perceptual epenthesis is modulated by transitional probability. *Lang. Speech* 21, 203–223. doi: 10.1177/00238309209 30042

Kiparsky, P. (1982). "Lexical phonology and morphology," in *Linguistics in the Morning Calm, Vol. 2*, ed I.-S. Yang (Seoul: Hanshin), 3–91.

Kubozono, H. (2006). Where does loanword prosody come from?: a case study of Japanese loanword accent. *Lingua* 116, 1140–1170. doi: 10.1016/j.lingua.2005.06.010

Lassiter, D., and Goodman, N. D. (2013). "Context, scale structure, and statistics in the interpretation of positive-form adjectives," in *Semantics and Linguistic Theory, Vol. 23* (Santa Cruz, CA), 587–610. doi: 10.3765/salt.v23i0.2658

Maekawa, K., and Kikuchi, H. (2005). "Corpus-based analysis of vowel devoicing in spontaneous Japanese: an interim report," in *Voicing in Japanese*, eds J. van de Weijer, K. Nanjo, and T. Nishihara (Berlin: Mouton de Gruyter), 205–228.

Martin, A., Utsugi, A., and Mazuka, R. (2014). The multidimensional nature of hyperspeech: evidence from Japanese vowel devoicing. *Cognition* 132, 216–228. doi: 10.1016/j.cognition.2014.04.003

Mattingley, W., Hume, E., and Hall, K. C. (2015). "The influence of preceding consonant on perceptual epenthesis in Japanese," in *Proceedings of the 18th International Congress of Phonetics Sciences, Vol. 888* (Glasgow: The University of Glasgow), 1–5.

Mattys, S. L., and Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition* 78, 91–121. doi: 10.1016/S0010-0277(00)00109-8

Mattys, S. L., White, L., and Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: a hierarchical framework. *J. Exp. Psychol.* 134:477. doi: 10.1037/0096-3445.134.4.477

Maye, J., Werker, J. F., and Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82, B101–B111. doi: 10.1016/S0010-0277(01)00157-3

Mohanan, K. P. (1982). *Lexical phonology* (Ph.D. thesis). Distributed by IULC Publications, Cambridge, MA, United States.

Monahan, P. J., Takahashi, E., Nakao, C., and Idsardi, W. J. (2009). Not all epenthetic contexts are equal: differential effects in Japanese illusory vowel perception. *Jpn. Kor. Linguist*. 17, 391–405.

Morgan, J. L., and Saffran, J. R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Dev*. 66, 911–936. doi: 10.2307/1131789

Norris, D. (2006). The Bayesian reader: explaining word recognition as an optimal Bayesian decision process. *Psychol. Rev*. 113:327. doi: 10.1037/0033-295X.113.2.327

Ogasawara, N. (2013). Lexical representation of Japanese high vowel devoicing. *Lang. Speech* 56, 5–22. doi: 10.1177/0023830911434118

Ogasawara, N., and Warner, N. (2009). Processing missing vowels: allophonic processing in Japanese. *Lang. Cogn. Process*. 24, 376–411. doi: 10.1080/01690960802084028

Pierrehumbert, J. (1993). "Dissimilarity in the Arabic verbal roots," in *Proceedings of the North East Linguistic Society, 23* eds A. Schafer and A. Schafer (Amherst, MA: GLSA Publications).

Rose, Y., and Demuth, K. (2006). Vowel epenthesis in loanword adaptation: representational and phonetic considerations. *Lingua* 116, 1112–1139. doi: 10.1016/j.lingua.2005.06.011

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

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Techn. J*. 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Shaw, J., and Kawahara, S. (2018). The lingual articulation of devoiced /u/ in Tokyo Japanese. *J. Phonet*. 66, 100–119. doi: 10.1016/j.wocn.2017.09.007

Shaw, J., and Kawahara, S. (2019). Effects of surprisal and entropy on vowel duration in Japanese. *Lang. Speech* 62, 80–114. doi: 10.1177/0023830917737331

Shih, S., and Inkelas, S. (2014). "A subsegmental correspondence approach to contour tone (dis)harmony patterns," in *Proceedings of the 2013 Meeting on Phonology*, eds J. Kingston, C. Moore-Cantwell, J. Pater, and A. Rysling (Washington, DC: Linguistic Society of America). doi: 10.3765/amp.v1i1.22

Smith, J. (2006). "Loan phonology is not all perception: evidence from Japanese loan doublets," in *Japanese/Korean Linguistics*, ed T. J. Vance (Stanford: CSLI Publications), 14.

Sonderegger, M., and Yu, A. (2010). "A rational account of perceptual compensation for coarticulation," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, eds S. Ohlsson and R. Catrambone (Portland), 375–380.

Steriade, D. (1993). "Closure, release, and other nasal contours," in *Nasals, Nasalization, and the Velum*, eds M. K. Huffman and R. A. Krakow (San Diego, CA: Academic Press), 401–470. doi: 10.1016/B978-0-12-360380-7.50018-1

Steriade, D. (2001). *The Phonology of Perceptibility Effects: The P-Map and Its Consequences for Constraint Organization*. University of California Los Angeles.

Tesar, B., and Prince, A. (2007). "Using phonotactics to learn phonological alternations," in *CLS 39, Vol. 2* (Chicago, IL), 209–237.

Uffmann, C. (2006). Epenthetic vowel quality in loanwords: empirical and formal issues. *Lingua* 116, 1079–1111. doi: 10.1016/j.lingua.2005.06.009

Vitevitch, M., Luce, P., Charles-Luce, J., and Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Lang. Speech* 40, 47–62. doi: 10.1177/002383099704000103

Werker, J., and Lalonde, C. (1988). Cross-language speech perception: initial capabilities and developmental change. *Dev. Psychol*. 24, 672–683. doi: 10.1037/0012-1649.24.5.672

Werker, J., and Tees, R. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behav. Dev*. 7, 49–63. doi: 10.1016/S0163-6383(84)80022-3

Whang, J. (2018). Recoverability-driven coarticulation: acoustic evidence from Japanese high vowel devoicing. *J. Acoust. Soc. Am*. 143, 1159–1172. doi: 10.1121/1.5024893

Whang, J. (2019). Effects of phonotactic predictability on sensitivity to phonetic detail. *Lab. Phonol*. 10:8. doi: 10.5334/labphon.125

Zuraw, K. (2000). *Patterned exceptions in phonology* (Ph.D. thesis). University of California Los Angeles, Los Angeles, CA, United States.

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 in Psychology

ORIGINAL RESEARCH
published: 06 September 2021
doi: 10.3389/fpsyg.2021.679278

Mishearing as a Side Effect of Rational Language Comprehension in Noise

Marjolein Van Os[1]*, *Jutta Kray*[2] *and Vera Demberg*[1,3]

[1] *Department of Language Science and Technology, Saarland University, Saarbrücken, Germany,* [2] *Department of Psychology, Saarland University, Saarbrücken, Germany,* [3] *Department of Computer Science, Saarland University, Saarbrücken, Germany*

Language comprehension in noise can sometimes lead to mishearing, due to the noise disrupting the speech signal. Some of the difficulties in dealing with the noisy signal can be alleviated by drawing on the context – indeed, top-down predictability has shown to facilitate speech comprehension in noise. Previous studies have furthermore shown that strong reliance on the top-down predictions can lead to increased rates of mishearing, especially in older adults, which are attributed to general deficits in cognitive control in older adults. We here propose that the observed mishearing may be a simple consequence of rational language processing in noise. It should not be related to failure on the side of the older comprehenders, but instead would be predicted by rational processing accounts. To test this hypothesis, we extend earlier studies by running an online listening experiment with younger and older adults, carefully controlling the target and direct competitor in our stimuli. We show that mishearing is directly related to the perceptibility of the signal. We furthermore add an analysis of wrong responses, which shows that results are at odds with the idea that participants overly strongly rely on context in this task, as most false answers are indeed close to the speech signal, and not to the semantics of the context.

Keywords: speech comprehension, background noise, mishearing, false hearing, predictive context, aging

OPEN ACCESS

Edited by:
Rory Turnbull,
Newcastle University, United Kingdom

Reviewed by:
Rüdiger Christoph Pryss,
Julius Maximilian University of Würzburg, Germany
Anita Mehta,
University of Oxford, United Kingdom

****Correspondence:***
Marjolein Van Os
vanos@coli.uni-saarland.de

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

Received: *11 March 2021*
Accepted: *23 July 2021*
Published: *06 September 2021*

Citation:
Van Os M, Kray J and Demberg V (2021) Mishearing as a Side Effect of Rational Language Comprehension in Noise. Front. Psychol. 12:679278. doi: 10.3389/fpsyg.2021.679278

INTRODUCTION

Noisy Channel Model of Rational Communication

When listening to speech, there are usually at least two sources of information available to decode the speaker's message: There is the sensory information in the form of the acoustic speech signal, and there is also contextual information that can help guide predictions (Boothroyd and Nittrouer, 1988; Nittrouer and Boothroyd, 1990). We rarely listen to other people speaking in perfectly quiet surroundings. Often, there is a lot of noise going on in the background, for example, other people speaking, traffic noise, or working machinery. The noise puts extra strain on our speech comprehension processes, something especially older adults can struggle with (Li et al., 2004).

Comprehenders take into account uncertainty in the perceptual input (for example, due to background noise). The noisy channel model (Shannon, 1949; Levy, 2008; Levy et al., 2009) proposes that language comprehension is a rational process, where we make use of all available sources of information. Bottom-up information from the speech signal is supplemented with top-down predictions of what the speaker is likely to say. Combining these two sources of information are a sensible strategy to maximize comprehension. Let's take, for example, the sentence "He buys the bar," In background noise, the listener might comprehend this as "He buys the car," where *car* might be more probable given the context of buying than *bar,* while sounding similar. The actual comprehended word w' is determined as $w' = \text{argmax}_i\ P(w_i|\text{context}) * P(s|w_i)$ where $P(w_i|\text{context})$ is the probability of the word given the preceding context, i.e., the top-down probability of a word, and $P(s|w_i)$ is the probability of the perceived signal for that word w_i. The task of the listener consists of identifying candidate word w_i for which this probability given context and probability of signal fitting that word is maximal. This means that there is a trade-off between top-down and bottom-up information, where the probability distribution is shaped differently depending on the clarity of the acoustic signal. A noisier signal leads to a flatter distribution: There are more words w_i for which the perceived signal s has a relatively high probability, compared to a situation in which signal s is clearly intelligible. In cases where we therefore have a relatively flat probability distribution for $P(s|w_i)$, the top-down probability $P(w_i|\text{context})$ will dominate what comes out as the most likely word w_i in the argmax calculation (besides words like *car*, also other words that frequently occur in a context of buying that share some overlap with the signal are probable based on the context). Under high noise, the top-down information will hence count more than the uncertain bottom-up information due to the stronger peaks in its distribution, leading to stronger reliance on prediction. In most cases, this will be beneficial to language comprehension, as it means that likely words can still be deciphered under noisy conditions. However, these predictions can also come at the cost of *mishearing*, where speech is misunderstood due to strong expectations (Rogers et al., 2012; Sommers et al., 2015; Failes et al., 2020). Rogers et al. (2012) explained this mishearing effect through general deficits in cognitive control for the older adults. They additionally report that older adults do not only show increased levels of mishearing compared to younger adults, but that they also report higher confidence in having heard a word which was not actually spoken.

In the present article, we argue that the larger mishearing effect observed in older adults compared to younger adults may be a simple consequence of rational integration of the bottom-up and top-down information, i.e., that their performance is not necessarily an effect related to deficits in cognitive control, but may reflect a combination of stronger top-down expectations due to increased linguistic experience, and lower confidence in the bottom-up input, due to first experiences of hearing loss. We have controlled our stimuli in such a way that in case of general cognitive causes, we should find no difference between our items (cognitive control should not be affected), while if the mishearing effect depends on clarity of the signal, we will find differences in comprehension performance. Different sound types have different signals that are easier or more difficult to distinguish in background noise, and the noisy channel model predicts that even minor changes in how well the acoustic signal can be perceived, can lead to a difference in the trade-off between top-down and bottom-up information. For example, the short burst of plosives is harder to distinguish in background noise than the steadier signal of a vowel. The noisy channel model would predict that in stimuli with plosives, listeners rely more on top-down prediction than in stimuli with vowel contrasts.

In the present study, we aim to investigate how background noise affects speech comprehension in younger and older adults, in situations where there is a predictive sentence context available that might facilitate or hinder speech recognition. Comparing younger and older adults is interesting, as older adults have more language experience and hence should have better expectations (Pichora-Fuller, 2008; Sheldon et al., 2008), while at the same time, they may already be subject to some hearing loss and know to trust the incoming signal less. Given both of these factors, we would predict based on the noisy channel model that older adults show larger effects of the top-down predictions on interpretation, and thus be subject to stronger mishearing effects than younger adults.

In the following sections, we will describe the effect of background noise on speech understanding in general ("Effect of Background Noise on Speech Understanding"), and in older adults more specifically ("Age Differences in Language Comprehension Under Noise"). We will discuss false hearing in more detail ("Age-Related Differences in False Hearing") and introduce the aims of the current study ("The Present Study").

Effect of Background Noise on Speech Understanding

Background noise has a negative effect on speech comprehension in younger as well as in older adults. It can lead to energetic masking, where both the speech signal and the competing noise have energy in the same frequency bands at the same time (Brungart, 2001). The acoustic cues that listeners need for sound identification are masked by the noise, or if the background noise is competing speech, its acoustic cues can "attach" themselves to the target speech (Cooke, 2009). The type of noise, for example, white noise, babble noise, or competing speech from a single speaker, might have different effects on the target speech. The present study uses multi-speaker babble noise, where none of the speakers are understandable.

Relevant for the current study is also the distinction that can be made between consonants, in particular plosives, and vowels. These different types of sound might be affected in different ways by various types of background noise. Plosives, on the one hand, consist of a closure of some part of the vocal tract, followed by a short burst of energy. This burst can easily be masked by noise, if that happened at the same time. On the other hand, vowels generally have a longer, more steady signal with a higher intensity, that can be easier to distinguish in background noise. Their energy primarily lies between 250 and 2000 Hz (first and second formant, Flanagan, 1955), thus lower than that of consonants, which have information also in higher formants (Edwards, 1981; Alwan et al., 2011). Spectral frequency information is in particular important for identifying the place of articulation in plosives (Liberman et al., 1954; Edwards, 1981).

When it comes to background noise, not only the type of background noise matters, but also the level of the noise, the level of background noise is commonly measured in Signal to Noise Ratio (SNR). It quantifies the relation between the amplitude of the speech signal and the amplitude of the background noise. A negative SNR means that the background noise is stronger than the speech signal (which is thus more difficult to understand), and a positive SNR means that the speech signal is stronger than the background noise. In the case of 0 SNR, both the noise and the speech are equally strong. In the present study, the noise levels have been set at 0 SNR and −5 SNR, so that we can investigate whether mishearings change as a function of the difficulty of the listening condition.

Age Differences in Language Comprehension Under Noise

There are differences between younger and older adults even in quiet situations. With increasing age, there are changes in auditory processing (Gordon-Salant et al., 2010; Helfer et al., 2020). In particular, changes in the inner ear and neural pathways can lead to age-related hearing loss, presbycusis, in which the highest frequencies (4–8 kHz) are most affected and continue to get worse in older adults (Gates and Mills, 2005). When the hearing loss progresses to frequencies of 2–4 kHz, this affects speech comprehension, and in particular understanding of voiceless consonants. Older adults also often have reduced ability to differentiate between different frequencies, to discriminate spectral and temporal transitions in the speech signal, and to localize sound sources (Schuknecht and Gacek, 1993; Chisolm et al., 2003; Tun et al., 2012; Helfer et al., 2020). These declines lead to greater difficulty understanding speech in adverse listening conditions (Pichora-Fuller et al., 1995; Li et al., 2004; Schneider et al., 2005; Pichora-Fuller et al., 2017). Additionally, there are cognitive changes with increasing age. Older adults have been found to show decreased attention, working memory, executive functions, and processing speed (Salthouse, 1990, 1996; Lindenberger and Ghisletta, 2009; Tun et al., 2012; Tucker-Drob et al., 2019). These abilities all play a role in speech comprehension, which will thus be negatively impacted as well.

General language abilities are well preserved in old age, and older adults are able to compensate for their reduced auditory and cognitive abilities by using knowledge-based factors, such as supportive sentence context (Stine and Wingfield, 1994; Wingfield et al., 1995, 2005). Studies compared groups of younger adults with groups of older adults to determine how noisy environments and informative contexts might affect the latter group differently than the former (Hutchinson, 1989; Pichora-Fuller et al., 1995; Sommers and Danielson, 1999; Dubno et al., 2000; Benichov et al., 2012). The results showed that older adults are generally more adversely affected by background noise than younger adults and that older adults rely more heavily on the provided sentence context than younger adults. In fact, older adults have been shown to rely on contextual prediction to such an extent that the predictions can make up for the adverse effect of noise (Wingfield et al., 2005) and other adverse listening conditions (Wingfield et al., 1995; Lash et al., 2013). Older adults might be particularly adept at using contextual information as a compensation mechanism, because every day they are exposed to challenging listening situations. They may have come to rely on using contextual cues to support speech comprehension processes, so that with age and experience, increased attention is allocated to higher-order knowledge structures (Steen-Baker et al., 2017). Koeritzer et al. (2018) investigated how background noise and ambiguous words in sentences affect recognition memory for spoken sentences. They presented the sentences in SNRs of +5 and +15, thus with an increased acoustic challenge, but with intelligible speech. Results showed that recognition memory was worse for acoustically challenging sentences and sentences containing ambiguous words, and older adults performed worse than younger adults in the ambiguous sentences in noise. Koeritzer et al. concluded that in particular older listeners rely on domain-general cognitive processes in challenging listening conditions, even when the speech is highly intelligible. Rogers et al. (2012) concluded that older adults are more biased to respond consistently with the context than younger adults, due to general deficits in cognitive control. However, other studies have argued that older adults' reliance on context is due to predictions and more language experience (Wingfield et al., 2005; Sheldon et al., 2008).

Age-Related Differences in False Hearing

Predictions made based on context might come at a cost. Older adults have been found to show higher rates of "false hearing" than younger adults (Rogers et al., 2012, p. 33). Here, false hearing is defined as a "mistaken high confidence in the accuracy of perception when a spoken word has been misperceived". In their study, Rogers and colleagues used a priming paradigm in which they paired semantically related words (*barn*/*hay*). In a training phase, participants were familiarized with these associations. In a subsequent testing phase, the cue word (*barn*) was presented in clear listening conditions, and subsequently the target word was presented in noise. There were three conditions: (1) congruent, where the target word was the same as in the training phase (e.g., *hay*); (2) incongruent, where the target word was a phonological neighbor that formed a minimal pair with the word in training (e.g., *pay*); and (3) baseline, where the target word was unrelated to the training word (e.g., *fun*). Both younger and older adults indicated which words they had heard and how confident they were that they had identified the word correctly. The results of the study showed that older adults made use of the trained context more often and with more confidence than younger adults, even when the presented words were not matched in the training phase. Thus, older adults showed a larger false hearing effect than the younger adults. Comparable results using a similar priming paradigm have been found by Rogers and Wingfield (2015) and Rogers (2017). In a follow-up study, Rogers (2017) investigated whether the false hearing effect is caused by semantic priming or repetition priming, by manipulating the number of exposures to the training cue-target pairs. The results showed that an increased number of exposures did not increase the effect of false hearing, but that this effect was strongest when the cue-target pair was not presented at all during the training phase. These observations indicate that the false hearing effect is caused by semantic priming rather than repetition priming, suggesting that false hearing relies on top-down semantic associations in the context.

More recent studies have investigated false hearing using a more naturalistic paradigm than the priming paradigm used in previous studies. Sommers et al. (2015) and Failes et al. (2020) used sentences rather than word pairs, in three conditions. A neutral carrier phrase formed the baseline condition, and there were congruent (e.g., "The shepherd watched his sheep.") and incongruent ("The shepherd watched his sheath.") sentences. Here, the sentence-final target items differed in the first or last phoneme, while controlling for frequency and neighborhood density. Participants listened to the sentence in quiet, and the target item embedded in babble noise. Identification accuracy and confidence ratings were analyzed, showing that older adults performed better than younger adults on congruent trials, but had a higher false alarm rate for the incongruent trials. Older adults were more confident of these false alarms than younger adults, showing the increased false hearing effect for older participants.

Like these two previous studies, the present study investigated the predictability of the target word based on the context, but in German instead of English. While we are mainly interested in mishearings, we do collect confidence ratings of the participants' responses to also investigate false hearing. Unlike previous studies, we systematically vary the sound type change between the target and distractor item so that only one phonetic aspect of the phoneme is changed, in order to investigate whether different types of sounds are affected by false hearing to a similar extent. Finding any differences between sound types (vowel quality vs. place of articulation in plosives) can help distinguish between accounts explaining the mishearing and false hearing effects, as this would mean listeners behave optimally based on the perceived information. If mishearing and false hearing in older adults is based on general deficits in cognitive control (Rogers et al., 2012), we should find the same effect for the different sound types.

Besides false hearing, larger effects for older adults compared to younger adults have been found for false memories (Hay and Jacoby, 1999) and false seeing (Jacoby et al., 2012). These processes seem to share a common mechanism, as Failes et al. (2020) found that participants who showed more false hearing, also were more likely to have false memories, and Jacoby et al. (2012) link false seeing to false hearing. In all cases, there seem to be top-down processes that lead to the false perceptions by overriding bottom-up signals (Bruner, 1957; Balcetis and Dunning, 2010 for false seeing; Roediger and McDermott, 1995 for false memory).

The Present Study

Our study investigates how bottom-up auditory processes and top-down predictive processes interact in speech comprehension. We tested both younger and older adults in our experiment, as we expect age differences in the quality of top-down and bottom-up processes. Participants completed a word recognition task, where sentences were either presented in quiet or in background noise, and where the sentence context could be used to predict the sentence-final target word or not. These sentence-final target words were designed to be minimal pairs with respect to pronunciation, so that in the low predictability context, the word sounded very similar to the word that in fact did fit the sentence semantically. This allowed us to investigate whether listeners are able to rely on small acoustic cues for word recognition, even in background noise, while keeping sentence contexts equal across conditions.

The main question that the present study aims to address is the replicability of mishearings in German. Like previous studies (Sommers et al., 2015; Failes et al., 2020), we use a paradigm of word recognition in sentences, where the context is predictive or unpredictive of the target word. We add a quiet condition without added background noise as a baseline condition, which will allow us to make sure that hearing ability between groups is comparable with respect to our materials. It is also possible that we will observe a general increase of mishearing in older adults compared to younger adults, even in the quiet condition. This would be an interesting finding, as it would show that older adults rely more on context than the acoustic signal even if the acoustic signal is easily accessible, comparable to the finding that older adults rely more on domain-general cognitive processes in challenging listening conditions with high intelligibility (Koeritzer et al., 2018). Like previous studies, we will collect confidence ratings to investigate false hearing as a second point of interest.

To be able to distinguish between different accounts that explain the mishearing effect, we investigate the effect of noise on different types of speech sounds, and how these are affected by false hearing. We constructed our stimuli such that the minimal pairs in our experiments differed in just one feature: either vowel quality or place of articulation in plosives. The acoustic properties of our manipulation in vowels vs. plosives differ in various ways. First, vowel sounds have a longer and steadier signal compared to the relatively short burst of the plosives. Second, higher frequencies are more informative for plosives than for vowels, in particular for place of articulation (Liberman et al., 1954; Edwards, 1981; Alwan et al., 2011), which is the contrast in our minimal pairs. Based on the noisy channel model, we expected to find that the top-down predictions play a larger role in the case of plosives, as here the signal of the target and distractor are more similar to each other compared to the vowel condition, and thus will have more flat probability distributions (where both the target and the distractor have a similar probability of leading to the observed acoustic signal) based on the bottom-up processes. Listeners try to overcome this by relying more on the contextual information that is more easily accessible and gives distinguishing information. Furthermore, we expected that this difference between vowels and plosives may also be more pronounced in older adults, as hearing ability in high-frequency ranges is known to degrade during aging (Gates and Mills, 2005). Listeners optimally combine bottom-up and top-down probabilities, leading to mishearing in difficult listening conditions where the choice of the most likely word is mostly determined by the top-down prediction, an effect that is stronger for older adults as they compensate for age-related reductions in auditory and cognitive processing, but still rationally combine the acoustic and top-down information that is available to them.

MATERIALS AND METHODS

Participants

A total of 93 native German speakers participated in the present experiment, for which we used the recruitment platform Prolific (prolific.co). We excluded seven older participants based on their performance in the quiet condition, because their number of distractor responses exceeded that of the younger adults. In this way, we ensured equal hearing abilities with respect to our stimuli across ages, as we were not able to collect hearing thresholds for our participants (because in-lab experiments were not possible at the time of conducting this study). The high number of unexpected responses in this relatively easy condition without background noise might also have been due to difficulty playing the audio or doing the task. The mean age of our final group of participants was 40 years (age range = 18–68 years), 43 were male. While all participants were self-reported native speakers of German, their current countries of residence varied as: 55 lived in Germany, 12 in the United Kingdom, 4 in Austria, 3 in Ireland and Spain, 2 in the United States, 1 in each of France, Israel, Portugal, Poland, and South Korea. Three did not list their country of residence. Three out of our 87 participants reported to not speak other languages besides German, all three were older adults. From the remaining 84 participants, the languages spoken besides German were most often English (reported by 82 participants), French (reported by 21), and Spanish (reported by 14). In the post-experimental questionnaire, most participants reported no hearing issues or use of hearing aids. One participant (age 29) reported tinnitus, and one reported reduced hearing in his right ear (age 48, 60% hearing left). In order to check for any effects of education, we computed Spearman's correlation between participants' age and education level. This correlation was small ($\rho = 0.2$, $p = 0.08$), indicating that the older participants in our study were slightly more highly educated than the young participants. All participants gave informed consent, and the study was approved by the Deutsche Gesellschaft für Sprachwissenschaft Ethics Committee. The experiment lasted approximately 20 min and all participants received 3.12 Euro as compensation for their participation.

Materials and Task

German minimal pairs were selected from the CELEX lexical database (Baayen et al., 1995), based on their phonetic transcription. These minimal pairs were chosen so that the contrast was in the middle of the word, rather than word-initial or word-final, as there were most pairs available for this position for the sound contrasts. In order to test the hypothesis that the effect of noise may be more detrimental to understanding the spoken target word for pairs that differed in a plosive than for those that differed in a vowel, we included both vowel contrasts (tense/lax: i/ɪ, y/ʏ, u/ʊ, ɛ/ɶ, o/ɔ, ɐ/ə) and plosive contrasts (paired on place of articulation contrasts: p/t, p/k, t/k, b/d, b/g, d/g). First, all pairs were inspected, and we excluded those that were not true minimal pairs (usual pronunciation differs from transcription), that had one or two too infrequent words (regionally used or technical terms), or those of which the words differed in gender or part of speech so that constructing stimuli for them was not possible. By controlling the phonetic contrast and part of speech of the words, we were not able to control for word frequency or neighborhood effects. Sentences were constructed around the minimal pairs, so that the target word appeared in sentence-final position and the word would be predictable from the sentence context. All stimuli were subjected to cloze testing using native German speakers on the Prolific platform. Cloze probabilities for each item were calculated based on the answers of 10 participants. We aimed for high cloze probabilities. Therefore, all stimuli that were still scoring too low on cloze probability were revised. Three rounds of cloze testing were completed, until we had 120 high-predictability sentence pairs (240 items in total). The cloze values ranged from 0.5 to 1 (mean = 0.72) for the 136 items constructed under strict conditions. In 104 cases, the cloze was still quite low. We relaxed the high cloze requirement when even after multiple revisions, there was a high cloze competitor that differed only in the prefix (*laden* vs. *aufladen* for "to charge"*)* or that was too highly frequent to allow us to improve the sentence (*sieden* vs. more frequent *kochen* for "to boil"), and included these items even though they had a lower cloze probability than 0.5. The average cloze for all items, including those with the relaxed requirements, was 0.52. None of the participants took part in more than one of the rounds of cloze testing, and none of them participated in the main experiment.

To make the unpredictable stimuli, we swapped the two sentence-final target words, aiming for unpredictable but grammatically correct swaps wherever possible. In practice, this meant that all swapped sentences were unpredictable and implausible. Almost all sentences were still grammatically correct after swapping the target word, but two out of 240 swapped sentences became grammatically incorrect (for example, an argument was missing for a transitive verb). This resulted in 120 sets of four sentences, with two predictable and two unpredictable sentences of the minimal pair ($N = 480$). Plausibility ratings were collected for all 480 items, again using the Prolific environment. Each item was rated 10 times, and ratings were averaged. Again, none of the participants took part in the main experiment. Plausibility was rated on a scale from

TABLE 1 | Example Stimuli.

1A	Am Pool im Hotel gab es nur noch eine freie **Liege** *At the pool in the hotel there was only one free **lounger** left*	HP
1B	Nach vier Jahren heiratete Paul seine große **Liebe** *After four years, Paul married his big **love***	HP
1C	Am Pool im Hotel gab es nur noch eine freie **Liebe** *At the pool in the hotel there was only one free **love** left*	LP
1D	Nach vier Jahren heiratete Paul seine große **Liege** *After four years, Paul married his big **lounger***	LP

Highly predictable sentences (HP) were made based on minimal pairs (Liebe/Liege) in 1A and 1B; then, sentence-final target words were swapped to make low-predictability items (LP) with the sentence frames of 1A and 1B, resulting in 1C and 1D. English translations have been given in italics.

1 (completely implausible) to 5 (completely plausible). The predictable sentences had a mean plausibility rating of 4.60 (*SD* = 0.41), and the unpredictable sentences had a mean plausibility rating of 1.73 (*SD* = 0.59). Example stimuli can be found in **Table 1**.

Recordings were made of all predictable sentences (240 in total). The sentences were read by a female speaker, who was a native speaker of German. The speaker was instructed to read slowly and to pay attention to not include any slips of the tongue or hesitations. Sentences that were not read as intended or included slips of the tongue were repeated until each sentence was recorded in a clean version suitable for testing.

Unpredictable sentences were constructed via cross-splicing of the recordings of predictable sentences, in order to make sure that the intonation and stress patterns were identical across conditions and not indicative of the unpredictable items. The splicing was performed using Praat (Boersma and Weenink, 2020, version 6.1.05) and resulted in the total of 480 sentences. All cross-spliced unpredictable items were listened to carefully, to identify any problems related to cross-splicing, and corrected by adapting the slicing boundary or adapting the pitch contours. This was done by the first author as well as two native German student assistants. The final 480 sentences all sounded natural for the purposes of the experiment.

All sentences were embedded in background noise, which was café noise (BBC Sound Effects Library, Crowds: Interior, Dinner-Dance[1]), a multi-speaker babble noise where none of the speakers was intelligible. This was done in two different Signal to Noise (SNR) ratios, namely, 0 (meaning the target sound and the background noise were equally loud) and −5 (meaning that the background noise was 5 dB louder than the target sound). These values were chosen by the authors as challenging but not impossible to understand. As we are interested in the effect of background noise and sentence context on the intelligibility of the sentence-final target word, we took the mean intensity of each target word and calculated the SNR values based on this value, rather than the mean intensity of the sentence. Because the intensity of a spoken sentence tends to drop toward the end (Vaissière, 1983), it would mean the SNRs were actually lower for the target word, in case the mean sentence intensity was to be used. The noise was the same level throughout the sentence and started 300 ms before sentence-onset and continued for 300 ms after sentence-offset. This way, we gave participants a chance to focus on the speech in the noise. This was also the reason not to keep the sentence clear and embed only the target word in noise: We feared participants would not have time to get used to the added noise and it would be a less natural way of presenting the stimuli. Besides the noise conditions, there was a quiet condition, resulting in three different noise levels, quiet, 0 SNR, and −5 SRN.

Design

The experimental items were arranged in a Latin Square design. Twenty-four different lists were constructed, consisting of 60 items each. These lists were constructed in such a way that each noise level and each predictability level occurred the same number of times and that each item appeared only once per list (same target pair or same predictability sentence). This was done in a crossed design, so that out of the 60 items, 30 were predictable and 30 were unpredictable. Out of each set of 30 items, 10 were presented in quiet, 10 in 0 SNR noise, and the remaining 10 in −5 SNR noise. The items were blocked by noise level, starting with 0 SNR, followed by −5 SNR, and ending with the quiet condition. This blocking was chosen to give participants a chance to maximally adapt to the noise and the task, starting with the relatively easy noise condition before being presented with the relatively hard noise condition. The quiet condition was presented at the end, so as not to give away the goal of the experiment at the start. Each list was preceded by a practice block, consisting of four items. This short practice block made the participants familiar with the task and online testing environment. All noise levels (quiet, 0SNR, and −5 SNR) were presented during the practice block.

Procedure

The experiment was hosted on Lingoturk, a crowdsourcing client (Pusse et al., 2016). Participants completed the experiment on a computer in a quiet room and using the Chrome web browser. They were instructed to use either headphones or speakers. In the experiment, participants had to listen to the sentence and report the final word they had heard. Before the start of the main experiment, the participant saw a series of instructions detailing the task. Participants were asked to listen carefully and report what they heard. We did not explicitly state that the sentences could be misleading. These screens included a sound check as well, so that the participant had the opportunity to make sure sound was being played correctly. In the main task of the experiment, the sentence, minus the target word, was presented on the screen in written form. We opted to include the written sentence up until the target word to make sure participants were able to use the context also in noisy conditions. A text box was provided for the participant to type their answer. Additionally, they rated their confidence in having given the correct answer on a scale from 1 (completely uncertain, guessed) to 4 (completely certain). At the start of a trial, the sound played automatically while

[1] http://bbcsfx.acropolis.org.uk/

the screen showed a fixation cross. Next, a screen with the two questions appeared after the recording had finished playing. The next item started playing as soon as the participant clicked to go to the next trial. As mentioned above, the experiment started with a short practice session consisting of four items, which were presented after the participant had seen all instructions. A schematic overview of the experiment is presented in **Figure 1**.

Analysis

After data collection had been completed, all received answers were first classified automatically on whether it was the *target*, the word that was played in the audio (e.g., in example 1A in **Table 1** "Liege"/"lounger"), the similar sounding *distractor* (e.g., in 1A "Liebe"/"love"), or were a different word entirely (e.g., in 1A "Platz"/"space," *wrong*). The list of answers that had been classified as *wrong* was then checked by the first author and a native German-speaking student assistant, to correct misclassifications because of typos. In our statistical analyses, we included the trial number of each block as a control variable to check for any learning effects. We analyzed the high-predictability and low-predictability items separately due to ceiling effects in the high-predictability condition. To determine whether participants relied on the sentence context or on the speech signal, we coded the semantic fit of the incorrect responses (fitting or not fitting), as well as the phonetic distance between the incorrect responses and target and distractor items. We made phonetic transcriptions based on the Deutsches Aussprachewörterbuch (German Pronunciation Dictionary; Krech et al., 2009) and calculated the weighted feature edit distance using the Python package *Panphon* (Mortensen et al., 2016). This distance was normalized by dividing it by the longest of the two compared words. The normalized distance fell between 0 and 1.

RESULTS

In the first part of the result section, we will report the results on age differences in response accuracy in the high- and low-predictability conditions investigating mishearing. In the second part, we will analyze confidence ratings and investigate age differences in the false hearing effect. We used general linear mixed models (GLMM; Quené and Van den Bergh, 2008, for a tutorial see Winter, 2019), implemented in the lme4 package (Bates et al., 2015) in R (R Development Core Team, 2020) to analyze our data. These models allow both fixed and random effects, letting us control for variation on the participant- and item-level (Baayen et al., 2008; Barr et al., 2013). To improve convergence, all models were run using the bobyqa optimizer and increased iterations to $2 \cdot 10^5$. Model comparisons were made to guide model selection based on the Akaike Information Criterion (AIC), and models with the lowest AIC are reported below.

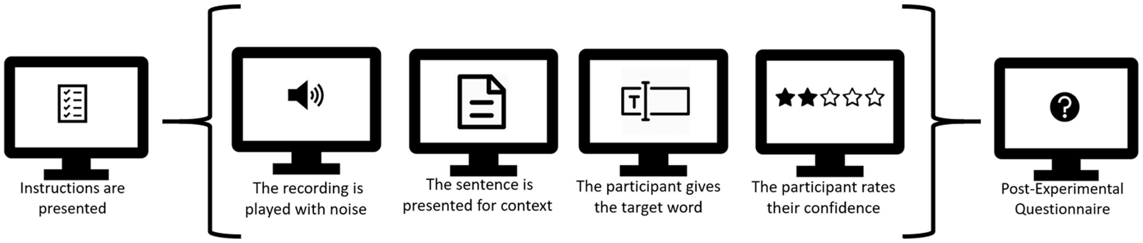

Instructions are presented
The recording is played with noise
The sentence is presented for context
The participant gives the target word
The participant rates their confidence
Post-Experimental Questionnaire

FIGURE 1 | This figure shows the different stages of the experiment, with a single trial between brackets. Participants completed four practice trials and sixty experimental trials.

TABLE 2 | Model outcomes for high- and low-predictability items.

	High-predictability items subset					Low-predictability items subset				
	Estimate	***SE***	***z*-value**	***value of p***		**Estimate**	***SE***	***z*-value**	***value of p***	
Intercept (quiet, P)	5.77	0.61	9.53	< 0.001	***	2.16	0.31	7.03	< 0.001	***
Noise −5SNR	−1.61	0.64	−2.48	< 0.05	*	−6.32	0.39	−16.09	< 0.001	***
Noise 0SNR	−1.38	0.65	−2.11	< 0.05	*	−4.87	0.32	−15.22	< 0.001	***
Age	−0.25	0.27	−0.93	0.35		−0.25	0.21	−1.18	0.24	
Trial No	−0.12	0.20	−0.61	0.54		0.47	0.08	5.97	< 0.001	***
ContrastVP V	0.24	0.40	0.61	0.54		1.55	0.34	4.63	< 0.001	***
Age: ContrastVP V	0.23	0.40	0.57	0.57		−0.33	0.14	−2.29	< 0.05	*
Age: Trial No	0.10	0.20	0.47	0.64		−0.004	0.08	−0.06	0.95	

This table presents the analyses for the subsets of high and low predictability items. The response variable is the participants' answer type, distractor (0), or target (1).

High Predictability Helps Comprehension in Noise

We are interested in whether listeners are able to pick up on small acoustic cues identifying words in minimal pairs, in quiet but also in background noise. For the initial analyses, we used a subset of our data consisting of the participants' target and distractor answers, thus disregarding the wrong responses. We tested the participants' binomial responses (0 = distractor and 1 = target) using a GLMM with a logistic linking function. First, we analyze the subset of the high-predictability items. In this analysis, all confidence ratings are collapsed. The model included fixed effects of Noise (categorical predictor with three levels using dummy coding and mapping the quiet condition to the intercept), Age (continuous predictor and scaled to improve convergence), ContrastVP (categorical predictor with two levels using dummy coding and mapping plosive to the intercept), and Trial Number (continuous predictor with trial number within each block and scaled to improve convergence). Additionally, the model included the interaction of ContrastVP and Age (scaled) and the interaction of Trial Number and Age (both scaled). The model included no random effects, since this led to non-converging models or singular fit. The model revealed a significant effect of Noise, where participants more often give the distractor answer noise compared to the quiet condition ($\beta = -7.12$, $SE = 1.70$, $z = -4.18$, $p < 0.001$ for 0 SNR and $\beta = -6.27$, $SE = 1.78$, $z = -3.55$, $p < 0.001$ for −5SNR). As can be seen in **Table 2**, all other effects were not significant (all *value of p*s > 0.35). The noise effects are relatively small, and overall participants score close to ceiling, where most responses are target responses. These effects can also be seen in the two left-hand panels in **Figure 2**.

Effects of Noise and Phoneme Change on Comprehension

The model for the low-predictability subset of the data included the same fixed effects as the high-predictability subset but included by-Participant and by-Item random intercepts and a random slope for Noise for the by-Item random intercept. The model revealed a significant effect of Noise, where the noise conditions had more distractor responses than Quiet ($\beta = -4.87$, $SE = 0.32$, $z = -15.22$, $p < 0.001$ for 0SNR and $\beta = -6.32$, $SE = 0.39$, $z = -16.09$, $p < 0.001$ for −5SNR). Additionally, the model revealed a significant effect of Trial Number ($\beta = 0.47$, $SE = 0.08$, $z = 5.97$, $p < 0.001$), meaning that participants slightly increased the amount of target responses with practice. The interaction of Age and Trial Number was not significant ($p = 0.95$), suggesting older adults also showed this learning effect. The model also revealed that items of minimal pairs differing in the vowel had more target responses than items of minimal pairs differing in the plosive ($\beta = 1.55$, $SE = 0.34$, $z = 4.63$, $p < 0.001$). This was in line with the expectation that words differing in the plosive contrast would be harder to identify correctly than words differing in the vowel. Finally, the interaction of ContrastVP and Age was significant as well ($\beta = -0.33$, $SE = 0.14$, $z = -2.29$, $p < 0.01$), showing that with increasing age, there was a larger decrease in the proportion of target responses for vowel contrasts than plosive contrasts. These effects in general are presented in **Table 2** and illustrated in the two right-hand panels of **Figure 2**. The interaction effect in particular is shown by the steeper downward slope of the lines in the LP vowel plot compared to the LP plosive plot.

Semantic Fit and Phonetic Distance

We coded the semantic fit and phonetic distance to the target of the wrong responses, to see whether participants relied more on the acoustic signal (low distance) or on the provided context (wrong response fits semantically). This gives more insight in the participants' strategies and allows us to tease apart whether participants relied on top-down (predictions based on context) or bottom-up (acoustic signal) information. **Figure 3** presents the normalized phonetic distance and semantic fit for the wrong responses in each of the three noise conditions. Lower normalized phonetic distance scores mean that the participant's response sounded more similar to the target word. Responses with a distance score of 1 were empty responses. **Figure 3** also shows that a majority of the wrong responses in each of the noise conditions, the participant's response did not fit the sentence semantically (76 vs. 12 for Quiet; 177 vs. 73 for 0 SNR; and 341 vs. 208 for −5 SNR). The peaks of the phonetic distance distributions seem to lie more to the right (meaning larger distance to the target) in the semantically fitting responses, suggesting a trade-off between acoustic fit and semantic fit. Participants made their response based on what they heard at a cost of fitting the semantic context.

Confidence Ratings

We calculated the mean confidence for each of the three response types, namely, targets ($M = 3.494$, $SD = 0.806$), distractors ($M = 2.997$, $SD = 0.994$), and wrong responses ($M = 1.756$, $SD = 0.988$), finding similar confidence for targets and distractors overall, and lower confidence for wrong responses. We transformed the participants' confidence responses to a binary variable of low confidence (confidence ratings 1 and 2) and high confidence (confidence ratings 3 and 4). This binary response variable was tested using a GLMM with logistic linking function. Equivalent results are found with ordinal regression analyses. Because of better interpretability, we present the binomial regression here, while results from the ordinal regression can be found in the **Supplementary Material**. For these analyses, we have taken three subsets of the data: one with the target responses ($N = 4{,}161$), one with the distractor responses ($N = 1{,}438$), and one with the wrong responses ($N = 881$). We expected to find different patterns of confidence ratings for these subsets, because in the wrong responses, participants relied mostly on the sentence context, while in the distractor responses, there was some supporting evidence from the acoustic signal as well. As such, we expected participants to be more certain in general of their distractor items, than of their wrong items, as they realized that the wrong items were not presented to them in the speech signal. These analyses will shed light on how confident participants were in the different response types overall. Subsequently, we will turn to the distractor

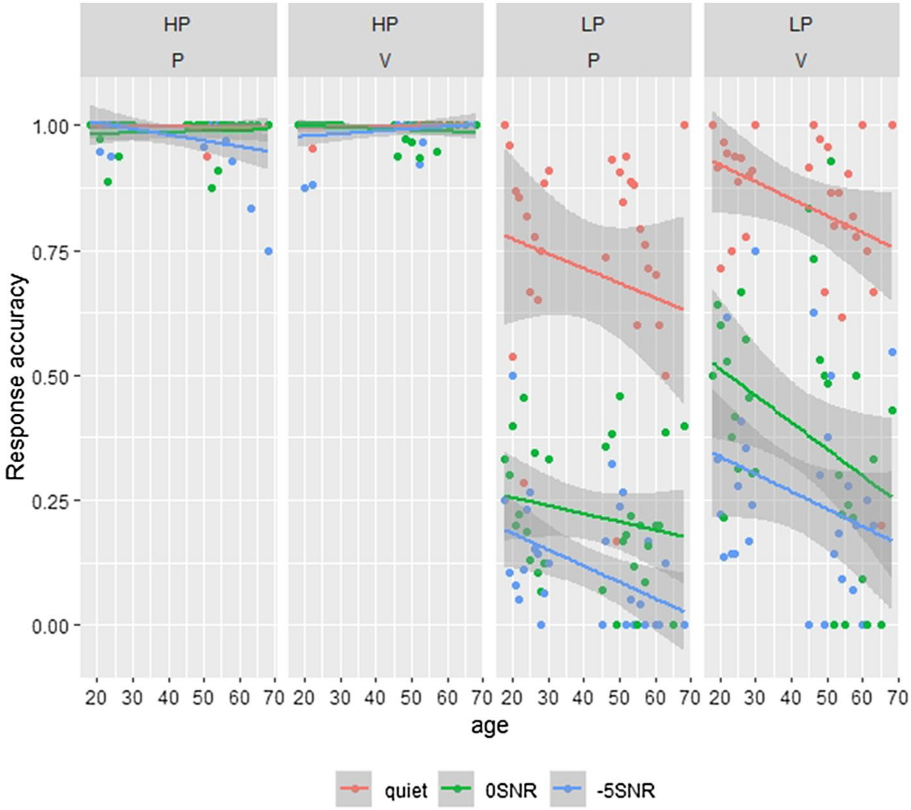


FIGURE 2 | This figure shows the participant's answers; split for target and distractor items, with age plotted on the x-axis and answer type on the y-axis. Here 0 denotes the distractor response and 1 the target response. Different line colors show different noise conditions. The different plots show the high (HP)- and low-predictability (LP) items for stimuli differing in a plosive (P) or vowel (V).

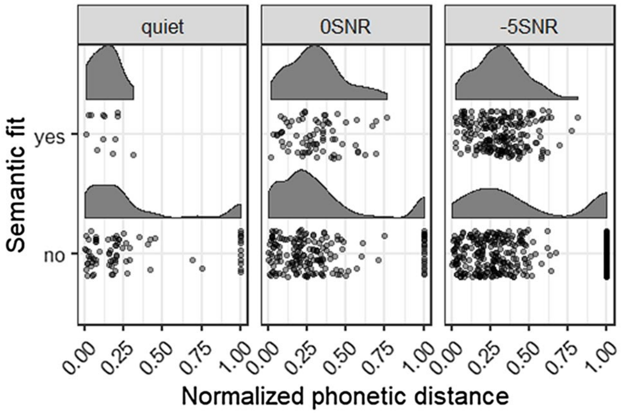


FIGURE 3 | This figure shows the wrong responses that semantically fit or did not fit the sentence, plotted with the normalized phonetic distance, in each of the three noise conditions. Lower phonetic distance means more similar to the target item. A distance of 1 means an empty response.

responses in the three noise conditions, as this was the condition most likely to elicit false hearing. **Figure 4** presents the participants' confidence ratings from uncertain (1) to certain (4), split for each of the predictability conditions, noise levels, and response types.

The model for the subset of target responses included fixed effects of Predictability (categorical predictor with two levels using dummy coding and mapping the high-predictability condition on the intercept), Noise, Age, Trial Number, and ContrastVP, as well as the three-way interaction of Predictability, Noise, and Age. All were coded and scaled as before. A by-Participant random intercept was included with random slopes for Noise and Predictability, and a by-Item random intercept with a random slope for Predictability. There was a significant effect of Predictability, with lower confidence in LP vs. HP ($\beta = -2.17$, $SE = 0.51$, $z = -4.28$, $p < 0.001$). The model revealed lower confidence in Noise compared to Quiet ($\beta = -1.71$, $SE = 0.46$, $z = -3.70$, $p < 0.001$ for 0SNR and $\beta = -4.10$, $SE = 0.47$, $z = -8.78$, $p < 0.001$ for −5SNR). The interaction of Noise and Age was significant, with lower confidence for older participants in noise ($\beta = -1.07$, $SE = 0.36$, $z = -3.02$, $p < 0.01$ for 0SNR and $\beta = -0.85$, $SE = 0.34$, $z = -2.52$, $p < 0.05$ for −5SNR). Finally, the three-way interaction of Predictability, Noise, and Age was significant for the 0SNR condition, with higher confidence ratings with age in LP ($\beta = 0.99$, $SE = 0.41$, $z = 2.42$, $p < 0.05$). The other effects were not significant (all values of $p > 0.08$), and all effects can be found in **Table 3**.

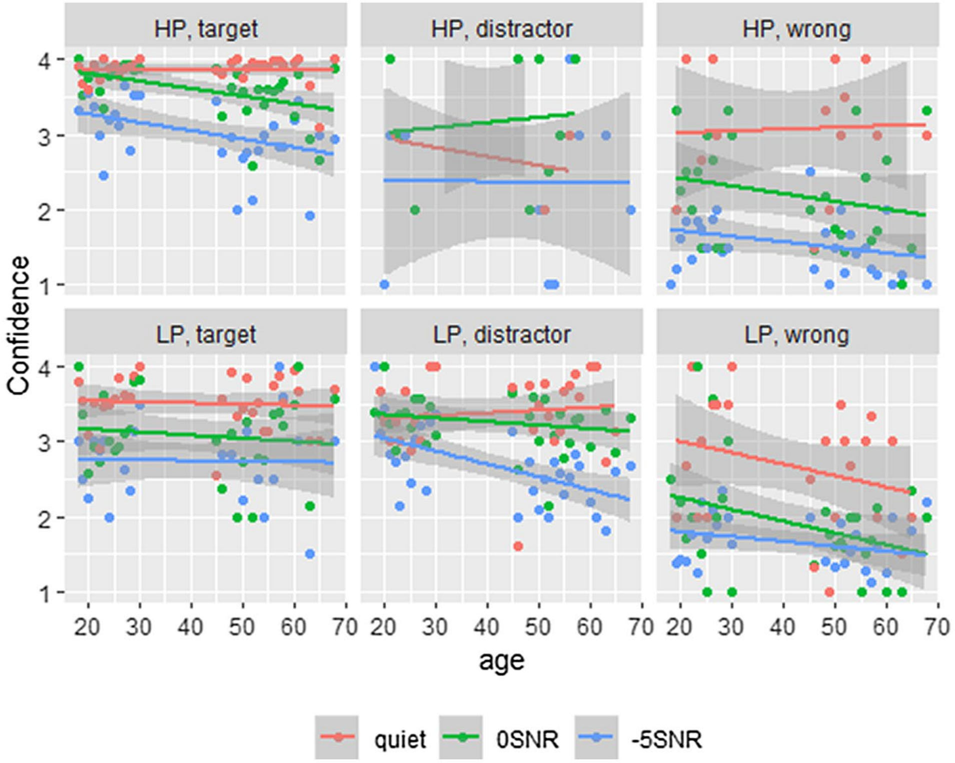


FIGURE 4 | This figure shows the participants' confidence ratings; split for the predictability conditions, with HP at the top row and LP at the bottom, as well as the three answer types. Age plotted on the x-axis and confidence on the y-axis. Here 1 denotes the lowest confidence and 1 the highest confidence. Different line colors show different noise conditions.

TABLE 3 | Model outcomes for the confidence rating analysis (target subset).

	Estimate	*SE*	*z*-value	*value of p*	
Intercept	5.65	0.48	11.69	< 0.001	***
Predictability LP	−2.17	0.51	−4.28	< 0.001	***
Noise −5SNR	−4.10	0.47	−8.78	< 0.001	***
Noise 0SNR	−1.71	0.46	−3.70	< 0.001	***
Age	0.15	0.31	0.48	0.63	
Trial No	−0.02	0.07	−0.34	0.73	
ContrastVP V	0.35	0.20	1.77	0.08	
Predictability LP: Noise −5SNR	0.76	0.55	1.38	0.16	
Predictability LP: Noise 0 SNR	−0.77	0.49	−1.56	0.12	
Predictability LP: Age	0.05	0.36	0.13	0.90	
Noise −5SNR: Age	−0.85	0.34	−2.52	< 0.05	*
Noise 0SNR: Age	−1.07	0.34	−3.02	< 0.01	**
Predictability LP: Noise −5SNR: Age	0.42	0.44	0.97	0.33	
Predictability LP: Noise 0SNR: Age	0.99	0.41	2.42	< 0.05	*

This table shows the analysis for the subset of target items. The response variable is the participants' confidence (high or low).

The model for the subset of distractor responses included the same fixed effects as the model on the subset of target responses. A by-Participant random intercept was included, as well as a by-Item random intercept with a random slope of Predictability. Inclusion of other random slopes led to models with a singular fit. The model revealed a significant effect of vowel/plosive contrast ($\beta = -0.46$, $SE = 0.20$, $z = -2.29$, $p < 0.01$), suggesting that participants were less confident about their answers on items that had a vowel contrast, rather than those with a plosive contrast. Additionally, there was a significant effect of Trial Number, where participants are less confident in later trials ($\beta = -0.19$, $SE = 0.08$, $z = -2.43$, $p < 0.01$). The other effects were not significant (all values of $p > 0.40$). All effects can be seen in **Table 4**.

The model for the subset of wrong answer items included the same fixed effects as the previous two models, except that this model did not include a three-way interaction, but only an interaction of Predictability and Noise. A by-Participant random intercept was included, as well as a by-Item random intercept. Inclusion of random slopes led to models with a singular fit. The model revealed a significant effect for both noise conditions. In 0SNR noise, participants were less confident than in quiet ($\beta = -1.53$, $SE = 0.55$, $z = -2.78$, $p < 0.01$), an effect that was also found for −5SNR noise ($\beta = -3.04$, $SE = 0.56$, $z = -5.46$, $p < 0.001$). These findings show that generally confidence ratings reflect the amount of noise that was presented. None of the other effects were significant (all values of $p > 0.20$), and all effects are presented in **Table 5**.

TABLE 4 | Model outcomes for the confidence rating analysis (distractor subset).

	Estimate	*SE*	*z*-value	*value of p*	
Intercept	1.73	2.77	0.62	0.53	*
Predictability LP	0.31	2.78	0.11	0.91	
Noise −5SNR	−1.51	2.90	−0.52	0.60	
Noise 0SNR	0.33	2.94	0.11	0.90	
Age	−2.16	3.37	−0.64	0.52	
Trial No	−0.19	0.08	−2.43	< 0.05	
ContrastVP V	−0.46	0.20	−2.29	< 0.05	*
Predictability LP: Noise −5SNR	−0.10	2.90	0.04	0.97	
Predictability LP: Noise 0 SNR	−0.32	2.96	−0.11	0.91	
Predictability LP: Age	2.63	3.38	0.78	0.44	
Noise −5SNR: Age	1.02	3.23	0.31	0.75	
Noise 0SNR: Age	2.89	4.29	0.67	0.50	
Predictability LP: Noise −5SNR: Age	−1.87	3.25	−0.58	0.56	
Predictability LP: Noise 0SNR: Age	−3.57	4.30	−0.83	0.41	

This table shows the analysis for the subset of distractor items. The response variable is the participants' confidence (high or low).

Finally, we want to investigate directly the false hearing effect in the noise conditions, thus focusing on the confidence ratings in mishearings. We take subsets of the data of all distractor items produced in 0SNR ($N = 646$), −5SNR ($N = 618$), and quiet ($N = 174$). Based on previous findings, we expect to find a false hearing effect in the noise conditions, where participants show high confidence in their incorrect responses as these distractor responses were supported by the sentence context. We expect to find an effect of age, so that older participants are more confident of their response than younger adults. All outcomes from the three GLMMs are presented in **Table 6**.

The model on the subset of 0SNR trials included fixed effects of Predictability, Age, Trial Number, and ContrastVP (all coded and scaled as before). The model also included random intercepts for Subject and Item (random slopes led to non-convergence or singular fit). The model showed significantly lower confidence as the trials went on ($\beta = -0.35$, $SE = 0.12$, $z = -2.85$, $p < 0.01$). Additionally, confidence ratings were significantly lower for items with a vowel contrast compared to items with a plosive contrast ($\beta = -0.91$, $SE = 0.28$, $z = -3.29$, $p < 0.01$). The other effects were not significant (all values of $p > 0.22$).

The model on the subset of −5SNR trials consisted of the same fixed and random effects as the 0SNR model. We find only a significant effect of Age, where older participants are less confident of their responses than younger adults ($\beta = -0.44$, $SE = 0.15$, $z = -2.99$, $p < 0.01$). This is the opposite of what we would expect for false hearing based on previous findings (Rogers et al., 2012; Failes et al., 2020), where older participants are *more* confident of their responses. None of the other effects were significant (all values of $p > 0.31$).

The model on the quiet subset of the data again included the same fixed and random effects as the previous two models.

TABLE 5 | Model Outcomes for the confidence rating analysis (wrong subset).

	Estimate	*SE*	*z*-value	*value of p*	
Intercept	0.92	0.52	1.77	0.08	***
Predictability LP	−0.08	0.58	−0.13	0.90	
Noise −5SNR	−3.04	0.56	−5.46	< 0.001	
Noise 0SNR	−1.54	0.55	−2.78	< 0.01	**
Age	−0.17	0.13	−1.28	0.20	
Trial No	−0.07	0.10	−0.71	0.48	
ContrastVP V	−0.28	0.25	−1.11	0.27	
Predictability LP: Noise −5SNR	0.29	0.64	0.45	0.65	
Predictability LP: Noise 0 SNR	−0.42	0.67	−0.62	0.53	

This table shows the analysis for the subset of wrong items. The response variable is the participants' confidence (high or low).

None of the effects were significant (all values of $p > 0.15$). These models together show no evidence for false hearing in our data, although mishearings were frequent.

DISCUSSION

In the present study, we investigated word recognition in background noise in younger and older adults, analyzing to what extent listeners rely on the acoustic speech signal or on top-down predictions made based on the sentence context. In our experiment, participants typed in the last word of the sentence that was played in quiet or embedded in background noise at 0SNR and −5SNR. Additionally, participants rated their confidence in giving the correct answer. The results showed that in quiet listening conditions, listeners of all ages and in both high- and low-predictive contexts, mainly make use of the information in the acoustic speech signal. However, they turn more to the sentence context than the acoustic signal as a guide when there is some level of background noise. This effect is stronger for older adults than for younger adults, and it is more pronounced in higher levels of background noise, in line with our hypotheses. Generally, we find that words with a vowel contrast are easier to recognize than words with a plosive contrast, a benefit that lessens with age, presumably due to floor effects. With regard to the confidence ratings, we generally find lower confidence ratings that reflect more difficult listening conditions and incorrect answers. Words with vowels get lower confidence ratings when the response is incorrect compared to items with a plosive contrast. In none of the conditions in our experiment do, we find a false hearing effect where participants rate their incorrect responses with higher confidence, even though mishearings were very common.

Sound Contrast

We carefully controlled the phonetic contrasts of our minimal pairs to investigate how the sound difference of the minimal pair might have an effect on recognition scores. Our pairs differed either in a plosive (place of articulation) or in a vowel (tense/lax). We expected that the items differing in

TABLE 6 | Model outcomes for the false hearing analysis.

	Quiet subset				0SNR subset				−5SNR subset			
	Estimate	***SE***	***z*-value**	***value of p***	**Estimate**	***SE***	***z*-value**	***value of p***	**Estimate**	***SE***	***z*-value**	***value of p***
Intercept	0.95	1.77	0.54	0.59	1.89	0.93	2.03	< 0.05 *	0.29	0.84	0.35	0.73
Predictability LP	1.47	1.79	0.82	0.41	0.36	0.92	0.40	0.69	0.22	0.84	0.26	0.79
Age	0.50	0.35	1.43	0.15	−0.17	0.15	−1.22	0.22	−0.44	0.14	−2.99	< 0.01 **
Trial No	−0.36	0.29	−1.23	0.22	−0.36	0.13	−2.85	< 0.01 **	< 0.001	0.11	0.01	0.99
ContrastVP V	0.34	0.58	0.59	0.55	−0.91	0.28	−3.29	< 0.01 **	−0.14	0.27	−0.53	0.59

This table shows the analysis for the subset of distractor items in quiet, 0SNR, and −5SNR. The response variable is the participants' confidence (high or low).

the plosive were more difficult to recognize correctly than the items differing in a vowel. Plosives consist of a relatively short sound, especially compared to vowels that have a longer duration and greater amplitude. Thus, plosives are more likely to get lost in the noise, in which case the listener would make use of the provided sentence context and report having heard the distractor item. This expectation was confirmed by our data. Other studies that looked at a wider range of plosives and vowels also found that, especially in more difficult listening situations, vowels led to easier recognition than plosives (Fu et al., 1998; Cutler et al., 2004).

Our results showed an interaction with age: The facilitative effect of a vowel contrast over a plosive contrast decreased as participants were older. The direction of this interaction is unexpected at first glance, as we had hypothesized that older adults would have increased difficulty identifying plosives, as for these sounds the higher frequencies are more informative than for vowels (Edwards, 1981; Alwan et al., 2011). These high frequencies are lost first in age-related hearing loss (Gates and Mills, 2005). We believe however that the observed interaction is the result of a floor effect: Older adults have a lot of trouble understanding the plosive correctly in noisy conditions, and almost always mistake the distractor for the target item in this condition. As there is already a substantial number of distractor responses for plosives even in the quiet condition, the decline in noise cannot be as steep as the one observed for vowels, for which comprehension is a lot better in quiet. Another possible explanation for the interaction effect might be that the older adults might have had age-induced hearing loss, in which they struggle, among other things, to discriminate spectral transitions in noise (Tun et al., 2012). This difference in mishearing between plosives vs. vowels suggests that even minor changes in how well the acoustic signal can be perceived affects the probability distribution of the bottom-up information and can lead to a more dominant top-down probability, as predicted by the noisy channel model. If, as suggested by Rogers et al. (2012), mishearing is caused by general deficits in cognitive control, we would expect to find no differences between the two sound types.

When looking at the confidence ratings, we find an effect of ContrastVP in the subset of distractor responses. This suggests that participants were less confident of their response if the target word was part of a minimal pair containing a vowel contrast, than when the word came from a pair with a plosive contrast. Most distractor responses were made in the low-predictability condition, where the sentence context supported the distractor word, while the acoustic information did not. We also found that in the low-predictability condition, words from a pair differing in the vowel generally were easier to identify correctly (participants responding with the target word more often than the distractor). When participants responded incorrectly (with the distractor rather than the target), they were less confident of this, suggesting that they were more aware that they misheard the word than they were for plosive contrasts.

We did not choose our sound contrasts with any models of speech perception in mind. In hindsight, our contrasts might not all be processed in the same way. For example, studies suggest that the coronal place of articulation for consonants is not specified and that it can vary freely for coronal consonants (Friedrich et al., 2006; Lahiri and Reetz, 2010; Roberts et al., 2013). We used the coronal sounds /t/ and /d/ in our consonant minimal pairs, contrasted with other plosives differing in place of articulation. Testing whether these sounds led to more distractor responses due to unspecified coronal place of articulation is outside the scope of this article, but would be an interesting question for future research.

Bottom-Up and Top-Down Processes

This study investigated how bottom-up auditory processes and top-down predictive processes interact in speech comprehension, in particular in noisy conditions and while looking at differences between younger and older adults. In the high-predictability condition of our experiment, we found an effect of noise, so that there were more distractor responses in the conditions with background noise compared to quiet. This effect was small, and most responses were in fact correct, suggesting a ceiling effect, in particular in quiet. In our paradigm, we presented the sentence context on the screen in written form, which will have led to these ceiling effects. Both the information provided by the speech signal and the information provided by the sentence context pointed to the target word. Participants could thus use information from both sources to recognize the correct word, there was no conflict between them. Especially in the quiet condition, there was no expectation that participants would identify the word incorrectly. The fact that we found this ceiling effect shows that our participants were paying attention to the task. The lack of an age effect in the high-predictability condition regarding the number of distractor responses even in noise shows that older adults can make up for difficult listening conditions by making use of the predictability of the message (Wingfield et al., 2005). As this is arguably the most frequent situation in normal language comprehension – i.e., words fit the context – this is a helpful strategy in everyday listening.

We found different results in the low-predictability condition, where the participants' answers depended greatly on the condition the items were presented in. In the low-predictability condition, the information provided by the acoustic signal is contradicted by the information given by the sentence context, as both point to different lexical items. On the one hand, the word supported by the context is also partially supported by the speech signal. Because we used minimal pairs, these two words only differed in one single phonetic feature. On the other hand, the word supported by the information from the speech signal is not supported by the sentence context at all. In the quiet condition, participants identified the sentence-final word for the most part correctly. In conditions with background noise, however, participants do rely more on the sentence context to guide word recognition, as shown by the shift to a large proportion of distractor answers. The increased rates of mishearing in noise are observed for both younger and older adults, but the effect is substantially stronger for older adults. This is in line with previous work that has shown that older adults tend to rely more heavily on the sentence context (Hutchinson, 1989; Pichora-Fuller et al., 1995; Sommers and Danielson, 1999; Dubno et al., 2000). Due to the presence of noise, it is more difficult to identify all the sounds in the speech signal, and here listeners turn to the other source of information they have available. This was an expected finding, as in previous studies, also younger adults do rely more on context when listening conditions get harder (Hutchinson, 1989; Dubno et al., 2000; Pichora-Fuller, 2008). We also observed a significant learning effect in our data: As the trials in a block proceed, participants are slightly more likely to get the target item correct. This holds for participants irrespective of age. One possible explanation for this is that they became aware of the manipulation and the fact that the context could be misleading, thus paying more attention to the sound signal than they did before. Listeners have been found to be able to re-weight cues based on their statistical properties (Bushong and Jaeger, 2019). It also shows that older adults are able to adapt to the task, unlike in Rogers et al. (2012). In the present study, they learned over the course of the experiment that context might be misleading and weighing the acoustic information more than the top-down predictions. Adaption suggests that older participants are behaving rationally when showing false hearing.

Analyses of semantic fit and phonetic distance to the target word show that the majority of the wrong responses did not fit the sentence semantically, while distances were smaller in the semantically incongruent responses. This suggests that participants did try to rely on the acoustic signal rather than the provided context, somewhat against our expectations. It might be the case that they had noticed the sometimes misleading sentence context and relied less on this information. Even though we already find high rates of mishearing in our study, it is likely that this underestimates the amount of mishearing that would occur for these materials in a more naturalistic setting. Participants were aware of the possible semantic mismatches in the presented audio and sentence context, and our analyses show that participants in fact paid considerable attention to the acoustic signal rather than the sentence context.

According to the noisy channel model (Levy, 2008), information from both sources are combined rationally. However, older adults have been found to rely more on top-down predictive processes than younger adults, which can lead to mishearing in cases when the target is not predicted by the context. A study by Gibson et al. (2013) showed that human language processing relies on rational statistical inference in a noisy channel. Their model predicts that semantic cues should point the interpretation in the direction of plausible meanings even when the observed utterance differs from this meaning, that these non-literal interpretations increase in noisier communicative situation, and decrease when the semantically anomalous meanings are more likely to be communicated. The findings from the present study are in line with the predictions based on the model by Gibson et al.: In more adverse listening conditions, i.e., the conditions with more background noise, listeners rely more on the sentence context to compensate for the difficulties introduced in auditory processing. In these cases, listeners respond that they heard a word that fits the sentence context (plausible meaning), rather than the word that was actually presented to them (implausible meaning). There is contextual information, as well as some sensory information (the shared sounds of the presented word, as these words form a minimal pair) to support the word favored by the sentence context. However, following Gibson et al.'s final prediction, over the course of the experiment participants noticed that the sentence context is not always reliable and showed a learning effect. They came to expect low-predictability sentence-final items, which led to less mishearing.

Rationally combining bottom-up and top-down information in speech comprehension is sensible, in particular in cases of a noisy channel, where the bottom-up signal is partially obscured. However, when the top-down predictions form a mismatch with the information being transferred in the signal, a too strong reliance on top-down processes can lead to problems in communication, in the form of mishearing. These are a side effect of rationally combining bottom-up and top-down information.

False Hearing

We also tested the replicability of the false hearing effect in German that was reported for English in previous literature (Rogers et al., 2012; Sommers et al., 2015; Failes et al., 2020). This effect generally has been found to be stronger for older adults than younger adults. Unlike previous studies and against our expectations, we do not find an age effect for false hearing in our study, i.e., while there was a substantial amount of mishearing, older participants were not more confident about their responses than younger participants. We also do not find an effect of age on confidence in distractor responses overall. While Rogers et al. (2012) do report a smaller false hearing effect in the condition with loud noise compared to the condition with moderate noise, they do still find a false hearing effect. In the present study, we do not find a significant effect of age at all for the 0 SNR subset, while in −5 SNR the effect is opposite to our expectations: With age, participants become less confident. One possible explanation for this failure to replicate the false hearing effect in noise is the age of the participants: The participants in previous studies were generally older than those in the present study, and thus perhaps more likely to show the false hearing effect due to age-related cognitive declines on top of the effects of mishearing predicted by the noisy channel model. Instead of false hearing, we find that our participants' confidence ratings reflect the difficulty of the listening condition: They tended to be lower in noisy conditions and in low-predictability sentences.

Limitations

One of the limitations of this study is that, due to collecting the data via the web, we were not able to collect hearing thresholds of our participants nor were we able to carefully control the sound levels at which the stimuli were presented. We excluded older participants with a large number of incorrect responses in quiet, so that we make sure that the performance in that condition was equated to younger adults. In hindsight, there is another option for controlling hearing levels among our participants. We could have used an alternative control condition where no context cues are available. These stimuli could have been filler sentences in which participants could only rely on the speech signal to make their response. In this way, auditory performance could be equated among our groups of younger and older adults. Peelle et al. (2016) showed that for intelligibility ratings, online testing is a feasible method to replace laboratory testing as it gave comparable results as testing in the laboratory. This suggests that careful control of participants' listening conditions and software used like in laboratory settings is not necessary to obtain reliable results. Additionally, previous studies have equated overall audibility for older and younger adults using individual speech recognition thresholds, and still found larger false hearing effects for older adults, suggesting it is not directly caused by differences in hearing acuity (Rogers et al., 2012; Sommers et al., 2015; Failes et al., 2020).

We constructed the items in our low-predictability condition by swapping the two words from the minimal pairs we had selected. It should be noted that this lead to sentences that, while unpredictable, also were implausible. In fact, in the low-predictability condition, the sentences provided a context that was strongly biased for the distractor word. This could have led to larger amounts of mishearing compared to when we would have used sentences that were unpredictable but plausible, in particular for older adults who tend to rely more on context. Due to the strong bias for the distractor and the implausibility of the target word, relying on the context would strongly favor the distractor response. Other studies investigating false hearing using sentences varied in whether their low-predictability items were plausible or not. Sommers et al. (2015) used unpredictable sentences that were still meaningful (LP: *The shepherd watched his sheath*), but Failes et al. (2020) had implausible items. They constructed their unpredictable items by changing one phoneme in the sentence-final target word in the predictable item (HP: *She put the toys in the box*; LP: *She put the toys in the fox*). Both these studies found a larger false hearing effect for older adults, and therefore, this effect seems to be independent of the plausibility of the low-predictability items. It therefore seems unlikely that our lack of an effect can be explained by having used implausible sentences. The false hearing effect has also been found using a word priming paradigm (Rogers et al., 2012; Rogers, 2017), which suggests that the effect does not depend on the use of a particular paradigm.

Another limitation of the present study is the age of our older adults, which is relatively young. Our oldest participant was 68 years old, and mean age of the older group was 53. Compare this to the ages of the older participants in Failes et al. (2020), which ranged from 65 to 81, with a mean of 71. This might explain the lack of an age-related false hearing effect in the present study. For our sample, we find that rational processes better explain our results of differences between vowel contrasts and plosive contrasts, but of course it could be the case that in an older sample, general cognitive decline plays a part as well (Rogers et al., 2012).

The present results are based on a restricted set of minimal pairs, namely, pairs of plosives only differing in place of articulation, and tense vs. lax vowels, and were tested in multi-speaker babble noise. More research is needed to investigate how these findings generalize to other sound combinations and other types of noise. Future studies could also test at different SNR levels, to prevent in particular the floor effects we found in the plosives as noise, as this can shed light on the true nature of the interaction effect of age and sound contrasts in noise. Currently, the noisy channel model does not incorporate metacognitive measures like confidence ratings. Confidence could be formulated in terms of the probability distribution between different lexical candidates. If, on the one hand, the probability of one candidate is a lot higher than that of another candidate, high confidence in the response should be reported. On the other hand, if the probabilities of different candidates are more similar, the confidence rating should be lower. The exact modeling of false hearing based on confidence ratings in the noisy channel model can be explored in future research.

CONCLUSION

Previous studies have investigated the mishearing effect, where listeners understand a word different from the one that was spoken. These effects are particularly prevalent in situations where the speech signal is noisy, and the word that is actually understood fits well with the semantic context, indicating that top-down predictability of the word may have overpowered the bottom-up auditory signal. Previously, this effect has been attributed to general deficits in cognitive control, in particular inhibition (Rogers et al., 2012; Sommers et al., 2015; Failes et al., 2020).

In the present article, we argue that the effect is a natural consequence of rational language processing in noise and thus does not require to be attributed to deficits in cognitive control. To test this idea, we designed a study which carefully controls the way in which the target and the distractor words differ from one another. Specifically, we constructed target-distractor pairs which only differed in the articulatory position in a plosive, and another set of target-distractor pairs that differed only in vowel quality. We conducted an online study in German, in which participants listened to sentences in quiet and two levels of background babble noise, and reported the sentence-final word they heard, as well as rated their confidence in this response. Our findings show that participants accurately report the actually spoken word in quiet listening conditions, but that they rely more on sentence context in the presence of background noise (both babble and white noise), leading to incorrect responses in particular in the low-predictability condition. While listeners thus do profit from high-predictability in noise (as they do correctly understand the words in this condition), they also suffered the downside of mishearing in the low-predictability condition. The mishearing effect was found to be larger in older adults compared to younger adults, replicating previous findings. We explain this within the noisy channel account in terms of increased language experience of older adults, possibly compounded with first experiences of hearing loss.

For our critical phonetic manipulation, we found that stimuli pairs with a vowel contrast were generally easier to identify correctly than pairs with a plosive contrast, although this benefit lessened with age. These different effects for vowels vs. plosives suggest that mishearing depends on the quality of the acoustic signal, rather than general deficits in cognitive control or inhibition. We also find a learning effect that suggests that participants of all ages were able to adapt to the task. We think that this finding also underscores the rational account and is not consistent with an account that relates age differences to a difference in cognitive control. Our findings also add to the literature by replicating the earlier mishearing effects in a different language, German.

Earlier work had however also reported an effect of false hearing, meaning that participants are very confident of their answer even though it is in fact incorrect (Rogers et al., 2012; Sommers et al., 2015; Failes et al., 2020). In particular, the false hearing effect was found to be increased in older adults. While our experiment was also set up to assess false hearing, we did not find any significant effects of false hearing in the older participants compared to the younger ones. Instead, confidence was related to the level of noise.

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 Deutsche Gesellschaft für Sprachwissenschaft (Ethics Committee of the German Linguistic Society). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MV, VD and JK were involved in planning and designing of the study. MV analyzed the data and wrote all parts of the manuscript. VD and JK made suggestions to the framing, structuring, and presentation of findings as well as on the interpretation of findings. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

ACKNOWLEDGMENTS

We would like to thank Sarah Mameche, Katharina Stein, Valentin Kany, Amélie, Leibrock, and Lisa-Marie Riedel for their help in constructing the stimuli and assistance with the phonetic transcriptions and semantic fit coding.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.679278/full#supplementary-material

REFERENCES

Alwan, A., Jiang, J., and Chen, W. (2011). Perception of place of articulation for plosives and fricatives in noise. *Speech Comm.* 53, 195–209. doi: 10.1016/j.specom.2010.09.001

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Baayen, R.H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX Lexical Database (Release 2) [CD-ROM]*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Balcetis, E., and Dunning, D. (2010). Wishful seeing: more desired objects are seen as closer. *Psychol. Sci.* 21, 147–152. doi: 10.1177/0956797609356283

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, 255–278. doi: 10.1016/j.jml.2012.11.001

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

Benichov, J., Cox, L. C., Tun, P. A., and Wingfield, A. (2012). Word recognition within a linguistic context: effects of age, hearing acuity, verbal ability and cognitive function. *Ear Hear.* 32, 250–256. doi: 10.1097/AUD.0b013e31822f680f

Boersma, P., and Weenink, D. (2020). Praat: doing phonetics by computer (Computer program) Version 6.1.05. Available at: http://www.praat.org/ (Accessed December 16, 2020).

Boothroyd, A., and Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *J. Acoust. Soc. Am.* 84, 101–114. doi: 10.1121/1.396976

Bruner, J. S. (1957). On perceptual readiness. *Psychol. Rev.* 64, 123–152. doi: 10.1037/h0043805

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Am.* 109, 1101–1109. doi: 10.1121/1.1345696

Bushong, W., and Jaeger, T. F. (2019). Dynamic re-weighting of acoustic and contextual cues in spoken word recognition. *J. Acoust. Soc. Am.* 146, EL135–EL140. doi: 10.1121/1.5119271

Chisolm, T. H., Willott, J. F., and Lister, J. J. (2003). The aging auditory system: anatomic and physiologic changes and implications for rehabilitation. *Int. J. Audiol.* 42(Suppl. 2), 3–10. doi:10.3109/14992020309074637

Cooke, M. (2009). "Discovering consistent word confusions in noise," in *Tenth Annual Conference of the International Speech Communication Association*; September 6–10, 2009.

Cutler, A., Weber, A., Smits, R., and Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *J. Acoust. Soc. Am.* 116, 3668–3678. doi: 10.1121/1.1810292

Dubno, J. R., Ahlstrom, J. B., and Horwitz, A. R. (2000). Use of context by young and aged adults with normal hearing. *J. Acoust. Soc. Am.* 107, 538–546. doi: 10.1121/1.428322

Edwards, T. J. (1981). Multiple features analysis of intervocalic English plosives. *J. Acoust. Soc. Am.* 69, 535–547. doi: 10.1121/1.385482

Failes, E., Sommers, M. S., and Jacoby, L. L. (2020). Blurring past and present: using false memory to better understand false hearing in young and older adults. *Mem. Cogn.* 48, 1403–1416. doi: 10.3758/s13421-020-01068-8

Flanagan, J. L. (1955). A difference limen for vowel formant frequency. *J. Acoust. Soc. Am.* 27, 613–617. doi: 10.1121/1.1907979

Friedrich, C. K., Eulitz, C., and Lahiri, A. (2006). Not every pseudoword disrupts word recognition: an ERP study. *Behav. Brain Funct.* 2, 1–10. doi: 10.1186/1744-9081-2-36

Fu, Q. J., Shannon, R. V., and Wang, X. (1998). Effects of noise and spectral resolution on vowel and consonant recognition: acoustic and electric hearing. *J. Acoust. Soc. Am.* 104, 3586–3596. doi: 10.1121/1.423941

Gates, G. A., and Mills, J. H. (2005). Presbycusis. *Lancet* 366, 1111–1120. doi: 10.1016/S0140-6736(05)67423-5

Gibson, E., Bergen, L., and Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proc. Natl. Acad. Sci.* 110, 8051–8056. doi: 10.1073/pnas.1216438110

Gordon-Salant, S., Frisina, R. D., Fay, R. R., and Popper, A. (2010). *The Aging Auditory System. Vol. 34*. New York: Springer.

Hay, J. F., and Jacoby, L. L. (1999). Separating habit and recollection in young and older adults: effects of elaborative processing and distinctiveness. *Psychol. Aging* 14, 122–134. doi: 10.1037/0882-7974.14.1.122

Helfer, K., Bartlett, E., Popper, A., and Fay, R. R. (eds.). (2020). *Aging and Hearing*. Cham: Springer.

Hutchinson, K. M. (1989). Influence of sentence context on speech perception in young and older adults. *J. Gerontol.* 44, P36–P44. doi: 10.1093/geronj/44.2.P36

Jacoby, L. L., Rogers, C. S., Bishara, A. J., and Shimizu, Y. (2012). Mistaking the recent past for the present: false seeing by older adults. *Psychol. Aging* 27, 22–32. doi: 10.1037/a0025924

Koeritzer, M. A., Rogers, C. S., Van Engen, K. J., and Peelle, J. E. (2018). The impact of age, background noise, semantic ambiguity, and hearing loss on recognition memory for spoken sentences. *J. Speech Lang. Hear. Res.* 61, 740–751. doi: 10.1044/2017_JSLHR-H-17-0077

Krech, E., Stock, E., Hirschfeld, U., and Anders, L. (2009). *Deutsches Aussprachewörterbuch*. Berlin, Boston: De Gruyter Mouton.

Lahiri, A., and Reetz, H. (2010). Distinctive features: phonological underspecification in representation and processing. *J. Phon.* 38, 44–59. doi: 10.1016/j.wocn.2010.01.002

Lash, A., Rogers, C. S., Zoller, A., and Wingfield, A. (2013). Expectation and entropy in spoken word recognition: effects of age and hearing acuity. *Exp. Aging Res.* 39, 235–253. doi: 10.1080/0361073X.2013.779175

Levy, R. (2008). "A noisy-channel model of human sentence comprehension under uncertain input," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*; October 25–27, 2008; 234–243.

Levy, R., Bicknell, K., Slattery, T., and Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proc. Natl. Acad. Sci.* 106, 21086–21090. doi: 10.1073/pnas.0907664106

Li, L., Daneman, M., Qi, J. G., and Schneider, B. A. (2004). Does the information content of an irrelevant source differentially affect spoken word recognition in younger and older adults? *J. Exp. Psychol. Hum. Percept. Perform.* 30, 1077–1091. doi: 10.1037/0096-1523.30.6.1077

Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychol. Monogr. Gen. Appl.* 68, 1–13. doi: 10.1037/h0093673

Lindenberger, U., and Ghisletta, P. (2009). Cognitive and sensory declines in old age: gauging the evidence for a common cause. *Psychol. Aging* 24, 1–16. doi: 10.1037/a0014986

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., and Levin, L. (2016). "Panphon: A resource for mapping IPA segments to articulatory feature vectors" in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*; December 11–16, 2016. 3475–3484.

Nittrouer, S., and Boothroyd, A. (1990). Context effects in phoneme and word recognition by young children and older adults. *J. Acoust. Soc. Am.* 87, 2705–2715. doi: 10.1121/1.399061

Peelle, J. E., Zhang, T., Patel, N., Rogers, C. S., and Van Engen, K. J. (2016). Online testing for assessing speech intelligibility. *J. Acoust. Soc. Am.* 140:3214. doi: 10.1121/1.4970121

Pichora-Fuller, M. K. (2008). Use of supportive context by younger and older adult listeners: balancing bottom-up and top-down information processing. *Int. J. Audiol.* 47(Suppl. 2), S72–S82. doi:10.1080/14992020802307404

Pichora-Fuller, M. K., Alain, C., and Schneider, B. (2017). "Older adults at the cocktail party," in *The Auditory System At the Cocktail Party*. eds. J. Middlebrooks, J. Simon, A. N. Popper and R. R. Fay (Berlin: Springer), 227–259.

Pichora-Fuller, M. K., Schneider, B. A., and Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *J. Acoust. Soc. Am.* 97, 593–608. doi: 10.1121/1.412282

Pusse, F., Sayeed, A., and Demberg, V. (2016). "LingoTurk: managing crowdsourced tasks for psycholinguistics," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*; June 12–17, 2016. 57–61.

Quené, H., and Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *J. Mem. Lang.* 59, 413–425. doi: 10.1016/j.jml.2008.02.002

R Development Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roberts, A. C., Wetterlin, A., and Lahiri, A. (2013). Aligning mispronounced words to meaning: evidence from ERP and reaction time studies. *The Mental Lexicon* 8, 140–163. doi: 10.1075/ml.8.2.02rob

Roediger, H. L., and McDermott, K. B. (1995). Creating false memories: remembering words not presented in lists. *J. Exp. Psychol. Learn. Mem. Cogn.* 21, 803–814.

Rogers, C. S. (2017). Semantic priming, not repetition priming, is to blame for false hearing. *Psychon. Bull. Rev.* 24, 1194–1204. doi: 10.3758/s13423-016-1185-4

Rogers, C. S., Jacoby, L. L., and Sommers, M. S. (2012). Frequent false hearing by older adults: the role of age differences in metacognition. *Psychol. Aging* 27, 33–45. doi: 10.1037/a0026231

Rogers, C. S., and Wingfield, A. (2015). Stimulus-independent semantic bias misdirects word recognition in older adults. *J. Acoust. Soc. Am.* 138, EL26–EL30. doi: 10.1121/1.4922363

Salthouse, T. A. (1990). Working memory as a processing resource in cognitive aging. *Dev. Rev.* 10, 101–124. doi: 10.1016/0273-2297(90)90006-P

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychol. Rev.* 103, 403–428. doi: 10.1037/0033-295X.103.3.403

Schneider, B. A., Daneman, M., and Murphy, D. R. (2005). Speech comprehension difficulties in older adults: cognitive slowing or age-related changes in hearing? *Psychol. Aging* 20, 261–271. doi: 10.1037/0882-7974.20.2.261

Schuknecht, H. F., and Gacek, M. R. (1993). Cochlear pathology in presbycusis. *Ann. Otol. Rhinol. Laryngol.* 102(Suppl. 1), 1–16. doi:10.1177/0003489493102 0S101

Shannon, C. E. (1949). Communication in the presence of noise. *Proc. IRE* 37, 10–21. doi: 10.1109/JRPROC.1949.232969

Sheldon, S., Pichora-Fuller, M. K., and Schneider, B. A. (2008). Priming and sentence context support listening to noise-vocoded speech by younger and older adults. *J. Acoust. Soc. Am.* 123, 489–499. doi: 10.1121/1.2783762

Sommers, M. S., and Danielson, S. M. (1999). Inhibitory processes and spoken word recognition in young and older adults: the interaction of lexical competition and semantic context. *Psychol. Aging* 14, 458–472. doi: 10.1037/0882-7974.14.3.458

Sommers, M. S., Morton, J., and Rogers, C. (2015). "You are not listening to what I said: false hearing in young and older adults," in *Remembering: Attributions, Processes, and Control in Human Memory (Essays in Honor of Larry Jacoby)*. eds. D. S. Lindsay, C. M. Kelley, A. P. Yonelinas and H. L. Roediger III (New York, NY: Psychology Press), 269–284.

Steen-Baker, A. A., Ng, S., Payne, B. R., Anderson, C. J., Federmeier, K. D., and Stine-Morrow, E. A. L. (2017). The effects of context on processing words during sentence reading among adults varying in age and literacy skill. *Psychol. Aging* 32, 460–472. doi: 10.1037/pag0000184

Stine, E. A., and Wingfield, A. (1994). Older adults can inhibit high-probability competitors in speech recognition. *Aging and Cognition* 1, 152–157. doi: 10.1080/09289919408251456

Tucker-Drob, E. M., Brandmaier, A. M., and Lindenberger, U. (2019). Coupled cognitive changes in adulthood: a meta-analysis. *Psychol. Bull.* 145, 273–301. doi: 10.1037/bul0000179

Tun, P. A., Williams, V. A., Small, B. J., and Hafter, E. R. (2012). The effects of aging on auditory processing and cognition. *Am. J. Audiol.* 21, 344–350. doi: 10.1044/1059-0889(2012/12-0030)

Vaissière, J. (1983). "Language-independent prosodic features," in *Prosody: Models and Measurements*. eds. A. Cutler and R. Ladd (Berlin: Springer), 53–66.

Wingfield, A., Tun, P. A., and McCoy, S. L. (2005). Hearing loss in older adulthood: what it is and how it interacts with cognitive performance. *Curr. Dir. Psychol. Sci.* 14, 144–148. doi: 10.1111/j.0963-7214.2005.00356.x

Wingfield, A., Tun, P. A., and Rosen, M. J. (1995). Age differences in veridical and reconstructive recall of syntactically and randomly segmented speech. *J. Gerontol. Ser. B Psychol. Sci. Soc. Sci.* 50B, P257–P266. doi: 10.1093/geronb/50B.5.P257

Winter, B. (2019). *Statistics for Linguists: An Introduction Using R*. New York, Routledge.

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.

HYPOTHESIS AND THEORY
published: 08 September 2021
doi: 10.3389/fpsyg.2021.678020

A Theoretical Framework for a Hybrid View of the N400

Ralf Naumann and Wiebke Petersen*

Department of Computational Linguistics, Institute for Language and Information Science, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany

In this study, we present a novel theoretical account of the N400 event-related potential (ERP) component. Hybrid views interpret this ERP component in terms of two cognitive operations: (i) access of information, which is related to predictions (predictability component), and (ii) integration of information, which is related to plausibility (plausibility component). Though there is an empirical evidence for this view, what has been left open so far is how these two operations can be defined. In our approach, both components are related to categorization. The critical word and the argument position it is related to are associated with categories that have a graded structure. This graded structure is defined in terms of weights both on attributes and values of features belonging to a category. The weights, in turn, are defined using probability distributions. The predictability component is defined in terms of the information gain with respect to non mismatched features between the two categories. The plausibility component is defined as the difference in the degree of typicality between the two categories. Finally, the N400 amplitude is defined as a function of both components.

Keywords: N400, hybrid view, categorization, entropy, predictions, frame theory, probability

OPEN ACCESS

Edited by:
Gina Kuperberg,
Tufts University, United States

Reviewed by:
Mante Sjouke Nieuwland,
Max Planck Institute for Psycholinguistics, Netherlands
Megan Ann Boudewyn,
University of California, Santa Cruz, United States

****Correspondence:***
Ralf Naumann
naumann@hhu.de

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

Received: *08 March 2021*
Accepted: *13 July 2021*
Published: *08 September 2021*

Citation:
Naumann R and Petersen W (2021) A Theoretical Framework for a Hybrid View of the N400. Front. Psychol. 12:678020. doi: 10.3389/fpsyg.2021.678020

1. THE N400: FUNCTIONAL CHARACTERIZATIONS AND EMPIRICAL MEASURES

The N400 is a centroparietally negative-going waveform that is largest between 300 and 400 ms after the onset of an incoming word. It was first investigated by Kutas and Hillyard (1980). They found that relative to a coherent control word (e.g., "butter") a semantic anomalous word (e.g., "socks") in the final position of the sentence elicited an N400 effect: "He spread the warm bread with butter / socks." In Kutas and Hillyard (1984), it was observed that the N400 effect does not depend on a semantic violation (see also Hagoort and Brown, 1994). For example, in "Don't touch the wet dog," the critical word (CW) "dog" elicited a larger N400 amplitude than the CW "paint" in the corresponding sentence "Don't touch the wet paint" though both words satisfy the semantic restrictions imposed by the verb "touch" and the adjective "wet." Later on, it was investigated how the N400 depends on the wider discourse context. For example, van Berkum et al. (1999) used target sentences like "Jane told the brother that he was exceptionally slow / quick." If these sentences were embedded in the wider (discourse) context "As agreed upon, Jane was to wake her sister and her brother at five o'clock in the morning. But the sister had already washed herself, and the brother had even got dressed," the discourse-coherent word "quick" elicited a smaller N400 amplitude than the discourse-anomalous word "slow" in the target sentence. Without this preceding context, this N400 effect was not observed.[1] Nieuwland and van Berkum (2006) showed that discourse context can overrule lexical properties assigned by a verb to its arguments. The influence of world knowledge in relation to word meanings on the N400 was investigated, e.g., in Hagoort et al. (2004).

Basically, there are two main strands in the debate on the interpretation of the N400 component.[2] The first one centers on the functional interpretation of this component: does N400 activity correlate with accessing information from semantic memory (access view) or does it correlate with integrating (the representation of) the CW into (the representation of) the preceding context? The most serious problem underlying this debate is that neither "access" nor "integration" has so far been defined in a precise and formal way (for a comprehensive overview see Kuperberg, 2016). For example, "access" has at least been used to refer to (i) lexical access, (ii) semantic access/retrieval, (iii) the effects of lexical prediction on access, and (iv) the effects of semantic prediction on lexical access. How "integration" is interpreted depends, in general, on the underlying theoretical framework. For example, Baggio and Hagoort (2011) use the term to refer to the linguistic operation of unification that combines the linguistic representation of the context with the linguistic representation of the CW. On this view, the N400 correlates with a compositional operation. This is contrasted with an access view according to which retrieving information from semantic memory is a non-compositional operation (see also Lau et al., 2008). Instead, Van Petten and Luka (2012) use the term to simply refer to any effects of context that start to impact as the form features of the incoming word become available (distinguishing this bottom-up primacy from pre-activation). Finally, other approaches, like the computational approach by Rabovsky and McRae (2014), do not assume separate stages for lexical access and subsequent integration.

The second debate centers on which (combinations of) empirical measures underlie N400 activity. Three such measures have been used: predictability, semantic similarity, and plausibility. Predictability of a word is mostly quantized as cloze probability: the percentage of participants in a cloze reading study that used this word to continue a sentence or a text (cloze probability was introduced in Taylor, 1953). Semantic similarity is related to memory-based models of text processing. Such models are based on the assumption that simple lexico-semantic relationships within the internal representation of context interact with lexico-semantic relationships stored in long-term memory and prime upcoming lexical information through spreading activation, called "resonance" (cf. Kuperberg and Jaeger, 2016). On this approach, the context is taken as a bag of words and, therefore, as a lower level representation that is distinct from higher-level representations of the event structure that are based on combinatorial operations, linking the objects (discourse referents), e.g., by thematic roles ("who does what to whom") (cf. Kuperberg and Jaeger, 2016). Semantic similarity is often quantized by means of latent semantic analysis (LSA, see the articles in Landauer et al., 2007 for details). On this account, pairwise term-to-document semantic similarity values (SSV) are extracted from corpora by calculating the cosine similarities between the vectors corresponding to the critical words and "pseudo-document vectors" that correspond to the prior context up to the critical word (see Kuperberg et al., 2020 for an application). Finally, plausibility can be quantized by offline rating or norming tasks in which participants evaluate the plausibility of the target sentence including the critical word.

A further question that is heavily debated concerns the relation between the functional characterizations (access vs. integration) and the three empirical measures. In this debate too, there is no consensus. For example, some researchers link access to prediction quantized by cloze probabilities (Federmeier and Kutas, 1999; Lau et al., 2008; Kuperberg et al., 2020) while others do not. An example of the latter strategy is the Retrieval-Integration model of Brouwer and colleagues in which access is related to semantic similarity though the similarity is not quantized by LSA (for details see Delogu et al., 2019). Integration is often linked to plausibility. The less plausible the critical word is in relation to its context, the higher is the cost of integrating the word into this context (see e.g., Nieuwland et al., 2019 for discussion). This cost is reflected in the size of the N400 amplitude. However, the correlation between the N400 and predictability and the N400 and plausibility is not necessarily evidence for an access or an integration view, respectively. For example, in the context of "You never forget how to ride a ..." "bicycle" is both a more predictable and a more plausible continuation than "elephant" (Nieuwland et al., 2019). The overall greater plausibility of the sentence with the completion "bicycle" can, therefore, also be taken as reflecting facilitated access.[3]

In this study, we will sidestep the issue of how access and integration should or could be defined and the question of how these two theoretical notions can be related to predictability, plausibility, and semantic similarity. The empirical starting point of our account is two important empirical findings about the N400. First, some studies have found that CWs with the same cloze probability differ in N400 activity (see e.g., Federmeier and Kutas, 1999; Kuperberg et al., 2020 and the discussion below in section Predictability, Plausibility, and Semantic Features). Second, there are studies that found a temporal dissociation between a predictability and a plausibility component (in that order) during the N400 time window (see Nieuwland et al., 2019 and section Temporal Dissociations Between Predictability and Plausibility below). Basically, two strategies have been proposed for dealing with these empirical findings. The first strategy takes predictability as central and tries to explain away plausibility by analyzing "same-cloze-different-N400" examples in terms of either differences in the overlap of pre-activated and actually found features (Federmeier and Kutas, 1999) or the number of non-pre-activated features that need to be activated upon encountering the CW (Kuperberg et al., 2020). On the negative side, one has that this strategy fails to give an account of how

[1] More specifically, the authors still observed a slightly larger N400 for "slow" compared to "quick." However, as noted by the authors, inspecting the grand average ERPs clearly showed that a substantial part of the N400 effect elicited by "slow" was eliminated if the target sentence was presented without the embedding context. This was confirmed by a joint ANOVA on mean amplitude in the 300 to 500 msec latency range (van Berkum et al., 1999, p.661).

[2] The following paragraphs owe a lot to comments from our editor Gina Kuperberg.

[3] We are indebted to one reviewer for this observation.

the plausibility component in the temporal dissociation examples can be reduced to predictability. The second strategy is hybrid views that are mostly based on temporal dissociation examples and in which N400 activity is functionally characterized by both a predictability and a plausibility component (see Nieuwland et al., 2019 and section Temporal Dissociations Between Predictability and Plausibility below). On the negative side, one has that these views do not provide a theoretical model in which predictability and plausibility are given formal definitions except in terms of cloze probability (predictability) and offline ratings (plausibility).

Given these strategies, the two central questions in this debate are whether plausibility can be reduced to predictability and how the temporal dissociation can be accounted for. One strategy for answering this question is to first provide a theoretical model in which both notions are formally defined. Given such a model, one way to proceed is to prove that the definition of plausibility can be reduced to that of predictability and then to show how the relevant data can be accounted for by this definition (reductive strategy). An alternative way is to stay with the two definitions and explain the data in terms of both definitions (hybrid view). In this study, we will adopt the second way. The theoretical model will be based on the notion of a frame (Barsalou, 1992), which is closely related to the notion of a script from cognitive science, Schank and Abelson (1977). The definition of both the predictability and the plausibility component is related to the cognitive operation of categorization.

Similar to prototype theory, we assume that categories have a graded structure. This structure is defined by assigning weights to both attributes and their values. Weights, in turn, are defined by probabilities. This graded structure allows for the definition of typicality, i.e., a binary relation between categories. Having the notion of typicality, it becomes possible to distinguish between information gain and typicality. This can be seen as follows: Given a context built upon the interpretation of the words $w_1 \ldots w_t$, a partial representation of a scenario or a script and an event have been construed. For the current event, particular argument positions *arg*, are still open in the sense that none of the words w_i are assigned to this position. With each *arg* a category C_{arg} is associated. If a CW w_{CW} is encountered that fills the open argument position *arg*, *arg* is discharged. The word w_{CW} expresses a category C_{CW}. The found category C_{CW} must be combined with the categorical information C_{arg} required by the event. This combination will be modeled as an update operation: C_{arg} is updated with C_{CW}. This update operation is the composition of two operations that are related to categorization in the following way. The first operation determines the information gain that is got by C_{arg} given C_{CW} by computing the features in C_{arg} that are not disconfirmed by C_{CW}. This operation is related to predictability: which information in C_{arg} is retained after the combination of C_{arg} and C_{CW}? The second operation computes the typicality of C_{CW} relative to C_{arg}. This computation correlates with plausibility because typicality can be taken as answering the question of how plausible are the features in C_{CW} relative to those in C_{arg}. Hence, whereas predictability focuses on C_{arg} (which features in this category are not disconfirmed?), plausibility focuses on C_{CW} (how typical are the features in this category in relation to C_{arg}?).

The reminder of this study is organized as follows. In section 1.1 we discuss feature-based approaches with special attention to studies focusing on animacy and the question whether plausibility can be reduced to predictability. In section 1.2, we discuss studies that found a temporal dissociation between predictability and plausibility during the N400 time window. The topic of 1.3 is the question whether plausibility in the N400 time window refers to whole event structures or to concepts related to objects participating in such structures. In section 2, we define our hybrid view in an informal manner by relating plausibility to typicality and by relating predictability to information gain. Finally, in section 3 an outline of the formal framework is presented together with a discussion of some relevant examples from the previous sections in this framework. This section closes with the sketch of an extension of the framework to script knowledge.

1.1. Predictability, Plausibility, and Semantic Features

As mentioned in the introduction, the most prominent way of operationalizing predictability is by cloze probability. The correlation between word predictability so defined and the amplitude of the N400 is well established with correlations of $r = 0.8$ or even higher for some studies (for details see Nieuwland et al., 2019).

One kind of counterexample to this dependency is cases in which two CWs with the same low cloze probability elicit N400 amplitudes of different size. For example, Federmeier and Kutas (1999) compared BC (best completions, i.e., highest cloze probability) with two other types of completions: those that came from the same semantic category as the best completion (within-category violations) and those that did not (between-category violations).

(1) They wanted to make the hotel look more like a tropical resort. So along the driveway they planted rows of palms / pines / tulips.

Though both "pines" and "tulips" in (1) have the same low cloze probability, the N400 amplitude for "pines" is smaller than that for "tulips." Federmeier and Kutas explain this pattern by assuming that semantic memory has a categorical structure such that categories are represented by interrelated sets of features instead of atomic units. Objects belonging to the same category share, in general, many features, namely those that are common to all members of the category. Given a particular context, specific features of a category are pre-activated. The greater the overlap between these pre-activated features and the features associated with the category expressed by the CW, the more the N400 amplitude is attenuated. For example, the context prior to the CW in (1) pre-activates features like "habitat = tropics" and "height = tall." The best completion "palms" satisfies all these pre-activated features. Though "pines" fails to satisfy "habitat = tropics," it satisfies "height = tall" and all features of the category "tree," which is the category of the BC "palms." Hence, "pines" is a within-category violation. By contrast, "tulips" is a between-category violation because tulips are flowers and not trees although there is a common supercategory, namely "plant." Hence, "tulips" does not satisfy the features that are specific of trees, and in addition, it fails to satisfy "habitat = tropics" and "height = tall."

A second kind of studies in which the N400 amplitude differed despite identical (low) cloze probabilities involves animacy violations. One example is the study by Kuperberg et al. (2020). They used examples like those in (2) where contexts where either categorized as high constraint (HC) as in (2-a) or as low constraint (LC) as in (2-b).

(2) a. The lifeguards received a report of sharks right near the beach. Their immediate concern was to prevent any incidents in the sea. Hence, they cautioned the swimmers / trainees / drawer . . . .
b. Eric and Grant received the news late in the day. They mulled over the information and decided it was better to act sooner rather than later. Hence, they cautioned the trainees / drawer . . . .

In the four conditions (HC vs. LC and "trainees" vs. "drawer") predictability quantized by cloze and semantic similarity quantized by LSA were held constant. The authors found that the N400 amplitude for "trainees" as well as that of "drawer" were independent of whether the context described a HC or a LC scenario. However, "drawer" elicited a slightly, but significantly, larger N400 amplitude than "trainees" (in both scenarios).

The authors interpret N400 activity as reflecting access to the semantic features associated with new bottom-up input that has not already been predicted (Kuperberg et al., 2020, p. 3). For example, in the LC scenario (2-b) only features that are characteristic of animate objects like SENTIENT and CAN_MOVE are pre-activated. By contrast, in the HC scenario (2-a) additional features like IN_WATER and AFLOAT are pre-activated. The CW "swimmers" in the HC scenario satisfies all of these features so that no new semantic information needs to be activated. As a result, "swimmers" only elicits a small N400 amplitude. The CW "trainees" satisfies the features related to animacy: SENTIENT and CAN_MOVE in both scenarios. However, in the HC scenario it fails to satisfy the additional features imposed by the context. Hence, a comprehender must retrieve additional features that more specifically characterize trainees like LEARNING and NOVICE in that context. Since more features need to be activated, the amplitude of the N400 for "trainees" is larger than that for "swimmers." Finally, in both scenarios "drawer" matches none of the pre-activated features. Therefore, a comprehender must retrieve all of its properties including features like STORAGE and CAN_OPEN. Hence, the N400 amplitude for "drawer" is the largest. The example provides evidence that even if the context is low-constraining the verb can already activate features of the upcoming word that are related to animacy (compare LC scenario). By contrast, the activation of other features depends on other factors like contextual information and context strength.

Empirical evidence for this distinguished role of animacy features comes from the study Wang et al. (2020). The authors exploited the inherent difference in the semantic similarity structure of animate and inanimate nouns. Objects denoted by animate nouns share more co-occurring features than objects denoted by inanimate nouns. This difference shows up in the fact that the category "inanimate" has a larger number of subcategories than the category "animate." In the brain, semantic features are thought to be represented within widely distributed networks (see, for example, Huth et al., 2016). These differences in the way features are stored can give rise to differences in similarity among the spatial patterns of neural activity associated with the processing of words that are related to these categories (Wang et al., 2020). The authors used representational similarity analysis (RSA) (Kriegeskorte et al., 2008), which is one way of detecting such neural differences, in combination with magnetoencephalography and electroencephalography (MEG and EEG). Their hypotheses were as follows: (i) If comprehenders can use the animacy constraints of verbs to predict the semantic features associated with the animacy of an upcoming noun, then the similarity in spatial patterns should be greater following animate-constraining than inanimate-constraining verbs; (ii) if these animacy predictions are generated regardless of being able to predict specific words, this effect should be independent of context strength, i.e., it should be the same in HC and in LC scenarios. They used three sentence scenarios like those in (2) and (3).

(3) a. Judith was working on the origami project for her office fundraiser. She was starting to get frustrated because it was her third attempt at making a crane. Nevertheless, she unfolded the . . . HC
b. Judith was nearing the end of her rope. She didn't think she could keep going. Nevertheless, she unfolded the . . . LC

Verbs in the final sentences constrained for either an animate [e.g., (2)] or an inanimate theme [e.g., (3)] and the broader discourse constrained for either a specific noun (HC scenario) or multiple nouns belonging to the same animacy category (LC scenario). The authors found that the spatial pattern of neural activity for animate-constraining verbs was significantly more similar than for inanimate-constraining verbs in both datasets (MEG/EEG). Furthermore, this effect was independent of context strength: It was just as large following HC as following LC scenarios. This effect began after the peak of the N400 component evoked by the verb and, therefore, past the stage at which comprehenders are likely to have accessed the lexico-semantic features of the verb and well before the direct object (theme) was actually encountered.[4]

Given examples like (2) and the results of Wang et al. (2020), the following hypothesis can be put forth.

(4) If two CWs w_1 and w_2 have the same low cloze probability and w_1 satisfies the animacy constraints imposed by the

[4] The authors suggest that this was the first time point at which comprehenders were able to infer the full high-level event structure (e.g., "agent cautioned animate noun") and that they used this structure to predict animacy features of the (upcoming) theme (Wang et al., 2020, 3289f.).

verb whereas w_2 does not, the N400 amplitude of w_2 is larger than that of w_1.

Though this hypothesis may seem to be unrelated to the account of Federmeier and Kutas, there is the following relationship. Consider (2-a). Both "swimmers" and "trainees" denote animate objects whereas "drawer" denotes inanimate ones. Hence, "swimmers" and "trainees" share a common supercategory, "animate," whereas "drawer" is a between-category violation because the common supercategory is "material object." The difference between the two examples is the level in the categorical hierarchy at which (dis-)similarities are located. Whereas, in (1) this is a very concrete level ("tree," "flower," and "plant"), and it is a more abstract level in (2). On this modeling, "trainees" shares with the set of pre-activated features (or the best completion "swimmers") the animacy features that are specific to all objects falling under the corresponding category, whereas this does not hold for "drawer." Since the activation of animacy features is independent of context strength, this argument applies mutatis mutandis also to the LC scenario in (2-b). Further evidence for the distinguished role of the animacy features comes from Paczynski and Kuperberg (2011). For example, the authors used examples like those in (5).

(5) a. At headquarters the manager interviewed the applicant for thirty minutes.
b. At headquarters the manager surprised the applicant after thirty minutes.
c. At headquarters the manager interviewed the application for thirty minutes.
d. At headquarters the manager surprised the application after thirty minutes.

The CW in the violating conditions differs from the CW in the non-violating conditions primarily in its animacy features. Furthermore, all CWs had the same low cloze probability. The authors found that the N400 amplitude in the non-violating cases (5-a) and (5-b) did not differ, i.e., it was not modulated by the thematic role (experiencer for "surprise" vs. patient for "interview"). Similarly, the N400 amplitude showed no difference in the two violating conditions though it was larger than in the non-violating conditions.

However, there are a number of studies that provide counterexamples to the claim that animacy features play the role attributed to them in hypothesis (4). The study Szewczyk and Schriefers (2011) shows that this need not be the case. The target language used in this study was Polish. The authors used scenarios in which the target sentence had a canonical subject, verb, object (SVO) order and in which subjects were unambiguously marked by nominative case and direct objects were unambiguously marked by accusative case. In all examples, either an animate or an inanimate object was highly expected. In one condition, the direct object satisfied all constraints, i.e., all selection restrictions imposed by the verb and contextual constraints. In a second condition the (in-)animacy constraint was violated and in a third condition, the (in-)animacy constraint was satisfied, but either another selection restriction or a contextual constraint was violated. Both violation conditions had a cloze probability of 0, whereas cloze probability was 0.44 in the non-violating condition. Below the English translation of two examples used in the study is given.[5]

(6) a. Although it was late autumn and bitter cold, little John was running in the backyard with his neck bare. His worried grandma prepared some wool and knitted a scarf (nv) / a medicine (sv) / an employee (av)
b. A young RAF pilot was returning to his base when he suddenly notices a Messerschmitt. The pilot fought a duel shooting down the airplane (nv) / the scarf (sv) / the patient (av).

The authors found that both kinds of violations elicited an N400 effect relative to the non-violating condition. Most importantly, the N400 amplitudes did not differ, i.e., both kinds of violations elicited an amplitude of the same size. If animacy violations were worse than others, than "medicine" in (6-a) should elicit a smaller N400 amplitude than "employee" because it satisfies the (in-)animacy constraints whereas "employee" does not. Following the same argument "scarf" in (6-b) should elicit a smaller N400 than "patient." Similar results have been reported in Quante et al. (2018). Two examples from this study are given in (7).

(7) a. Peter stand bei Morgendämmerung auf, fuhr den ganzen Tag Traktor und fütterte abends seine Kühe. An manchen Tagen wäre er aber lieber kein Bauer / Trick sondern ein unbekümmertes Kind.
Peter gets up at dawn, drives the tractor all day and feeds his cows in the evening. On somw days he would rather not be a farmer / trick but a carefree child.
b. Luisas neues WG-Zimmer war sehr klein, hatte aber hohe Decken. Um Platz zu sparen, kaufte sie sich deshalb ein Hochbett / Schwein im Baumarkt.
Luisa's new room was very small but had high ceiling. To save space, she bought herself a loft bed / pig in the store.

Though "Trick" violates the animacy constraint while "Schwein" does not (pigs can be bought), there was no difference in the N400 amplitude between the two conditions.

What these counterexamples show is that violations of constraints that are not related to animacy violations can have the same effect on N400 activity: An N400 amplitude of the same size is elicited. This provides evidence against the hypothesis in (4). More generally, one has the following. At least implicitly, the hypothesis (4) is based on the following assumption. Features are related to a particular level in a categorical hierarchy. The higher this level, the greater is the set of violated features and the higher is the corresponding N400 amplitude. For example,

[5] The following abbreviations are used: nv: no violation; av: animacy violation; sv: semantic violation. By a semantic violation Szewczyk and Schriefers (2011) understand any violation of a non-animacy selection restriction or a contextual constraint.

animacy features are related to the distinction at the (high) level of material objects. A violation of an animacy feature results in a violation of many features and a pronounced N400. Other types of features are related to lower levels in the hierarchy. Two types of such features related to N400 activity that need to be distinguished are (a) selection restrictions imposed by the verb that are not related to animacy and (b) features that are imposed by the context. As an example of the former type, consider the verb "caution." It requires its theme argument to be in danger. The verb "knit" imposes the constraint that its theme argument can be manufactured by this type of action. These constraints are lower in the hierarchy because they can be failed to be satisfied while the animacy features are still satisfied. For example, in (6-a) "medicine" is inanimate but it cannot be manufactured by a knitting process. Constraints imposed by the context are even lower in the hierarchy. For example in (2-a), the objects cautioned are most likely persons (and, therefore, human) who happen to be afloat and in water. What the counterexamples discussed above show on this modeling is that at least in particular contexts the failure to satisfy a particular feature that is not related to an animacy violation and that is therefore lower in the categorical hierarchy can have the same effect on N400 activity as a violation of an animacy constraint, contrary to the hypothesis in (4). More importantly, these counterexample provide evidence that differences in N400 activity for CWs with the same low cloze probability can be explained solely in terms of differences at the level of predictability. For example as discussed above, Kuperberg et al. (2020) explicitly adopt the strategy that the difference between "trainees" and "drawer" is a difference in pre-activated and features activated upon encountering the CW and not as a difference in plausibility over and above predictability. As a result, plausibility is "explained away" in favor of predictability. What is left open by these counterexamples is, of course, whether this additional component in N400 activity is in the effect plausibility of an event structure (or a sentence). Before discussing this question, we will discuss a second problem for strategies that are based on "explaining away" plausibility.

Critical words can be preceded by prenominal elements like determiners and adjectives that provide information about the category expressed by this CW. These prenominal elements can either confirm pre-activated features in the preceding context (matching condition) or not (mismatching condition). If N400 activity can be characterized solely in terms of the size of the set of correctly pre-activated features, the question arises on how the effect of mismatching features can be explained in this approach. Before tackling this question, we will present the results of the study Boudewyn et al. (2015) that examined the influence of prenominal adjectives on N400 activity. More specifically, the authors investigated the pre-activation of features by the ERP response to adjectives that are not themselves predictable but denote features of objects that are denoted by highly predictable not yet presented nouns. To this end, they constructed two-sentence stories in which a noun in the second (target) sentence was highly predictable (e.g., "cake") and was preceded by an adjective that denotes either a typical or atypical feature of objects denoted by the critical noun. An example story is given in (8).

(8) Frank was throwing a birthday party, and he had made the dessert from scratch. After everyone sang, he sliced up some sweet/healthy and tasty cake/veggies that looked delicious.

Event-related potentials were examined at two points during the second sentence. The first time lock was to the unpredictable adjective and the second time lock was to the critical noun. For the noun, there were four different conditions: (i) locally consistent and globally predictable noun ("sweet and tasty cake"), (ii) locally inconsistent and globally predictable noun ("healthy and tasty cake"), (iii) locally consistent and globally unpredictable noun ("healthy and tasty veggies"), and (iv) locally inconsistent and globally unpredictable noun ("sweet and tasty veggies"). Predictability of the noun was established by a cloze test (cloze for BC : 78% and 0% for non-BC). All adjectives were unexpected, regardless of whether norming participants were asked to provide a single-word continuation (cloze : 0.01%) or a multiple-word continuation (cloze : 1.81%).

For the adjectives, the authors found a reduced N400 amplitude for adjectives denoting features consistent with the best completion compared to adjectives denoting inconsistent ones. The authors conclude that semantic features of objects denoted by highly predictable nouns are accessible before the predictable noun is encountered. At the critical noun, they found a graded effect of global predictability and local consistency, with (i) the smallest N400 amplitude to globally predictable, locally consistent nouns ("sweet and tasty cake"), followed by globally predictable, locally inconsistent nouns ("healthy and tasty cake") with a slightly, but significantly, larger amplitude than for "sweet and tasty cake." then follows the globally unpredictable, locally consistent nouns ("healthy and tasty veggies") and finally one has the globally unpredictable and locally inconsistent nouns ("sweet and tasty veggies").

Consider first the N400 at the (mismatching) prenominal element. Before the prenominal element is encountered, the context raises expectations about the theme of "slice up." For example, it can be sliced, served as a dessert, and served at a birthday party. Hence, features that are typical of cake-like "sweet" are pre-activated. What happens if "healthy" is encountered instead? This feature applies to different sorts of food that can be served and sliced up. However, in general, "healthy" is not a defining property of a category in the sense that it either applies to all exemplars belonging to the category or to none. Thus, the question arises whether "healthy" is a feature of cake or veggies or not. If one assumes that it is a features of the latter (because veggies are normally healthy) but not of the former (because the cake is rarely healthy), "healthy" contributes to the feature overlap if the CW is "veggies" but not if the CW is "cake." A second problem is related to correlations between features. For example, "healthy" correlates with "sweet." Knowing that some food is "healthy" will, in general, lower the expectation that it is in addition sweet because healthy food is, in general, not sweet. Applied to (8), one has the following: encountering "healthy" will lower the expectation for "sweet," which is a consequence of the fact that predicting is, in general, a non-monotonic process.[6] Does this have the effect that "sweet" no longer belongs to the set of pre-activated features? If the answer is "yes," this suggests that "veggies" is more expected because veggies are more likely to be healthy and not sweet (= both features are an element of the feature overlap) whereas cake is more likely to be sweet and not healthy (= both features don't belong to the feature overlap). As a result, "veggies" is more likely to be the CW than "cake" so that the N400 amplitude elicited by the former should be smaller than that elicited by the latter. However, this is not compatible with the results of Boudewyn et al.

The above discussion calls into question the assumption that N400 activity can be characterized by a single operation on features based solely on criteria like "confirmed" (or "matched") vs. "disconfirmed" (or "mismatched"). What seems to be missing is the possibility of expressing the condition that a confirmed or disconfirmed feature is, in addition, a feature that normally or typically belongs (or does not belong) to a category like this that is the case for "healthy" and "sweet" in relation to cake and veggies. The reason for this is that categories are not defined in terms of definitional properties, i.e., a particular set of features that together provide necessary and sufficient conditions for membership in this category. Rather, categories are defined as graded structures that allow for the definition of typicality (see e.g., Rosch and Mervis, 1975). If viewed from the debate on predictability vs. plausibility, the above discussion can be interpreted in the following way. The relation between N400 activity and prenominal elements suggests that in addition to confirmed vs. disconfirmed the distinction between typical vs. non-typical plays a role for N400 activity. If one correlates "confirmed/disconfirmed" with predictability, "typical/non-typical" is correlated with plausibility using the results from above on the role of animacy features. When taken together, the discussion in this section suggests the following three hypotheses related to N400 activity.

HT1: Categories have a graded (or prototypical) structure which allows for distinguishing between typical and atypical features (e.g., sweet vs. healthy for cake).

How should typicality be defined? One ingredient (component) is the (subjective) probabilities of a comprehender that a category has a particular feature. These probabilities are based on both world and linguistic knowledge. For example, the probability that veggies are healthy is higher than that for cakes, whereas for the feature "sweet" the opposite holds. A second ingredient is the relevance (weight and diagnosticity) of an attribute in a particular context. For example, in (7-b) the buying is carried out with the particular goal to save space. Any objects that are not conducive reaching this goal are excluded on this occasion, independently of whether they satisfy the selection restrictions imposed by the verb. Hence, features related to the goal of saving space are more relevant than other features though they also hold of the object, e.g., inanimacy features in (7-b). Relevance need not be related to a goal. In scenario (6-a), attributes related to the way the object is manufactured are more relevant than other attributes related to inanimacy.

HT2: The graded structure of categories is context-dependent. The context-dependency shows up in weights on attributes. Typicality is defined in terms of weights on attributes and weights on values.

Typicality defined in terms of weights on attributes and weights on values must be distinguished from (correct) predictions. Consider again the scenario of the birthday party. Upon encountering the prenominal element "healthy," the corresponding feature becomes pre-activated. It provides evidence for "veggies" and evidence against "cake." However, this evidence can be counterbalanced by typicality. In this particular context, the feature "healthy" has a low relevance (weight) because other features like "sweet" and "served_at_a_birthday-party" are more relevant. This has the effect that the overall contribution of this pre-activated feature to N400 activity is lower than that of a pre-activated feature with higher relevance. As a result, one has that the contribution of a pre-activated feature to N400 activity cannot be reduced to a difference in confirmed or disconfirmed prediction ("healthy" is confirmed by "veggies" but disconfirmed by "cake"). Rather, it also matters how typical these features are relative to the pre-activated features. Hence, two CWs may not differ with respect to prediction "accuracy" though they differ with respect to how typical they are relative to the set of pre-activated features. When taken together, we get the following further hypothesis.

HT3: The contribution of a feature to N400 activity is a function of both its pre-activation and its typicality.

According to the above three hypotheses, differences in N400 amplitude are not reduced to differences in pre-activated features but in addition also reflect differences in the graded structure of categories. Hence, plausibility is not "explained away" as in the approaches by Federmeier and Kutas and that of Kuperberg and colleagues. Two principle assumptions of an account based on the three hypotheses above are as follows: (i) N400 activity is correlated to two different components: information gain (prediction) and (context-sensitive) typicality and (ii) plausibility is, in effect, typicality between two concepts and not the plausibility of an event structure (or of a sentence). In the next two sections, we will review evidence for these two assumptions.

1.2. Temporal Dissociations Between Predictability and Plausibility

The so-called hybrid views (see Nieuwland et al., 2019, and references cited therein) claim that N400 activity does not index a single process but a cascade of semantic activation and integration processes. Whereas, the (non-compositional) activation component is correlated to predictability, the

[6] This is also noted by Boudewyn et al. (2015) who take their results as showing that the occurrence of an adjective denoting a feature that is atypical of the objects denoted by the expected noun leads a comprehender to dynamically adjust her expectations in such a way that the noun no longer receives the same level of facilitation as in the case of the occurrence of an adjective that denotes a typical feature, and that the presence of a local consistent feature (e.g., healthy) can raise expectations for a noun that denotes objects for which this feature is typical (e.g., veggies).

TABLE 1 | Example items in the Lau et al. study.

Predictability manipulation	Plausible predictable	Plausible unpredictable
	runny nose	dainty nose
	mashed potato	shredded potato
Plausibility manipulation	Plausible unpredictable	Implausible unpredictable
	yellow bag	innocent bag
	healthy cat	empty cat

(compositional) integration component is correlated to plausibility. Furthermore, effects of predictability and plausibility can both be observed in the N400 time window, but effects of predictability precede and may even be functionally distinct from those of plausibility (Nieuwland et al., 2019). The more general point of these approaches is a functional interpretation of ERP components according to which they most likely reflect "the combined activity of multiple subcomponents that are associated with related yet distinct cognitive processes" (Nieuwland et al., 2019, p. 20).

The main empirical evidence for the hybrid view comes from studies in which predictability and plausibility are independently varied, and a temporal dissociation between effects of these two factors is observed in the N400 time window. Lau et al. (2016) examined modulations of the N400 amplitude associated with independent manipulations of predictability. They used an adjective-noun paradigm that allowed for contrasting the effects of contextual predictability and semantic plausibility on the N400 amplitude by holding one of the two factors constant. In particular, they compared implausible adjective-noun combinations to plausible adjective-noun combinations in which the predictability of the noun given the adjective was very low ($p < 0.005$). To create balanced plausible and implausible sets, they crossed animate nouns and inanimate nouns with adjectives that must modify animate nouns and with adjectives that usually modify inanimate nouns. Example combinations are given in **Table 1**.

Predictability was computed using corpus counts instead of cloze probabilities. Plausibility was computed in an offline rating study using a scale from 1 to 7 according to what degree the adjective-noun combination made sense. Plausible items were rated much higher than implausible ones (mean: 6.59 vs. 1.75). The authors found a large effect of predictability (runny nose vs. dainty nose) with a central posterior distribution and a small effect of plausibility (yellow bag vs. innocent bag) with a leftward distribution. Furthermore, they observed a temporal dissociation of the two effects. Whereas the predictability effect appeared to onset by around 200 ms, the N400 difference due to implausibility appeared to onset substantially later.

A second study is that by Brothers et al. (2015). They used moderately constraining (cloze BC : 50%) two-sentence passages like the following.

(9) The author was writing another chapter about the fictional detective. To date, he thinks it will be his most popular <u>novel</u> / <u>book</u>.

The context before the critical word was constructed in such a way to moderately constrain toward two alternative completions that were equally likely given this preceding context, e.g., "novel" and "book" in the above example. The second set of passages was moderately constraining toward an unrelated target, e.g., "dish," but formed a low-cloze context for the actual final word, e.g., "novel," that was unpredictable (cloze : $<$ 1%), though semantically coherent.

(10) Everyone congratulated the chef on all his hard work. To date, he thinks it will be his most popular <u>dish</u> / <u>novel</u>.

Participants were instructed to actively predict the final word of each passage and to respond after each trial whether their prediction was correct. By separately averaging ERP trials for predicted ("novel") and unpredicted ("book") targets in the first passage, the authors isolated processing differences at the final CW that were uniquely driven by prediction accuracy [prediction effect (accuracy)]. The second, control, passage was used to compare unpredicted target words in the first passage (predicted: book, found: novel) with unpredicted targets in low-cloze contexts (predicted: dish, found : novel). Any differential activity between these two conditions should index the amount of semantic or discourse-level facilitation provided by the preceding context (contextual support).

For the N400 amplitude, the authors found that predicted CWs had the smallest amplitude, followed by unpredicted CWs in medium-cloze contexts, and finally CWs in low-cloze contexts. Most importantly, there was a strong temporal dissociation between effects of prediction and context facilitation. In the N400 time window, the peak of the prediction effect occurred earlier (380 ms) than that of the context effect (around 480 ms). The authors used a multiple regression analysis to single out which factors of the context were responsible for the context effect. Possible candidates were as follows: plausibility, semantic similarity, and semantic feature overlap. Plausibility was computed using offline plausibility ratings. Semantic similarity was calculated using LSA. For semantic feature overlap, the authors used first the results of the cloze norming procedure to determine the most likely completions of each low-cloze passage and the next best completion of each medium-cloze passage. They then used LSA to compute the degree of semantic overlap between each alternate completion and the actual final word, e.g., book-novel = 0.50 and dish-novel = 0.04. The result of this regression analysis showed that the N400 amplitude approximately 100 ms after the onset of the prediction effect was strongly correlated with (i) the degree of shared semantic overlap between the CW and the next best completion of the passage and (ii) the rated plausibility of the passage as a whole. The authors conclude that this analysis suggests that for unpredicted lexical items both coherence (plausibility) with the preceding discourse and activation of overlapping semantic features reduced the amplitude of the N400 and that the time difference suggests that there is no single point during lexical processing when all potential constraints affecting word processing simultaneously come to bear.

Common to all studies discussed above is that they looked at the effects of plausibility (or semantic similarity) on unpredictable, "low-cloze" words. As noted by Nieuwland et al. (2019, p. 5), these studies, therefore, do not directly address the question of whether or to what extent the well-established, graded relationship between predictability and N400 activity is confounded by other contextual semantic factors. For example, possible correlations between predictability, plausibility, and semantic similarity can make it difficult to establish their effects on semantic processing. To overcome this weakness, Nieuwland et al. (2019) examined the effects of predictability, plausibility, and semantic similarity across a full range of cloze values.[7] They simultaneously modeled variance associated with the three measures allowing them to investigate the effects of one variable (measure) while controlling for the others. Predictability was determined using a cloze test, and plausibility was computed using a norming test based on a 7-point scale. On average, high predictable nouns were rated as plausible whereas low predictable nouns were rated as neither plausible nor implausible. The authors found that effects of predictability and plausibility both occurred in the N400 time window, but the former dominated the rise of N400 (i.e., upward flank), while the latter set in at its fall (i.e., its downward flank). By contrast, semantic similarity [calculated using both LSA and Snout, a word2vec-compatible 'continuous bag of words' (CBOW) prediction-model] did not have a strong effect on N400 activity over and above the effects of predictability and plausibility. Importantly, they found that even when accounting for the possibility that plausibility and semantic similarity have stronger effects for relatively unexpected words, plausibility modulated activity of the N400 after the peak effect of predictability (Nieuwland et al., 2019, p.18).

1.3. Plausibility of Event Structures or Typicality Between Categories?

If N400 activity is not only characterized by predictability but also by plausibility, the question arises how the plausibility component can be defined. In order to answer this question, the following two questions have to be answered: (i) do pre-activated features play a role, and (ii) what concepts are involved? Pre-activated features are related to an (undischarged) argument of the current event structure. The corresponding concept is C_{arg}. The event structure is related to the concept C_e (which is of type "event"). Finally, C_{CW} is the concept expressed by CW. If pre-activated features play no role, this means that C_{arg} is not involved in the definition of the plausibility component. Plausibility is defined as the plausibility of the update of C_e with C_{arg}. This way of defining the plausibility component will be called the Strict Plausibility Hypothesis. If pre-activated features play a role, C_{arg} is involved. Two possibilities must be distinguished. According to the first possibility, plausibility is computed in two steps. C_e is first updated by C_{arg} to C_e' and than the plausibility of C_e' with C_{CW} is computed. Updating C_e with C_{arg} possibly changes the probabilities of which nouns are expected and hence which nouns yield a (most) plausible event structure. This will be called the Plausibility-cum-Prediction Hypothesis. Common to this hypothesis and the first one is the assumption that it is the plausibility of an event structure that is computed. This is in contrast to the third hypothesis that corresponds to the second possibility. Plausibility is defined in terms of an operation on C_{arg} and C_{CW}. On this account, pre-activated features act directly through semantic memory without an intermediate step relating them to C_e (for a similar view see Paczynski and Kuperberg, 2011). As a result, plausibility of an event structure plays no role.

One way of testing the three hypotheses is to introduce a feature or a set of features of the CW before this word is encountered. Importantly, this feature (or set of features) is not in accordance with features that have already been pre-activated so that a mismatch between the newly and the previously activated features results. In the study two different strategies have been used to test these hypotheses. The first strategy uses an induced prediction. Before the target sentence, the comprehender is told that a particular word will occur in the continuation, and unbeknown to her, this word is the CW. The second strategy uses prenominal elements in an NP of which the CW is the head noun. Examples of prenominal elements are adjectives and determiners.

The first strategy was used by Szewczyk and Schriefers (2018). They used two types of scenarios. The context for both scenarios was the same. In the first type, this context was followed by the target sentence in which the CW was either plausible or implausible given the preceding context. In the second type, the target sentence was preceded by a sentence in which an explicit prediction was introduced. A comprehender was told that the particular word X would be used in the following text. This word was identical with the CW. Hence, there were four conditions by crossing induced vs. non-induced prediction with the factor "(im-)plausible." An example is given below in (11).

(11) a. Context: My uncle loves to make practical jokes. During the last summer he mounted a triangle fin on his back, jumped into the water and approached the swimming area with his fin only above the water.
b. Induction of prediction: In the upcoming sentence you will see the following word: "shark" / "doctor."
c. Target sentence: There was terrible fuss and everybody thought they saw a <u>shark</u> / <u>doctor</u> approaching them.

The authors found an N400 only in the no-induced-non-plausible condition. In the other three conditions, no N400 was observed. These results are incompatible with the Strong Plausibility Hypothesis. According to this thesis, there should be a difference in N400 amplitude in the two induced prediction conditions. The prediction component yields the same results because $C_{arg} = C_{CW}$ in both conditions. Since the induced prediction does have no effect on the plausibility of the resulting event structure, the CW "shark" results in an event structure that is more plausible than the event structure that results if "doctor" is encountered. However, the N400 amplitudes did not differ in the two conditions. The results are compatible with the other two hypotheses. Let us start with the Plausibility-cum-Prediction

[7] This study re-analyzed data from the large scale replication study Nieuwland et al. (2018), which is based on the data in DeLong et al. (2005).

Hypothesis. Processing the explicit prediction in the incongruent condition changes the expectations with respect to which event structure is described. Prior to the prediction, an event structure was expected in which sharks participate, e.g., that a shark is approaching the swimming area. This expectation is changed by the induced prediction "doctor," which has the effect of raising the probability of an event structure in which a doctor participates (and lowers the probability of an event structure in which a shark occurs). Compatibility with the Typicality Hypothesis is shown as follows. According to this thesis, the plausibility component is modeled as an operation on C_{arg} and C_{CW}. Since one has $C_{arg} = C_{CW}$, no N400 is expected.

Let us next turn to studies that allow for distinguishing between the two other hypotheses. Recall from section 1.1 that Boudewyn et al. (2015) found for examples like those in (12) the following ranking of N400 amplitudes: "sweet and tasty cake" < "healthy and tasty cake" < "healthy and tasty veggies" < "sweet and tasty veggies."

(12) Frank was throwing a birthday party, and he had made the dessert from scratch. After everyone sang, he sliced up some sweet/healthy and tasty cake/veggies that looked delicious.

The results are incompatible with the Plausibility-cum-Prediction Hypothesis. Encountering "healthy," changes the expectations of the kind of birthday party that is being described. Now a comprehender expects a birthday party that is atypical at least with respect to some food that is served. Healthy food becomes the most expected food in this context. As a result, "healthy and tasty veggies" should elicit an N400 amplitude that is not smaller than that for "healthy and tasty cake." By contrast, the results are compatible with the Typicality Hypothesis. The context pre-activates features of food that is typically served at a birthday party. Encountering "healthy," C_{arg} is updated because the corresponding feature is added to this concept. Cake is still an expected food. However, it is now not the most typical kind of this sort because being healthy is an atypical property of cakes.

According to the Typicality Hypothesis, a "mismatching" prenominal element targets only C_{arg}. Before the prenominal element is encountered, a particular set of objects falling under this concept is expected most. The effect of a mismatching prenominal element is to change this expectation to a different set. As a result, nouns that were unpredictable before become (more) predictable afterward. According to this thesis, the effect of a mismatching element is, therefore, purely prediction-driven and not related to the plausibility of event structures. By contrast, according to the Plausibility-cum-Prediction Hypothesis, not only C_{arg} is changed but also C_e. This latter change is related to the plausibility of the event structure. Hence, it is, at least in part, plausibility-driven. This raises the question of whether there is neural evidence that allows for distinguishing between the two hypotheses. According to the Plausibility-cum-Prediction Hypothesis, a mismatching pre-nominal element should trigger a revision that is driven by the overall plausibility of the continuing text and, therefore, of the resulting event structure. By contrast, according to the Typicality Hypothesis, the revision should be driven by a revision that only targets C_{arg} and, hence, the predictability of an upcoming noun, independently of the plausibility of the resulting event structure. This question was investigated in Fleur et al. (2020). The authors investigated pre-nominal effects in Dutch definite NPs. In Dutch, definite articles ("de / het") are marked for gender. One hypothesis tested by the authors was the "noun prediction revision hypothesis." According to this hypothesis, comprehenders predict the noun (with or without its gender) and then use article gender, once available, to revise the noun prediction. They used scenarios that strongly predicted a definite NP as its best continuation, followed by a definite NP with the expected noun or an unexpected, different gender NP. An example is given in (13).

(13) Het is zondagochtend. De gehele gelovige familie gaat zoals altijd naar de kerk / het gebedshuis in het dorp.
It is Sunday morning. The whole religious family goes, as always, to the church / the worship place in the village.

The authors found that gender-mismatching articles elicited increased N400 activity compared to matching articles, consistent with several other studies (see Fleur et al., 2020 for references). A second question that was addressed by the authors was whether mismatching articles caused comprehenders to revise their noun prediction instead of simply dropping it. Such a revision process could be correlated with the contextual constraint toward one alternative continuation. For example, encountering "het" in (13) instead of "det" a comprehender may revise his prediction to "gebedshuis." This revision should show up in two effects. First, there should be an effect in the neural response to gender-mismatching articles, and second, a successful revision should facilitate the processing of the corresponding noun that should be reflected in an attenuated N400 amplitude. Prediction revision at the article was quantized as next-word entropy on article-elicited ERPs in the 500–700 ms time window. Revised predictability of nouns was quantized as cloze probability of the prediction mismatching nouns given a gender-mismatching article. The authors found that next-word entropy on article-elicited ERPs correlated with revised predictability, i.e., more predictable nouns elicited smaller N400 amplitudes. Importantly, since other factors like semantic similarity to the (originally) predicted noun and plausibility of the resulting sentence were controlled for, the reduction in the N400 amplitude can be attributed to a revision of a prediction and not to semantic similarity to the initially predicted noun or the overall plausibility of the sentence.[8] When taken together, the studies Boudewyn et al. (2015) and Fleur et al. (2020) provide evidence for the Typicality Hypothesis and against the Plausibility-cum-Prediction Hypothesis.

[8] It is important to note that the authors underline the exploratory character of the analysis of prediction revision. In particular, the ERP effect associated with revised constraint, i.e., next-word entropy, reached the traditional level of statistical significance only in a subset of the analyses performed by the authors.

2. A THEORETICAL ACCOUNT OF THE HYBRID VIEW FOR THE N400

The preceding sections have provided evidence that (i) N400 activity is functionally characterized by two different components: predictability and plausibility; (ii) both components are operations on C_{arg} and C_{CW}; (iii) the predictability component cannot be defined in terms of feature overlap between C_{arg} and C_{CW}; and (iv) from (ii) it follows that the plausibility component is not related to the plausibility of an event structure.

In order to give a theoretical model of a hybrid account, both the predictability and the plausibility components have to be defined. For the predictability component, the central question is as follows: how exactly is retrieving features from long-term memory linked to the modulation of the N400 amplitude? Since retrieving information is related to prediction, the link to N400 activity should be defined in terms of a function of pre-activated and actually found features. For the plausibility component, the corresponding question is as follows: what is the target into which pre-activated and non-pre-activated features get integrated and how is this operation defined? An answer to this question must take into account that the N400 is only *one* ERP component that is linked to semantic processing. More specifically, one has to distinguish the integration operation related to N400 activity from that (or those) related to brain activity in the post-N400 time window, in particular to late positivities.

2.1. Plausibility and Typicality

One, if not the most important cognitive role of categorization, is to allow for generating (default) inferences. As Holland et al. (1986) put it: "To know that an instance is a member of a natural category is to have an entry point into an elaborate default hierarchy that provides a wealth of expectations about the instance." This can be illustrated with an (in-)famous example from Artificial Intelligence (AI). If someone learns that Tweety is a bird, then using her knowledge that birds normally fly, she will (defeasibly) infer that Tweety can fly. Such default inferences not only apply to categories expressed by common nouns like "bird" but also to the categories associated with argument positions in event structures and scenarios. More specifically, one has that each critical word expresses a category. Similarly, each argument position of a verb is associated with a (most specific) category and in each scenario each event or state denoting expression is associated with a (most specific) category. For example, in scenario (1) two default inferences for the theme of the planting event is that its habitat are the tropics and that it is tall. Default inferences are, at least in general, context-dependent and, hence, non-monotonic. If a comprehender later comes to know that Tweety is in effect a penguin, the default inference that he can fly will be given up. Similarly, if she learns that, in effect, pines and not palms were planted, she has to withdraw the inference that the habitat are the tropics. What triggers such inferences is the graded structure of categories. Features that belong to a category are not equivalent with respect to category membership in the sense that they represent necessary and sufficient conditions for objects to belong to the category but are assigned weights that reflect their discriminative value for the category. Hence, objects falling under a category vary in how good an example or how typical they are of the category (see Barsalou, 1985 for discussion and references). For example, the ability to fly is a typical property of birds and the property of being found in the tropics is a typical property of objects that do or should look tropical. A direct consequence of this difference in typicality is that features in the representation of the CW differ in the way they fit into the feature structure given by the pre-activated features. Even if they match with one of those features, the typicality of the feature must be taken into account.

This graded structure is not invariant but is highly dependent on constraints inherent in specific situations and contexts, (Barsalou, 1987, p. 107). As an effect, not all features of objects are relevant in a particular scenario but only a particular subset. One reason for this partial character of categories in contexts is that objects are usually used to achieve particular goals or are involved in prerequisites or consequences of actions that are undertaken to achieve such goals. For example in (14) taken from Chwilla et al. (2007), the paddles or Frisbees are used to dislocate water in order to move a canoe in the water. Hence, the important similarity between Frisbees and paddles is that they are typically made of a solid material. By contrast, the fact that pullovers share with paddles the property of being prototypically made of some biological material (wool and wood) plays no role. This relation between a goal and the relevant properties for achieving it is reflected in the N400 amplitude. It is larger for "pullovers" than for "Frisbees."

(14) The boys found a canoe in the spare room. With this, they wanted to go canoeing on the canal whatever the costs. The fact that they could not find the paddles did not lead them to make up their mind. According to the boys, you do not at all need them. They let the canoe into the water and paddled with Frisbees / pullovers.

In the scenario (1), the objects planted along the driveway are chosen in such a way that they have the effect of making the resort look tropical because this was the ultimate intention of the owners. Consider as a further illustration the following example from Roth and Shoben (1983). The authors let participants read pairs of sentences like those in (15).

(15) a. 1st sentence: Stacy volunteered to milk the animal whenever she visited the farm.
b. 1st sentence: Fran pleaded with her father to let her ride the animal.
c. 2nd sentence: She was very fond of the cow / horse.

In order to understand the second sentence of the scenario in (15-c), a comprehender must establish an anaphoric link between "cow" or "horse" and "animal" in the first sentence. Both expressions are co-referential, i.e., they refer to the same object (animal). Using reading times on the CW, the authors found that in the context of (15-a) "cow" was facilitated compared to "horse." By contrast, in the context of (15-b) the facilitation effect was reversed. One way of explaining these findings is to assume that "animal" had a different graded structure in the two examples. Whereas, cows and goats are typical examples of animals in the first context, horses and mules are typical examples in the second one. This is the case because different properties of the category "animal" are activated on two occasions. In the case of (15-a), features like MILKABLE and LIVES_ON_FARM are activated, whereas in the context of (15-b) RIDEABLE is activated. One way of modeling this context dependency of categories was suggested in Barsalou (1983). On a given occasion of use, only a subset of the properties associated with a category is usually activated. This active subset contains the following: (i) context-independent properties that are active on all occasions the concept is processed, and (ii) context-dependent properties that are activated only in relevant contexts. Such context-dependent uses of concepts will be called category concepts. In this study, we use "context-independent" as synonymous with "selection restrictions" imposed by a verb. For example, "caution" imposes on its theme both animacy constraints and the constraint that this object be in (some kind of) danger.

An example from the ERP study that was already discussed above further illustrates this distinction. For the verb "caution," animacy features are context-independent both for the actor and the theme argument. In addition, the theme argument satisfies the further selection restriction "in_danger," which too is context-independent because it is activated in every context in which this verb is used. Depending on the context in which the verb is used, additional context-dependent features can be imposed. For example, in the context of a seaside scenario like that in (2-a) features like IN_WATER and AFLOAT are added to those pertaining to animacy and other selection restrictions like "in_danger" to the theme argument. By contrast, in the LC scenario (2-b) these context-dependent features are not added.[9]

Let us relate the above considerations to N400 activity. C_{arg} is a category concept. The (pre-activated) features of C_{arg} are default inferences that are licensed either by the category underlying C_{arg} (context-independent) or by the context in which *arg* occurs (context-dependent features). C_{arg} extends the information about the current scenario and the current event, more specifically, adding C_{arg} to C_{event} leads to an extension of C_{event}, say C'_{event}, in which C_{arg} is embedded. C_{CW} is not a category concept because so far it has not yet been situated in the sense that it has been combined with the current context. This is carried out by the update operation that combines C_{arg} with C_{CW}. During this update process, the typicality of the features in C_{CW} that corresponds to features in C_{arg} is computed. The more typical these features are to those in C_{arg}, the more attenuated is the N400 amplitude. Hence, on this definition of the plausibility component, plausibility is, in effect, typicality. The computation of typicality can be seen as locating C_{CW} in the graded structure of C_{arg}. One way of viewing this "localizing" is to take it as an operation that (partially) "integrates" C_{CW} into C_{arg}. The refined thesis about the plausibility component is given in (16).

(16) The plausibility component of N400 activity is related to a typicality computation: how typical are the features in C_{CW} that correspond to a feature in C_{arg} to those in C_{arg}?

2.2. Predictions and Information Gain

Given a context c consisting of the words $w_1 \ldots w_n$ a set of pre-activated features belonging to C_{arg} related to $w_{arg} \notin c$ is given. Before the CW is encountered, the information in C_{arg} is not confirmed by bottom-up information. If CW or a prenominal element related to CW is encountered, the information in C_{arg} is so to speak tested against the empirical bedrock in form of bottom-up information. The result of this testing operation is the information gain (or prediction error) relative to C_{arg}.

The question arises of how this test operation can be defined. If categories are based on a bi-valent taxonomic hierarchy, the answer is simple. Given a feature f in C_{arg}, it is confirmed (success of prediction) if it is also in C_{CW} and it is disconfirmed (prediction error) if it is not in C_{CW}. However, this model does not take into account the situated and partial character of predictions. What gets predicted is only a small subset of the set of features that are appropriate for objects falling under a category. Hence, for most of the features neither f nor its negation is an element of C_{arg}. Let us make this precise. For a given category concept and a feature f, three cases must be distinguished: f is an element of the category concept, its negation is an element of the category concept, or neither f nor its negation is an element of this category concept. If f (the negation of f) is an element both of C_{arg} and C_{CW}, f is said to be confirmed by C_{CW}. If f is an element of C_{arg} whereas its negation is an element of C_{CW}, f is said to be disconfirmed by C_{CW}. Similarly, if the negation of f is in C_{arg} but f is in C_{CW}, f is said to be disconfirmed by C_{CW}.

The interesting case arises if there is a default inference in C_{arg} but no corresponding inference in C_{CW}. If there is a default inference in C_{arg}, this means that in the particular context that gives rise to this category concept it is likely that the object has this property. Consider, e.g., (2-a). In this scenario upon processing the verb "caution," there is a default inference for the attribute IN_DANGER and the value "swimmers" because in this particular context it is highly likely that the swimmers will be cautioned by the lifeguards. To put it differently, C_{arg} can be taken as a (situated) category concept for swimmers that extends (or situates) $C_{swimmers}$. In this case too, the feature f is said to be confirmed by C_{CW}. What happens for sorts like "trainees" for which there is no corresponding default inference in C_{arg} for the attribute IN_DANGER, i.e., for which neither f nor its negation is an element of the category concept? Given the context, it is not likely that trainees are in danger. However, there is an extension of C_{arg}, say C^{*}_{arg}, in which the inference is licensed, which can be taken as a category concept of C_{CW}. For example, if scenario

[9] We do not assume that a low-constraining context always only pre-activates context-independent features. This is the case only if a comprehender interprets the sentence in a literal way. However, a comprehender may also apply background knowledge or information that is given by the non-linguistic context. Consider the following example: "John is drinking a glass of ...." If this sentence is given in a study, it will be low-constraining because many beverages will be mentioned. However, if "John" denotes a particular person who is a strict anti-alcoholic, a comprehender who knows John will pre-activate a category concept with features that only apply to non-alcoholic beverages. An example of a non-linguistic factor is information in spoken language about the age of the speaker. If the sentence "I always read the newspaper before I leave" is uttered by the voice of a young child, an N400 is elicited on the CW, van Berkum et al. (2008). If, by contrast, this sentence is read in silence by a comprehender, this will not be the case.

(2-a) is continued by the information that the sharks were seen in a location close to that in which trainees were bathing, this probability will be high for this sort of object. Hence, one not only considers the current C_{arg} but also possible extension of it. If there is an extension that licenses the inference for C_{CW} because this extension is a (situated) category concept of C_{CW}, the default inference in C_{arg} will be said to be compatible with C_{CW}.

Hence, we arrived at a three fold distinction: confirmed, disconfirmed, and compatible. The information gain relative to C_{arg} can be defined in two different ways. One can take only those features that are confirmed by C_{CW}. This excludes both mismatched and (only) compatible features. Alternatively, this gain can include in addition to the confirmed features also the compatible ones. We suggest that for N400 activity the latter definition is correct. There are at least two reasons for this. First, as shown above, compatible features can be confirmed at a later stage of the discourse and given the fact that the speaker introduced them into the discourse, it is likely, provided she is reliable (rational). The second argument is related to a peculiarity of the N400. It is not a direct index of prediction violation. Its amplitude for "trainees" is the same in the HC scenario (2-a) and in the LC scenario (2-b). As will be shown below in section 3, in order to account for this sameness, compatible features need to be part of the information gain. Our hypothesis about the predictability component of N400 activity is given in (17).

(17) The predictability component of N400 activity is related to the information gain relative to C_{arg} defined as the set of pre-activated features in this category concept that are not disconfirmed by C_{CW}.

Whereas predictability focuses on C_{arg}: which features in this category concept are not disconfirmed?, plausibility focuses on C_{CW}: how typical are the features in this category in relation to C_{arg}? Hence, on our view of a hybrid approach to N400 activity, the whole process comprises three steps, two of which characterize N400 activity. In the first step, the context determines a category concept C_{arg} to which belong both context-independent features determined by the underlying category and context-dependent features that provide information about the category in this particular context. If the CW is encountered, C_{arg} and C_{CW} must be combined with each other. This update operation comprises two steps that are related to N400 activity. First, the information gain in terms of non-disconfirmed features of C_{arg} is computed (predictability component) and next the typicality of features in C_{CW} that have corresponding features in C_{arg} is computed (plausibility component).

3. OUTLINE OF A FORMALIZATION

Pre-activated features in C_{arg} represent default inferences that are either licensed by the underlying category (context-independent) or by the embedding context (context-dependent). Let this set be Ω. Encountering C_{CW} triggers an update operation that combines the two concepts, yielding a combined category concept. This resulting category concept is computed in two steps. In the first step, the set Ω is split into three disjoint sets: the set of confirmed features Σ_{conf}, the set of compatible features Σ_{comp}, and the set of disconfirmed features $\Sigma_{disconf}$. In the second step, the resulting category concept is construed using the result of the first step. The first step is related to the predictability component, and the second step to the plausibility component. N400 activity is functionally characterized by the properties of the two operations. For the first step, this is the entropy reduction triggered by Σ_{conf} and Σ_{comp}, and for the second step, this is the typicality of C_{CW} relative to C_{arg}. In this section, we will sketch how these ideas can be made formally precise.[10]

3.1. Concepts as Frames

The first task is to find an appropriate representational format for categories. From what has been said so far it follows that there are three principle constraints that such a format must account for: (i) the internal structure in terms of features; (ii) the graded structure in order to allow for the definition of similarity (of values) and salience (of attributes); and (iii) the context-dependent use of categories. An appropriate representational format that allows for the satisfaction of these constraints is frames. Frames are built out of attribute-value pairs. Such pairs have been called features or properties in the sections above. The value space of an attribute is sorted, i.e., an attribute can take values only in a particular set which is the sort of the attribute. The structure of frames is recursive, i.e., the value of a frame can be a frame so that this value can be specified by further attributes. Each frame is of a particular sort. Sorts are not restricted to those associated with common nouns like "fruit," "apple," or "dog" but also include sorts associated with verbs and their arguments (e.g., theme or actor) as well as sorts for scenarios (or scripts) like "seaside" or "going to a restaurant." The relation between a frame and the chains of attributes belonging to it is captured by a function θ. One has $\theta(f) = \Sigma$ if Σ is the set of features, i.e., the set of chains of attributes together with their values belonging to f. In this study, we will denote a feature consisting of an attribute (chain) A with value V as V^A. On frames of a particular sort σ, an information ordering $\sqsubseteq_\sigma$ is defined. One has $f \sqsubseteq_\sigma f'$ if each chain of attributes that is defined for f is also defined for f', and the value of the chain in f subsumes the value of the chain in f'. The information ordering and the frame hierarchy it induces can be used to account for the context-dependent use of categories. A category of a particular sort can be represented by the whole hierarchy of that sort. The use of a category in a particular context, i.e., a category concept, is represented by an element in this hierarchy so that only a particular subset of the (chains of) attributes is activated, (for a more detailed presentation of the underlying frame theory, see Naumann and Petersen, 2019). Frames in which a (chain of) attributes is assigned its value space together with a probability distribution on this space are stochastic frames (cf. Naumann et al., 2018).

3.2. Weights on Values, Probabilities, and Default Inferences

One strategy of defining weights on values is to assume that in category concepts attributes are not assigned a particular

[10] A more detailed formalization can be found in Naumann and Petersen (2021).

value but a data structure containing values that are weighted by typicality (see Cohen and Murphy, 1984 and the approach by Smith et al., 1988 for a similar proposal). One way of making this idea of a data structure formally precise in a frame theory has been suggested in Schuster (2016) (see also Schurz, 2012). Instead of assigning an attribute a particular value, it is assigned its value space together with a probability distribution on this space. In particular, each value V in the value space of an attribute A that belongs to a category C is assigned a (conditional) probability $P(V^A \mid C)$, i.e., the probability of V^A given C. These conditional probabilities can be taken to reflect subjective conditional probabilities of a comprehender that are based on his world knowledge and linguistic knowledge based on statistical regularities in texts and discourses.

Having weights on values, one can define which features belong to a category or a category concept. Recall that default inferences belong to a category concept. What is required, therefore, is a link between probability distributions and default inferences. One way to relate default inferences to probabilities was suggested by Schurz (2012). A default inference or normic conditional of the form “Cs normally have P” or “Cs are normally Ps” (formally $C \Rightarrow P$), e.g., “Birds (can) normally fly” or “Cake is normally sweet and unhealthy” only holds if the corresponding conditional probability is high. This is summarized in the statistical consequence hypothesis (SC), (Schurz, 2012, p. 531).

(18) SC: $C \Rightarrow P$ implies that the conditional statistical probability of P given C, $P(P \mid C)$ is high.

In our framework, one has $C \Rightarrow V^A$ if $P(V^A \mid C) := max(P(V_1^A \mid C), \ldots, P(V_n^A \mid C))$ and $P(V^A \mid C) > r$. Thus, a normic conditional holds for a feature V^A in a category C if its conditional probability is the maximum of the conditional probabilities of the n values of the value space. The constraint that the probability is high is defined by the requirement that $P(V^A \mid C) > r$ for some threshold value r, e.g., $r > 0.5$.[11]

How does the SC hypothesis relate to categories and category concepts? Recall that to a category concept belong both context-independent and context-dependent features. For example, for the category concept associated with the theme of “caution” context-independent features are BE_IN_DANGER and features related to animacy. These features are determined by the underlying category because they do not depend on the context. For this reason, they always belong to a category concept, (see Barsalou, 1983 for discussion). Context-dependent features result from correlations in the following way. In a category of a scenario or an event, the values of attributes are, in general, not independent of each other. Rather, there are correlations between these features [or the values of (chains of) attributes]. For example, in the holiday resort scenario in (1) the information that the resort should look tropical and that something was planted along the driveway triggers the default inference that the habitat of the objects planted is most likely the tropics and that they are tall in order to be visible. Hence, the inference has the form $C_{script} \Rightarrow V^A$ or $C_{event} \Rightarrow V^A$. In our application, V^A is always a feature in the category concept C_{arg} associated with an argument of the scenario or the event that has not yet been discharged. C_{script} or C_{event} provides the context in which the category concept C_{arg} is processed.

3.3. The Predictability Component and Entropy Reduction

Recall that we hypothesize that the predictability component is related to the information gain of pre-activated features in C_{arg} that are not disconfirmed by C_{CW}. Since the first step is input to the second step, this first step is defined in such a way that it yields three sets of features: Σ_{conf}, Σ_{comp} and $\Sigma_{disconf}$. We formalize this first step as the operation `update_set`, which takes two categories and returns a triple of sets of features. `update_set`(C_{arg}, C_{CW}) is a partial function; it is defined only if the chain of attributes for every feature in C_{arg} is also defined for C_{CW}. If this function is defined, the update operation is defined as follows: `update_set`$(C_{arg}, C_{CW}) = \langle \Sigma_{conf}, \Sigma_{comp}, \Sigma_{disconf} \rangle$ iff for each $V^A \in C_{arg}$: if $V^A \in C_{CW}$, then $V^A \in \Sigma_{conf}$; if $V \neq \bar{V}$ and $\bar{V}^A \in C_{CW}$, then $V^A \in \Sigma_{disconf}$; if $V^A \notin C_{CW}$ and $V^A \notin \Sigma_{disconf}$, then $V^A \in \Sigma_{comp}$. One has: $V^A \in C_{arg} \wedge V^A \in C_{CW}$ iff $P(V^A \mid C_{arg}) > r$ and $P(V^A \mid C_{CW}) > r$; an example is HABITAT = tropics in the holiday resort in (1) for the CW “palms.” $V^A \in C_{arg} \wedge \bar{V}^A \in C_{CW}$ iff $P(V^A \mid C_{arg}) > r$ and $P(\bar{V}^A \mid C_{CW}) > r$ for two different values V and $\bar{V}$. An example is HABITAT = tropics in C_{arg} and HABITAT = moderate in $C_{\texttt{pine}}$. An example where a feature is in C_{arg} but not in C_{CW} is LOCATION = water in the seaside scenario in (2-a) for the CW “trainees.”

The computation of the three sets fails if an attribute in C_{arg} is encountered that is not defined for C_{CW}. An example is “drawer” in scenario (2-a), as for its associated category, animacy attributes are not defined. In our approach, this failure has the effect that typicality is not computed. We will come back to this point below. Though Σ_{conf}, Σ_{comp}, and $\Sigma_{disconf}$ are sets, they are uniquely related to frames (categories). For example, one has $\theta(f_{\Sigma_{conf}}) = \Sigma_{conf}$, i.e., $f_{\Sigma_{conf}}$ is the (unique) frame to which the features in Σ_{conf} belong.

An alternative view on $\Sigma_{conf} \cup \Sigma_{comp}$ is as a measure of prediction error. The smaller this set is, the higher is the prediction error, or, using the gain in information: the smaller the gain in information in non-disconfirmed features, the higher is the prediction error. By itself, $\Sigma_{conf} \cup \Sigma_{comp}$ does not measure the information gain of non-disconfirmed features in C_{arg}. We suggest that this gain can be measured by entropy reduction, which is related to the information-theoretic measure of entropy. More generally, two theoretical metrics that have been used to measure information are surprisal and entropy. Surprisal quantifies how unexpected a state s is given a context c. Entropy quantifies how uncertain a system is about what comes next so that it derives from the probabilities of all future states (Willems et al., 2016). From these characterizations, it follows that surprisal is backward-looking: given a context c, how likely is it to encounter a state s or how likely is the updated context $c \sqcap s$? By contrast, entropy is forward-looking. It quantifies the

[11] The determination of r is an empirical question and will in general also depend on the context.

reduction in uncertainty about the current state the system is in. This difference is also reflected in the definition of the two metrics. Given a state s_t with predecessors $s_1 \ldots s_{t-1}$, surprisal at t is defined as the negative logarithm of the conditional probability of s_t given $s_1 \ldots s_{t-1}$.

(19) $\texttt{surprisal}(t) := -\log P(s_t \mid s_1 \ldots s_{t-1})$.

By contrast, entropy is not a function of the probability of the state at t but of the distribution of probability of all future states.

(20) $H(t) := -\sum_{s_{t+1} \in S} P(s_{t+1} \mid s_1 \ldots s_t) \log P(s_{t+1} \mid s_1 \ldots s_t)$.

Given this forward-looking character of entropy, one often is interested in entropy reduction. Given two states s_{t_1} and s_{t_2} at t_1 and t_2, respectively, entropy reduction triggered by state s_{t_2} is defined as the difference in entropy between t_1 and t_2.

(21) $\triangle H(s_{t_2}) := H(t_1) - H(t_2)$.

The higher entropy reduction is, the more information is gained relative to non-disconfirmed features. The relation to prediction error is the following. The lower entropy reduction is, the higher is the prediction error.

We hypothesize that there is the following relation between the two measures and the two processing components. Entropy reduction is related to the predictability component, whereas surprisal is related to the plausibility component. According to our account, the predictability component is related to the gain in the information of non-disconfirmed pre-activated features. This information gain can be taken to be given by the reduction in uncertainty about the category that is expressed by the CW. By contrast, the computation of typicality, i.e., the location of C_{CW} in the graded structure of C_{arg}, can be taken as reflecting how (un)expected the features in C_{CW} are given C_{arg}, which comprises the influence of the context on C_{CW}.

In our approach, entropy is defined on the frame hierarchy of frames of a particular sort, e.g., "swimmer" or the theme of caution events. Due to the recursive character of frames, there is, at least in principle, no upper bound on the length of chains in a frame, though for a particular frame there always exists such a bound. Hence, considering arbitrary extensions would make the computation of entropy reduction intractable. We suggest to consider n-step extensions, i.e., frames in which the maximal length of chains is n. The minimal case are frames in which all chains have length 0. These are minimal frames in the sense that only information about the sort of the frame is supplied but no relational information that links the referent of the frame to other objects. In our application, n will, in general, be low, say $n = 2$ or $n = 3$. Let us next define entropy in our approach. For a given frame hierarchy $\sqsubseteq_\sigma$ of sort σ, let $F^n_{\sqsubseteq_\sigma}$ be the set of frames in which the maximal length of chains is n and let f_t be the frame at t. Entropy at t is then defined as given in (22).

(22) $H(t) := -\sum_{f^n \in F^n_{\sqsubseteq_\sigma}} P(f^n \mid f_t) \log P(f^n \mid f_t)$.

According to this definition, entropy is 0 if f_t singles out a unique frame of length n, i.e., a unique category concept of this length. This is the case if f_t specifies values for all chains of length less or equal n. This will most likely never be the case for the simple reason that it goes against the context dependence of category concepts. For entropy reduction, one considers $f_{\Sigma_{conf} \cup \Sigma_{comp}}$ at t_2, i.e., one has $f_{t_2} = f_{\Sigma_{conf} \cup \Sigma_{comp}}$, i.e., the frame (category) that corresponds to the set of confirmed and compatible features. What is the frame (category) at t_1? This is the frame containing the features already got for the argument position before the CW is encountered. An example is the scenario of the birthday party involving sweet or healthy cake in (8) where one or more features of the category concept associated with the CW "cake" are determined by preceding adjectives. If no bottom-up information is given, one possibility is to consider a minimal frame of the given category that only contains sortal information. However, this does not account for the fact that there are possible discourse-independent default inferences in the category like those related to animacy in the category of the theme of caution events. We, therefore, suggest that at t_1 one uses a minimal frame that is closed under such context-independent default inferences.

3.4. The Plausibility Component and Typicality

The second step is executed only if the operation associated with the predictability component did not yield failure. If the first step was successful, the second step consists in building up the final category concept. This is formalized by an operation $\texttt{update}_t(\Sigma_{conf}, \Sigma_{comp}, \Sigma_{disconf})$, which takes three sets of features and returns a frame (category). This operation is defined as follows: $\texttt{update}_t(\Sigma_{conf}, \Sigma_{comp}, \Sigma_{disconf}) = f_{\Sigma_{conf} \cup \Sigma_{disconf}}$. Note that features in Σ_{conf} are features that are default inferences in both C_{arg} and C_{CW} and are thus taken over to $\texttt{update}_t(\Sigma_{conf}, \Sigma_{comp}, \Sigma_{disconf})$ because the feature in C_{arg} is confirmed by the corresponding feature in C_{CW}. However, features in Σ_{comp} that are only compatible, i.e., which are in C_{arg} but not in C_{CW} are not taken over to $\texttt{update}_t(\Sigma_{conf}, \Sigma_{comp}, \Sigma_{disconf})$ because, at least at this stage, they are not confirmed by the bottom-up information given by C_{CW}. An example is BE_IN_DANGER in the scenario (2-a) for "trainees." If instead of "trainees" "swimmers" is encountered, the situation is different. In this case the default inference is licensed for this sort of objects so that it is taken over to the resulting category concept. Elements of $\Sigma_{disconf}$ are taken over. They are context-independent default inferences in C_{CW} that disconfirm the corresponding feature in C_{arg}. An example is the features HABITAT = moderate and HEIGHT = small in scenario (1) if the CW is "tulip."

The property of this operation by which the plausibility component of N400 activity is functionally characterized is typicality of C_{CW} relative to C_{arg}. Typicality is defined in terms of weights on values (similarity) and weights on attributes (diagnosticity).

3.5. The Definition of Similarity and Diagnosticity

Similarity between features of the category concept and the representation of the CW is defined in terms of the weights on values, i.e., the conditional probabilities. We follow Schuster

(2016) whose definition is inspired by that of Smith et al. (1988) and define this similarity for a feature as the minimum probability of this feature for C_{arg} and for the corresponding value in the representation C_{CW} of the CW.

(23) $sim(C_{CW}, C_{arg} \mid V^A) = min(P(V^A \mid C_{arg}), P(V^A \mid C_{CW}))$.

Using the minimum of the value in the category concept and the value in the representation of the CW ensures that probabilities are considered at least as strongly as the values of the category concept and at least as much as the value of CW. Next, we turn to the definition of diagnosticity.

Statistical frequency does not account for the fact that features with the same (high) frequency can differ in the way they can contribute to the categorization process. What is required, therefore, is a measure that specifies the discriminative value of an attribute for the categorization process. One measure that has been proposed is cue-validity, which is defined in (24).

(24) $\texttt{cue-validity}(C, V^A) := P(C \mid V^A) = \frac{P(C \wedge V^A)}{P(V^A)} = \frac{P(V^A \mid C) \cdot P(C)}{\sum_{i=1}^{n} P(V^A \mid C_i) \cdot P(C_i)}$.

The C_i's in (24) are contrast classes, i.e. siblings of a common superordinate category. One example of such sibling categories is fruit and vegetable. Let us illustrate this definition by an example taken from Schuster (2016). Both fruit and vegetable have similar values in the COLOR attribute, e.g., "green," "red," "yellow," and "orange." In general, only knowing that an object has the value "green" for this COLOR attribute, the probability to categorize it as a vegetable is high, that is, one has that (24) is high whereas the corresponding cue-validity values for other values of the COLOR attribute are lower and, say, equally probable. Given this fact that one cue-validity value is high, the COLOR attribute should receive a high discriminative value for vegetable because peaks in the probability distribution of attribute values indicate a high discriminative strength for this attribute relative to other contrast classes or categories.

For the definition of diagnosticity, we follow Schuster (2016) and define the diagnosticity of an attribute A for a category C in terms of the maximum of the cue-validity of each of the n values of the attribute. Suppose there are m attributes $A_1 \dots A_m$. Let attribute A_i, $1 \leq i \leq m$, have n values $V_{i1} \dots V_{in}$. One then first defines $d(C, A)$ as the maximum of the reversed conditional probabilities of each of the n values of A.

(25) $d(C, A) := max(P(C \mid V_1^A), \dots, (P(C \mid V_n^A))$
$= max(\texttt{cue-validity}(C, V_1^A), \dots,$
$\texttt{cue-validity}(C, V_n^A))$.

Given d, diagnosticity of attribute A_i in category C is defined as follows.

(26) $diag(A_i, C) := \frac{d(C,A_i)}{\Sigma_{j=1}^{m} d(C,A_j)}$

One question that needs to be answered is how contrast classes (or categories) are determined. In general, diagnosticity is highly context dependent so that the determination of contrast classes depends on the context. We suggest that the contrast classes are category concepts of the same category. In particular, we suggest that the contrast classes are those category concepts in which the values of attributes are changed that give rise to a correlation, i.e., to a context-dependent default inference in the category concept. For example, in the seaside scenario (2-a) with the lifeguards and the swimmers/trainees there are correlations involving the values of the DANGER attribute and the value of the LOCATION attribute of the theme of the caution event: DANGER.LOCATION $=$ *water* $\wedge$ DANGER.CAUSE $=$ *sharks* $\Rightarrow$ CAUTION.THEME.LOCATION $=$ *water*. By varying the values of the chains of attributes in the antecedent, one gets the set of contrast classes. Contrast classes, therefore, are in effect category concepts of sort "seaside" in which the danger is located elsewhere, say, on the beach or some other location different from the water and in which the cause of the danger is not sharks. In the scenario in (2-a) the conditional probability for the value "water" of the LOCATION attribute is (almost) 1, whereas it is (almost) 0 for other locations of the danger like the beach because the danger is related to sharks.

3.6. Typicality

Finally, typicality is defined in terms of the diagnosticity of attributes and similarity of values. A preliminary definition for the typicality of category C_{CW} with respect to a category C is given in (27) (see also Smith et al., 1988; Schuster, 2016).

(27) $typicality(C, C_{CW}) = \sum_{j=1}^{m} diag(A_j, C) \sum_{i=1}^{n} sim(C_{CW}, C \mid V_i^{A_j})$.

For each attribute A_j and each value V_i of this attribute, the product of its diagnosticity with the similarity of the value is computed and the sum of these products is taken. Since both the number of pre-activated features and diagnosticity depend on the strength of the context (HC vs. LC), the typicality value is dependent on this distinction, For this reason, the computation of typicality must be adapted to these dependencies. We, therefore, suggest to use (28).

(28) $typicality(C, C_{CW}) = \frac{\sum_{j=1}^{m} diag(A_j,C) \sum_{i=1}^{n} sim(C,C_{CW} \mid V_i^{A_j})}{\sum_{j=1}^{m} diag(A_j,C) \sum_{i=1}^{n} sim(C,C \mid V_i^{A_j})}$.

(28) reflects the fact that the typicality value for C itself is lower in an LC scenario than in an HC scenario. In particular, comparing C with itself in the denominator yields the maximal value of typicality in the given context. The nominator then computes the degree of typicality of C_{CW} relative to C in that particular context. Typicality is computed for each feature in C_{arg}, provided the corresponding attribute is defined in C_{CW}. The similarity value of the feature in C_{CW} is its probability in this category, independently of whether a default inference is licensed or not. Hence, typicality is computed only for features that are context independent or that are default inferences licensed by the context. This accounts for the fact that the N400 amplitude is modulated only by a subset of the admissible features of a category, e.g., those related to achieving a particular goal.

Typicality is computed during the computation of $\texttt{update}_t(\Sigma_{conf}, \Sigma_{comp}, \Sigma_{disconf})$. In particular, one has that if feature V^A is checked in the above operation, its typicality is computed. Hence, typicality is computed for all features in C_{arg}, provided the corresponding features are defined for C_{CW}.

3.7. The Interplay Between Predictability and Plausibility

The N400 amplitude is a function of both the predictability and the plausibility component. How do the two components contribute to this amplitude? For both components, one has to distinguish between HC and LC scenarios. Let us begin with the predictability component. In an HC scenario, more features are pre-activated than in an LC scenario because there are, in general, more context-dependent default inferences in the former than in the latter. One reason for this difference is correlations between features in C_{script} and C_{arg}, i.e., context-dependent default inferences. If an HC scenario extends an LC scenario, e.g., by providing an embedding context, one has that the set $\Sigma_{conf} \cup \Sigma_{comp}$ in the HC scenario is larger than in the LC scenario. As a result, entropy reduction is larger in the former kind of scenario. This difference in context-dependent features yields a higher gain in information for the following kinds of critical words. First, there are CWs for which Σ_{conf} is large, i.e., the default inferences in C_{arg} and C_{CW} are (almost) the same. This is the case for "swimmers" in (2-a) and it also holds for "Frisbees" in (14). Though the associated category has only a few features in common with that of "paddles," what counts are the features that are activated in the particular context of (14), yielding C_{arg}, and in this respect the overlap with C_{CW} is large. A second case is CWs for which most features in C_{CW} are compatible with those in C_{arg}, i.e., for which Σ_{comp} is large. An example is "trainees" in (2-a). Here, the context-dependent features like IN_WATER and AFLOAT can possibly apply to trainees so that they are part of Σ_{comp}. As a result, there is no difference between "swimmers" and "trainees" at this level. In an LC scenario, the influence of the context on C_{arg} is weaker in the sense that less context-dependent default inferences are licensed (if any such inferences are licensed at all). As a result, the gain in information is, in general, lower than in an HC scenario.

Let us next turn to the plausibility component. In the predictability component, compatible features lead to a gain in information because they can potentially be verified. This is a consequence of the forward-looking character of this component. By contrast, in the plausibility component it is tested how typical the features in C_{CW} are relative to the graded structure of C_{arg}, i.e., with respect to the information that is predicted by the underlying category together with the preceding context. Hence, this test is based on information that solely derives from given information. This difference shows up in particular for features that are only compatible and, therefore, for non-best non-anomalous completions. Though they positively contribute to the gain in information, they have a low typicality value. There are two reasons for this. First, in an HC scenario diagnosticity for context-dependent features is high because they are discriminative of a particular role in the scenario. This high value boosts the difference in typicality between a best completion and a non-best completion because for these two CWs the difference in similarity is high. Consider again the scenario (2-a). The discourse-dependent default inference that the theme of the caution event is in water and afloat has a high diagnosticity in the category concept whereas diagnosticity is low for other locations of the danger. These features have a low similarity value for "trainees" so that typicality is low though the gain in information is high. This is in contrast to "swimmers." For this CW, the similarity values for these features are high. As a result, both the gain in information and the typicality are high for this CW.

Let us next consider an LC scenario. In this kind of scenario, less features are pre-activated compared to a corresponding HC scenario. In particular, most features pre-activated are context independent. Since these features belong to both C_{arg} and C_{CW}, typicality is high for low cloze words like "trainees" in these scenarios. Hence, the two components show contrary behavior for low cloze words. Whereas, in an LC scenario the gain in matched and compatible features is lower, the degree of typicality is higher because there are less pre-activated context-dependent features with a high diagnosticity value for which the similarity value is low. In an HC scenario, the gain in information is higher, but the degree of typicality is lower. This has the effect that for low cloze CW the N400 amplitude can be the same. The gain at the level of the predictability component in the HC scenario, compared to the LC scenario, is compensated by the lower degree of typicality, again compared to the LC scenario. Consider "trainees" in (2). In the HC scenario, the gain in pre-activated features is higher because (i) more features are pre-activated and (ii) most of these features are compatible with the information in the category concept expressed by this word in this context. However, the features are only compatible, which has the effect that they are not typical or even atypical of the category $C_{\texttt{trainees}}$. As an effect, the typicality of $C_{\texttt{trainees}}$ relative to C_{arg} is greater in the LC scenario than in the HC scenario.

If only context-independent features are pre-activated, it does not follow that there are no differences in typicality. Recall that in this case there are no other contrast classes. Hence, there are only minor differences in diagnosticity so that differences in similarity play the major role. By way of example, consider the context-independent feature BE_IN_DANGER of the theme of caution events. Without a context, there is no peak in the probability distribution that singles out a particular sort. Yet, the probability that swimmers are in danger is higher than that for trainees.

Let us illustrate our approach by discussing two examples in more detail. We start with the holiday resort scenario in (1). Recall that in this scenario there are relations between categories. Palms and pines are both trees, whereas tulips are flowers, and all three sorts of objects are plants. These categorial relations are also reflected at both the predictability and the plausibility component. Context-dependent default inferences for C_{arg} include HABITAT = tropics and HEIGHT = tall.[12] Given these context-dependent default inferences, one has for the set of non-disconfirmed features $\Sigma_{conf} \cup \Sigma_{comp}$ that it is largest for "palms," followed by that for "pines," which is followed by

[12] For simplicity, we do not state the complete chains of attributes but only the last attribute together with the value.

TABLE 2 | Similarity values for the three CWs.

	Palms	Pines	Tulips
Tropics	High	Low	Low
Moderate	Low	Low	Low

that for "tulips." As a result, entropy reduction is largest for "palms," followed by that for "pines" and that for "tulips" last. The computation of typicality yields the same ordering. For example, since tulips are small and are from a moderate habitat, the similarity values are those of tulip for those features, which are very low. Let us link this example to the definition of typicality. We use the attribute HABITAT for illustration. We make the simplifying assumption that this attribute has only two values, "tropics" and "moderate" (see **Table 2**).

Remember that the similarity values are defined as the minimum of $P(V^{\text{HABITAT}}|C_{CW})$ and $P(V^{\text{HABITAT}}|C_{arg})$. Thus, the value "high" requires that the probability is high in both C_{arg} and C_{CW}. The value "low" is got if the probability is low in either of the two categories. This is the case whenever a feature is confirmed in one category but not in the other. For "palms," the value "tropics" is high in both categories and, therefore, has a high similarity value. Since the value "moderate" has a low probability in C_{arg}, the similarity value is low too because for similarity the minimum of the similarity values in C_{arg} and $C_{\texttt{pine}}$ is taken. For "pines," one gets the following. The value "tropics" has a high probability in C_{arg} and a low probability in $C_{\texttt{pine}}$. Since the minimum is taken for similarity, the similarity value is low. The argument for the value "moderate" is similar with the roles of C_{arg} and $C_{\texttt{pine}}$ switched. Now the probability is low in C_{arg} and high in $C_{\texttt{pine}}$. Again, one gets a low similarity value for this value of the HABITAT attribute. For "tulips," the argument is the same as that for "pines." What about the value of diagnosticity? Recall that contrast classes result by modifying the value of a chain of attributes that license a context-dependent default inference. In the scenario in (1), this is the value "tropical" for the way the hotel should look like. Other values yield different looks. Let us assume that there is only one other value that is "moderate" so that there is only one contrast class. The probability $P(C_{arg} \,|\, \texttt{tropics})$ gets a high value in the scenario (1), say 0.9, whereas its value in the contrast class is low, say 0.1. For the value "moderate," the opposite values can be assumed. One, therefore, has that the cue-validity value for "tropics" is $\frac{0.9}{0.9+0.1} = 0.9$ and for "moderate" $\frac{0.1}{0.1+0.9} = 0.1$.[13] As a result, HABITAT has a high discriminative value and this high value even boosts the differences in typicality between "palms" on the one hand and "pines" and "tulips" on the other hand. When taken together, one has that "palms" has a high typicality value relative to HABITAT, whereas "pines" and "tulips" get a low value.

Let us next illustrate entropy reduction. For the sake of simplicity, we assume that the frame extensions that are considered are directly related to C_{arg}, i.e. they contain the context-dependent features in this category concept. In the scenario in (1), there are two such features: HABITAT and HEIGHT. Crossing these two features with the two values for these attributes assumed above yields four frames: f_{tt} (HABITAT = tropics and HEIGHT = tall), f_{ts} (HABITAT = tropics and HEIGHT = small), f_{mt} (HABITAT = moderate and HEIGHT = tall), and f_{ms} (HABITAT = moderate and HEIGHT = small). The frame f_{t_1} at t_1 is the category concept for the theme of a planting event that only contains context-independent default inferences. The frame f_{t_2} at t_2 is the extension of f_{t_1} that in addition contains the features from $\Sigma_{conf} \cup \Sigma_{comp}$. For the scenario with the CW "palms," both features belong to f_{t_2}, i.e., $f_{t_2} = f_{tt}$. For the scenario with the CW "pines", only the HEIGHT feature belongs to f_{t_2}, i.e., $f_{t_2} = f\_t$ because HEIGHT = tall is confirmed. Finally, for the scenario with the CW "tulips," one has $f_{t_1} = f_{t_2}$ because all predicted context-independent features are disconfirmed, i.e., prediction error is maximal. Hence, at t_1 all four frames are possible extensions (maximal entropy) whereas at t_2 the extensions depend on which features were not disconfirmed. For "palms," there are no such extensions because both features were confirmed. By contrast, for "tulips" all four extensions are still possible. In the case of "pines," there are two extensions because the HEIGHT feature is confirmed whereas the HABITAT feature is disconfirmed so that it is not an element of $\Sigma_{conf} \cup \Sigma_{comp}$.

Let us suppose the following conditional probabilities at t_1: $P(f_{\text{HABITAT}=tropics} \,|\, f_{t_1}) = 0.9$, $P(f_{\text{HABITAT}=moderate} \,|\, f_{t_1}) = 0.1$, $P(f_{\text{HEIGHT}=tall} \,|\, f_{t_1}) = 0.6$, and $P(f_{\text{HEIGHT}=small} \,|\, f_{t_1}) = 0.4$. Hence, the conditional probabilities for the four extensions are $P(f_{tt} \,|\, f_{t_1}) = 0.54$, $P(f_{ts} \,|\, f_{t_1}) = 0.36$, $P(f_{mt} \,|\, f_{t_1}) = 0.06$, and $P(f_{ms} \,|\, f_{t_1}) = 0.04$. Entropy at t_1 is 0.433. If "palms" is encountered, there are no extensions because the frame at t_2 contains both features. Hence, entropy at t_2 is 0 so that entropy reduction is 0.433. If "tulips" is encountered, one has that $f_{t_1} = f_{t_2}$ because all predicted context-dependent features are disconfirmed. Hence, entropy remains the same so that there is no reduction in entropy. For "pines," the situation is different. In this case, there are two frame extensions, adding either HABITAT= tropics or HABITAT= moderate. One has $P(f_{\text{HABITAT}=tropics} \,|\, f\_t) = 0.9$ and $P(f_{\text{HABITAT}=moderate} \,|\, f\_t) = 0.1$. Entropy at t_2 is 0.14. As a result, entropy reduction between t_1 and t_2 is 0.29 if "pines" is encountered.

The second example is the scenario of a birthday party from above and repeated below in (29), which uses prenominal elements.

(29) Frank was throwing a birthday party, and he had made the dessert from scratch. After everyone sang, he sliced up some sweet/healthy and tasty cake/veggies that looked delicious.

The context prior to the adjective raises expectations (pre-activates features) that are related to a birthday party, in particular, features that belong to categories expressed by food that is typically served on such an occasion. For our example,

[13] For the sake of simplicity it is assumed that the probabilities of the two contrast classes are the same. Furthermore, the attribute HABITAT is assumed to form a domain of its own so that in the denominator of (26) the sum is only over this attribute.

we assume that C_{arg} contains default inferences for the attribute TASTE with the values "sweet" and "non sweet," the attribute NUTRITION_VALUE with the values "healthy" and "non healthy," and the attribute SERVED_AT with values "birthday party" and "non birthday party."[14] Let the probabilities be TASTE = "sweet" : 0.95, TASTE = "not sweet" : 0.05, NUTRITION_VALUE = "healthy" : 0.05 and NUTRITION_VALUE = "non healthy" : 0.95, SERVED_AT = "birth party" : 0.98, and SERVED_AT = "not birthday party" : 0.02. The values for the attribute SERVED_AT reflect the fact that it is known that the context is a birthday party. For diagnosticity, the following assumption is made. Being a birthday, the attributes TASTE and SERVED_AT are more diagnostic (relevant) than the attribute NUTRITION_VALUE. Hence, the weight on the former two attributes is higher than on the latter one. Let us assume that it is 0.45 for the former two and 0.1 for the latter.[15]

C_{arg} before prenominal element:

$$\left[\begin{array}{cccc} \textit{feature} & \textit{diag.} & \textit{sim.} & \textit{diag.} * \textit{sim.} \\ \textit{sweet} & 0.45 & 0.95 & 0.4275 \\ \textit{not sweet} & 0.45 & 0.05 & 0.0225 \\ \textit{healthy} & 0.1 & 0.05 & 0.005 \\ \textit{not healthy} & 0.1 & 0.95 & 0.095 \\ \textit{served\_bp} & 0.45 & 0.98 & 0.441 \\ \textit{not served\_bp} & 0.45 & 0.02 & 0.009 \end{array}\right]$$

Encountering "sweet" confirms these expectations and raises the probability of TASTE = "sweet" to 1 because it is bottom-up information and it lowers the expectation for NUTRITION_VALUE = "healthy" to, say, 0.02.[16] Hence, one gets TASTE = "sweet" : 1, TASTE = "not sweet" : 0, NUTRITION_VALUE = "healthy" : 0.02, and NUTRITION_VALUE = "not healthy" : 0.98 in C_{arg}.

C_{arg} after prenominal element "sweet":

$$\left[\begin{array}{cccc} \textit{feature} & \textit{diag.} & \textit{sim.} & \textit{diag.} * \textit{sim.} \\ \textit{sweet} & 0.45 & 1 & 0.45 \\ \textit{not sweet} & 0.45 & 0 & 0 \\ \textit{healthy} & 0.1 & 0.02 & 0.002 \\ \textit{not healthy} & 0.1 & 0,98 & 0.098 \\ \textit{served\_bp} & 0.45 & 0.98 & 0.441 \\ \textit{not served\_bp} & 0.45 & 0.02 & 0.009 \end{array}\right]$$

If eventually "cake" is encountered, typicality is high because the probabilities (similarity values) in C_{arg} are of the same magnitude as those in C_{CW}. By contrast, if "veggies" is encountered instead, typicality is much lower because now the probabilities for all features go in the opposite direction. Whereas "sweet," ‘non healthy," and "birthday party," all have a high probability in C_{arg}, and the probabilities in $C_{veggies}$ are low.

$$C_{cake}: \left[\begin{array}{cccc} \textit{feature} & \textit{diag.} & \textit{sim.} & \textit{diag.} * \textit{sim.} \\ \textit{sweet} & 0.45 & 0.9 & 0.405 \\ \textit{not sweet} & 0.45 & 0.1 & 0.045 \\ \textit{healthy} & 0.1 & 0.2 & 0.02 \\ \textit{not healthy} & 0.1 & 0,8 & 0.08 \\ \textit{served\_bp} & 0.45 & 0.98 & 0.441 \\ \textit{not served\_bp} & 0.45 & 0.02 & 0.009 \end{array}\right]$$

$$C_{veggies}: \left[\begin{array}{cccc} \textit{feature} & \textit{diag.} & \textit{sim.} & \textit{diag.} * \textit{sim.} \\ \textit{sweet} & 0.45 & 0.2 & 0.09 \\ \textit{not sweet} & 0.45 & 0.8 & 0.36 \\ \textit{healthy} & 0.1 & 0.9 & 0.09 \\ \textit{not healthy} & 0.1 & 0,1 & 0.01 \\ \textit{served\_bp} & 0.45 & 0.02 & 0.009 \\ \textit{not served\_bp} & 0.45 & 0.98 & 0.441 \end{array}\right]$$

If instead of "sweet" "healthy" is encountered, this raises the probability of this feature, i.e., of NUTRITION_VALUE = "healthy," to 1 because it is bottom-up information and it lowers the probability for the feature "sweet" due to the correlation between the two features. Let us assume that the values for TASTE are updated to TASTE = "sweet" : 0.4 and TASTE = "not sweet" : 0.6.

C_{arg} after prenominal element "healthy":

$$\left[\begin{array}{cccc} \textit{feature} & \textit{diag.} & \textit{sim.} & \textit{diag.} * \textit{sim.} \\ \textit{sweet} & 0.45 & 0.4 & 0.18 \\ \textit{not sweet} & 0.45 & 0.6 & 0.27 \\ \textit{healthy} & 0.1 & 1 & 0.1 \\ \textit{not healthy} & 0.1 & 0 & 0 \\ \textit{served\_bp} & 0.45 & 0.95 & 0.4275 \\ \textit{not served\_bp} & 0.45 & 0.05 & 0.0225 \end{array}\right]$$

This has the effect that the typicality for "cake" is lowered (compared to encountering "sweet") and that the typicality of "veggies" is raised. However, one also has to consider diagnosticity. As already said above, being a birthday, the attributes TASTE and SERVED_AT are more diagnostic (relevant) than the attribute NUTRITION_VALUE. Hence, the weight on the former two attributes is higher than on the latter one.

Using (27)[17], one gets the following typicality values where C_x is C_{arg} after prenominal element x: $typicality(C_{sweet}, C_{cake}) = 0.937$, $typicality(C_{healthy}, C_{cake}) = 0.6815$, $typicality(C_{healthy}, C_{veggies}) = 0.4815$, and $typicality(C_{sweet}, C_{veggies}) = 0.12$. See **Figure 1** for an overview of the results for this example.

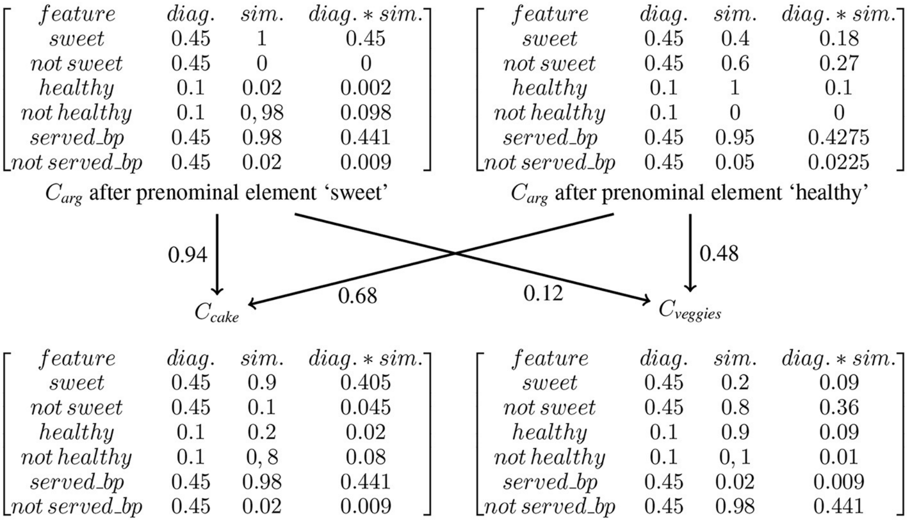
For the predictability component, one gets the following. For "sweet (and tasty) cake," all three default inferences are confirmed, while "sweet (and tasty) veggies" disconfirms all three inferences. The interesting cases are "healthy (and tasty) cake" and "healthy (and tasty) veggies." Since encountering the prenominal element "healthy" leads to a change for the feature "sweet," now "not sweet" is expected, "sweet (and tasty) cake" confirms only one default inference in C_{arg}, whereas "healthy (and tasty) veggies" confirms two. However, one has to bear in mind that we have restricted the examples to three attributes. If more attributes are considered, e.g., the way the food is prepared, which ingredients are used, etc, "healthy (and tasty) cakes" will confirm more inferences.

[14] For the sake of simplicity, we leave out the prenominal element "tasty."

[15] Diagnosticity is determined for the attributes of C_{arg} because it is the typicality of C_{CW} relative to C_{arg} that is computed. Below we include these diagnosticity values in the frames of C_{cake} and $C_{veggies}$ to ease understanding of how the typicality values are computed.

[16] Using standard Bayesian update, the new probability of a feature f given the feature f_{pe} associated with the prenominal element pe is given by $p(f \mid f_{pe})$. In the given case, one has f_{pe} = TASTE = "sweet" and an example of f is NUTRITION_VALUE = "healthy."

[17] We use (27) and not (28) because we do not compare HC and LC scenarios.



feature	diag.	sim.	diag. ∗ sim.
sweet	0.45	1	0.45
not sweet	0.45	0	0
healthy	0.1	0.02	0.002
not healthy	0.1	0, 98	0.098
served_bp	0.45	0.98	0.441
not served_bp	0.45	0.02	0.009

C_{arg} after prenominal element 'sweet'

feature	diag.	sim.	diag. ∗ sim.
sweet	0.45	0.4	0.18
not sweet	0.45	0.6	0.27
healthy	0.1	1	0.1
not healthy	0.1	0	0
served_bp	0.45	0.95	0.4275
not served_bp	0.45	0.05	0.0225

C_{arg} after prenominal element 'healthy'

feature	diag.	sim.	diag. ∗ sim.
sweet	0.45	0.9	0.405
not sweet	0.45	0.1	0.045
healthy	0.1	0.2	0.02
not healthy	0.1	0, 8	0.08
served_bp	0.45	0.98	0.441
not served_bp	0.45	0.02	0.009

feature	diag.	sim.	diag. ∗ sim.
sweet	0.45	0.2	0.09
not sweet	0.45	0.8	0.36
healthy	0.1	0.9	0.09
not healthy	0.1	0, 1	0.01
served_bp	0.45	0.02	0.009
not served_bp	0.45	0.98	0.441

FIGURE 1 | Overview of the diagnosticity, similarity, and typicality values for the four examples in (29). The arrows are labeled by the typicality values.

The last two examples show how the two components underlying N400 activity interact with each other in our approach. In particular, they show how the two components depend on world knowledge and how features that are given by prenominal elements before the CW is encountered modulate the N400 amplitude.

3.8. The N400 and Schema-Based Knowledge

Our approach assumes that the N400 amplitude is sensitive to both the number of disconfirmed features and the typicality of C_{CW} relative to C_{arg}. However, the examples below in (30) seem to be counterexamples to our approach.

(30) a. A huge blizzard swept through town last night. My kids ended up getting the day off from school. They spent the whole day outside building a big snowman / jacket / towel in the front yard.

b. The prescription for the mental disorder was written by the psychiatrist / schizophrenic / guard / pill / fence ....

Though both "jacket" and "towel" in (30-a) taken from Metusalem et al. (2012) share few features with the best completion "snowman" and both words are highly atypical given the partial event structure built up preceding the CW in the target sentence, "jacket" elicited a reduced N400 amplitude compared to "towel." However, this attenuation only occurred when the target sentence was embedded in the wider context given in (30-a) and not when it was used in isolation. A similar argument holds for (30-b) taken from Vega-Mendoza et al. (2021), which is a replication study of Paczynski and Kuperberg (2012). The authors found the following pattern of N400 amplitude per condition: plausible control (psychiatrist) < animate-related (schizophrenic) < animate-unrelated (guard) < inanimate-related (pill) < inanimate-unrelated (fence). Furthermore, this pattern followed the pattern of plausibility judgments with larger N400 found for increasingly implausible conditions.[18]

So far, we assumed that C_{arg} is related to one particular argument position, e.g., the theme of the event of sort "build" in (30-a) or the actor of the writing in (30-b). We take the examples in (30) as evidence that this need not always be the case. Rather, instead of a unique C_{arg} several such category concepts can be determined by the scenario that is described by the context. This raises the question of which arguments or objects can be targeted by category concepts that are related to other arguments or objects. Recall that C_{arg} contains pre-activated features. One kind of attributes is the properties of the expected object, e.g., whether it is sweet (TASTE), is healthy (NUTRITIONAL_VALUE), or is in water (LOCATION). The second kind of attribute relates the expected object to other objects. For example, the prescription can be related to its recipient (e.g., a schizophrenic) and the prescribed medicine (e.g., some kind of pills). Let us call such attributes *object-related*. Now the thesis is that C_{arg} can be related to an attribute that is object-related. On this generalized, view a C_{arg} is related to an extension of the information (frame) about an object that is going to be introduced or that has already been introduced into the scenario or the event structure. In the previously discussed examples,

[18] Two caveats are in order. First, this pattern was found only when participants performed a plausibility judgment task but not when they only passively read the sentences. In this case, the authors found that inanimate nouns elicited N400 effects compared to the control nouns and compared to animate nouns whereas the N400 amplitudes for the animate nouns did not differ from the control nouns. Second, in the original study Paczynski and Kuperberg (2012) the authors found (a) an interaction wherein related words elicited smaller N400 amplitudes than unrelated words when these words were animate, but not when they were inanimate and (b) animate-related words like "schizophrenic" did not elicit a reliable N400 effect compared to control words. For a discussion of the differences in the methodological design of the two studies, see Vega-Mendoza et al. (2021).

C_{arg} is related to the event denoted by the verb and the object itself has not yet been introduced. The problematic cases in (30) are instances in which C_{arg} is linked to an object that has already been introduced. In (30-a), there is a C_{arg} that is linked to the clothes of the children and in (30-b) there is a C_{arg} that is linked to the recipient of the prescription and another one that is linked to the medicine prescribed. If several C_{arg} can be activated, the question has to be answered how their typicality can be computed. Recall that each C_{arg} is related to a particular argument position, which, in turn, expresses a particular thematic role that links the event denoted by the verb to the object denoted by the argument. One possibility, therefore, is to make this dependency explicit by relating each pre-activated feature to a particular thematic role. Thus, if the feature π is an element of a C_{arg}, it is replaced by $tr \bullet \pi$ for tr the thematic role and $\bullet$ denoting the operation of chain concatenation. Hence, tr is an attribute. How is the diagnosticity of these attributes defined? We hypothesize that the diagnosticity is the expectation that CW provides information about this role. This expectation is highest for thematic roles that are related to undischarged arguments. For example, in (30-a) information about the theme is most expected whereas in (30-b) it is information about the actor. For the values of the thematic role attributes tr, diagnosticity is computed as defined above in section 3.6. The typicality value of a chain $tr \bullet \pi$ is computed by multiplying the diagnosticity of tr with the typicality of π. For the latter value, this means that selection restrictions imposed by the verb have to be taken into account, in particular, the animacy constraints. For example, for (30-b), this has the effect that features related to animate objects have higher typicality than features that are related to inanimate objects. Hence, the diagnosticity (expectancy) of tr and the animacy constraint interact with each other. For example, one has that a feature for an undischarged argument that is related to an object that satisfies the animacy constraint has higher typicality than a feature for a discharged argument that is related to an object that fails to satisfy the animacy constraint because in the latter case diagnosticity for tr is lower and the similarity value will be very low due to the violation. This is the case for "fence" in (30-b). For features for discharged arguments that satisfy the animacy constraint and that have both a high diagnosticity and a high similarity value, the overall typicality can be high even if the diagnosticity for tr is lower than in the case of an undischarged argument. This is the case for "schizophrenic" in (30-b). It is related to the recipient of the prescription (high diagnosticity) and due to the information that it is for a mental disorder this sort of recipient has a high similarity. The difference between "schizophrenic" and "guard" is that the latter has a low similarity value both for the actor role and for the recipient role in the frame related to the prescription.

4. OUTLOOK AND CONCLUSIONS

The theoretical account of a hybrid view of the N400 developed in this study has so far not been empirically tested. An important question, therefore, is to design experimental tests that provide evidence for or against it. The first, and important, strategy is related to the theoretical dimension. Given that we interpret N400 in terms of two operations, it must be possible to define a (monotone) function taking these two operations as arguments that correlates with the N400 amplitude (see Werning et al., 2019 for a similar strategy in a different theoretical setting). At the empirical dimension, two interesting strategies are the following. Our approach assumes that predictions are related to particular concepts or category concepts. As already mentioned above, Wang et al. (2020) have shown, using RSA in combination with EEG/MEG, that animacy features related to an argument position of a verb are pre-activated upon processing the verb before the argument is encountered. Such predictions should not be restricted to animacy features but should also include finer-grained categories that are related to other selection restrictions or the context. For example, upon encountering "cautioned" in (2-a) not only animacy features but also features that are related to the concept "danger" should be activated (see also Wang et al., 2020 for a similar argument in relation to other categories). A more general question is whether it is possible to detect differences between animacy, other selection restrictions, constraints imposed by the event structure, and constraints imposed by the scenario (script knowledge). A second empirical test is related to revisions that are triggered by mismatching prenominal elements. According to our approach, such revisions should not index the plausibility of the resulting event structure but a shift in the probability distribution of which kinds of objects are expected. This should result in a different set of features that are pre-activated. Hence, an interesting question is to combine the methods used in Wang et al. (2020) and Fleur et al. (2020). On a mismatching prenominal element, the activation pattern (measured using RSA with EEG/MEG) should change.

On a more theoretical side, one has that the definitions of similarity, diagnosticity, and typicality given here are only one option among others. What are alternatives and how can the choice between them be empirically tested? Furthermore, the approach must be extended to additional data. Of particular interest are data that seem to provide evidence that plausibility does play no role in the modulation of the N400 amplitude. An example is the results from Delogu et al. (2019).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This article grew out of work of the authors in the CRC 991 "The Structure of Representations in Language, Cognition, and Science," funded by the German Science Foundation (DFG).

ACKNOWLEDGMENTS

We would like to thank our editor and our reviewers for their detailed and critical comments and their bringing to our attention some recent relevant studies for our approach that led to a substantial revision of the original manuscript. A preliminary version of our approach was presented at the Rails conference in Saarbrücken in October 2019. We thank the audience, in particular, Harm Brouwer, for valuable comments, and criticism.

REFERENCES

Baggio, G., and Hagoort, P. (2011). The balance between memory and unification in semantics: a dynamic account of the N400. *Lang. Cogn. Process*. 26, 1338–1367. doi: 10.1080/01690965.2010.542671

Barsalou, L. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *J. Exp. Psychol. Learn. Mem. Cogn*. 11, 629–654. doi: 10.1037/0278-7393.11.1-4.629

Barsalou, L. W. (1983). Ad hoc categories. *Mem. Cogn*. 11, 211–227. doi: 10.3758/BF03196968

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, Vol. 1, Emory Symposia in Cognition*, ed U. Neisser (Cambridge: Cambridge University Press), 101–140.

Barsalou, L. W. (1992). "Frames, concepts, and conceptual fields," in *Frames, Fields and Contrasts. New Essays in Semantic and Lexical Organization*, eds A. Lehrer and E. F. Kittay (Hillesdale, NJ: Lawrence Erlbaum Associates, Inc.), 21–74.

Boudewyn, M. A., Long, D. L., and Swaab, T. Y. (2015). Graded expectations: predictive processing and the adjustment of expectations during spoken language comprehension. *Cogn. Affect. Behav. Neurosci*. 15, 607–624. doi: 10.3758/s13415-015-0340-0

Brothers, T., Swaab, T. Y., and Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: prediction takes precedence. *Cognition* 136, 135–149. doi: 10.1016/j.cognition.2014.10.017

Chwilla, D. J., Kolk, H. H. J., and Vissers, C. T. W. M. (2007). Immediate integration of novel meanings : N400 support for an embodied view of language comprehension. *Brain Res*. 1183, 109–123. doi: 10.1016/j.brainres.2007.09.014

Cohen, B., and Murphy, G. L. (1984). Models of concepts*. *Cogn. Sci*. 8, 27–58. doi: 10.1207/s15516709cog0801_2

Delogu, F., Brouwer, H., and Crocker, M. W. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain Cogn*. 135:103569. doi: 10.1016/j.bandc.2019.05.007

DeLong, K. A., Urbach, T., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci*. 8, 1117–1121. doi: 10.1038/nn1504

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

Fleur, D., Flecken, M., Rommers, J., and Nieuwland, M. (2020). Definitely saw it coming? the dual nature of the pre-nominal prediction effect. *Cognition* 204:104335. doi: 10.1016/j.cognition.2020.104335

Hagoort, P., and Brown, C. (1994). "Brain responses to lexical-ambiguity resolution and parsing," in *Perspectives on Sentence Processing*, eds C. Clifton Jr., L. Frazier, K. and Rayner (Hillesdale, NJ: Earlbaum), 45–80.

Hagoort, P., Hald, L., Bastiaansen, M., and Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science* 304, 438–441. doi: 10.1126/science.1095455

Holland, J. H., Holyoak, K. J., Nisbett, R. E., and Thagard, P. R. (1986). *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA: MIT Press.

Huth, A., de Heer, W., Griffiths, T., Theunissen, F., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532, 453–458. doi: 10.1038/nature17637

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci*. 2:4. doi: 10.3389/neuro.06.004.2008

Kuperberg, G. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Lang. Cogn. Neurosci*. 31, 602–616. doi: 10.1080/23273798.2015.1130233

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

Kuperberg, G. R., Brothers, T., and Wlotko, E. (2020). A tale of two positivities and the N400: distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *J. Cogn. Neurosci*. 32, 12–35. doi: 10.1162/jocn_a_01465

Kutas, M., and Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* 207, 203–205. doi: 10.1126/science.7350657

Kutas, M., and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307, 161–163. doi: 10.1038/307161a0

Landauer, T. K., McNamara, D. S., Dennis, S., and Kintsch, W. (eds.). (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lau, E., Phillips, C., and Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nat. Rev. Neurosci*. 9, 920–933. doi: 10.1038/nrn2532

Lau, E. F., Namyst, A., Fogel, A., and Delgado, T. (2016). A direct comparison of N400 effects of predictability and incongruity in adjective-noun combination. *Collabra Psychol*. 2, 13. doi: 10.1525/collabra.40

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., and Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *J. Mem. Lang*. 66, 545–567. doi: 10.1016/j.jml.2012.01.001

Naumann, R., and Petersen, W. (2019). "Bridging inferences in a dynamic frame theory," in *Language, Logic, and Computation*, eds A. Silva, S. Staton, P. Sutton, and C. Umbach (Berlin; Heidelberg: Springer), 228–252.

Naumann, R., and Petersen, W. (2021). "Bridging the gap between formal semantics and neurolinguistics: the case of the N400 and the LPP," in *Proceedings Tbilisi Conference* (Berlin; Heidelberg).

Naumann, R., Petersen, W., and Gamerschlag, T. (2018). "Underspecified changes: a dynamic, probabilistic frame theory for verbs," in *Proceedings of Sinn und Bedeutung 22, Vol. 2, ZASPiL 61*, eds U. Sauerland and S. Solt (Berlin: Leibniz-Centre General Linguistics), 181–198.

Nieuwland, M., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., et al. (2019). Dissociable effects of prediction and integration during language comprehension: evidence from a large-scale study using brain potentials. *Philos. Trans. R. Soc. B Biol. Sci*. 375:20180522. doi: 10.1098/rstb.2018.0522

Nieuwland, M., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., et al. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *Elife* 7:e33468. doi: 10.7554/eLife.33468.024

Nieuwland, M. S., and van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *J. Cogn. Neurosci*. 18, 1098–1111. doi: 10.1162/jocn.2006.18.7.1098

Paczynski, M., and Kuperberg, G. (2012). Multiple influences of semantic memory on sentence processing: distinct effects of semantic relatedness on violations of real-world event/state knowledge and animacy selection restrictions. *J. Mem. Lang*. 67, 426–448. doi: 10.1016/j.jml.2012.07.003

Paczynski, M., and Kuperberg, G. R. (2011). Electrophysiological evidence for use of the animacy hierarchy, but not thematic role assignment, during verb argument processing. *Lang. Cogn. Process* 26, 1402–1456. doi: 10.1080/01690965.2011.580143

Quante, L., Bölte, J., and Zwitserlood, P. (2018). Dissociating predictability, plausibility and possibility of sentence continuations in reading: evidence from late-positivity ERPs. *PeerJ* 6:e5717. doi: 10.7717/peerj.5717

Rabovsky, M., and McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition* 132, 68–89. doi: 10.1016/j.cognition.2014.03.010

Rosch, E., and Mervis, C. B. (1975). Family resemblances: studies in the internal structure of categories. *Cogn. Psychol.* 7, 573–605. doi: 10.1016/0010-0285(75)90024-9

Roth, E. M., and Shoben, E. J. (1983). The effect of context on the structure of categories. *Cogn. Psychol.* 15, 346–378. doi: 10.1016/0010-0285(83)90012-9

Schank, R., and Abelson, R. (1977). *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Earlbaum Assoc.

Schurz, G. (2012). "Prototypes and their composition from an evolutionary point of view," in *The Oxford Handbook of Compositionality*, eds W. Hinzen, F. Machery, and M. Werning (Oxford: Oxford University Press), 530–554.

Schuster, A. (2016). *Prototype frames: Theories of concepts and their empirical evidence* (Master's thesis), University of Düsseldorf.

Smith, E. E., Osherson, D. N., Rips, L. J., and Keane, M. (1988). Combining prototypes. *Cogn. Sci.* 12, 485–527. doi: 10.1207/s15516709cog1204_1

Szewczyk, J. M., and Schriefers, H. (2011). Is animacy special?: ERP correlates of semantic violations and animacy violations in sentence processing. *Brain Res.* 1368, 208–221. doi: 10.1016/j.brainres.2010.10.070

Szewczyk, J. M., and Schriefers, H. (2018). The N400 as an index of lexical preactivation and its implications for prediction in language comprehension. *Lang. Cogn. Neurosci.* 33, 665–686. doi: 10.1080/23273798.2017.1401101

Taylor, W. L. (1953). 'cloze procedure': a new tool for measuring readability. *J. Q.* 30, 415–433. doi: 10.1177/107769905303000401

van Berkum, J. J. A., van den Brink, D., Tesink, C. M. J. Y., Kos, M., and Hagoort, P. (2008). The neural integration of speaker and message. *J. Cogn. Neurosci.* 20, 580–591. doi: 10.1162/jocn.2008.20054

van Berkum, J. J. A. V., Hagoort, P., and Brown, C. M. (1999). Semantic integration in sentences and discourse: evidence from the N400. *J. Cogn. Neurosci.* 11, 657–671. doi: 10.1162/089892999563724

Van Petten, C., and Luka, B. J. (2012). Prediction during language comprehension: benefits, costs, and ERP components. *Int. J. Psychophysiol.* 83, 176–190. doi: 10.1016/j.ijpsycho.2011.09.015

Vega-Mendoza, M., Pickering, M., and Nieuwland, M. (2021). Concurrent use of animacy and event-knowledge during comprehension: evidence from event-related potentials. *Neuropsychologia* 152:107724. doi: 10.1016/j.neuropsychologia.2020.107724

Wang, L., Wlotko, E., Alexander, E., Schoot, L., Kim, M., Warnke, L., and Kuperberg, G. (2020). Neural evidence for the prediction of animacy features during language comprehension: evidence from MEG and EEG representational similarity analysis. *J. Neurosci.* 40, 3278–3291. doi: 10.1523/JNEUROSCI.1733-19.2020

Werning, M., Unterhuber, M., and Wiedemann, G. (2019). Bayesian pragmatics provides the best quantitative model of context effects on word meaning in EEG and cloze data. *Proc. Cogn. Sci.* 41, 3085–3091. Available online at: https://cogsci.mindmodeling.org/2019/papers/0517/0517.pdf

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., and van den Bosch, A. (2016). Prediction during natural language comprehension. *Cereb. Cortex* 26, 2506–2516. doi: 10.1093/cercor/bhv075

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 Communication

ORIGINAL RESEARCH
published: 14 September 2021
doi: 10.3389/fcomm.2021.661862

The Evolutionary Dynamics of Negative Existentials in Indo-European

Shahar Shirtz[1]*, *Luigi Talamo*[2] *and Annemarie Verkerk*[2]

[1]Independent Researcher, New Jersey, NJ, United States, [2]Language Science and Technology, Saarland University, Saarbrücken, Germany

Where in earlier work diachronic change is used to explain away exceptions to typologies, linguistic typologists have started to make use of explicit diachronic models as explanations for typological distributions. A topic that lends itself for this approach especially well is that of negation. In this article, we assess the explanatory value of a specific hypothesis, the Negative Existential Cycle (NEC), on the distribution of negative existential strategies ("types") in 106 Indo-European languages. We use Bayesian phylogenetic comparative methods to infer posterior distributions of transition rates and parameters, thus applying rational methods to construct and evaluate a set of different models under which the attested typological distribution could have evolved. We find that the frequency of diachronic processes that affect negative existentials outside of the NEC cannot be ignored—the unidirectional NEC alone cannot explain the evolution of negative existential strategies in our sample. We show that non-unidirectional evolutionary models, especially those that allow for different and multiple transitions between strategies, provide better fit. In addition, the phylogenetic modeling is impacted by the expected skewed distribution of negative existential strategies in our sample, pointing out the need for densely sampled and family-based typological research.

Keywords: linguistic typology, negation, existential predication, diachronic typology, phylogenetic comparative methods

Edited by:
Gerhard Jäger,
University of Tübingen, Germany

Reviewed by:
Taraka Rama Kasicheyanula,
University of North Texas,
United States
Gerd Carling,
Lund University, Sweden

****Correspondence:***
Shahar Shirtz
shahar.shirtz@gmail.com

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

Received: *31 January 2021*
Accepted: *17 June 2021*
Published: *14 September 2021*

Citation:
Shirtz S, Talamo L and Verkerk A (2021) The Evolutionary Dynamics of Negative Existentials in Indo-European. Front. Commun. 6:661862. doi: 10.3389/fcomm.2021.661862

1 INTRODUCTION

The negative existential domain, the expression of negated existential statements, may appear to be a simple, unremarkable, area of grammar. In many languages, it simply involves the deployment of the usual means used to express the affirmative existential with the standard verbal negation marker. This leads to clauses such as (1), from Swedish (Indo-European, Germanic), where the structural coding means involved in expressing existence are deployed alongside the Swedish standard verbal negation marker *inte*.

Swedish (Germanic; Veselinova 2013: 115)

(1) *Det finns inte ost i kylskap-et*
3SG.NTR be.at NEG cheese in fridge-DEF
'There isn't any cheese in the fridge'

A thorough look at the negative existential domain in the languages of the world, however, suggests that it is expressed by a variety of construction types. In many languages, one finds a dedicated negative existential marker used as a negative existential copula. This marker is not used to negate verbal predicates, but may be used to negate other domains of nominal predication such as predicate location. This is illustrated in (2) from Turkish, where the negative existential marker *yok* is deployed, but not the Turkish standard verbal negation marker, a suffix. In another construction type, the standard negation marker is used as the only marker of negative existence, without the existential marker used in affirmative existential clauses. This is illustrated by the Tongan examples in (3a-b). The negation marker *'ikai* is used in (3a) to negate the main verbal predicate, and in (3b) as the sole marker of negative existence, without another existential marker.

Turkish (Turkic; own knowledge)

(2)	*bahçe-de*	*kedi*	*yok*
	garden-LOC	cat	NEG.EX
	'there are no cats in the garden'		

Tongan (Austronesian, Polynesian, Veselinova 2014: 1342)

(3a)	*Na'e*	*'ikai*	*ke*	*kata*	*'a*	*Pita*
	PST	NEG	SUB	laugh	ABS	Pita
	'Pita did not laugh (lit. It was not that Pita laugh[ed])					

(3b)	*'oku*	*'ikai*	*ha*	*me'a*
	PRS	NEG	NSP	thing
	'There is not a thing'			

This cross-linguistic variation led Croft (1991) to propose a typology for the negative existential domain with synchronic and diachronic components (fully explained in the following section). In the synchronic portion, Croft identifies the construction types illustrated by (1–3) above as well as three intermediate construction types. The six types are defined based on a comparison of the negation marker(s) used to negate verbal predicates and the expression of negation in negative existential clauses: are they identical? Distinct? Is the verbal negation marker used as the negative existential copula, or do we find some intermediate situation? Croft's synchronic typology has been successful as a cross-linguistic taxonomy of negative-existential constructions and fits well with other variables in the typology of negation (e.g., articles in Veselinova and Hamari, Forthcoming).

The dynamic component of Croft's typology, the Negative Existential Cycle (NEC), connects these six construction types in a cycle where each construction is the source for another. Elaborating on the NEC, Veselinova (2013, 2014, 2016; see also Verkerk and Shirtz, forthcoming) showed that while the diachronic transitions proposed by Croft are indeed attested, other transitions are also involved in the rise of innovative negative existential constructions or in changes to the typological classification of old constructions. It is unclear, however, how widespread these non-NEC transitions are, and how much of the attested cross-linguistic variation in the negative existential domain arises out of, or can be explained by, the transitions in Croft's NEC.

This article directly tackles this by asking: how likely is it that the attested variation in the negative existential domain resulted from the transitions that compose the NEC? To do so, we use Bayesian phylogenetic comparative methods to analyze data from 106 Indo-European languages, collected by consulting grammars and published texts, as well as questionnaires filled by language experts. This article tests the likelihood that the NEC is the main set of transitions behind the cross-linguistic variation in the Indo-European negative existential domain, and compares it to the likelihood of other potential sets of transitions. We use the results of this modeling to illustrate our answer to another question: what is the relationship between rational quantitative and statistical modeling approaches and "traditional," analytic approaches in studies of morphosyntactic change and diachronic typology? In a way, we additionally explore the feasibility of phylogenetic diachronic typology; how many languages, or how big of a language family does one need to investigate a complex typological hypothesis? Of course, our answer to these questions is limited to the Indo-European family and to our current sample, as **Section 2** elaborates.

The rest of this section further defines the negative existential domain, describes Croft's typology and NEC in more detail, and sketches some of the major issues with the NEC. In **Section 2** we turn to describe the data and the methods used in this study, and turn to present some of our findings in **Section 3**. There, we give an overview of the negative existential domain in the different subfamilies of Indo-European, and illustrate two types of transitions that are not included in the NEC. Our phylogenetic modeling of the Indo-European negative existential domain is presented in **Section 4**, and our findings are summarized and discussed in **Section 5**.

1.1 Negative Existentials: Definitions and Synchronic Typology

A cross-linguistic study of the negative existential domain requires that it be defined without reference to any language-specific property, i.e., as a comparative concept (Haspelmath 2010, 2016; Croft 2016). Furthermore, the different types of negative existential constructions need to be defined as what Croft (2016) calls "hybrid comparative concept": a combination of functional and formal properties defined without reference to any language-specific properties so they are identifiable in different languages. Following Croft (1991) and Veselinova (2013, 2014), we view existential constructions as expressing the existence or presence of a particular figure constituent relative to a specific location (the ground) or generally "in the world." Its negated counterpart expresses the fact that a particular figure constituent does not exist or is not present generally "in the world" or relative to some ground location. As this definition does not refer to any language-specific grammatical devices, it can be and has been successfully deployed cross-linguistically, and qualifies as a functional comparative concept.

The definition adopted here is largely compatible with the approach of Creissels (2013, 2019; see also Clark 1978) who views what is usually referred to by the term existential predication as an inverse-locative predication[1]. In this type of predication, the important information is the presence of the figure against some locative ground, rather than the location of the figure, and thus it is the inverse of clauses expressing a predicate location. Creissels' (2019) definition of inverse-locative predication, however, focuses on constructions that are not formally related to constructions expressing locative predication. Instances where the difference between clauses expressing locative and inverse-locative semantics has to do with information packaging devices such as word order or topic/focus markers (as in some Indo-Iranian languages; Shirtz 2019), are not included in Creissels' inverse-locative.

This article's focus is on clauses expressing the negative existential domain regardless of their structural similarity to clauses expressing locative predication. We do include here instances where the main distinction between predicate location and existential constructions has to do with the relative order of the figure and the ground or with other information-packaging devices such as articles or so-called topic/focus markers. In this sense, our approach to the negative existential domain, as well as the approach of Croft (1991) and Veselinova (2014), is compatible with Creissels' criteria for inverse-locative predication, except for his requirement that it be structurally unrelated to locative predication.

Croft's (1991) typology of the negative existential domain is composed of six language types. Their identification rests on comparing the negation marker in clauses expressing the negative existential domain to the negation markers used in standard verbal negation, and marginally on the existence of other negative existential constructions in the same language. The constructions on which the typology rests, then, are bundles of functional and abstract formal properties and as instances of Croft's (2016) "hybrid comparative concepts" are cross-linguistically identifiable. We classify our languages in terms of Croft's (1991) typology, but instead of classifying in terms of language types, we use construction types (as in Veselinova's approach and in Verkerk and Shirtz, forthcoming). This implies that languages may have more than one type of negative existential construction, and that attested constructions may undergo change that is in part independent of other constructions that may be attested in that language.

The six construction types in Croft's typology are divided into three major types, called Type-A, Type-B, and Type-C, and three transitional types, called Type-A~B, Type-B~C, and Type-C~A[2]. In Type-A, the same marker used in standard verbal negation accompanies the affirmative existential marker in clauses expressing negative existence. This was illustrated by the Swedish clause in (1) above, where the standard verbal negation marker *inte* is used to negate the verbal locative copula *finns* "be.at." In Type-B, a special marker distinct from the standard verbal negation marker is used in negative existential clauses. This was illustrated by the Turkish clause in (2) above, where the negative existential marker *yok* is used.

In the intermediate Type-AB one finds instances of Type-A and instances of Type-B that may be diachronically related, each in its own functional niche. This situation is very common in Iranian languages, where an innovative negative copula often emerges from a reduction of the verbal negation marker and the present-tense copula or some other copular element, thus leading to the innovation of a Type-B construction. But a similar reduction does not occur with the past-tense copula, thus a conservative Type-A construction is retained. This is illustrated by (4a-b), from Sivandi. In (4a) the past tense copula, also used in affirmative existential clauses, is negated by the Sivandi Standard Verbal Negation marker *na*=. The negative existential in (4b) is expressed by *nūnd*, a negative copula that resulted from the reduction of the standard verbal negation marker *na*= with some other element.

Sivandi (Iranian, Lecoq, 1979: 89, 150)

(4a) *albatta barqa=m na=bi*
evidently electricity=TOP NEG=be.PST.3SG
'(someone lit a candle), evidently there wasn't any electricity'

(4b) *vāllāh, me či tū das=em nūnd*
by.God 1SG what in hand=1SG NEG.COP
'By God, there's nothing in my hand'

In Type-C, the standard verbal negation marker is used as a negative existential marker, without an affirmative existential marker. This was illustrated above by the Tongan examples in (3a-b), where the standard verbal negation marker *'ikai* is also used as the negative existential marker in (3b). In the intermediate Type-BC a special negative existential marker is also used as a verbal negation marker under some circumstances, but other verbal negation markers also exist. That is, the domain of verbal negation includes several markers, one of which also functions as the negative existential marker. This is illustrated by (5a-c), from Darai (Indo-Aryan). The clauses in (5a-b) illustrate two Darai verbal negation constructions: the *nai*- prefix in (5a) and in (5b) the particle *nidzə*. This particle is also used in (5c) as the negative existential marker, without the Darai affirmative existential copulas. Thus, in Darai a special negative existential marker is also used as one of the verbal negation markers.

Darai (Indo-Aryan; Dhakal 2012: 134, 134, 137)

(5a) *nai-dza-m g^{h}ərə*
NEG-go-1SG house
'I shall not go home'

(5b) *u b^{h}ot^{h}i nidzə mor-lə*
DEM bhothi.fish NEG die-PST.3SG
'the bhothi fish didn't die'

(5c) *tərə hame-rə səskriti-jə bãsi pəhile nidzə*
but 1PL-GEN culture-LOC flute early NEG.EX
'but there was no flute in our culture then'

[1] See https://dlc.hypotheses.org/2516 for a recent blog post by Martin Haspelmath on the nature of existentials.

[2] To save space, we use AB, BC, and CA to designate Croft's (1991) transitional types.

The sixth type in Croft's typology, Type-CA, includes a negative existential marker which also functions as the verbal negation marker, but is optionally used alongside an affirmative existential marker. This is illustrated here by (6a-b), from Marathi, where the verbal negation marker *nāhi* in (6a) is also used as the negative existential marker in (6b), optionally co-occurring with the existential marker *āhe*.

Marathi (Indo-Aryan, Croft 1991: 12; his glosses and translation)

(6a) *koɳi tithə dzāt* [*əts*] *nāhi*
anyone there goes [EMPH] NEG
'Nobody goes there'

(6b) *tithə koɳi nāhi* [*āhe*]
there anyone NEG [EX]
'There isn't anyone there'

While using Croft's (1991) typology to classify negative existential constructions is not always straightforward, it often is so, and it is an illuminating taxonomy for the IE data analyzed here. What this paper sets out to test, then, is how well does the diachronic component of Croft's typology, the negative existential cycle (NEC), explain the attested variation in negative existential construction types across the family.

1.2 Negative Existentials: Dynamic Typology

The diachronic component of Croft's typology arranges the six construction types in a unidirectional cycle such that each type is the source of one other type of negative existential. The cycle is presented in (7), arbitrarily beginning with Type-A, and cycling through the different types until we return to Type-A.

(7) Type-A > Type-AB > Type-B > Type-BC > Type-C >
Type-CA > Type-A

Croft's dynamization of his typology is appealing. It is simple and unidirectional, and each transition on the cycle is described and illustrated by Croft as an instance of an internal mechanism of morphosyntactic change: reanalysis + actualization[3] or extension. The emergence of Type-B negative existential markers, for example, often involves a reanalysis of the relationship between a negation marker and a copula as a single unit, actualized by a phonological reduction of the two or changes in the distribution of the negated copula. Whether the reanalysis occurs with all copular forms or only with some, it is clear that Type-AB will likely occur at some point in such transitions from Type-A to Type-B, leading to a Type-A > Type-AB > Type-B pathway.

Croft illustrates the transition from Type-B to Type-C with two processes. First, an extension of the negative existential marker into the domain of verbal negation results in competition between the negative existential marker and the standard verbal negation marker. This can occur, for example, when a new main clause construction emerges from the reanalysis of a nonfinite form of the verb and an existential marker. When the two forms are in competition, or when each form is used in its own functional niche, the system is an instance of Croft's Type-BC and if the negative existential marker overtakes the entire domain of verbal negation, we arrive at Type-C. The second type of process involves an extension of the negative existential marker to reinforce the standard verbal negation marker under some conditions. As predicted in Jespersen's cycle (Jespersen 1917; van der Auwera 2009, see also van Gelderen forthcoming), where novel negation markers arise out of older negation-reinforcing elements which end up replacing the older markers, the negative existential marker may be reanalyzed as the main verbal negation marker. At first, the complete replacement will occur only in certain situations, and the system will be an instance of Type-BC, but after the complete loss of the old verbal negation marker, the system will be best classified as Type-C. The negative existential marker of Type-C, then, may optionally combine with the affirmative existential marker, often for information management purposes, innovating a Type-CA construction. When the combination of the old negative existential marker and the affirmative existential marker becomes obligatory, we arrive back at Type-A.

1.3 Issues With Croft's Cycle

Croft's proposal, then, includes both a synchronic typology of six types and a set of diachronic transitions between them that results in a cycle. Veselinova (2013, 2014, 2016 see also Verkerk and Shirtz, forthcoming) further explored the different negative existential types in Slavic and Oceanic languages, and the different transitions between these types as proposed by Croft. Doing so, she identified several issues with the dynamic portion of Croft's proposal.

First, Veselinova notes that languages may have two (or more) negative existential constructions of different types. She illustrates this with the co-existence of Type-B and Type-C in Tahitian and the coexistence of Type-B and Type-BC in Kapingamarangi. Verkerk and Shirtz (forthcoming) further illustrate these patterns with the data from the Eastern Indo-Aryan languages Kupia and Standard Oriya. They also note that pairs of negative existential types may differ in whether their coexistence is at all possible given the way Croft's typology is set up. There is nothing prohibiting the coexistence of multiple Type-A or multiple Type-B negative existential constructions, but multiple Type-C constructions cannot coexist. In such a situation, by definition, the deployment of each particular negation marker will be limited in some way (as there are two verbal negation markers), and hence the two constructions are an instance of multiple Type-BC constructions. This entails that in situations where a Type-B and a Type-C negative existential constructions are found in the same language, predictable changes to one construction, Type-B > Type-BC, entail changes to the classification of the other against the NEC direction (Type-C > Type-BC).

Veselinova (2014, 2016) also identifies transitions that are not represented on Croft's cycle. These include a transition from Type-AB directly to Type-BC (without an intermediate Type-B), potentially documented in Russian and Hawai'ian, and a

[3] The term "actualization" for the morphosyntactic changes that follow reanalysis is due to Timberlake (1977); See also Harris and Campbell (1995) and de Smet (2012) for a discussion of the relationship between reanalysis and actualization.

transition from Type-B directly to Type-C (without an intermediate Type-BC). The directionality of these changes is the same overall direction of the NEC, but a stage is "skipped." These transitions, together with the Type-C > Type-BC proposed by Verkerk and Shirtz (forthcoming), form a set of transitions that are outside of the set of transitions proposed by Croft. This suggests that there is more to the diachronic changes that negative existential constructions undergo than the NEC. The diachronic processes involved are summarized by Veselinova (2016: 155) as follows: "They include (i) subordination processes; (ii) the reanalysis of an external negator into a negator external to the proposition; (iii) a direct inheritance of a construction; (iv) the use of negative existentials with nominalized verb forms." van Gelderen (forthcoming) discusses the NEC in relation to two other negative cycles, the Jespersen Cycle (see the previous section) and the Givón Cycle[4] (Jespersen 1917, Givón 1978; see also van Gelderen forthcoming, van der Auwera et al., forthcoming) as well as the Copula Cycle[5] (Katz 1996), demonstrating how other diachronic processes impact the NEC.

The attested transitions that are not a part of the NEC, by virtue of being in the opposite direction to the NEC or by virtue of "skipping" an NEC stage or two, seem to depend on very specific configurations in the grammar of individual languages, such as the coexistence of Type-B and Type-C in a single language, and these configurations may be cross-linguistically rare. As a result, it seems that transitions that are not a part of the NEC are infrequent. But without a wider cross-linguistic survey, we simply do not know how rare these situations are, and it could very well be that their rarity is a result of diachronic instability.

2 MATERIALS AND METHODS

The previous section suggests that while the transitions in the NEC seem to be common, other transitions are attested as well, including transitions in the same general direction of the NEC and transitions in the opposite direction. The question, then, is how much of the attested variation in the negative existential domain does the NEC explain? This article answers this question for the Indo-European language family. This section describes the data, the theoretical assumptions, and the methods used in this article.

2.1 Data

The data used in this article is the classification of negative existential constructions in 106 Indo-European languages. The Indo-European language family is well suited for a study such as the one done here. First, it is a large family that includes many subfamilies with a wide geographic dispersal and deep historical records. While not a requirement, this dispersal raises the likelihood for variation in the typological classification of negative existential constructions. This variation is required for a meaningful quantitative testing of the NEC. In families with little to no variation to explain, it will be difficult to reject transition sets based on their low explanatory power. Further, the documentation of Indo-European is quite extensive, and while there are still several lacunae in the documentation of the family (e.g., the Indo-European languages of Pakistan and the Pamir region), many branches and sub-branches are well documented.

The data collection for this article relied on three types of data sources. First, as with many typological surveys, we made extensive use of published grammatical descriptions. The domain of existential clauses, affirmative or negative, however, is often not directly mentioned in such sources. This may be because of their marginal nature or low frequency, or because they sometimes do not have any unique or unpredictable grammatical properties (e.g., in compositional Type-A constructions). To have as wide a coverage as possible, then, we also used translation questionnaires and analyzed published textual data. The questionnaire is based on the one used by Veselinova (2014; see **Supplementary Information S1**) and was filled by language experts. It includes questions about affirmative and negative verbal clauses, affirmative and negative existential clauses, and other types of nonverbal predication. We also made use of published textual data (not necessarily computer-readable corpora) which accompanied documentation projects. This was required when the reference or sketch grammars did not include an explicit discussion or illustration of existential and negative existential predication, but were clear and enabled us to go through textual data published as a part of a documentation project.

2.2 Typology—Historical Morphosyntax—Phylogenetic Modeling?

This article is situated at the intersection of linguistic typology, historical morphosyntax, and phylogenetic modeling. This requires the article to adhere to the main assumptions of each of the three fields. As this is a typological study, we approach the negative existential domain here as a functional comparative concept, and define it without reference to any language-specific construction. Further, Croft's definition of the six negative existential types are instances of hybrid comparative concepts, based on both form and function (Croft 2016, see **Section 1.1**).

More controversial is the relationship between the more traditional, analytic approaches to historical morphosyntax and newly adopted statistical, phylogenetic approaches. These two approaches are often viewed as competing, or even contradictory in their assumptions (see, for example, replies to Dunn et al., 2011 in *Linguistic Typology 15.2*; especially Dryer 2011 and Plank 2011). We argue, however, that when it comes to the study of morphosyntactic change, the two do not contradict each other, but rather highlight different aspects of language change. As such, they are best viewed as complementing each other by answering slightly different questions so that each of them may inform the hypotheses and the work done in the

[4] Givón's Cycle (Givón 1978; see also van Gelderen forthcoming) is a diachronic hypothesis on the origin of negators, stating that these most commonly derive from negative verbs with meanings such as *fail*, *lack*, and *deny*.

[5] The Copula Cycle (Katz 1996; Lohndal 2009: 239) described how copulas emerge from demonstratives or pronouns, and change to grammatical markers, such as special negative existential markers.

other (see again *Linguistic Typology 15.2* for Levinson et al., 2011 and Croft et al., 2011; as well as Levinson and Gray 2012 and Dunn 2014: Sect. 5.2). We use the domain of the negative existential in Indo-European to illustrate this point.

The field of historical morphosyntax in general, and morphosyntactic reconstruction in particular, has been rife with controversy about the very plausibility of its goals. The identification of the mechanisms of morphosyntactic change (Harris and Campbell 1995) and the introduction of a constructional interpretation for morphosyntactic change (Barðdal and Eythórsson 2012; Barðdal 2013; Barðdal and Gildea 2015), however, enable an explicit statement of the methods used and assumptions required in the analysis of morphosyntactic change. These assumptions include the identification of cognate constructions using a set of principles that are parallel to those used in the identification of lexical cognates, and the identification of the plausible mechanism of change involved (see Gildea et al., 2020 for a detailed survey and discussion). These principles are also applied in diachronic typology (Bybee 1988; Bybee et al., 1994; Hendery 2012; Sansò 2017).

The goal of phylogenetic comparative modeling of the type pursued here, focusing on change in a typological variable, involves the estimation of the likelihood of a set of transitions, or historical changes between construction types, in a set of observed data (Pagel 1999). This estimation depends on the topology of the family tree, which should be arrived at independently (e.g., using the Comparative Method), and the length of its branches, estimating time elapsed since the diversification of two languages (Pagel 1999: 878, Dunn 2014: Sect. 5.2). The cognate status of the attested constructions, as well as the specific mechanisms involved in the rise of each of these constructions, matters less for such phylogenetic modeling. That is, the modeling pursued here treats transitions between construction types with no regard to the actual process of change “on the ground.” Several different processes can often lead from one construction type to another, but which of these actually occurred is not a part of the model.

The fact that analytic, “traditional” methods in historical morphosyntax and phylogenetic comparative modeling of morphosyntactic change highlight different aspects of the data may be taken to suggest that these are competing methods. We however believe the exact opposite: the fact that different aspects of the historical record are highlighted by these methods allows them to complete and inform each other. Croft’s NEC and Veselinova’s critique of the NEC were arrived at using analytic morpho-syntactic methods. Both were arrived at without taking into account a specific family tree topology (although family relations must have been taken into account implicitly), and without testing the NEC against other plausible pathways using quantitative tools. Both of these obviously motivate and inform the current study. Testing how much of the cross-linguistic variation the NEC can explain will either fortify it as the main set of diachronic transitions in the negative existential domain, or propose other (sets of) transitions active in this domain that can then be further explored by a more direct analysis of language data.

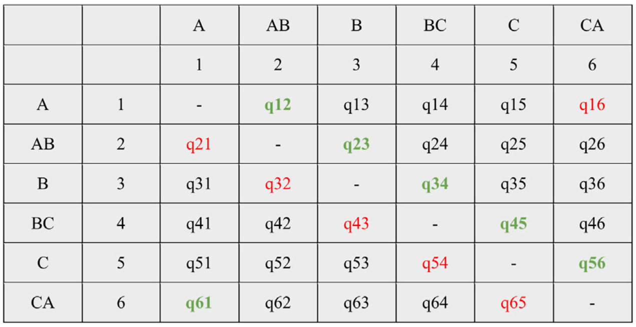
2.3 Phylogenetic Comparative Methods

We model the type of negative existential strategy that each language in our sample has in terms of an explicit phylogenetic process, i.e., as the outcome of evolutionary processes that take place on the branches of a phylogenetic tree. The phylogenetic tree set is given, we are not inferring phylogenies, but rather using them to do quantitative diachronic typology: testing an influential hypothesis using phylogenetic methods. Phylogenetic comparative models have been used to estimate what typological strategy the ancestors of sampled languages must have had (Maurits and Griffiths 2014), and the rates of evolutionary change (Cathcart et al., 2020). We will focus on which transition parameters are most relevant for explaining the distribution of strategies attested in our sample (see also Dunn et al., 2017). Here, we test whether the transition parameters associated with the NEC are essential for explaining the diachrony of the distribution of negative existential strategies in the current sample of 106 Indo-European languages.

Doing this requires three components: data on negative existential strategies, a tree sample of phylogenies of the languages under investigation, and a set of models, grounded in a particular way of thinking about evolutionary change. This section describes the sample of phylogenetic trees we use, and sketches the relevant model of evolutionary change. Our dataset is covered in **Section 3**. More details regarding specific phylogenetic comparative testing are given in **Section 2.3**.

Since none of the currently available Bayesian tree sets (Bouckaert et al., 2012; Chang et al., 2015; Heggarty et al. in review) sample all of the languages in our dataset, we use trees from Glottolog (Hammerström et al., 2014) which have been given necessary branch lengths by Dediu (2018). Dediu (2018) takes cladogram-like trees from four different sources, and adds branch length, vital for phylogenetic comparative analysis, using nine different methods. We describe in **Supplementary Information S2** how we opted for two of these trees. We used the function multi2di() in the R package ape (Paradis et al., 2004; R Core Team 2020) to create 250 trees in which the polytomies (nodes in the tree which lead to more than two clades or languages) present in Glottolog were resolved in a random fashion. Subsequently, branch lengths were added to these newly created branches in a random fashion corresponding to the distribution that the branch lengths for each of these trees have. This resulted in a sample of 500 phylogenetic trees.

There are different models for the evolution of different types of characters (Pagel 1999; Meade and Pagel 2019): *binary characters* (a language either has a characteristic, like having one or more click consonants, or it does not), *continuous characters* (a real number, such as the entropy of object-verb word order in a parallel corpus; Levshina et al. in review; or Greenberg’s 1960 morpheme-word ratio), or *multistate characters* in (Comrie’s 2013 WALS chapter on the alignment of case marking of full noun phrases, a language may



		A	AB	B	BC	C	CA
		1	2	3	4	5	6
A	1	-	q12	q13	q14	q15	q16
AB	2	q21	-	q23	q24	q25	q26
B	3	q31	q32	-	q34	q35	q36
BC	4	q41	q42	q43	-	q45	q46
C	5	q51	q52	q53	q54	-	q56
CA	6	q61	q62	q63	q64	q65	-

FIGURE 1 | Q matrix of the six states negative existential constructions may be in. Type-A has been coded as 1, AB as 2, B as 3, BC as 4, C as 5, and CA as 6. The first number refers to the state that is left (rows), and the second number refers to the state that is entered (columns). Thus qij is the transition parameter from Type-i to Type-j (example: q12 is Type-A > AB; q21 is Type AB > A, etc). The changes between the states described by Croft's (1991) Negative Existential Cycle have been marked in **green color**. The set of opposite transitions is marked in **red color**. Note that, with the exception of q61 and q16, changes from smaller to bigger numbers (for example, q25; Type-AB > C) designate changes in line with the direction of the NEC, while changes from bigger to smaller numbers (for example, q52; Type C > AB) go against the direction of the NEC.

have one of six possible alignment patterns). In this study, we are concerned with a multistate character with exactly six states that detail the interaction between existential and standard negation (see **Section 1**).

The standard model to account for multistate characters is the *continuous-time Markov* process of character evolution (Pagel et al., 2004). This model describes the probability of change between *states* of a character (here, negative existential strategies) in terms of a set of *transition rate parameters*. Here, "continuous-time" implies the character can change its state at any instant of time rather than at fixed intervals; "Markov," from "Markov chain," indicates that the probability of changing from one state to another depends only on the current state, and not on any earlier states. The changes that take place in the continuous-time Markov model are summarized using a *transition rate matrix* or a Q matrix, where the individual transition rate parameters are designated by q, followed by codes for two states. **Figure 1** illustrates this matrix for the negative existential domain. It is the set and values of transition rate parameters captured by the Q matrix, as well as the likelihoods that are associated with different states at the internal nodes of the tree, that are tracked during analysis.

Because there are six types in Croft's (1991) typology, there are 6×6–6 = 30 transition rate parameters. States cannot change into themselves (these dependencies are marked in **Figure 1** by "-"), hence these represent the probabilities that the state stays the same. In the Q matrix, the diagonal "no change" probabilities and the off-diagonal transition rate parameters (q's) sum to zero. Croft's (1991) NEC proposes a diachronic typology using only six of these changes between types, as indicated by **Figure 1**. Croft's (1991) NEC is very ambitious given the possible transition-rate parameter space: modeling change between six types using only six diachronic pathways between types.

The large number of types and correspondingly large number of transition rate parameters, together with the rarity of some types (see **Section 3**), pose a practical problem. The "one in ten" rule in statistics also applies to phylogenetic comparative analysis, i.e., having ten data points (species or languages) per free parameter is an aim during data collection, despite the fact that actual sample size is reduced through phylogenetic dependencies (Mundry 2014). This implies needing a sample size of 300+ languages to run the model in which all 30 transition rate parameters are included. Ideally, construction types would be distributed evenly across that ideal sample, but this is not realistic (see **Section 3.1**, Croft 1991; Veselinova 2016). To have a reliable estimate, for instance, whether change to Type-CA is more likely to come from Type-C (as predicted by the NEC) or any other type, we would probably need even more data. In other words, our Indo-European dataset is still too small, and the distribution of types is too skewed, to comprehensively test Croft's (1991) NEC. However, we will try regardless of these issues and report on the results in **Section 4**.[6]

We aim at model optimization that will 1) test which set of transition-rate parameters explains best the distribution of negative existential types in our Indo-European dataset and 2) compare the best fit models to the NEC. The Bayesian model used here allows several options for model optimization. The first option we have is simply excluding certain transition-rate parameters manually. This is also how we test the NEC model, by excluding the 24 transition-rate parameters in black and red typeface in **Figure 1** by setting them to zero. The second option, Reverse Jump MCMC (RJ MCMC, Green 1995; Pagel and Meade 2006), automatically turns on and off transition-rate parameters while at the same time estimating and reducing the number of different rates. In the posterior models, transition-rate parameters that do not contribute to the model are excluded and the number of individual transition-rate parameters is typically reduced such that a small number of rates is shared across parameters, optimizing the model and making it "more elegant." Excluding transition rate parameters manually and doing RJ MCMC can also be done at the same time.

Again, the large number of types and correspondingly large number of transition rate parameters poses a practical problem. If, for example, we want to compare a model with ten transition rate parameters (perhaps the six

[6] One solution to the large transition rate parameter space problem that we have tried for an earlier version of this paper is to exclude Croft's (1991) transitional types AB, BC, and CA from the dataset and model. The three possible types are then A, B, and C, AB would be re-coded as A&B, BC as B&C, and CA as A&C. This reduces the transition rate parameter space to 3 × 3–3 = 6; and three of these transition rate parameters, i.e., change from A > B, B > C, and C > A are implied by Croft's (1991) NEC; the other three transition rate parameters, A < B, B < C, and C < A, are the reverse of Croft's (1991) NEC. However, while this effectively eliminates the large transition rate parameter space problem, it is not viable because 1) it does not do justice to the data, for example, split languages have to be coded as A&B&C; 2) because of the resulting prevalence of having two states (we have especially many AB languages which would be coded as A&B), including/excluding specific transition rate parameters is not informative, as any combination of transition rate parameters becomes equally likely.

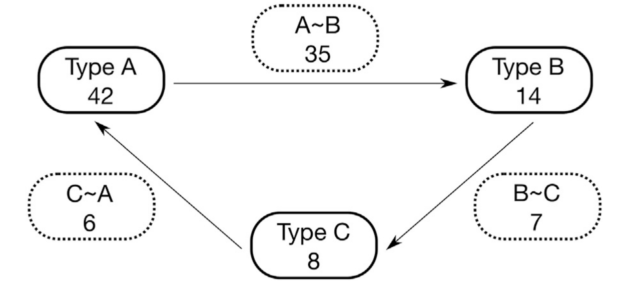


FIGURE 2 | Frequency diagram of attested types in our 106 language sample. Note that six languages have more than one type; Oriya is Type-A and Type-BC, Kumzari is Type-A and Type-C, Kupia, Chitpavani Goan Konkani, Goan Konkani, and Varhadi-Nagpuri are Type-B and C.

NEC parameters plus four more) with ten *random* transition-rate parameters picked out of the 30 in our Q matrix, we have to face the fact that there are 30,045,015 possible sets of 10 parameters out of 30 transition-rate parameters. It is very likely that some combinations of ten parameters fit our data better than others; but identifying these combinations is exactly the problem. Testing so many models is not feasible: a single model takes four to seven hours to run on a normal desktop computer, and the number of models grows exponentially when we add models with 11, 12, 13 transition-rate parameters out of our set of 30. To improve our chances of finding the model with the best fit, then, we have to rely on RJ MCMC. In **Section 4**, we introduce the models we tested and discuss their fit.

We used *BayesTraits V3.0.2* (Meade and Pagel 2019) to conduct phylogenetic Bayesian MCMC analysis, specifically its component *MultiState* (Pagel, Meade, and Barker 2004). We construct various models we want to test, focusing on the transition-rate parameters; these are covered in **Section 4**. We conducted a single MCMC analysis for each model, which was run for 2×10^7 iterations, with a burn in of 1×10^7, sampling every 10^5 iteration, resulting in a sample of 1,000 posterior estimates. Convergence was assessed by checking the absence of a correlation between the posterior likelihood and the iteration number. Lack of autocorrelation between samples was assessed visually. When used, Reverse Jump was used on all transition-rate parameters, with a default exponential prior with mean 50. When Reverse Jump was not used, the default uniform prior with distribution 0–100 was used for the transition rate parameters. *BayesTraits V3.0.2*'s built-in stepping stone sampler was used to estimate log marginal likelihoods after the MCMC analysis was concluded. The log marginal likelihoods were used to assess the fit of the various models in **Section 4**.

3 NEGATIVE EXISTENTIALS IN INDO-EUROPEAN: A TYPOLOGICAL SURVEY

This study surveys the expression of the negative existential domain in 106 Indo-European languages. Data on 42 languages come from Verkerk and Shirtz (forthcoming), data on 13 Slavic languages come from Veselinova (2014), and data on 51 additional languages are added to this article (see **Supplementary Information S3**). We aimed to sample as extensively as possible, but were constrained by both available resources and time during a global pandemic. This section first briefly summarizes the results of our survey, and highlights some noteworthy areal tendencies (see Verkerk and Shirtz, forthcoming for a more detailed discussion): the relative typological stability of the negative existential in some branches or areas and its relative instability in other branches or areas. Then, we provide a brief historical and phylogenetic overview of the data, and following these two sections we discuss morphosyntactic innovations in the negative existential domain that do not involve a change in the typological classification of a construction and innovations that lead to transitions that are not included in the NEC.

3.1 General and Areal Overview of the Typology

The grammatical expression of the negative existential domain in Indo-European is varied, with each of Croft's six types attested somewhere in the family. This variation is not homogenous and some types are quite frequent while others are rare. Furthermore, the variation is not equally distributed across Indo-European, and some families exhibit a rather uniform typology of the negative existential domain while other families are diverse. The diagram in **Figure 2**, constructed based on the NEC itself, indicates the raw counts of each attested construction type.[7]

In our Indo-European sample, Type-A (37.5%) and Type-AB (31%) are the most common types. The biggest difference between **Figure 2** and the world-wide surveys in Veselinova (2014, 2016: 147, 150) is that in our Indo-European sample, Type-AB (current paper: 31%; Veselinova: 8.9%) is far more common than type-B (current paper: 12.5%; Veselinova: 29.7%), but our Indo-European sample resembles Veselinova's findings for Berber and Uralic. Aside from these differences, the Indo-European data, just like Veselinova (2014, 2016) worldwide sample, mostly confirms Croft's (1991) remark that types A and B are more common than Type-C, and the transitional types AB, BC, and CA are uncommon, with the caveat that Type-AB is quite common across Indo-European. However, we found six CA languages, much more than the two instances in Veselinova (2016: 150) world sample and her family-based studies (1 CA language out of 109 languages).

The areal distribution of the different construction types is illustrated by the map in **Figure 3**. It illustrates how the unequal distribution of construction types in general is magnified when focusing on certain areas and certain families. For a first impression, one can simply contrast the relative color

[7] It is not unheard of that languages have more than one type, Veselinova (2016: 154) discusses 9 other cases. Interestingly, the types do not have to be (sometimes cannot be) consecutive in the NEC, showing that different parts of the negative existential domain in an individual language can undergo different transitions.

uniformity across Europe, especially in the Romance and Germanic-speaking areas, to the diversity found in the Indo-European languages of Western, Central, and South Asia.

The relative uniformity across the European part of the map in **Figure 3** is the result of two larger Indo-European families, Romance and Germanic, exhibiting little to no typological variation in the expression of the domain. The Romance languages in our sample uniformly negate existential statements using the same negation marker used in standard verbal negation. Thus, they consistently exhibit a Type-A negative existential construction. This is illustrated by (8a-b) below from Piedmontese (Turinese):

Piedmontese (Turinese: Romance; Emanuele Miola p.c.)

(8a) *A Maria a-j=piasu nen i=film*
to Maria SBJ-3SG.OBL=like.PRS.3PL NEG ART=movies
'Maria does not like movies'

(8b) *i=gat sarvaj a=esistu nen*
ART=cats wild SBJ=exist.PRS.3PL NEG
'There are no wild cats'

Typological uniformity is also attested across Germanic, where alongside Type-A constructions (illustrated in (1) above, from Swedish), existential statements may also be negated using a negative indefinite pronoun or determiner. In (9), from Swedish, the existential statement is negated by the Swedish Negative indefinite pronoun *inget*, and in the English translation of the example, the statement is negated by the English marker *no.*

Swedish (Bordal 2017: 6; their glosses and translation)

(9) *Det finns inget vatten i kran-en*
there exists NEG water in tap-DEF
'There is no water in the tap'

The uniformity of the Romance and Germanic families stands in contrast to the diversity attested in Iranian and Indo-Aryan, and to a lesser degree in Slavic. The factors involved in this difference may include borrowing, diachronic replication, substrate factors, or universal tendencies (e.g., Nichols 1992). It is beyond the scope of the current paper to argue which of these factors led to the typological uniformity of negative existential construction in Romance and Germanic on the one hand, and to the typological variation in Slavic, Iranian, and Indo-Aryan on the other hand. For now, suffice it to mention that the languages of Western Europe form a Sprachbund (e.g., Haspelmath 2001; van der Auwera 2011), and propose that the uniformity across Germanic involves some sort of diachronic replication. Finally, note that the pattern whereby the Iranian and Indo-Aryan families exhibit much more typological variation than the Germanic and Romance families is not limited to the negative existential domain. A

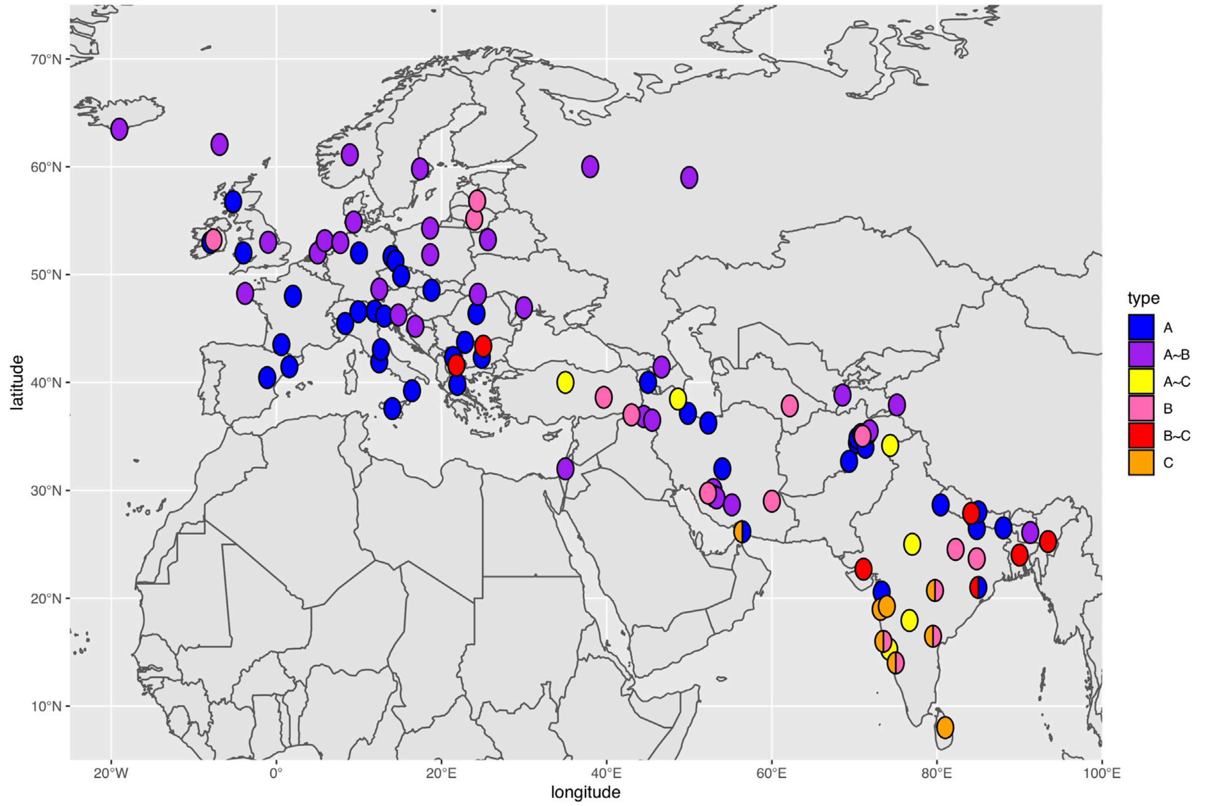


FIGURE 3 | An overview of negative existential construction types in Indo-European languages, overlaid on a map of western Eurasia.

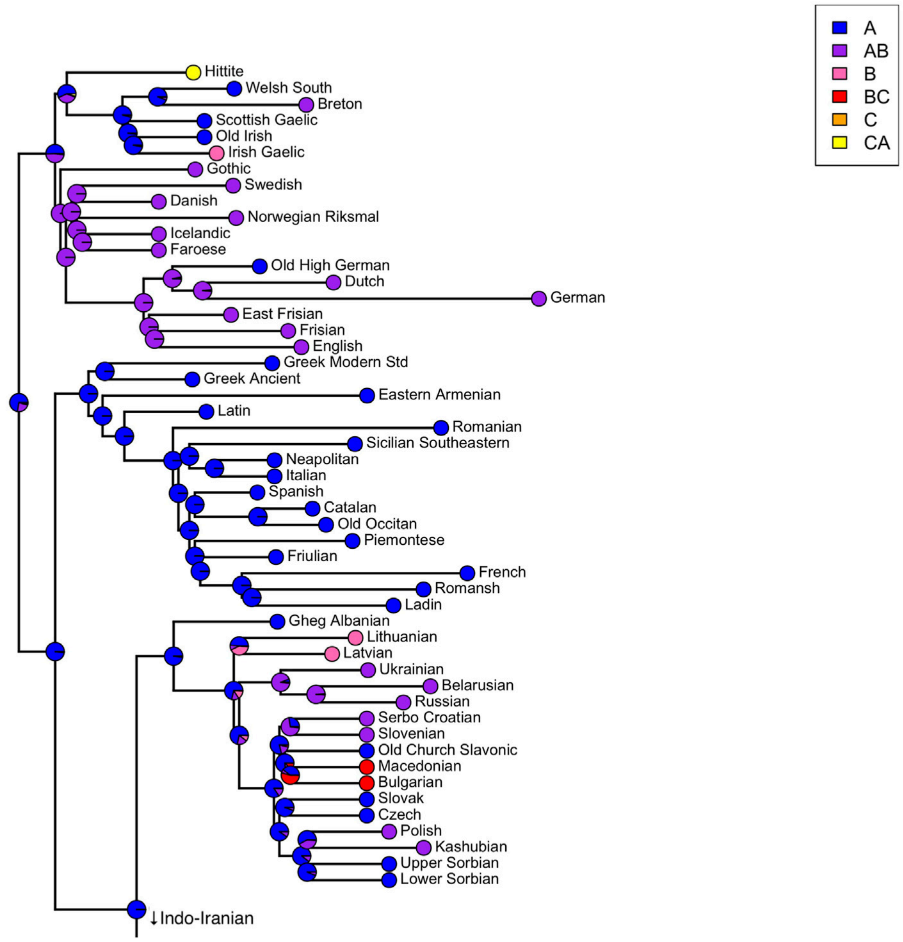


FIGURE 4 | An overview of negative existential construction types in non-Indo-Iranian languages, overlaid on a modified Indo-European Glottolog tree (Hammarström et al., 2014; Dediu 2018).

similar pattern is attested, for example, with the alignment of core arguments, where Iranian and Indo-Aryan are very diverse (e.g., Haig, 2008; Verbeke, 2013), while the Indo-European languages of Western Europe are rather uniform.

3.2 Historical and Phylogenetic Overview

Not surprisingly, the uneven areal distribution of variation in construction types goes hand-in-hand with uneven distribution across different Indo-European subfamilies. We illustrate this on a randomly chosen phylogenetic tree from our tree sample in **Figures 4** and **5**. As described in the methodology, the trees we used for phylogenetic comparative analysis were built by making the polytomies binary in a random way, and assigning branch lengths to these newly created branches on the basis of the distribution of existing branch lengths from Dediu (2018). This leads to sometimes unrealistic higher order groupings, such as the one we find here relating Hittite and Celtic. We do not argue that this is how the Indo-European languages actually evolved; this is simply one of many possibilities that was selected for display purposes only. Given the size of the sample and tree, we split it such that the non-Indo-Iranian part of the family is displayed in **Figure 4**, and Indo-Iranian is displayed in **Figure 5**. The pie plots on the internal nodes of the tree represent marginal ancestral state reconstructions conducted in the R package corHHM (Beaulieu et al., 2013; R Core Team 2020). These are again illustrations on a single tree; the analyses were conducted on the full tree sample and reconstructions will differ across trees (see **Section 4**). A simple parsimony reading can be misleading. For example, in Bihari (Chitwania Tharu, Ranna Tharu, Darai, Sadri, and Bhojpuri), Darai and Sadri have not changed Type-A > Type-BC or Type-B, but rather, the Tharu languages and Bhojpuri are likely to have finished the NEC cycle and reached Type-A again. Hence, the reconstructions are partly realistic, partly a consequence of the tree structure coupled with gaps in the diachronic record as intermediate stages are often not present in the dataset (or not recorded at all).

While much of this paper focuses on transitions that are not a part of the NEC, it should be mentioned that many instances of

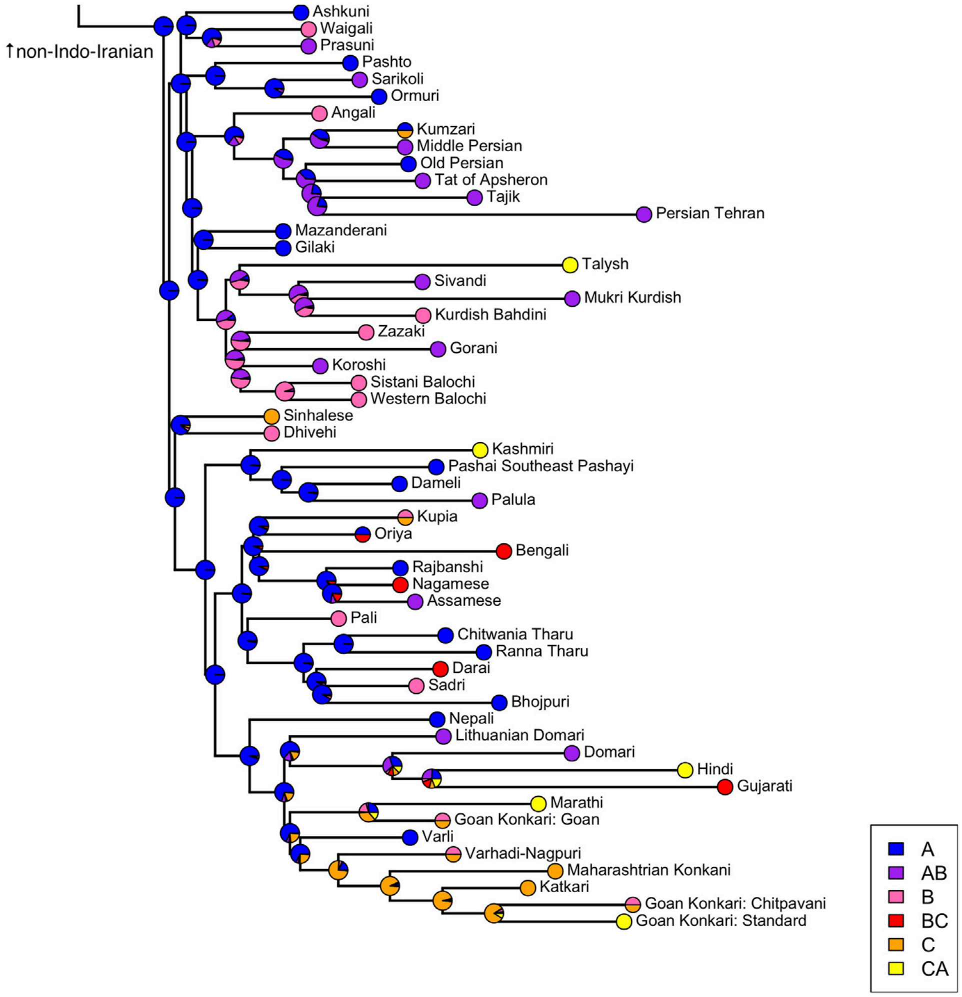


FIGURE 5 | An overview of negative existential construction types in Indo-Iranian languages, overlaid on a modified Indo-European Glottolog tree (Hammarström et al., 2014; Dediu 2018).

change in the negative existential domain are instances of NEC transitions. Here, we only briefly sketch some such transitions. Further details and references can be found in **Supplementary Information S3** and Verkerk and Shirtz (forthcoming). Note again that reading the transitions from the tree in **Figures 4** and **5** can be deceptive due to missing intermediary stages and languages we did not sample.

The transition from Type-A to Type-AB can be clearly seen in Palula (Indo-Aryan). The verbal negation marker in Palula is *na* (Liljegren 2016). Existential predicates may be negated by *na*, or by the special negative existential copula *náinu*, a reduction of *na* NEG + *hínu* COP (Liljegren 2016: 413). The Palula data, then, illustrates the NEC's A > AB transition. Two Kurdish languages in our sample illustrate change from Type-AB to Type-B. In Mukri (Central) Kurdish, standard negators are *nā-* for present tense, *ne-* for past tense (Öpengin 2016: 74). Negative existential strategies show another tense-based split, with standard negation used in the past tense, and in non-past tense, the negative copula negation *nī=* is used. Bahdini Kurdish (Kurmanji) is Type-B, with standard negation using the clitic or prefix *na-*, *ne-*, and a special negative existential copula *tun-* (Thackston 2006).

While there are several BC languages in our sample, it is not easy to find a clear example of the type B > BC transition. One tentative example is Darai (Type-BC), where the special negative existential marker *nidze* is used as one of two nonexistential negation markers (Dhakal 2012). Closely related Sadri has a Type-B construction with the potential cognate special negative existential verb *nʌkh* (*nʌkhe*, Jordan-Horstmann 1969). Perhaps an earlier stage of these languages was Type-B, with Sadri being conservative and all other languages in this subgroup being innovative (see below on the CA > A transition). A tentative example of the BC > C transition can be found between Dhivehi and Sinhalese. Dhivehi has a special negative existential copula *net* (<OIA *nā́sti*, Fritz 2002) and is Type-B, Sinhalese is Type-C, with the free standing negative existential *nææ* that also functions as postverbal predicate negator (Chandralal 2010).

The transition between type C > CA is attested twice in the Indo-Aryan Southern zone. In Standard Goan Konkani the verb-like negator *nay* is combined with existential predicates (Ghatage 1966). In closely related Chitpavani Goan Konkani the likely cognate *nãy* ~ *naĩ* is used as standard negator and as special negative existential, without an existential predicate (Bhide 1982). Similarly, Marathi (Type-CA, the special negative existential is *nāhi*, Croft 1991) can be contrasted with Katkari (Type-C), where the cognate negator *nahĩ* ~ *nay* does not yet combine with an existential verb (Kulkarni 1969).

Evidence for a CA > A transition is found in Chitwaniya Tharu (Indo-Aryan), for example, where the verbal negation marker *hoyne*, a combination of the old *h*-copula and a negation marker, is used to negate existential statements alongside an innovative, and obligatory, existential verb. This negation marker, however, is still deployed as a nonverbal negative copula, without a synchronic verbal copula, in some conservative nonverbal predication constructions, where it is often followed by an emphatic clitic marker (Paudyal, 2014). Varli (Abraham and Abraham 2012) has likewise undergone the CA > A transition, as the negator *nahĩ*: (likely cognate with Marathi (CA) *nāhi* and Katkari (C) *nahĩ* ~ *nay*) can no longer be used without an existential predicate (as is still optional in Marathi).

3.3 Illustrations of Innovations That do Not Involve a Change of Construction Type

The typological stability in Romance and Germanic, as well as in some sub-branches of other families, may lead one to believe in some extreme conservatism in the verbal and existential negation in these families. Reality, however, is more complex and across these families there are several instances of innovation in these domains, as well as innovations in the domain of existential constructions in general. These innovations do not lead to a change in the typological classification of the expression of negative existence in these languages.

Across Romance, innovations are attested in the expression of existential predication and in the expression of negation. Existential predicates in Catalan, illustrated in (10), are expressed by a combination of the locative adverbial clitic *hi* "there" and a third person form of *haver* "have." In Romanian, on the other hand, existential predicates can be expressed by *a se găsi*, the middle form of the verb "to find," as illustrated by (11), or by the verbs *a exista* "to exist" or *a fi* "to be."

Catalan (Romance, Wheeler et al., 1999: 460)

(10)	*Hi*	*ha*	*tres*	*possibilidades*
	there	have.PRS.3SG	three	posibility.PL

'There are three possibilities'

Romanian (p.c. Andreea Calude)

(11)	*Se*	*găsesc*	*pisici*	*sălbatice*
	MID.3SG	find	cat.PL	wild.PL

'There are wild cats'

As the innovative expression of existential predication involves verbs in both languages, it is only natural for them to be negated by the standard verbal negation marker (at least initially). Thus, an innovative existential verb in a language which already had a Type-A negative existential construction, the common situation in Romance, leads to a novel negative existential construction without a change in the typological classification of the domain.

The use of a negative indefinite pronoun or determiner to negate existential predication in Germanic also hides instances of innovation. This involves the rise of innovative negative indefinite forms, such as German *kein,* Dutch *geen*, and Swedish *ingen* when compared to English *no.* While these forms are related, they involve different types of syntactic and lexical innovations (German *kein* from **nih* "neither" and **aina*-"one"; Dutch *geen* from *neh* "and not" and *ein* "one" (Philipa et al., 2003-2009), Swedish *ingen* from *einn* "one" +-*gi*, privative suffix; the use of negative indefinite pronouns across different nonverbal predicates differs quite radically, see Verkerk and Shirtz, forthcoming). Thus, once a negative existential construction with a negative indefinite pronoun as the negation marker exists, an innovation in the domain of this marker would not alter the typological classification of the construction itself.

Similar innovations can be found in Greek, where innovations in the expression of negation occurred from time to time (e.g., Kiparsky and Condoravdi 2006). Across Indo-Iranian, locative verbs have often been co-opted into existential predication, and by extension also negative existential predication. This often results in an innovative Type-A construction that may or may not lead to a change in typological classification.

3.4 Illustrations of Innovations That Are "Outside" the NEC

Figures 4 and **5** use a phylogenetic tree to illustrate the synchronic variation in negative existential constructions in Indo-European. More than that, these figures also illustrate one proposal for the reconstruction of the type of negative existential in ancestral nodes on the tree. A closer look at the tree would suggest that there are several transitions that cannot be explained in terms of the NEC. We describe such transitions in this section on the basis of our own analysis, as the tree can mislead through data gaps. These transitions are of two types. First, we find transitions where a development within the domain of negative existence itself leads to a change in the classification of a negative existential construction. Second, we find transitions that involve innovations that occur *outside* of the negative existential domain but affect it. This includes innovative negative existential constructions entering the domain and innovations in the realm of verbal negation.

The first type of transition was illustrated by Macedonian and Bulgarian, where Veselinova (2014) shows a transition from Type-A (as illustrated by Old Church Slavonic data) directly to Type-BC in Bulgarian and Macedonian, without moving through the intermediary types. Another example can be found in Kumzari where we find a split between Type-A and Type-C, the latter evolving directly from an older Type-A construction. Verbal negation in Kumzari is expressed by a post-verbal *na* (van der Wal-Anonby 2015: 211–213; see also the main clause in (12a)). Affirmative existentials are expressed by clauses containing the figure NP, an optional locative ground, and a copula. Now, the source of the Kumzari enclitic copula is the Old Iranian **h*-copula, and this copula underwent a great deal of phonological reduction that resulted in its complete deletion in the Kumzari 3SG form (van der Wal-Anonby 2015). This resulted in clauses such as the subordinate clause in (12a) or the unipartite clause in (12b) where only the figure NP is expressed. The negative existential is simply expressed by clauses composed of the figure NP followed by the verbal negation marker *na*.

Kumzari (Iranian, van der Wal-Anonby, 2015: 184; 164; 140)
(12a) *mār, aqrab, inda yē a dām na*
snake scorpion in 3SG SUB know.1SG.IMPF NEG
'I don't know whether there was snake or scorpion in it'

(12b) *knār-e=ø*
jujube.tree-INDF=COP.3SG
'there was a jujube tree' (author's ø)

(12c) *iʃ ɣēla na*
any grain NEG
'there wasn't any grain

The Kumzari construction illustrated in (12c), then, is an instance of Type-C, as the standard verbal negation marker is used as the sole marker of existential negation. This Type-C construction did not arise out of a previous Type-BC, but given the source of the enclitic copula in the Old Iranian **h*-verbal copula with its subsequent extreme phonological lenition, arose out of an older Type-A construction (other instances of a Type-A > Type-C transition are illustrated by Croft, 1991).

The second type of change involves innovations outside the negative existence domain that affect it. One type of such an innovation has been proposed by Verkerk and Shirtz (forthcoming), where the rise of an innovative verbal negation marker may change the classification of Type-C negative existential construction in the opposite direction to the NEC into Type-BC, as now only one of several verbal negation markers is used in clauses expressing negative existence. An opposite scenario also occurs, where a loss of an older verbal negation marker may affect the classification of an existing construction. In Eastern Indo-Aryan, for example, there were (at least) a preverbal negation marker and a postverbal negation marker. In Kupia, the older preverbal negation marker was lost, but was fossilized in some lexical verbs including "not know," "be unable," and in the negative copula *nenj-* which is still used in negative existential clauses (Christmas and Christmas 1973: 310). Thus, the Kupia negative existential construction with *nenj-* went from being a Type-A construction to being a Type-B construction, without an intermediate Type-AB. Similarly, changes in the domain of verbal negation in Assamese seem to have led to a change from a previous Type-BC to a Type-B construction. As these constructions coexisted with a conservative Type-A construction, Assamese is now classified as Type-AB.

Another type of innovation involves a novel verbal existential marker that is negated by the standard verbal negation marker. Thus the novel negative existential construction in this case is a Type-A construction. This was illustrated above by Catalan and Romanian existential constructions. Such innovations are common in Iranian, where verbs translatable as "be.at" evolved to express predicate location and existence. This is illustrated by the Sivandi example below, where *dār* "be.at" is used as the existential copula and is negated by the Sivandi verbal negation marker, a preverbal *na*. When such a construction evolves in a language that already has a Type-B construction, and each construction settles in its own functional niche, the result would be Type-B > Type-AB transition, in the opposite direction of the cycle.

Sivandi (Iranian, Lecoq 1979: 15)
(13) *ke bār na=dār-e*
COMP grain NEG-be.at-3SG
'(he closed the mill) because there was no grain'

We conclude this section with an overview of changes discussed here and elsewhere that do not fit the NEC:

(14)	A > C	Kumzari, see also Croft (1991)
	A > B	Kupia
	A > BC	Macedonian/Bulgarian, Veselinova (2014)
	AB > A	Mazanderani, Gilaki, Verkerk and Shirtz (forthcoming)
	AB > BC	Russian, Hawaiian, Veselinova (2014)
	B > C	Polynesian, see Veselinova (2014)
	BC/A > AB	Assamese

The transitions that form the NEC, then, may account for many attested changes in the Indo-European negative existential domain, but there are other attested or plausible changes that are not a part of the NEC set of transitions. This was already mentioned in Croft's original description of the NEC (1991), developed by Veselinova (2013, 2014), and further systematized here.

4 PHYLOGENETIC COMPARATIVE MODELING

This section deals with diachronic change in strategies of negative existentials as modeled on the branches of a phylogenetic tree set. There are two main options for designing the set of transition rate parameters: leaving both the selection of parameters and the estimation of their (communal) rate to the RJ MCMC analysis, and selecting transition rate parameters to be included or excluded manually. We use both approaches, and additionally try to combine them. The RJ models and results are presented in **Section 4.1**, the manual models are presented in **Section 4.2**. All code and results can be found in **Supplementary Information S5**.

4.1 Reverse Jump MCMC Models

In this section, we report on a set (or rather, a chain) of Reverse Jump MCMC models where we exclude, step by step, transition

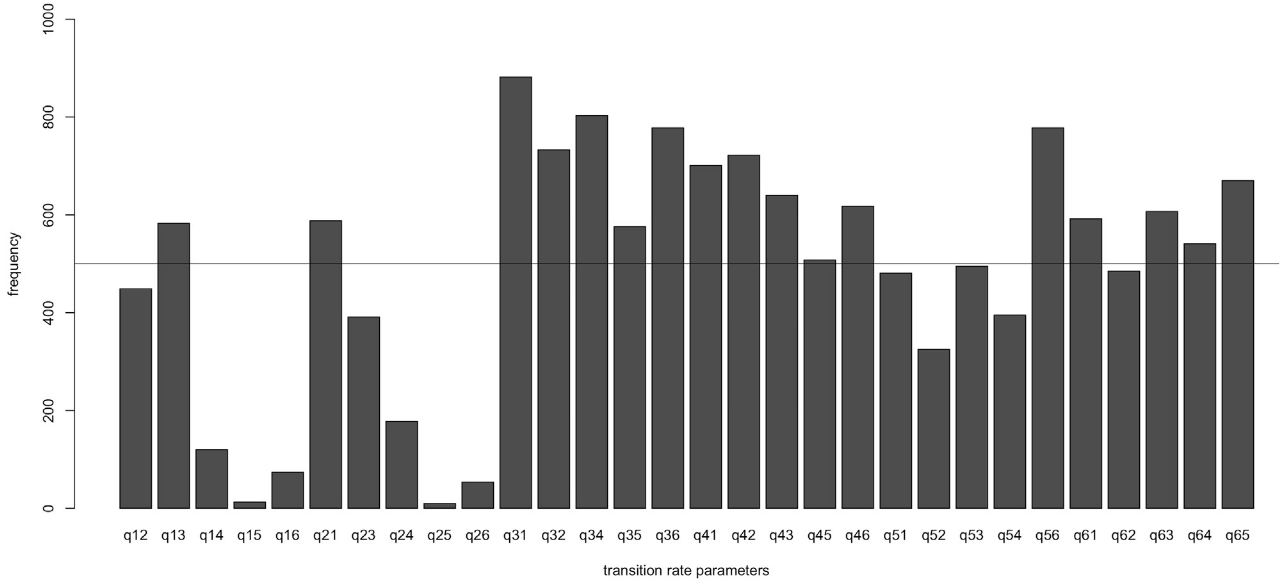
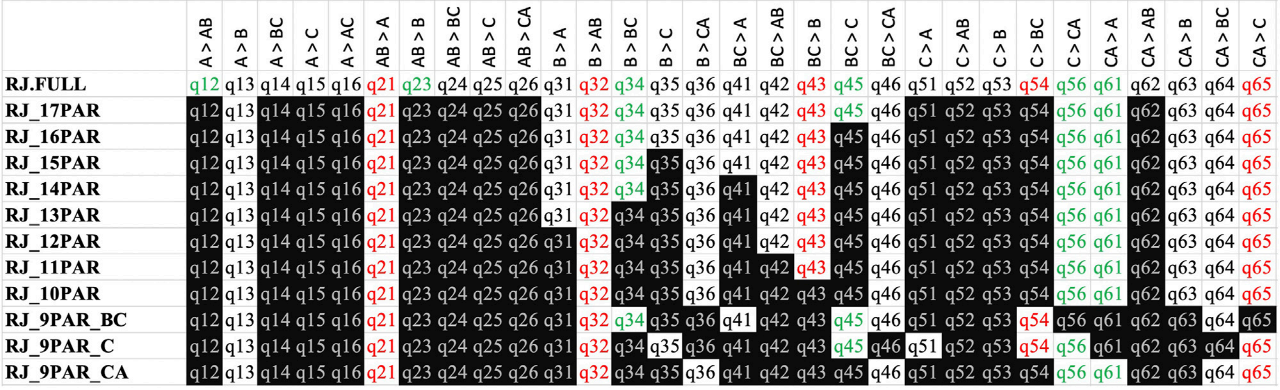


FIGURE 6 | Frequency of transition rate parameters being turned on in the RJ.FULL model. The *y* axis marks the number of posterior models in which a transition parameter is turned on, with the horizontal line marking being turned on in 50% of the posterior. Type-A has been coded as 1, AB as 2, B as 3, BC as 4, C as 5, and CA as 6. Hence, q12 refers to change A > AB; q13 refers to change A > B, etc. See **Figure 1** for the Q matrix.

rate parameters that were infrequently turned on in the set of posterior models. Our starting model is a full RJ model (exponential prior, mean 50), which generates posterior models by allowing transition rate parameters to be excluded and/or their rate to be set equal across parameters. The posterior distribution of models gives a sense of which transition rate parameters are of vital importance (i.e., often turned on) and whether rates differ across rate parameters. Note that *BayesTraits* (Meade and Pagel 2019) does not allow for a large number of free transition rate parameters (30) to be estimated *without* RJ. Similarly, too low an included number of free transition rate parameters, especially where one of the states ends up unreachable, are also disallowed by *BayesTraits*. The first model, RJ.FULL, is an RJ model with no prior restrictions on the transition rate parameters, so all transition rate parameters (also called qs, see **Figure 1** for the Q matrix) can be turned on. The number of times in which each transition rate parameter is indeed turned on in RJ.FULL's set of posterior models is given in **Figure 6**.

In 989/1,000 posterior models, there was a single rate estimated for all transition rate parameters that were turned on. This implies that there is little evidence for different types of changes occurring at different rates in models; however, this may also be due to the lack of constraints on excluded parameters. We could hypothesize, for example, that change *from* transitional types AB, BC, CA *to* nontransitional types A, B, and C would occur at a faster rate than vice versa, but there is no evidence for that in this model. In the RJ.FULL model, the mean transition rate is 0.39 (median 0.32).

Figure 6 shows that very few transition rate parameters are consistently turned off in the posterior models (only q15 (A>C) and q25 (AB>C)). Conversely, very few transition rate parameters are *consistently* turned on in posterior models (q31 (B>A) has the largest frequency, featured in over 80% of posterior models). Most transition rate parameters are turned on about 40–60% of the time. Hence, we do not observe a clear pattern of which transition rate parameters are relevant and which are not. There are several explanations for this pattern: 1) there is a multimodal distribution of well-fitting transition rate parameters that is dependent on the characteristics of the phylogenetic trees; 2) dependencies between transition rate parameters, such that they replace each other across models; we could, for instance, imagine that in some model, q12 (A > AB) and q23 (AB > B) are turned on, while in another model, these two are not needed, but only q13 (A > B) is turned on; 3) the sheer amount of transition rate parameters allows for a multitude of likely models, all of about an equal good fit. Unfortunately, it is difficult to tease apart the cause for this mixed pattern of turning on and off transition rate parameters. The lower and upper bounds for the number of transition rate parameters that were turned on for the RJ.FULL model were 9 and 26, compared to the six transitions of the NEC. However, there are millions and millions of options to create models with 9–26 parameters, so this information is not useful. It would further be an immense task to find out if there are indeed correlations between characteristics of the trees and the transition rate parameters that are turned on or off in subsets of the posterior models. Therefore, we have to conclude that RJ.FULL does not immediately point us toward an elegant, clear model of diachronic change. Nevertheless, we find the model informative because 1) it demonstrates how Reverse Jump MCMC models work in the context of character evolution;



	A > AB	A > B	A > BC	A > C	A > AC	AB > A	AB > B	AB > BC	AB > C	AB > CA	B > A	B > AB	B > BC	B > C	B > CA	BC > A	BC > AB	BC > B	BC > C	BC > CA	C > A	C > AB	C > B	C > BC	C > CA	CA > A	CA > AB	CA > B	CA > BC	CA > C
RJ.FULL	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
RJ_17PAR	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
RJ_16PAR	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
RJ_15PAR	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
RJ_14PAR	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
RJ_13PAR	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
RJ_12PAR	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
RJ_11PAR	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
RJ_10PAR	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
RJ_9PAR_BC	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
RJ_9PAR_C	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
RJ_9PAR_CA	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65

FIGURE 7 | Overview of RJ models. Shaded transition rate parameters are excluded; nonshaded transition rate parameters are included (but not necessarily turned on in the posterior models). Green indicates the six transition rate parameters that model the NEC; red indicates the transition rate parameters that revert the NEC.

2) it demonstrates the intrinsic difficulty in modeling a feature with six states; and 3) RJ.FULL gives us at least some sense of which transition rate parameters are relevant and which are not, albeit limited. In the remainder of this section, we build on RJ.FULL.

The way forward is to manipulate the set of transition rate parameters, such that we exclude from the prior those parameters that were not turned on in RJ.FULL often, in the hope of making the model more decisive and obtain a higher log marginal likelihood (log mLh). The mLh of a model "is the integral of the model likelihoods over all values of the models parameters and over possible trees, weighted by their priors" (Meade and Pagel 2019: 14). It is the main mechanism used by *BayesTraits* (and other software) to assess model fit. The mLh is computationally expensive, and is therefore estimated using stepping stone sampling (Xie et al., 2011) in *BayesTraits*, which provides an estimated log mLh. We can compare the log mLh of the two models, and see if there is evidence for a significantly better fit of the better fitting model by calculating log Bayes Factors (BF). The better fitting model is the one with a higher log mLh (because log likelihoods are negative, it makes sense to think about the better-fitting model being the one that is closer to zero).

(15) Log Bayes Factor = 2(log marginal likelihood better fitting model – log marginal likelihood worse fitting model)

The log Bayes Factor can be interpreted such that a BF > 2 constitutes positive evidence against the null hypothesis, the bigger the BF the more convincing the evidence (Kass and Raftery 1995: 777, their two $\log_e$ (B_{10})).[8]

We build smaller RJ models by excluding parameters that were not often turned on in RJ.FULL. In the first model, RJ_17PAR, all parameters that were absent in 50% or more of the posterior models of RJ.FULL (q12, q14, q15, q16, q23, q24, q25, q26, q51, q52, q53, q54, and q62) were excluded in the prior. What follows are models in which further infrequent transition parameters are excluded, each constructed on the basis of the preceding one. Model names such as RJ_17PAR, RJ_16PAR, etc. are constructed for these models such that they indicate the number of transition rates PARameters that is included. The models are displayed in **Figure 7**, and the results of all of them are reported in **Table 1**.

As indicated by the ordering in **Figure 7**, the RJ models were calculated consecutively, i.e., model RJ_16PAR was constructed by excluding an infrequent transition rate parameter from RJ_17PAR, etc. The figures that detail how often the transition rate parameters are turned on in each model, which served to frame this successive exclusion of parameters, are included in **Supplementary Information S5**. Note that many different choices could have been made in this successive exclusion of transition rate parameters and that we have not exhaustively sampled the set of possible models in any sense.[9] Doing this, we ultimately arrive at RJ_9PAR_CA and can no longer exclude any parameters from the RJ model (models with eight parameters out of the nine in RJ_9PAR_BC, _C, or _CA do not run). Hence, any RJ model has a minimum of nine transition rate parameters. RJ_9PAR_CA has two epicenters of change: cyclical change between Types-A, AB, and B, and then chance centering around Type-CA, with movement between the two epicenters through B > CA and CA > A. To investigate this specific RJ model with a "hub"

[8] log Bayes Factors (Kass and Raftery 1995: 777).

0 to 2	weak evidence against null hypothesis.
2 to 6	positive evidence against null hypothesis.
6 to 10	strong evidence against null hypothesis.
> 10	very strong evidence against null hypothesis.

[9] As an ad-hoc test, we constructed a RJ parallel to the RJ_16PAR model, excluding the transition rate parameters that were well attested in the RJ.FULL, such that only the following were left in the model: q12 q14 q15 q16 q23 q24 q25 q26 q35 q45 q51 q52 q53 q54 q62 q64. This model performed much worse than the RJ_16PAR, its log mLh was −159.45; log BF = 2(−143.99−−159.45) = 30.92, providing decisive evidence for RJ_16PAR over this alternative model with 16 transition rate parameters. In addition, RJ is of critical importance for the good fit of the models reported in **Table 1**. Without RJ, the alternative RJ_16PAR model has a log mLh was −171.99, so again much worse than the alternative RJ_16PAR model (log BF 25.08).

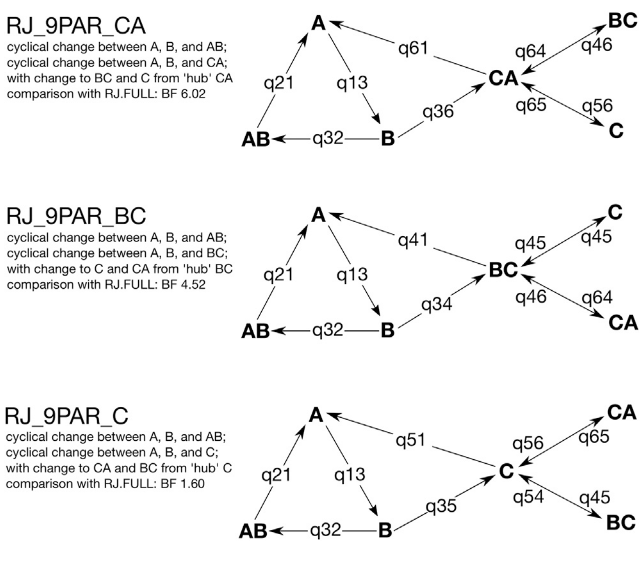

RJ_9PAR_CA
cyclical change between A, B, and AB;
cyclical change between A, B, and CA;
with change to BC and C from 'hub' CA
comparison with RJ.FULL: BF 6.02
A
q61
q64
BC
q46
q21
q13
CA
q36
q65
q56
C
AB
q32
B
RJ_9PAR_BC
cyclical change between A, B, and AB;
cyclical change between A, B, and BC;
with change to C and CA from 'hub' BC
comparison with RJ.FULL: BF 4.52
A
q41
q45
C
q45
q21
q13
BC
q34
q46
q64
CA
AB
q32
B
RJ_9PAR_C
cyclical change between A, B, and AB;
cyclical change between A, B, and C;
with change to CA and BC from 'hub' C
comparison with RJ.FULL: BF 1.60
A
q51
q56
CA
q65
q21
q13
C
q35
q54
q45
BC
AB
q32
B

FIGURE 8 | Visualizations of transition rate parameters included in models RJ_9PAR_CA, RJ_9PAR_BC, and RJ_9PAR_C.

TABLE 1 | Performance of the RJ models, ordered by log mLh. log BFs are calculated for each row using 2(log mLh current model–log mLh RJ.FULL). no. TRP = no. of transition rate parameters; no. 1 rate = no. of models with 1 rate.

Model	log(mLh)	Log BF	No. TRP	No. 1 rate
RJ.FULL	− 150.03		30	989
RJ_9PAR_C	− 149.23	1.60	9	984
RJ_10PAR	− 148.49	3.08	10	962
RJ_12PAR	− 148.21	3.64	12	895
RJ_9PAR_BC	− 147.77	4.52	9	993
RJ_11PAR	− 147.61	4.84	11	942
RJ_9PAR_CA	− 147.02	6.02	9	994
RJ_17PAR	− 144.56	10.94	17	1,000
RJ_13PAR	− 144.41	11.24	13	999
RJ_16PAR	− 143.99	12.08	16	999
RJ_14PAR	− 143.51	13.04	14	998
RJ_15PAR	− 142.95	14.16	15	998

more fully, we constructed RJ_9PAR_BC and RJ_9PAR_C, with the same amount of parameters, but change out of Type-B toward Type-BC (RJ_9PAR_BC) and toward Type-C (RJ_9PAR_C). These models are depicted in **Figure 8**.

Model fit assessment using log Bayes Factors is given in **Table 1**, where each model is compared to RJ.FULL. We can identify several "zones" of model fit. The best fitting RJ model is RJ_15PAR, with a log mLh of -142.95. However, the fit of RJ_14PAR is not significantly different from RJ_15PAR (log BF < 2), and the fit of RJ_13PAR, RJ_16PAR, and RJ_17PAR is only marginally worse than RJ_15PAR's. These models score much better than RJ.FULL, suggesting some manual restrictions on transition rate parameters help model fit. The next "zone" of model fit is that of RJ_12PAR through all RJ_9PARs, with log mLh between −147.02 and −149.23 (log BF 4.42). These scores are significantly worse than those of the first "zone," suggesting that RJ prefers a larger number of free parameters to choose from. Last, we can compare the three models with nine parameters, which differ in the "hub" through which change in types BC, C, and CA is directed. Here, RJ_9PAR_CA and RJ_9PAR_BC have similar fit (BF < 2), whereas RJ_9PAR_C performs significantly worse, log BF 4.42 and 2.92, respectively.

TABLE 2 | Probability of ancestral state estimation of Proto-Indo-European being each of the six states. ~0 are probabilities below 0.05.

Model	A	AB	B	BC	C	CA
RJ.FULL	0.39	0.25	0.19	~0	~0	~0
RJ_17PAR	0.41	~0	0.32	0.14	~0	0.12
RJ_16PAR	0.43	~0	0.29	0.16	~0	0.1
RJ_15PAR	0.45	~0	0.29	0.16	~0	~0
RJ_14PAR	0.59	~0	0.29	~0	~0	0.1
RJ_13PAR	0.55	~0	0.34	~0	~0	0.08
RJ_12PAR	0.77	~0	0.07	~0	~0	0.11
RJ_11PAR	0.87	~0	~0	~0	~0	0.07
RJ_10PAR	0.9	~0	~0	~0	~0	0.06
RJ_9PAR_CA	0.93	~0	~0	~0	~0	0.04
RJ_9PAR_BC	0.85	~0	~0	0.11	~0	~0
RJ_9PAR_C	0.85	~0	~0	~0	0.12	~0

As described above, RJ MCMC estimation can set the transition rate(s) to be equal or shared across q-parameters. This happened in RJ.FULL in 989/100 posterior models, showing there is no evidence for multiple rates even in the models including only 9 or 10 transition rate parameters. This is true for all other RJ models in **Table 1**, with RJ_12PAR being possibly the only exception. One might have expected some evidence for two or more rates as more transition rate parameters were excluded, and the space for model optimization shrank; this is not borne out by the results reported in **Table 1**.

The inference of the ancestral state for Proto-Indo-European is detailed in **Table 2**.[10] From RJ.FULL at the top to RJ_9PAR_C at the bottom, we observe a distinct tendency for Type-A to be reconstructed for Proto-Indo-European with increasing certainty. Type-AB, despite its frequency in the data set, is not estimated to be ancestral in the best-fitting models. Throughout the consecutive exclusion of the transition rate parameters, parameters leading away from Type-B are excluded, decreasing the probability that Proto-Indo-European was Type-B. Note that this shows that the differences in ancestral state estimation across models depend directly on the transition rate parameters that are included. The ancestral state estimations of the three models with nine parameters match the "hub" out of which change between Type-BC, C, and CA is directed.

[10] In another set of six models, we directly estimated the ancestral state of Proto-Indo-European by constraining Proto-Indo-European to be type A, AB, B, BC, C or CA. These are reported in **Supplementary Information S4**. The only model which does not perform worse than RJ.FULL (log BF < 2) is the model when Proto-Indo-European is constrained to be type A, providing additional support for the findings in **Table 2**.

4.2 Manual Models

Alongside using RJ MCMC to establish which transitions are relevant, we also tested models that we constructed manually, inspired by Croft's (1991) NEC and by the changes observed in our dataset (see **Section 3**). The minimum number of parameters that has to be included is six; otherwise, there are states/types that cannot be reached and *BayesTraits* will not run. Therefore, two minimal models are Croft's (1991) NEC or its reverse[11]:

(16) model NEC: A > AB; AB > B; B > BC; BC > C; C > CA;
CA > A

(17) model REV.NEC: A > CA; CA > C; C > BC; BC > B; B > AB,
AB > A

These two models perform worse than the RJ models (log BF 35.72 if we compare REV.NEC with RJ_15PAR, the best-fitting RJ model).[12] The log mLh of the NEC and the REV.NEC models are −169.90 and −160.81 respectively; hence, the REV.NEC model outperforms the NEC model by log Bayes Factor 18.18. **Figure 9**, illustrating the variable rates in the two models, shows that neither model makes a lot of sense given what we know about diachronic change in negative existentials. The NEC model suggests more or less comparable rates toward A, AB, B, BC, and CA, hardly any change toward Type-C, and a lot of change toward Type-A. Croft (1991) and Veselinova (2016) have pointed out that type CA is rare in the languages of the world, which is also true of our sample (6 instances of CA out of 106 languages). In the NEC model, the root of the tree, Proto-Indo-European, is estimated to be Type-CA with 0.99 probability, which explains the massive change away from CA. Given the results in **Section 4.1** (Proto-Indo-European was likely Type-A), the cross-linguistic rarity of CA, and its unstable nature, this result is probably false. REV.NEC shows even larger rate disparity across parameters, with change toward Type-C, BC, and B being far more common than toward the other parameters. Proto-Indo-European is estimated to be Type-A with 0.96 probability in REV.NEC, which does not explain this disparity in rates.

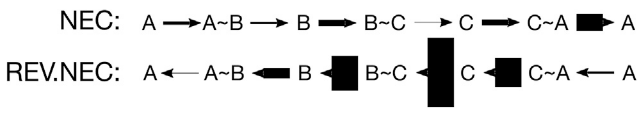


FIGURE 9 | Rates between the six negative existential types in NEC and REV.NEC, the widths of the arrows correspond to rate.

TABLE 3 | Performance of the manual models. Log BFs are calculated for each row using 2(log mLh RJ model–log mLh RJ.FULL). no. TRP = no. of transition rate parameters included (but not necessarily turned on in RJ analysis).

Model	Uniform prior	RJ	Log BF	No. TRP
RJ.FULL	-	– 150.03		30
ALL_THROUGH_B	– 156.41	– 152.35	– 4.64	10
REV.NEC + ALL_A	– 156.44	DNC	-	10
NEC + ALL_B	– 158.49	– 155.51	– 10.96	10
REV.NEC + extra	– 159.24	DNC	-	10
REV.NEC	– 160.81	– 588.8	– 21.56	6
NEC + PARSIMONY	– 167.67	– 156.28	– 35.28	14
NEC	– 169.9	– 588.66	– 39.74	6
NEC + extra	– 171.66	– 165.64	– 43.26	10
PARSIMONY	– 182.99	DNRiBT	-	11

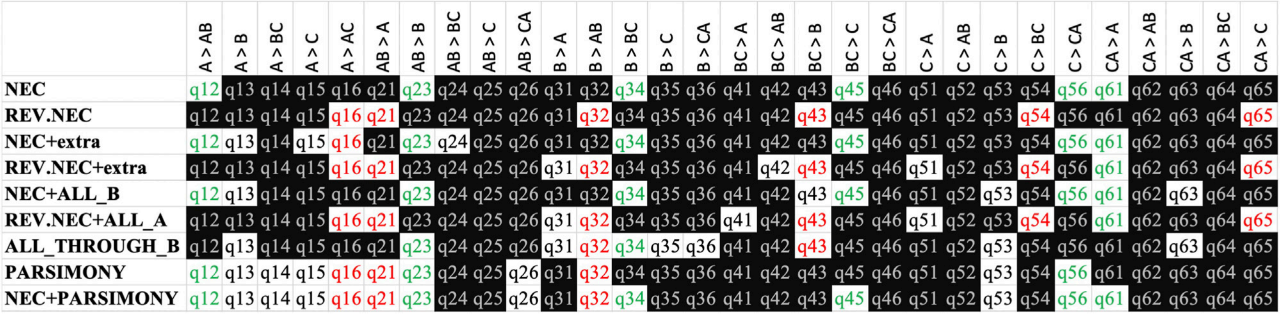
DNC - Does not converge; DNRiBT - Does not run in BayesTraits.

We further tested a range of models informed by the NEC, the set of changes outside NEC, presented in **Section 3**, and the prevalence of attested change toward Types A and AB:

1. **NEC + extra:** parameters included in the NEC plus other attested changes q13 (A > B), q15 (A > C), q16 (A > CA), and q24 (AB > BC) (see (14)).
2. **REV.NEC + extra:** exact reverse of NEC plus extra parameters from NEC (not reversed).
3. **NEC + ALL_X:** parameters included in the NEC and four parameters that lead to Type-X, with separate analyses for each type. For instance, model NEC + ALL_A includes the NEC + q21 (AB > A), q31 (B > A), q41 (BC > A), q51 (C > A).
4. **REV.NEC + ALL_X:** parameters included in REV.NEC and four parameters that lead to Type-X, with separate analyses for each type.
5. **PARSIMONY:** only parameters that can be observed on the tree when a strict parsimony analysis is conducted. This implies looking at the tree presented in **Figure 4** and **Figure 5**, and observing changes leading to languages we have data on, ignoring uncertainty in the ancestral state estimation. This model contains the following parameters:
   a. A > AB q12 attested throughout
   b. A > B q13 Irish, Baltic, Wailgali, Angali, Dhivehi, Pali, Sadri
   c. A > BC q14 Macedonian, Bulgarian, Bengali, Nagamese, Darai
   d. A > C q15 Sinhalese, Kumzari
   e. A > CA q16 Hittite, Kashmiri
   f. AB > A q21 Old High German, Old Persian
   g. AB > B q23 Bahdini Kurdish
   h. AB > CA q26 Talysh
   i. B > AB q32 Balochi, Zazaki
   j. C > B&C q53 Varhadi-Nagpuri, Goan Konkani: Chitpavani
   k. C > CA q56 Standard Goan Konkani
6. **NEC + PARSIMONY:** same as model PARSIMONY, but the missing three NEC parameters (q34 (B > BC), q45 (BC > C), q61 (CA > A)) are added.
7. **ALL_THROUGH_X:** this is a radically different, noncyclical model: all change moves through a single type. For instance, for model ALL_THROUGH_A, five parameters lead out of type A (q12 (A > AB), q13 (A > B), q14 (A > BC), q15 (A > C), q16 (A >

[11] NEC and REV.NEC are two models out of 6! = 720 possible models using six parameters.

[12] Adding RJ to model NEC and model REV.NEC rather than a uniform prior dramatically reduces their fit: RJ.NEC has a log mLh of −588.66; RJ.REV.NEC a log mLh of −588.80; both caterpillar plots look capped, suggesting that ∼−588 is the lowest log mLh possible for this data set and tree set.



	A > AB	A > B	A > BC	A > C	A > AC	AB > A	AB > B	AB > BC	AB > C	AB > CA	B > A	B > AB	B > BC	B > C	B > CA	BC > A	BC > AB	BC > B	BC > C	BC > CA	C > A	C > AB	C > B	C > BC	C > CA	CA > A	CA > AB	CA > B	CA > BC	CA > C
NEC	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
REV.NEC	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
NEC+extra	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
REV.NEC+extra	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
NEC+ALL_B	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
REV.NEC+ALL_A	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
ALL_THROUGH_B	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
PARSIMONY	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65
NEC+PARSIMONY	q12	q13	q14	q15	q16	q21	q23	q24	q25	q26	q31	q32	q34	q35	q36	q41	q42	q43	q45	q46	q51	q52	q53	q54	q56	q61	q62	q63	q64	q65

FIGURE 10 | Overview of the manual models. Shaded transition rate parameters are excluded; non-shaded transition rate parameters are included (but not necessarily turned on). Green indicates the six transition rate parameters that model the NEC; red indicates the transition rate parameters that reverse the NEC.

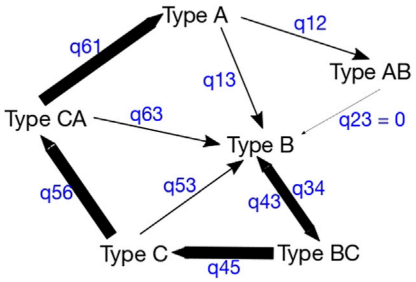


FIGURE 11 | The most frequent posterior transition rate parameter settings for the NEC + ALL_B model, attested in 594/1,000 posterior models. q23, AB > B is excluded from the model. There are two rates, a slow (0.39) and a fast (3.58) rate, marked by edge width.

CA)) and five parameters back into type A (q21 (AB > A), q31 (B > A), q41 (BC > A), q51 (C > A), q61 (CA > A)).

We include in **Figure 10** and **Table 3** only the best fitting models out of the sets above (see for a full description **Supplementary Information S4**). For convenience, models NEC and REV.NEC are also included in **Figure 10**, and **Table 3** lists RJ.FULL again. We ran the manual models twice: once using a uniform prior and once using RJ. Using uniform priors, each transition rate parameter is included and takes its own, individual rate (initially sampled from the uniform prior, 0–100). Using RJ, we again allow transition rate parameters to be turned off, and allow for a unified rate of change across included transition rate parameters. We use both uniform priors and RJ because we want to test 1) the manual models informed by our typological and historical analysis directly using the uniform prior, i.e., without further optimization by RJ; and 2) if the fit of the manual models improves by using RJ, most importantly considering whether a unified rate of change is supported. The RJ manual models also provide us with a better comparison to the RJ models discussed in **Section 4.1. Table 3** shows that RJ manual models outperformed models with a uniform prior by a positive to a large margin (except for NEC and REV.NEC, as discussed above).

The RJ models presented in **Table 1** perform better than the manually constructed models reported on in **Table 3**, regardless of the latter's prior. Three RJ manual models, REV.NEC + ALL_X, REV.NEC + extra, and PARSIMONY, did not converge or did not run. Most manual models outperform NEC and REV.NEC. The NEC + extra and the PARSIMONY models did not fit better than the NEC model (log BF > 2). This probably has to do with how both emphasize change away from Type-A. Type-A is the most common type attested, and including transition rate parameters toward it, especially q21 (AB < A, such as included in RJ models in **Section 4.1**, REV.NEC, REV.NEC + extra), improved model performance.

Out of all the models where we add parameters to those involved in the NEC (NEC + extra, NEC + ALL_X, NEC + PARSIMONY), the NEC + ALL_B model performs best (uniform: log mLh −158.49, RJ: log mLh -155.53). However, NEC + ALL_A and NEC + ALL_AB perform equally well (**Supplementary Information S4**). **Figure 11** illustrates the transition rate parameter settings for the most common posterior model of NEC + ALL_B (attested in 594/1,000 posterior models, with two rates). It shows that the transition rate parameters of the NEC are not all turned on as q23 (AB > B) is turned off. Type-AB becomes an endpoint type, where languages get stuck. Nevertheless, cyclicity still moves from A > B > BC > C > CA > A in this model, with additional parameters leading to type B, out of which only one (BC > B) takes the fast rate.

Out of all models compared in **Table 3**, the best performing one is ALL_THROUGH_B (log BF > 2 with all other models, an illustration is given in **Figure 12**). Again, however, **Supplementary Information S4** states that ALL_THROUGH_B, ALL_THROUGH_C, and ALL_THROUGH_CA perform equally well. We believe that this can be at least partially explained by the distribution of negative existential types in our sample. There is a very skewed distribution toward Type-A and AB and fewer instances of Type-B, BC, C, and CA (see **Figure 2**). Hence, it makes (mathematical) sense to have change leading out of an infrequent type to more frequent types, especially Type-A and AB. REV.NEC + ALL_A performs equally well as ALL_THROUGH_B, showing again that models which allow for transitions toward the commonly attested types are preferred. This result may be distinctive for Indo-

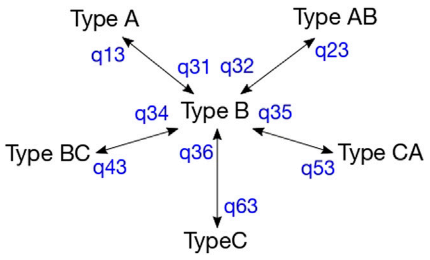


FIGURE 12 | Star-like model ALL_THROUGH_B where all diachronic change is led through a single "hub," type B.

European, with its marked prevalence of Type-A, AB, and B; or it may tentatively suggest that there is something different regarding Type-A, AB, and B, through which a special negative existential marker arises, and Type-BC, C, and CA, through which a negative existential marker replaces in part and takes over standard negation. This hypothesis is further fleshed out in **Section 5**.

None of the manual models we thought to be most relevant outperform the best-performing RJ models presented in **Section 4.1**. In the beginning of this section we already had to conclude that the NEC as formulated by Croft (1991), with six changes, does not suffice in a the modeling context. Regardless of the approach, RJ or manual plus RJ, the best fitting models are those that allow for a "cycle within a cycle," i.e., to have several ways to move between parameters, and not the unidirectional way implied by the NEC.

5 CONCLUSION

The results presented above further underline some of the claims put forth by Veselinova (2013, 2014, 2016; see also Verkerk and Shirtz, forthcoming): the NEC does not represent the entire set of historical changes in the domain of negative existence, and other transitions do occur. It is obvious, however, from the results presented here as well as from the work cited above that the six transitions of the NEC do occur both in Indo-European and in other language families (there is indication for a complete, or a nearly complete, cycle occurring across several subfamilies of Indo-Aryan). This means, we believe, that the NEC is neither "false" nor "unhelpful" for understanding historical changes in the negative existential domain. The NEC is simply not the complete story, and this explains, at least in part, the results of phylogenetic modeling presented in **Section 4**. Unlike what a simple, unidirectional, cyclic model would imply, the domain of negative existence in Indo-European is not easily modeled as a closed subsystem of grammar. Some transitions involve innovations in pre-existing negative existential constructions, but many transitions that were identified here and elsewhere (see also Veselinova 2016: 151ff) involve innovations in constructions *outside* the domain of negative existence that either lead to innovative (negative) existential constructions or influence the classification of already present negative existential constructions.

The changes from outside the domain of negative existence affect the different negative existential construction types in different ways. The nine parameter RJ models from **Section 4.1** suggest cyclical change AB > A > B > AB, with changes involving BC, C, and CA modeled differently. This may be because renewal of standard negation strategies outside of the negative existential domain and the emergence of new existential verbs impacts negative existential types A, AB, and B more directly than types BC, C, or CA. The six possible changes between types A, AB, and B are all attested, suggesting that change between A, AB, and B may in fact be bidirectional. The tipping point seems to be B > BC, marking whether the special negative existential makes its move into standard negation or not, but transitions to BC, C, and CA not predicted by the NEC are attested as well, caused by different types of diachronic changes (A > BC in Macedonian/Bulgarian, AB > BC in Russian/Hawaiian, Veselinova 2014; A > C in Kumzari as described above). This suggestion hence remains tentative at this point, because we have limited information on transitions outside of the NEC, and because types BC and CA (C less so) are rare in Indo-European. The skewed distribution of construction types poses a problem both for analytical work, as we do not yet have enough data to count transitions outside the NEC and categorize them in a sensible way, as well as for the phylogenetic models we constructed in **Section 4**. This is most clear from the ALL_THROUGH_X models (**Supplementary Information S4**), but also from the rest of the results: the rarer constructions (BC, C, and CA) are modeled as ancestral, with change toward common types A and AB. Extending the data set is an obvious solution here, both in terms of Indo-European languages and including other large language families (given Veselinova 2014 study of Polynesian, Oceanic/Austronesian seems to be an obvious candidate).

A separate issue for studying the NEC and negative existentials at large is the occurrence of multiple strategies in the same language (see **Section 1.3**). This is not a very common issue, but frequent enough across the languages of the world (Veselinova 2014, 2016) that we cannot ignore it, as is usually done in typology (see Dryer, 2013 and Comrie, 2013 for two different strategies to "do away with" this issue). Further analytical work should be devoted to finding common diachronic pathways in how multiple strategies arise, coexist, and resolve in the negative existential domain.

There is no simple historical scenario that explains the synchronic variation in the Indo-European domain of negative existence. The reason we propose here, following Veselinova (2016), is that functional domains such as the negative existential domain are not always closed ecosystems of constructions, and innovative constructions of different types may enter these domains. Constructions resulting from these different processes may coexist, each deployed in its own functional niche, or replace each other after some period of time. When there are many pathways leading into the domain from "outside," the source of many constructions will not be a construction "inside" the domain. The more pathways leading into the domain, the messier the historical process may seem and the more difficult it is to model.

Our proposal entails, then, that processes whose origin is "outside" the domain of negative existence result in transitions of a different nature from processes whose origin is "within" the domain, and this leads to difficulties in modeling changes with a unidirectional model. Innovations whose origin is "within" the negative existential domain involve a reanalysis (+actualization) or an extension of an older negative existential construction, which may lead to a change in the typological classification of a construction. This innovation type is the one assumed in the NEC and other unidirectional models. The (re-)classification of older constructions is less central for innovative constructions involving reanalysis (+actualization) or extension of some material which is "outside" the negative existential domain. The result, if these novel constructions end up replacing the older constructions, is a set of transitions with the same endpoint but with different starting points. Such transitions, when frequent enough, mean that unidirectional models are unlikely to have adequate explanatory power.

Despite the fact that some processes with an "outside" source have been illustrated above and elsewhere, confirming or rejecting our proposal requires a more direct analysis and systematic collection of such instances. If this interpretation is on the right track, we should be able to identify negative existential constructions whose source is clearly outside the negative existential domain. These constructions may be innovative on leaf level (e.g., Romanian Type-A construction with "to be found") or in some ancestral stage (e.g., innovative locative copulas in Iranian). Testing this proposal, then, would require further analytic work, highlighting the complementary relationship between phylogenetic comparative and analytic methods in historical morphosyntax. Subsequent phylogenetic modeling could be used to test the hypothesis that change "outside" and "inside" of the domain is dependent on the construction type.

As we have already mentioned, while our sample is sizable and non-sparse in a diachronic typological context, it is still not comprehensive enough to test the hypotheses here to the fullest extent. This will have to wait for a more comprehensive sample of Indo-European languages or another big language family. Further, our results should not be imposed onto other families: different language families may involve different tendencies and differ in the common transitions between construction types (e.g., Dunn et al., 2011). Our results suggest that the NEC is not an accurate general typological hypothesis, as it does not fully explain the distribution of negative existential construction types in Indo-European (see again Veselinova (2014) for problems raised for the NEC from a Polynesian perspective). We do not believe Indo-European is in any way special, and suspect the patterns we found here are attested throughout the languages of the world. Further investigation of different families will bear out this hypothesis.

Finally, we proposed that one reason for this may be the frequency of state transitions arising from "outside" the negative existential domain. Testing this hypothesis would involve "traditional," analytic, studies of language change in specific subgroups of Indo-European. This illustrates the relationship we envision between phylogenetic comparative studies in historical morphosyntax and more "traditional," analytic studies: they inform and complement one another. More generally, the results here suggest that the more the expression of a functional domain interacts with other domains, the more likely are changes that depend less on the current typological classification of the domain, and the more difficult it will be to model changes in it by a unidirectional model.

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.

AUTHOR CONTRIBUTIONS

SS: conceptualization, validation, data collecting, writing (original draft) sections Introduction, Materials & Methods, Typological survey, and Conclusion, writing (review & editing). LT: validation, data collecting, data curation, and writing (review & editing). AV: conceptualization, validation, data collecting, methodology, writing (original draft) sections Materials & Methods, Phylogenetic modeling, writing (review & editing). All authors contributed to the article and approved the submitted version.

FUNDING

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

ACKNOWLEDGMENTS

We would like to thank two reviewers for their feedback; Harald Hammarström for access to data; all the people that provided us with data on negative existentials in the languages of their expertise; and KiGa Unikate for remaining open during a global pandemic.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2021.661862/full#supplementary-material

REFERENCES

Abraham, G., and Abraham, H. (2012). *Varli Phonology and Grammar Sketches*. (*SIL Language and Culture Documentation and Description*. Dallas, Texas: SIL International).

Barðdal, J., and Eythórsson, T. (2012). "Reconstructing Syntax: Construction Grammar and the Comparative Method," in *Sign-Based Construction Grammar*. Editors H. C. Boas and I. A. Sag (Stanford: CSLI Publications), 257–308.

Barðdal, J., and Gildea, S. (2015). "Diachronic Construction Grammar: Epistemological Context, Basic Assumptions and Historical Implications," in *Diachronic Construction Grammar*. Editors J. Barðdal, E. Smirnova, L. Sommerer, and S. Gildea (Amsterdam: John Benjamins), 1–50.

Barðdal, J. (2013). "Construction Based Historical-Comparative Reconstruction," in *Oxford Handbook of Construction Grammar*. Editors G. Trousdale and T. Hoffman (Oxford: Oxford University Press), 438–457.

Beaulieu, J. M., O'Meara, B. C., and Donoghue, M. J. (2013). Identifying Hidden Rate Changes in the Evolution of a Binary Morphological Character: The Evolution of Plant Habit in Campanulid Angiosperms. *Syst. Biol.* 62, 725–737. doi:10.1093/sysbio/syt034

Bhide, V. V. (1982). *A Descriptive Study of Chitpavani: A Dialect of Marathi*. Poona: Deccan College: Doctoral dissertation.

Bordal, V. H. (2017). *Negation of Existential Predication in Swedish: A Corpus Study*. Stockholm, Sweden: Stockholm University. MA Thesis.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., et al. (2012). Mapping the Origins and Expansion of the Indo-European Language Family. *Science* 337 (6097), 957–960. doi:10.1126/science.1219669

Bybee, J., Perkins, R., and Pagliuca, W. (1994). *The Evolution of Grammar: Tense, Aspect and Modality in the Languages of the World*. Chicago: The University of Chicago Press.

Bybee, J. (1988). "The Diachronic Dimension in Explanation," in *Explaining Language Universals*. Editor John A. Hawkins (Oxford: Blackwell), 350–379.

Cathcart, C. A., Hölzl, A., Jäger, G., Widmer, P., and Bickel, B. (2020). Numeral Classifiers and Number Marking in Indo-Iranian. *Lang. Dyn. Change* 10, 1–53. doi:10.1163/22105832-bja10013

Chandralal, D. (2010). *Sinhala*. Amsterdam: John Benjamins.

Chang, W., Cathcart, C., Hall, D., and Garrett, A. (2015). Ancestry-Constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis. *Language* 91 (1), 194–244. doi:10.1353/lan.2015.0005

Christmas, R. B., and Christmas, J. E. (1973). "Clause Patterns in Kupia," in *Patterns in Clause, Sentence, and Discourse in Selected Languages of India and Nepal, Part 2*. Editor R. L. Trail (Dallas, Texas: Summer Institute of Linguistics), 257–343.

Clark, E. V. (1978). "Locationals: Existential, Locative, and Possessive Constructions," in *Universals of Human Language, Vol. 4*. Editor J. H. Greenberg (Stanford: SyntaxStanford University Press), 85–126.

Comrie, B. (2013). "Alignment of Case Marking of Full Noun Phrases," in *The World Atlas of Language Structures Online*. Editors M. S. Dryer and M. Haspelmath (Leipzig: Max Planck Institute for Evolutionary Anthropology).

Creissels, D. (2019). Inverse-locational Predication in Typological Perspective. *Ital. J. Linguistics* 31 (2), 38–106. doi:10.26346/1120-2726-138

Creissels, D.S. (2013). "Existential Predication in Typological Perspective," in *Paper Presented at the 46th Annual Meeting of the Societas Linguistica Europaea* (Split, Croatia), 18–21.

Croft, W., Bhattacharya, T., Kleinschmidt, D., Smith, D. E., and Jaeger, T. F. (2011). Greenbergian Universals, Diachrony, and Statistical Analyses. *Linguistic Typology* 15 (2), 433–453. doi:10.1515/lity.2011.029

Croft, W. (2016). Comparative Concepts and Language-specific Categories: Theory and Practice. *Linguistic Typology* 20 (2), 377–393. doi:10.1515/lingty-2016-0012

Croft, W. (1991). The Evolution of Negation. *J. Ling.* 27, 1–27. doi:10.1017/s0022226700012391

De Smet, H. (2012). The Course of Actualization. *Language* 88, 601–633. doi:10.1353/lan.2012.0056

Dediu, D. (2018). Making Genealogical Language Classifications Available for Phylogenetic Analysis. *Lang. Dyn. Change* 8 (1), 1–21. doi:10.1163/22105832-00801001

Dhakal, D. N. (2012). *Darai Grammar. (Languages of the World/Materials, 489)*. München: Lincom.

Dryer, M. S. (2013). "Order of Subject, Object and Verb," in *The World Atlas of Language Structures Online*. Editors M. S. Dryer and M. Haspelmath (Leipzig: Max Planck Institute for Evolutionary Anthropology).

Dryer, M. S. (2011). The Evidence for Word Order Correlations. *Linguistic Typology* 15 (2). doi:10.1515/lity.2011.024

Dunn, M., Greenhill, S. J., Levinson, S. C., and Gray, R. D. (2011). Evolved Structure of Language Shows Lineage-specific Trends in Word-Order Universals. *Nature* 473 (7345), 79–82. doi:10.1038/nature09923

Dunn, M., Kim Dewey, T., Arnett, C., Eythórsson, T., and Barđdal, J. (2017). Dative Sickness: A Phylogenetic Analysis of Argument Structure Evolution in Germanic. *Language* 3, 1–22. doi:10.1353/lan.2017.0012

Dunn, M. (2014). "Language Phylogenies," in *Routledge Handbook of Historical Linguistics*. Editors C. Bowern and B. Evans (London: Routledge), 190–211.

Fritz, S. (2002). *The Dhivehi Language. (Beiträge zur Südasienforschung, 191)*. Würzburg: Ergon.

Ghatage, A. M. (1966). *Kunabī of Mahāḍ. (A Survey of Marathi Dialects, III)*. Bombay: State Board for Literature and Culture.

Gildea, S., Luján, E. R., and Barðdal, J. (2020). "The Curious Case of Reconstruction in Syntax," in *Reconstructing Syntax* (Leiden: Brill), 1–44. doi:10.1163/9789004392007_002

Givón, T. (1978). "Negation in Language," in *Syntax and Semantics, Vol. 9*. Editor P. Cole (New York: Academic Press), 69–112.

Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika* 82 (4), 711–732. doi:10.1093/biomet/82.4.711

Greenberg, J. H. (1960). A Quantitative Approach to the Morphological Typology of Language. *Int. J. Am. Linguistics* 26 (3), 178–194. doi:10.1086/464575

Haig, G. L. J. (2008). *Alignment Change in Iranian Languages: A Construction Grammar Approach*. Berlin: Mouton de Gruyter.

Hammarström, H., Forkel, R., Haspelmath, M., and Nordhoff, S. (2014). Glottolog 2.3. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: http://glottolog.org (Accessed February 8, 2018).

Harris, A. C., and Campbell, L. (1995). *Historical Syntax in Cross-Linguistic Perspective*. Cambridge: Cambridge University Press.

Haspelmath, M. (2010). Comparative Concepts and Descriptive Categories in Crosslinguistic Studies. *Language* 86 (3), 663–687. doi:10.1353/lan.2010.0021

Haspelmath, M. (2016). The Challenge of Making Language Description and Comparison Mutually Beneficial. *Linguistic Typology* 20 (2), 299–303. doi:10.1515/lingty-2016-0008

Haspelmath, M. (2001). "The European Linguistic Area: Standard Average European," in *Language Typology and Language Universals*. Editor M. Haspelmath (Berlin: de Gruyter), 1492–1510.

Heggarty, P., Anderson, C., Scarborough, M., Bouckaert, R., Jocz, L., Jügel, T., et al. (in review). *Language Trees with Sampled Ancestors Support an Early Origin of the Indo-European Language Family*.

Hendery, R. (2012). *Relative Clauses in Time and Space: A Case Study in the Methods of Diachronic Typology*. Amsterdam: Benjamins.

Jespersen, O. (1917). *Negation in English and Other Languages*. København: A. F. Høst & Søn.

Jordan-Horstmann, M. (1969). *Sadani: A Bhojpuri Dialect Spoken in Chotanagpur. (Indologia Berolinensis, 1.)*. Wiesbaden: Otto Harrassowitz.

Kass, R. E., and Raftery, A. E. (1995). Bayes Factors. *J.Am. Stat. Assoc.* 90 (430), 773–795. doi:10.1080/01621459.1995.10476572

Katz, A. (1996). *Cyclical Grammaticalization and the Cognitive Link between Pronoun and Copula*. Doctoral dissertation, Rice University, Houston, Texas.

Kiparsky, P., and Condoravdi, C. (2006). "Tracking Jespersen's Cycle," in *Proceedings of the Second International Conference of Modern Greek Dialects and Linguistic Theory*. Editors M. Janse, B. D. Joseph, and A. Ralli (Patras: University of Patras), 172–197.

Kulkarni, S. B. (1969). *Descriptive Analysis of Kātkarī Dialect*. Poona: Deccan College. Doctoral dissertation.

Lecoq, P. (1979). *Le dialect du Sivand*. Weisbaden: Harrasowitz.

Levinson, S. C., and Gray, R. D. (2012). Tools from Evolutionary Biology Shed New Light on the Diversification of Languages. *Trends Cogn. Sci.* 16 (3), 167–173. doi:10.1016/j.tics.2012.01.007

Levinson, S. C., Greenhill, S. J., Gray, R. D., and Dunn, M. (2011). Universal Typological Dependencies Should Be Detectable in the History of Language Families. *Linguistic Typology* 15 (2), 35–71. doi:10.1515/lity.2011.034

Levshina, N., Namboodiripad, S., Allassonnière-Tang, M., Kramer, M. A., Talamo, L., Verkerk, A., et al. (in review). 'Why We Need a Gradient Approach to Word Order'. Available at: https://psyarxiv.com/yg9bf.

Liljegren, H. (2016). *A Grammar of Palula*. Berlin: Language Science Press.

Lohndal, T. (2009). "The Copula Cycle," in *Cyclical Change*. Editor E. van Gelderen (Amsterdam: Benjamins), 209–242.

Maurits, L., and Griffiths, T. L. (2014). Tracing the Roots of Syntax with Bayesian Phylogenetics. *Proc. Natl. Acad. Sci. USA* 111 (37), 13576–13581. doi:10.1073/pnas.1319042111

Meade, A., and Pagel, M. (2019). BayesTraits V3.0.2. Available at: http://www.evolution.rdg.ac.uk/BayesTraitsV3.0.2/BayesTraitsV3.0.2.html.

Mundry, R. (2014). "Statistical Issues and Assumptions of Phylogenetic Generalized Least Squares," in *Modern Phylogenetic Comparative Methods and Their 131 Application in Evolutionary Biology*. Editor L. Z. Garamszegi (Heidelberg: Springer-Verlag), 131–153. doi:10.1007/978-3-662-43550-2_6

Nichols, J. (1992). *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Öpengin, E. (2016). *The Mukri Variety of Central Kurdish*. Wiesbaden: Dr. Ludwig Reichert Verlag.

Pagel, M. (1999). Inferring the Historical Patterns of Biological Evolution. *Nature* 401 (10), 877–884. doi:10.1038/44766

Pagel, M., Meade, A., and Barker, D. (2004). Bayesian Estimation of Ancestral Character States on Phylogenies. *Syst. Biol.* 53, 673–684. doi:10.1080/10635150490522232

Pagel, M., and Meade, A. (2006). Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo. *The Am. Naturalist* 167 (6), 808–825. doi:10.1086/503444

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R Language. *Bioinformatics* 20, 289–290. doi:10.1093/bioinformatics/btg412

Paudyal, K. P. (2014). A Grammar of Chitoniya Tharu. München: LINCOM.

Philipa, M. F., Debrabandere, A., Quak, T., Schoonheimen, N., and van der, S. (2003-2009). *Etymologisch Woordenboek Van Het Nederlands*. Amsterdam.

Plank, F. (2011). Where's Diachrony? *Linguistic Typology* 15 (2). doi:10.1515/lity.2011.030

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/.

Sansò, A. (2017). Where Do Antipassive Constructions Come from? *Dia* 34 (2), 175–218. doi:10.1075/dia.34.2.02san

Shirtz, S. (2019). Isomorphic Co-expression of Nominal Predication Subdomains: An Indo-Iranian Case Study. *J.South Asian Languages Linguistics* 6 (1), 59–89. doi:10.1515/jsall-2019-2009

Thackston, W. M. (2006). Kurmanji Kurdish: A Reference Grammar with Selected Readings. Available at: https://sites.fas.harvard.edu/~iranian/Kurmanji/kurmanji_complete.pdf.

Timberlake, A. (1977). "Reanalysis and Actualization in Syntactic Change," in *Mechanisms of Syntactic Change*. Editor C. Li (Austin/London: University of Texas Press).

van der Auwera, J., Krasnoukhova, O., and Vossen, F. (forthcoming). "'Interwining the Negative Cycles'," in *The Negative Existential Cycle*. Editors, Arja Hamari and Ljuba Veselinova (Berlin: Language Science Press), 549–587.

van der Auwera, J. (2011). "Standard Average European," in *The Languages and Linguistics of Europe*. Editors B. Kortmann and J. van der Auwera (Berlin/Boston: Mouton de Gruyter).

van der Auwera, J. (2009). "The Jerspersen Cycles," in *Cyclical Change*. Editor E. van Gelderen (Amsterdam/Philadelphia: John Benjamins).

van der Wal-Anonby, C. (2015). *A Grammar of Kumzari. A Mixed Perso-Arabian Language of Oman*. Leiden, Netherlands: Doctoral Dissertation, Universiteit Leiden.

van Gelderen, E. (forthcoming). "The Negative Existential and Other Cycles: Jespersen, Givón, and the Copula Cycle," in *The Negative Existential Cycle*. Editors A. Hamari and L. Veselinova (Berlin: Language Science Press), 527–548.

Verbeke, S. (2013). *Alignment and Ergativity in New Indo-Aryan Languages*. Berlin: De Gruyter Mouton.

Verkerk, A., and Shirtz, S. (forthcoming). "Negative Existentials in Indo-European: A Typological and Diachronic Overview," in *The Negative Existential Cycle*. Editors A. Hamari and L. Veselinova (Berlin: Language Science Press).

Veselinova, L. (2013). Negative Existentials: A Cross-Linguistic Study. *Rivista di Linguistica* 25 (1), 107–145.

Veselinova, L. N. (2015). "Special Negators in the Uralic Languages," in *Negation in Uralic Languages*. Editors M. Miestamo, A. Tamm, and B. Wagner-Nagy (Amsterdam: John Benjamins), 547–600. doi:10.1075/tsl.108.20ves

Veselinova, L. N. (2016). "The Negative Existential Cycle Viewed through the Lens of Comparative Data," in *Cyclical Change Continued*. Editor E. van Gelderen (Amsterdam: John Benjamins), 139–188. doi:10.1075/la.227.06ves

Veselinova, L. (2014). The Negative Existential Cycle Revisited. *Linguistics* 52 (6), 1327–1389. doi:10.1515/ling-2014-0021

Veselinova, L., and Hamari, A., Eds. (Forthcoming). *The Negative Existential Cycle*. Berlin: Language Science Press.

Wheeler, M. W., Yates, A., and Dols, N. (1999). Catalan: A comprehensive grammar. London: Routledge.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. (2011). Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection. *Syst. Biol.* 60 (2), 150–160. doi:10.1093/sysbio/syq085

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.

ORIGINAL RESEARCH
published: 15 October 2021
doi: 10.3389/fpsyg.2021.674202

A Rational Model of Incremental Argument Interpretation: The Comprehension of Swedish Transitive Clauses

Thomas Hörberg[1,2]* *and T. Florian Jaeger*[3,4]

[1] *Department of Linguistics, Stockholm University, Stockholm, Sweden,* [2] *Department of Computational Science and Technology, KTH Royal Institute of Technology, Stockholm, Sweden,* [3] *Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, United States,* [4] *Department of Computer Science, University of Rochester, Rochester, NY, United States*

A central component of sentence understanding is verb-argument interpretation, determining how the referents in the sentence are related to the events or states expressed by the verb. Previous work has found that comprehenders change their argument interpretations incrementally as the sentence unfolds, based on morphosyntactic (e.g., case, agreement), lexico-semantic (e.g., animacy, verb-argument fit), and discourse cues (e.g., givenness). However, it is still unknown whether these cues have a privileged role in language processing, or whether their effects on argument interpretation originate in implicit expectations based on the joint distribution of these cues with argument assignments experienced in previous language input. We compare the former, *linguistic* account against the latter, *expectation-based* account, using data from production and comprehension of transitive clauses in Swedish. Based on a large corpus of Swedish, we develop a rational (Bayesian) model of incremental argument interpretation. This model predicts the processing difficulty experienced at different points in the sentence as a function of the Bayesian surprise associated with changes in expectations over possible argument interpretations. We then test the model against reading times from a self-paced reading experiment on Swedish. We find Bayesian surprise to be a significant predictor of reading times, complementing effects of word surprisal. Bayesian surprise also captures the qualitative effects of morpho-syntactic and lexico-semantic cues. Additional model comparisons find that it—with a single degree of freedom—captures much, if not all, of the effects associated with these cues. This suggests that the effects of form- and meaning-based cues to argument interpretation are mediated through expectation-based processing.

Keywords: language comprehension, argument interpretation, grammatical function assignment, expectation-based processing, Bayesian inference, self-paced reading, Swedish

OPEN ACCESS

Edited by:
Matthew W. Crocker,
Saarland University, Germany

Reviewed by:
Dave Kush,
University of Toronto Scarborough, Canada
Katja Haeuser,
Saarland University, Germany

***Correspondence:**
Thomas Hörberg
thomas_h@ling.su.se

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

Received: *28 February 2021*
Accepted: *21 September 2021*
Published: *15 October 2021*

Citation:
Hörberg T and Jaeger TF (2021) A Rational Model of Incremental Argument Interpretation: The Comprehension of Swedish Transitive Clauses. Front. Psychol. 12:674202. doi: 10.3389/fpsyg.2021.674202

INTRODUCTION

Language understanding requires comprehenders to integrate incoming information to form hypotheses about the intended structure and meaning of sentences. One of the central components of this process is argument interpretation: determining how the referents of the verb's arguments relate to the events or states expressed by the verb. This determines, for example, whether an argument refers to the *actor* of the event described by the verb, i.e., the most agent-like referent, or the *undergoer* of that event, i.e., the most patient-like referent (see e.g., Dowty, 1991; Primus, 2006). This way, argument interpretation informs us about *who did what to whom*.[1] This interpretation proceeds incrementally, with comprehenders changing their hypotheses about the intended argument role assignment as the sentence unfolds and more information becomes available. For example, upon hearing a sentence starting with "The policeman . . .", the policeman might initially be interpreted as the likely actor of an event to be described. This interpretation will change if the next words are ". . . was arrested . . .". Previous work has found that incremental argument interpretation is affected by a wide range of linguistic cues. This includes both form-based (e.g., case-making) and meaning- or discourse-based properties of the arguments (e.g., animacy, givenness), as well their interactions with verb semantics (e.g., Ferreira and Clifton, 1986; MacWhinney and Bates, 1989; MacDonald et al., 1994; Trueswell et al., 1994; McRae et al., 1998; Kamide et al., 2003; Gennari and MacDonald, 2008; Bornkessel-Schlesewsky and Schlesewsky, 2009; Wu et al., 2010).

While the effects of these cues are now well-attested, questions remain about their theoretical interpretation. Some accounts attribute a privileged role to argument properties that have been linked to increased "accessibility" of argument's referents in memory (Bornkessel and Schlesewsky, 2006; Kuperberg, 2007; Alday et al., 2014; see also Nakano et al., 2010; Szewczyk and Schriefers, 2011). This includes conceptual (e.g., animacy, number) and discourse-based (e.g., givenness, definiteness) properties of arguments (henceforth *prominence* cues) as well as arguments' morphosyntactic properties (e.g., case-marking). For example, some accounts consider prominence and morphosyntactic cues to argument interpretation to either be the *only* information that is taken into account during initial stages of processing (Bornkessel and Schlesewsky, 2006), or to be utilized by a separate combinatorial processing stream (Kuperberg, 2007: 37). On these *linguistic* accounts, other information—such as the plausibility of verb-argument combinations—is either taken into account only at a later stage of processing (Bornkessel and Schlesewsky, 2006), or processed in parallel but by other processing mechanisms (Kuperberg, 2007). Competing, expectation-based accounts attribute the effect of prominence and other cues to implicit expectations based on the distribution of cues in previously experienced language input (e.g., MacDonald et al., 1994; Trueswell et al., 1994; McRae et al., 1998; Narayanan and Jurafsky, 1998; Kempe and MacWhinney, 1999; Vosse and Kempen, 2000, 2009; Tily, 2010; MacDonald, 2013; Bornkessel-Schlesewsky and Schlesewsky, 2019; Rabovsky, 2020). Both linguistic and expectation-based accounts predict that prominence and other cues affect incremental argument interpretation. The two types of accounts differ, however, with respect to whether these effects are taken to be direct, or mediated through expectations. Previous work has found that expectation-based models provide a good fit against human data: across a variety of different structural contexts, expectation-based models correctly predict in which sentences, and where in those sentences, comprehenders will experience processing difficulty (e.g., Demberg and Keller, 2008; Levy, 2008; Boston et al., 2011; Frank and Bod, 2011; Frank et al., 2015). This includes—sometimes complex—interactions between cues that require additional explanations under the linguistic account (we provide examples in Section "Previous Work on Argument interpretation"), as well as qualitative differences in the effects of the same cue across languages (MacWhinney et al., 1984; MacWhinney and Bates, 1989; Desmet et al., 2002, 2006; Acuña-Fariña et al., 2009). This ability to correctly predict the data is particularly noteworthy since the expectation-based account is more parsimonious than the linguistic account: the expectation-based account allows linguistic cues to affect argument interpretation only to the extent that these cues affect the relative probability of different argument interpretations. Since researchers can determine the latter—the objective probabilities—from appropriate language databases, the expectation-based account has few degrees of freedom in predicting language comprehension. In short, previous work suggests that the expectation-based account provides a parsimonious, unifying explanation for a variety of otherwise puzzling processing behaviors. Direct comparisons to the linguistic account on the same data have, however, been lacking. This is the comparison we aim to provide here.

Our general approach to this question is illustrated in **Figure 1**. We develop a *rational expectation-based model* of incremental argument interpretation that links processing times to the Bayesian surprise over changes in argument interpretation as the sentence unfolds. To test this model, we draw on a corpus of transitive clauses in written Swedish (Panel A). The corpus is annotated for a large number of cues previously shown to affect argument interpretation, including morpho-syntactic (e.g., case), syntactic (e.g., clause embedding), prominence (e.g., animacy, definiteness, givenness, deixis) and verb-semantic cues (e.g., volitionality). We then use this corpus to estimate, at different points throughout the sentence, the probability of object-subject (OS) vs. subject-object word order (SO) (Panel B), the former order corresponding to an undergoer-initial interpretation, and the latter to an actor-initial interpretation.[2] These probabilities are taken to approximate

[1] We use the terms argument interpretation and argument role assignment to refer to the process of "assigning" or "linking" arguments to the argument-slots required by the verb (Van Valin, 2006). For example, the transitive verb "kick" requires an actor and an undergoer of the kicking action.

[2] Throughout this paper, we use the assignment of *grammatical functions*—specifically, subjects and (direct) objects—to operationalize the assignment of argument roles—specifically, actor and undergoer roles. The two processes are not

comprehender's expectations—based on previously experienced input—about the underlying argument assignment at different points in the sentence.

We operationalize the cognitive cost associated with changes in these expectations as *Bayesian surprise* (following Kuperberg and Jaeger, 2016; defined below). Once the model is introduced, we use it to derive predictions about comprehension. We use the rational model to design a moving window self-paced reading experiment over sentence stimuli that are predicted to exhibit a large range of Bayesian surprise across stimuli conditions and sentence regions. We test whether Bayesian surprise—derived from the rational model—provides a good *quantitative* and *qualitative* fit against human reading times from this experiment (Panel C). This brings us to the critical comparison that has been lacking in previous work. We compare the fit of the rational model against that of a much less constrained *linguistic model* that can accommodate any type of functional relation between linguistic cues and reading times. This comparison determines whether the rational model—with its hypothesized linear link between Bayesian surprise and reading times—constitutes a parsimonious theory of incremental argument interpretation, explaining effects of various linguistic properties on argument interpretation with a single degree of freedom (the linear effect of Bayesian surprise on RTs). Finally, we investigate how the effects of Bayesian surprise—capturing changes in expectations about *argument interpretation*—relate to effects of word surprisal—an estimate of expectations about *individual words* previously found to be a strong predictor of reading times (e.g., Levy, 2008; Frank and Bod, 2011; Smith and Levy, 2013).

Previous Work on Argument Interpretation

Previous support for an expectation-based account of argument interpretation has come from studies highlighting how the effects of and interaction between various cues on argument interpretation *qualitatively* match to the distribution of those cues in language use.

This tendency is perhaps most thoroughly attested with regard to the linguistic properties of NPs: linguistic properties that make NP arguments less likely to carry the intended argument assignment also tend to negatively affect processing, compared to linguistic properties that make NP arguments expected candidates for the argument assignment. For example, grammatical subjects are cross-linguistically more frequently animate, definite, 1st/2nd person, pronominal and given (i.e., high in prominence), while objects are more commonly inanimate, indefinite, 3rd person, lexical and new (i.e., low in prominence; e.g., in Dutch: Bouma, 2008; Swedish: Dahl and Fraurud, 1996; Dahl, 2000; German: Kempen and Harbusch, 2004; Norwegian: Øvrelid, 2004; for review, see Du Bois, 2003).[3] And, when given an implicit choice, speakers preferentially encode animate and previously mentioned referents as subject, rather than object (e.g., English: Bock and Irwin, 1980; Bock and Warren, 1985; German: Nice and Dietrich, 2003; Greek: Feleki and Branigan, 1999; Japanese: Ferreira and Yoshita, 2003; Tanaka et al., 2011; Tagalog: Sauppe, 2017; Chinese: Hsiao and MacDonald, 2016; for a cross-linguistic review, see Jaeger and Norcliffe, 2009). Prominence properties are thus statistically informative about argument assignment, so that expectation-based accounts predict that prominence properties should affect argument interpretation. In line with these qualitative predictions, subject arguments that are low in prominence (e.g., inanimate), and object arguments that are high in prominence (e.g., definite)—and thus unexpected—tend to cause processing difficulty (Kuperberg et al., 2003; Roehm et al., 2004; Philipp et al., 2008; Nakano et al., 2010; Paczynski and Kuperberg, 2011; Muralikrishnan et al., 2015; Czypionka et al., 2017; Philipp et al., 2017). Similarly, structures that are locally ambiguous with respect to argument functions are easier to process when the arguments are prototypical in animacy or referentiality (e.g., reduced relative clauses: Just and Carpenter, 1992; Trueswell et al., 1994; object-relative clauses: Weckerly and Kutas, 1999; Warren and Gibson, 2002; Traxler et al., 2005; Mak et al., 2006, 2008; Gennari and MacDonald, 2008; Hsiao and MacDonald, 2016; temporarily ambiguous transitive sentences: Kaiser and Trueswell, 2004; Frenzel et al., 2011; Kretzschmar et al., 2012).

Another domain for which this parallelism between patterns in the input and processing is now well-documented is the interaction between verb semantics and NP properties. For example, verbs of cognition and perception, expressing private knowledge and subjective experiences (e.g., *know*, *think*, *see*, or *feel*) and volitional verbs, referring to acts that are based upon intentions of an agent (e.g., *avoid*, *choose*, *steal,* or *seek*), most often require an actor referent that is sentient and/or volitionally acting, and therefore animate. The information that prominence cues carry about argument interpretation therefore to some extent depends on the semantics of the verb. Expectation-based accounts thus predict that comprehenders should take the interplay between NP properties and verb semantics into account during argument interpretation. Research on sentence processing suggests that this is indeed the case: NP arguments with prominence or other semantic properties that are unexpected based on the verb's semantics (Wang et al., 2020; see also Szewczyk and Schriefers, 2013) or that violate the verb's selectional restrictions result in neural signatures that reflect processing costs (e.g., as in *At the homestead the farmer penalized the *meadow for laziness*, Kuperberg et al., 2003; Kim and Osterhout, 2005; van Herten et al., 2005, 2006; Kuperberg et al., 2006, 2007; Bornkessel-Schlesewsky et al., 2011; Paczynski and Kuperberg, 2011, 2012). At the same time, comprehension is

the same, and it is unclear whether sentence understanding *requires* grammatical function assignment, depending on the grammatical system of the language (Van Valin and LaPolla, 1997; Van Valin, 2005; for review, see Bickel, 2010) or even the "depth of processing" (e.g., Ferreira, 2003). However, for the type of clauses we test the rational model against here (Swedish transitive clauses in active voice), grammatical function determines argument assignment (as defined here, and in more detail in Hörberg, 2016: 8–10). If one was to scale the rational model we present here beyond the scope of the present study, it is important to keep in mind the distinction between argument roles and grammatical functions.

[3] For Swedish transitive clauses, for example, Dahl (2000) found that 93.2% of the subjects and 9.9% of the objects were animate, and that 60.7% of the subjects but only 2% of the objects were 1st or 2nd person pronouns.

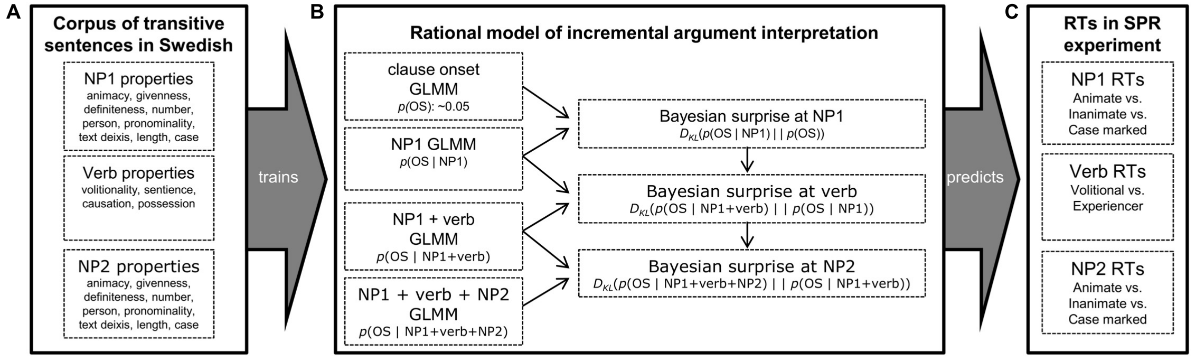

A
Corpus of transitive sentences in Swedish
NP1 properties
animacy, givenness, definiteness, number, person, pronominality, text deixis, length, case
Verb properties
volitionality, sentience, causation, possession
NP2 properties
animacy, givenness, definiteness, number, person, pronominality, text deixis, length, case
trains
B
Rational model of incremental argument interpretation
clause onset GLMM
p(OS): ~0.05
NP1 GLMM
p(OS | NP1)
NP1 + verb GLMM
p(OS | NP1+verb)
NP1 + verb + NP2 GLMM
p(OS | NP1+verb+NP2)
Bayesian surprise at NP1
D_KL(p(OS | NP1) | | p(OS))
Bayesian surprise at verb
D_KL(p(OS | NP1+verb) | | p(OS | NP1))
Bayesian surprise at NP2
D_KL(p(OS | NP1+verb+NP2) | | p(OS | NP1+verb))
predicts
C
RTs in SPR experiment
NP1 RTs
Animate vs. Inanimate vs. Case marked
Verb RTs
Volitional vs. Experiencer
NP2 RTs
Animate vs. Inanimate vs. Case marked

FIGURE 1 | Illustration of the approach taken in the present study. **Panel A:** A corpus of 16,552 transitive clauses in written Swedish is created and annotated for 16 different cues to argument interpretation. **Panel B:** From this corpus, the probability of OS vs. SO order is estimated at four different sentence regions, using all cues available up to that point in the sentence. The Bayesian surprise at each sentence region—quantifying the incremental change in expectations about argument interpretation—is then derived from these probabilities. **Panel C:** The rational model is tested by predicting human reading times at different sentence regions from the model-predicted Bayesian surprise.

facilitated when an NP argument is compatible with the semantic role assigned to it by the verb (e.g., in terms of its animacy, Czypionka et al., 2017; Philipp et al., 2017; or in terms of thematic fit, e.g., Trueswell et al., 1994; Garnsey et al., 1997; McRae et al., 1998).

For additional examples and discussion, we refer to Hörberg (2016). This review of the literature came to the conclusion that expectation-based accounts can in most cases explain the effects of cues to argument interpretation. As compelling as these results might be, however, they do not show whether expectations are sufficient to predict the effects of linguistics cues on argument interpretation.

This caveat also applies to previous computational modeling of argument interpretation: pioneering work showed that competition models trained on the statistical relations between linguistic cues and argument assignment can predict the qualitative patterning of, for example, reading times or eye-movements (e.g., MacDonald et al., 1994; Tabor et al., 1997; McRae et al., 1998; Spivey-Knowlton and Tanenhaus, 1998; Vosse and Kempen, 2000, 2009). The goodness of fit of these expectation-based models was not, however, compared against linguistic models that are *not* constrained by the statistics of the input. It is therefore still unclear how much of the variability in reading times associated with linguistic cues can be reduced to expectations. This is the question we seek to address here.

A RATIONAL MODEL OF INCREMENTAL ARGUMENT INTERPRETATION

We follow previous expectation-based models of sentence processing and assume that comprehenders incrementally update their implicit expectations about the underlying sentence interpretation as new input becomes available (Jurafsky, 1996; Narayanan and Jurafsky, 1998; Crocker and Brants, 2000; Hale, 2001; Levy, 2008). In rational expectation-based models, sentence interpretation involves continuously shifting from a prior to a posterior probability distribution over possible parses, a process known as *Bayesian belief-updating*. The processing cost associated with new input is in part determined by the amount of new information provided by the input—specifically, the degree of shift in expectations or beliefs about the underlying parse (Levy, 2008, 2011). Formally, this shift can be quantified in terms of Bayesian surprise. Bayesian surprise constitutes a principled measure of the prediction error experienced while processing new input (for review, Friston, 2010) and has been linked to attention (Itti and Baldi, 2009) and learning (Ranganath and Rainer, 2003). More recently, it has been proposed to reflect the amount of information gain at a specific level of linguistic representation incurred while processing new input (Kuperberg and Jaeger, 2016; Yan et al., 2017; for a related approach, see Rabovsky et al., 2018).

Bayesian surprise is equivalent to the Kullback-Leibler (KL) divergence of the posterior distribution with respect to the prior distribution. The KL divergence of probability distribution Q from probability distribution P is defined as:

$$D_{KL}\left(P||Q\right) = \sum_{i} \log_2 \left(\frac{P\left(i\right)}{Q\left(i\right)}\right) P\left(i\right) \tag{1}$$

The Bayesian surprise of encountering word w_i is therefore equal to the KL divergence between the posterior probability distribution over possible argument role assignment sARA after seeing w_i and the prior distribution of argument role assignments just prior to that on w_{i-1}:

$$D_{KL}\left(p\left(ARA|w_1 \ldots w_i\right)||p\left(ARA|w_1 \ldots w_{i-1}\right)\right) \tag{2}$$

To calculate the Bayesian surprise of a word, or sequence of words, it is necessary to estimate the relevant prior and posterior probability distributions. This can be done by estimating the relevant distributions from corpus data. Previous rational models have, for example, integrated lexical ngram contexts (e.g., Smith and Levy, 2013; Frank et al., 2015), syntactic (Hale, 2001; Levy, 2008, 2011; Linzen and Jaeger, 2014), or other latent structure (Frank and Haselager, 2006; Frank and Yang, 2018). These models have been found to predict word- or region-based reading times (e.g., Demberg and Keller, 2008; Roark et al., 2009; Boston et al., 2011; Frank and Bod, 2011; Smith and Levy, 2013; Linzen and Jaeger, 2014; Brothers and Kuperberg, 2021) or neural indices of processing costs (e.g., Frank et al., 2015; Willems et al., 2016; Rabovsky et al., 2018; Weissbart et al., 2020; Yan and Jaeger, 2020). The rational model presented here differs from those models in that it is intended to quantify the cognitive cost associated with specifically *argument interpretation*. We thus estimate the incremental Bayesian surprise caused by changes in the relative probability of different argument interpretations. We estimate these probabilities based on the corpus statistics of the types of cues found in previous work to affect argument interpretation.

The present focus on argument interpretation is shared with classic competition models (MacDonald et al., 1994; Tabor et al., 1997; McRae et al., 1998; Spivey-Knowlton and Tanenhaus, 1998; Vosse and Kempen, 2000, 2009; MacDonald and Seidenberg, 2006). In these models, processing cost is a function of the agreement between the relative change in activation of competing argument interpretations from one sentence region to another. This is conceptually closely related to Bayesian surprise, which measures the change in the relative support for competing interpretations. Compared to competition models, however, the rational model presented here is functionally less flexible, making it more parsimonious. Whereas competition models allow non-linear relations between changes in activation and RTs (e.g., mediated through the decision threshold, Δ_{crit}, in McRae et al., 1998), we assume that Bayesian surprise is a linear predictor of reading times (cf. the linear link between word surprisal and RTs demonstrated in Smith and Levy, 2013; but see Brothers and Kuperberg, 2021). This arguably makes the rational model an even stronger test of the expectations-based hypothesis.

We test the rational model against data from the reading of simple transitive clauses in Swedish. In such clauses, information regarding argument role assignment is provided by the grammatical functions of the NP arguments. The subject NP refers to the actor of the event and the object NP to the undergoer of the event. Argument interpretation in such sentences is thus equivalent to the assignment of grammatical functions. We specifically focus on canonical Swedish transitive clauses with subject-object (SO) order, and object-initial sentences with object-verb-subject (OS) order (see Hörberg, 2018). We make the simplifying assumption that comprehenders know—or at least strongly expect—that the sentence they are processing are a transitive clause. For the experiment we present below to test the model, this assumption is plausibly warranted since *all* sentences in the experiment are simple transitive clauses. Previous work has found that comprehenders are sensitive to the distribution of syntactic structures in experiments (e.g., Kaschak and Glenberg, 2004; Fine et al., 2013; Yan and Jaeger, 2020; but see also Harrington Stack et al., 2018). Under this simplifying assumption, the Bayesian surprise over argument interpretations associated with the processing of information available at constituent C_i is:[4]

$$D_{KL}\left(p\left(\mathrm{OS}|C_1 \ldots C_i\right)||p\left(\mathrm{OS}|C_1 \ldots C_{i-1}\right)\right) \quad (3)$$

The Bayesian surprise in Eq. 3 captures the change in expectations about argument interpretation—specifically, whether the first or the second NP is the subject—based on the cues available in constituent C_i (e.g., the second noun phrase, NP2) with respect to the cues available at the previous constituent C_{i-1} (e.g., NP1 and the verb). Here, we test whether this Bayesian surprise predicts the incremental processing difficulty associated with argument assignment during the comprehension of Swedish transitive sentences.

Corpus Data

The rational model is trained on a corpus of 16,552 transitive sentences (Panel A in **Figure 1**) from the Svensk Trädbank treebank (Nivre and Megyesi, 2007). This corpus consists of about 1.3 million words of syntactically annotated Swedish texts from various genres (a subset of the 13 billion word Korp collection, Borin et al., 2012). As described in more detail in the **Supplementary Section 1**, these sentences display a broad range of structural variation. They consist of canonical transitive sentences with SVO order, object-initial transitive sentences with OVS order, and adverbial-initial sentences with VSO or VOS order. They further vary with respect to NP length, number of auxiliary verbs, verb particles, and adverbials, etc. These sentences were annotated for morphosyntactic (e.g., case-marking, auxiliary verbs), syntactic (embedding, verb-initial vs. verb-medial word order), prominence (e.g., animacy, person, givenness, definiteness), and verb semantic cues (e.g., volitionality, sentience). In total, we annotated 16 different cues, each with two or more possible values (for a full list, see **Table 1** and **Supplementary Section 1.3**). The annotated corpus data is available at https://osf.io/rw5nf/.

Estimating the Distributions of Object-Subject vs. Subject-Object Orders

We use this corpus to estimate the Bayesian surprise at three sentence regions: at NP1, at the verb, and at NP2. These estimates are used below to test whether Bayesian surprise predicts reading times at these different sentence regions. As shown in **Figure 1** (Panel B), the Bayesian surprise at these three sentence regions is obtained by estimating the distribution of OS vs. SO order at four different points in the sentence: (i) at the clause onset prior to any sentence input, (ii) after NP1 has

[4] During natural reading, the information available at constituent C_i might include information available through parafoveal preview from upcoming constituents. Under a rational account of reading, this information is expected be weighted less strongly as parafoveal preview has less visual resolution, resulting in increased uncertainty about the input (for related discussion, see Bicknell and Levy, 2012; Kliegl et al., 2013; Bernard and Castet, 2019). In the self-paced reading experiment we present below, this possibility is severely limited since only information regarding the *length* of neighboring words is available parafoveally in a moving-window display.

TABLE 1 | All linguistic cues used to predict OS vs. SO order at four different points in the sentence through separate Bayesian mixed-effects regressions (GLMMs).

	Cue	**GLMM model**		
		NP1 (12 DFs)	**NP1 + verb (22 DFs)**	**NP1 + verb + NP2 (36 DFs)**
NP1	animacy (*animate* vs. *inanimate*)	×	×	×
	givenness (*given* vs. *new*)	×	×	×
	definiteness (*definite* vs. *indefinite*)	×	×	×
	number (*singular* vs. *plural*)	×	×	×
	person/egophoricity (*1st* and *2nd person* vs. *3rd person*)	×	×	×
	pronominal (*pronominal* vs. *lexical*)	×	×	×
	case (*unmarked* vs. *subject* vs. *object*)	×	×	×
	text deixis (*text deictic* vs. *other*)	×	×	×
	length (continuous)	×	×	
Verb	volitional (*volitional* vs. *not*)		×	×
	experiencer (*experiencer* vs. *not*)		×	×
	causative (*causative* vs. *not*)		×	×
	possessive (*possessive* vs. *not*)		×	×
	auxiliary (*auxiliary verb(s)* vs. *not*)		×	×
NP2	animacy (*animate* vs. *inanimate*)			×
	givenness (*given* vs. *new*)			×
	definiteness (*definite* vs. *indefinite*)			×
	number (*singular* vs. *plural*)			×
	person/egophoricity (*1st* and *2nd person* vs. *3rd person*)			×
	pronominal (*pronominal* vs. *lexical*)			×
	case (*unmarked* vs. *subject* vs. *object*)			×
	text deixis (*text deictic* vs. *other*)			×
	length (continuous)			×
Other	embedded (*main* vs. *embedded clause*)	×	×	×
	verb before S and O (*verb-initial* vs. *verb-medial*)	×	×	×
Interactions	animacy × volitional		×	
	animacy × causation		×	×
	person × experiencer		×	×
	givenness × possessive			×
	definiteness × possessive		×	×
	pronominality × possessive			×

The clause onset GLMM is not shown as it only contained the intercept (and the same random effects as the other three GLMMs). The procedure to determine which interactions of cues to include in the model is described in the ***Supplementary Section 2.1****. The total number of degrees of freedom (DFs) for each GLMM are shown at the top of each column. Text deixis concerns whether an NP is a neuter pronominal or demonstrative object (i.e., det and detta – "that") that refers back to a proposition in the immediate left context. Objects that consist of such NPs very frequently occupy the sentence initial position Swedish (Hörberg, 2016, 2018). Text deixis thus serves as a highly reliable cue to argument interpretation.*

been processed, (iii) after NP1 and the verb has been processed, and (iv) after NP1, the verb, and NP2 has been processed. The Bayesian surprise at NP1 is the KL divergence between the distribution of OS vs. SO after NP1 has been processed (ii) and the distribution of OS vs. SO at the clause onset prior to NP1 (i), etc.

These distributions of OS vs. OS order at (i)–(iv) was estimated by fitting four separate Bayesian mixed-effects logistic regressions (GLMMs). Each of these four GLMMs included all annotated cues available up to that point in the sentence. **Table 1** summarizes these cues and which GLMM included them. The predictors and why they were chosen are further motivated in the **Supplementary Section 1.3** (see also Hörberg, 2016).

The use of Bayesian GLMMs with regularizing priors makes it possible to model both cues with gradient effects on argument interpretation (e.g., definiteness) and cues that are fully disambiguating (e.g., case-marking). Regularizing priors "shrink" coefficient estimates toward zero, thereby reducing the chance of overfitting to the data, and facilitate model convergence. We used somewhat weaker priors than is standardly recommended for *data analysis* (e.g., Gelman, 2006; Gelman et al., 2008). *Post-hoc* analyses presented in the **Supplementary Section 2.3**, confirmed that our results do not change over a large range of prior strengths. For the intercept, we used a normal prior with mean −2.994 (the log-odds of the overall proportion of OS order, which is 0.05), and a scale of 2.5. For all other fixed effects, we used Student *t* prior centered at 0 with 30 degrees of freedom and a scale of 5. For the standard deviation of random effects (i.e., the by-genre intercepts), we use a Cauchy prior with location 0 and scale 2. All models were fit with the statistical package *brms* (Bürkner, 2017, 2018) in *R* (R Core Team, 2020). All analysis scripts are available at https://osf.io/rw5nf/.

The fitted GLMMs provide estimates of the probability of OS vs. SO order for any of the four sentence regions and all 16,552 sentences in our corpus. These estimated probabilities can then be plugged into Eq. 3, yielding the predicted Bayesian surprise for the three sentence regions NP1, verb, and NP2. Without refitting the GLMMs, the same procedure can also be used to calculate the predicted Bayesian surprise for any hypothetical combination of linguistic cues, including combinations that were never observed in the corpus. The NP1 + verb + NP2 GLMM, for example, makes predictions about OS vs. SO order for all 2^{36} hypothetically possible combinations of the 36 predictors in the GLMM (see **Supplementary Table 8**).

Illustrating the Model Predictions

To illustrate the predictions of the rational model, we focus on the subset of transitive sentences as well as the subset of NP and verb semantic properties for which the rational model predicts the greatest variation in Bayesian surprise. Predictions for a wider range of structures and properties are presented in the **Supplementary Section 3**. The qualitative predictions we illustrate here also inform the interpretation of the self-paced reading experiment we present below.

Since Swedish lacks case-marking on nouns, OVS sentences with pronominal subjects are morpho-syntactically ambiguous with respect to argument interpretation until the presentation of the post-verbal subject, which disambiguates the sentences toward OVS. These sentences are a perfect test case for investigating how the expectation for a particular argument interpretation varies as a function of the cues of NP1, the verb, and their interactions. Consider the following example sentences taken from the corpus:

1. [De levande $D_{KL} = 0.02$] [fångade $D_{KL} = 0.00$]
The living caught
[jag $D_{KL} = 7.76$].
I
'The living, I caught them.'

2. [En liknande strävan $D_{KL} = 0.29$]
A similar endeavour
[urskiljer $D_{KL} = 1.44$] [han $D_{KL} = 0.16$] också.
discerns he also
'He also discerns a similar endeavour.'

Figure 2 (Panel A) illustrates the Bayesian surprise as well as the probability of OS order at each constituent of example (1). Prior to the beginning of the sentence, SO order is much more likely than OS order, $p(\text{OS}) = 0.047$. The first NP (*De levande*) in (1) is high in prominence (*De levande* is animate and definite). These cues are predicted to make OS order even less likely after NP1 is processed ($p(\text{OS}) = 0.02$). This predicted change in beliefs is, however, small since OS order was unexpected to begin with. As a consequence, Bayesian surprise is close to zero at NP1. Similarly, the semantics of the verb in (1) do not conflict with the strong expectations for SO order either. As a consequence, the probability of an OS order remains low after processing the verb, $p(\text{OS}) = 0.02$, and Bayesian surprise on the verb is predicted to be close to zero ($D_{KL} = 0.00$). This changes, however, when NP2 (*jag*) is encountered. This NP consists of a personal pronoun with nominative case-marking, providing unambiguous evidence for OS order. The rational model thus predicts a large increase in the probability of OS order, $p(\text{OS}) = 0.99$, and correspondingly large Bayesian surprise ($D_{KL} = 7.76$).

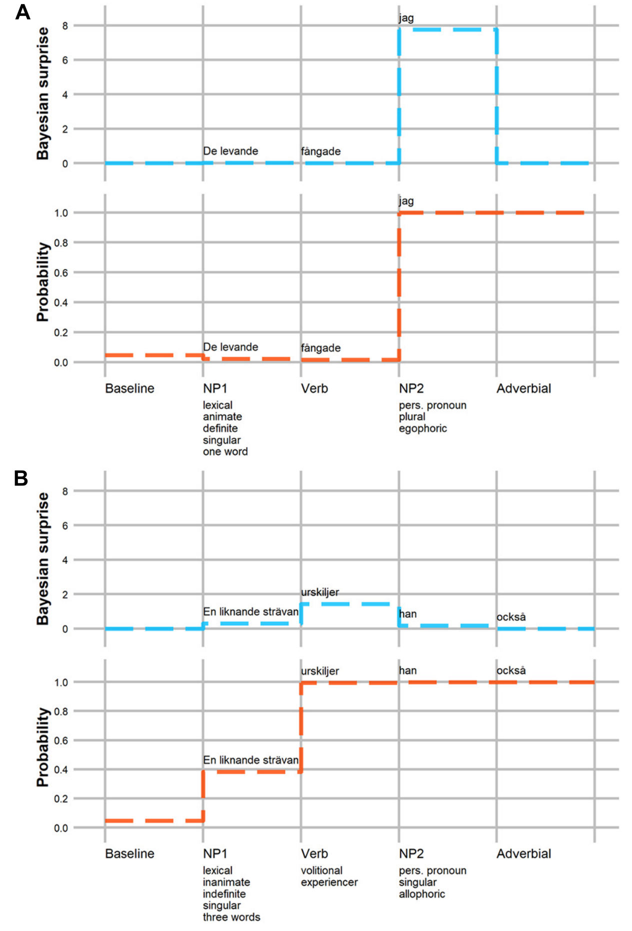


FIGURE 2 | Model-predicted probability of OS order (bottom of each panel) and Bayesian surprise (top) for two example sentences from the corpus. **Panel A:** information that disambiguates the sentence toward OVS order is provided at NP2. **Panel B:** information that speaks in favor of OVS order accrues over the sentence constituents. The relevant cues of each sentence are specified on the x-axis.

In example (2), on the other hand, NP1 is low in prominence (*En liknande strävan* is inanimate and indefinite), and therefore provides some initial evidence for an object-initial interpretation, $p(\text{OS}) = 0.38$. As illustrated in **Figure 2** (Panel B), this is reflected in a small but noticeable increase in Bayesian surprise at NP1 ($D_{KL} = 0.29$). In (2), the upcoming verb *urskiljer* is both volitional as well as experiencer. In combination with the preceding NP1, these verb semantics strongly bias for an object-initial interpretation, $p(\text{OS}) = 0.99$. This large increase in the probability of OS order results in large Bayesian surprise at the verb ($D_{KL} = 1.44$). In this context, the final NP2 (*han*)—a personal pronoun with nominative case-marking like in (1)—does *not* provide much additional evidence for an object-initial interpretation, $p(\text{OS}) = 0.99$. The rational model thus predicts little Bayesian surprise at NP2 ($D_{KL} = 0.16$).

Figure 3 illustrates the predicted effects of a wider range of cues to argument assignment. It shows changes in Bayesian surprise in sentences with a 3rd person lexical NP1 and a 1st person pronoun NP2 (as in example (1) as a function of NP1 and verb semantic cues. Panels A and B show the Bayesian surprise on NP1 and the verb, respectively. Panels C and D summarize Bayesian surprise on NP2 depending on whether that NP is a subject or object pronoun. The patterns in **Figure 3** further confirm that NP prominence cues (animacy, definiteness, number, etc.) interact with verb semantics in determining the probability of OS order, and thus Bayesian surprise. This is visible in Panels B–D, where the difference between the red and blue lines (indicating verb semantics) *strongly* depends on the specific properties of NP1. Also striking is that animacy is the NP1 cue that most strongly interacts with verb semantics. This is evident, for example, in Panel B in a jump in Bayesian surprise for experiencer verbs—but not for volitional verbs—when the preceding NP1 is inanimate, compared to when it is animate. Similarly strong interactions between NP1 animacy and verb semantics are also observed in Panels C-D, though the *direction* of that interaction depends on the case-marking of NP2. Finally, the overall differences between Panels C and D further illustrate how NP2 case-marking affects Bayesian surprise, and how these effects, too, depend on verb semantics (and NP1).

These strong interactions between NP1 animacy, verb semantics, and NP2 case-marking are in line with previous work on subject-object order preferences in Swedish (Rahkonen, 2006; Hörberg, 2018). They are also in line with the observation that animate subjects—in particular 1st/2nd person pronoun subjects—first and foremost occur with experiencer verbs, and secondly with volitional verbs (Dahl, 2000). The information that person (i.e., 1st and 2nd vs. 3rd person) and animacy provide about argument assignment is therefore expected to interact with the semantics of the verb: a 3rd person NP is more predictive of OS order when it co-occurs with an experiencer verb. This is reflected in the strong interplay between NP1 prominence cues and verb semantics, described in more detail below. These patterns of effects motivate the design of the self-paced reading experiment we present next.

TESTING THE PREDICTIONS OF THE RATIONAL MODEL AGAINST HUMAN READING TIMES

We test the predictions of the rational model in a self-paced reading experiment against Swedish transitive sentences with either SVO or OVS order. Sentence stimuli were designed to test the predicted main effects and interactions of constituent order, animacy and verb class shown in **Figure 3**. We chose to manipulate these specific cues—constituent order, animacy, and verb semantics—because we found them to have the strongest effects on Bayesian surprise (for additional details, see Hörberg, 2016). The design of our experiment thus holds constant all other cues to argument assignment listed in **Table 1**.[5] It is important to note, however, that the rational models' predictions are based on *all* cues present in the stimuli, i.e., all properties listed in **Table 1**. In the context of this experiment, it is thus only constituent order, verb semantics, and animacy that affect the predicted Bayesian surprise. The two questions we seek to address are (1) to what extent the differences in Bayesian surprise across items and sentence regions explain differences in reading times, and (2) whether Bayesian captures most (or even all) of the effects of constituent order, animacy, and verb semantics on RTs.

An example item is shown in **Table 2**. The design fully crosses the constituent order (SVO vs. OVS), verb class (volitional vs. experiencer verb) and the animacy of the direct object (inanimate vs. animate). In the critical sentences, the object is always a lexical NP and therefore lacks case-marking. The subject, on the other hand, is a case-marked pronoun. OVS sentences are therefore morpho-syntactically ambiguous with respect to argument interpretation until the presentation of the post-verbal subject, which disambiguates the sentences toward OVS. In SVO sentences, on the other hand, the pronominal subject is positioned sentence-initially, and morphosyntactic information regarding constituent order is provided directly. The Bayesian surprise of each sentence constituent as predicted by the rational model is illustrated in Panel A of **Figure 4**. The model predicts that constituent order and object animacy interact in determining Bayesian surprise on NP1: sentence-initial animate nouns lead to less Bayesian surprise than sentence-initial subject pronouns or inanimate nouns. At first, this might seem counter-intuitive, but the effect stems from a stronger bias *in favor* of an SVO interpretation by a subject pronoun than by an animate noun. Whereas the pronoun provides unequivocal support for SVO order, effectively reducing $p(\text{OS})$ to zero, the animate noun does not change $p(\text{OS})$ as much, keeping it close to the baseline probability of 0.047. An inanimate noun, on the other hand, provides a small effect in the opposite direction, thereby biasing *against* an SVO interpretation. Thus, the rational model predicts somewhat faster RTs for animate nouns in OVS sentences.

Except in sentences with animate objects and volitional verbs, the Bayesian surprise on the verb is somewhat higher

[5] One exception is that we varied the grammatical number (singular vs. plural) of NP1 and NP2 *between* items. This decision was made in order to avoid that all stimuli have identical structure. As confirmed in **Figure 3**, the effect of number on Bayesian surprise—and thus its predicted effect on RTs—is very small. We thus do not discuss it further. The negligible effect of number is also the reason why we do not take these cues into account in the *linguistic* model below, since inclusion of number (or additional predictors that do not vary across items) in the linguistic model would unfairly bias the model comparison *against* the linguistic account (making the linguistic model more complex without commensurate improvements in expected fit). Additionally, our design varied the person of the subject pronoun (1st vs. 2nd) between items. This, however, does not have *any* effect on the predictions of the model since our rational model only contrasts speech act participants (1st and 2nd person) against non-speech act participants (3rd person), as it is this difference that primarily differentiates subjects and objects (e.g., Dahl, 2000).

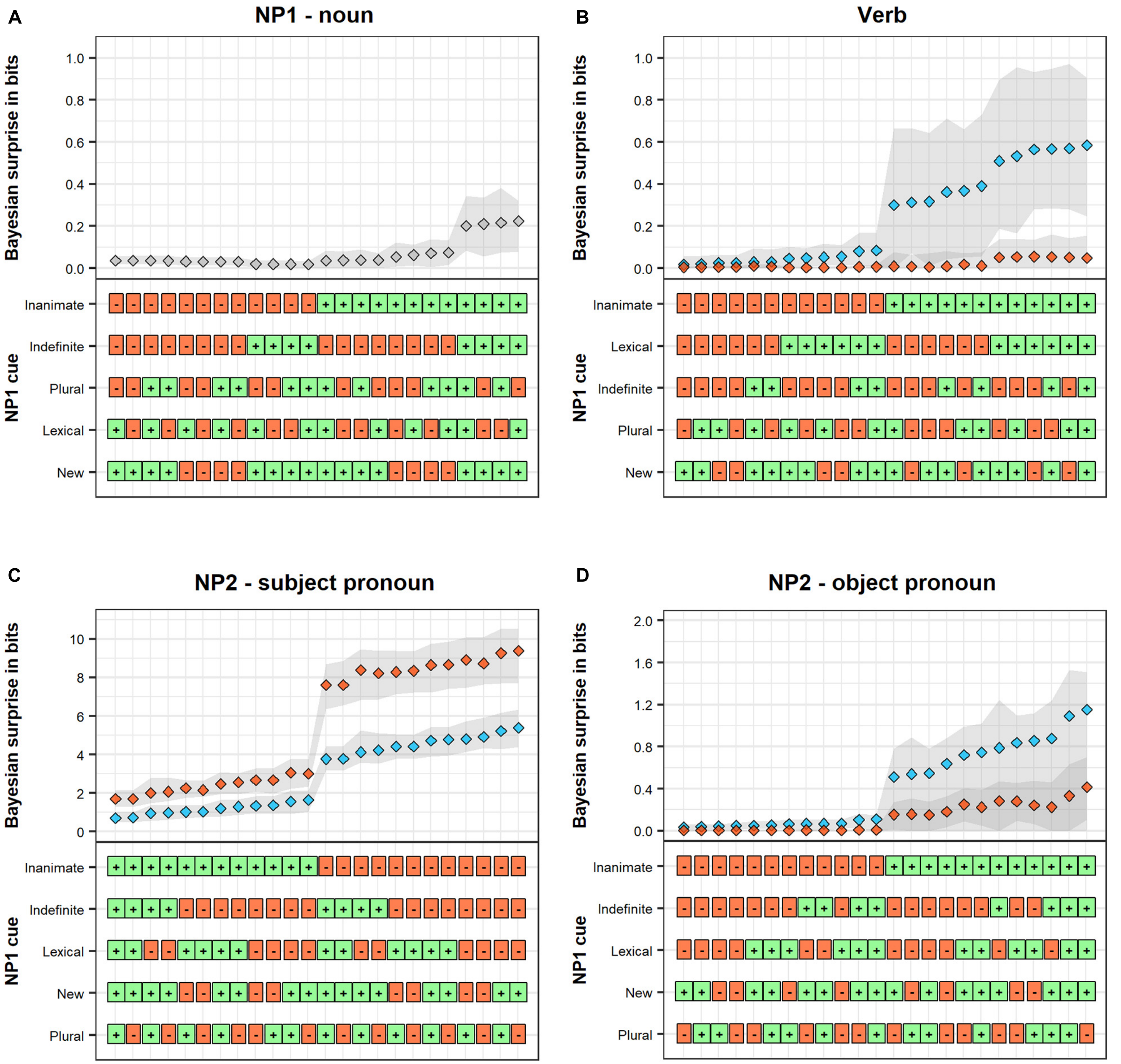


FIGURE 3 | Predicted Bayesian surprise of the NP1, verb, and NP2 constituents of a transitive sentence with a lexical NP1 and a 1st person pronoun NP2. Bayesian surprise is shown as a function of NP1 prominence cues (green square with plus indicates presence of feature) and verb semantics (red for volitional and blue for experience verbs). **Panel A:** Bayesian surprise on NP1 (verb semantic information not yet available). **Panel B:** Bayesian surprise on the verb. **Panel C:** Bayesian surprise for NP2 when NP2 is a case-marked object pronoun disambiguating toward SVO order. **Panel D:** Bayesian surprise for NP2 when NP2 is a case-marked subject pronoun disambiguating toward OVS order. Shaded areas illustrate 89% highest posterior density intervals (HPDIs) of predicted Bayesian surprise, calculated on the basis of the posterior predictions of the underlying GLMMs. Note that the range of the y-axis as well as the order of cues differ between plots. For each panel cues are ordered in decreasing importance from top to bottom.

in OVS than in SVO sentences. This difference is particularly pronounced when NP1 is inanimate: the combination of an inanimate NP and either a volitional or experiencer verb provides some additional support for an OVS interpretation, over and above what is provided by the inanimate NP by itself. However, Bayesian surprise is particularly high in OVS sentences with experiencer verbs when NP1 is inanimate in comparison to when it is animate. Here, the combination of an inanimate 3rd-person NP and an experiencer verb work in concert and provide a lot of support for the object-initial interpretation. The rational model thus predicts somewhat slower verb RTs in OVS compared to SVO sentences, particularly in sentences with

TABLE 2 | Example sentence stimuli of the critical sentences used in the self-paced reading experiment.

Constituent order	Verb	Object animacy	Example
SVO	Volitional	Animate	Jag sparkar killen mitt på smalbenet. 'I kick the guy in the middle of the shin.'
		Inanimate	Jag sparkar bollen mitt upp i krysset. 'I kick the ball right up into the top corner.'
	Experiencer	Animate	Jag glömmer killen sent på kvällen. 'I forget the guy late at night.'
		Inanimate	Jag glömmer bollen mitt på fotbollsplanen. 'I forget the ball in the middle of the soccer field.'
OVS	Volitional	Animate	Killen sparkar jag mitt på smalbenet. 'The guy I kick in the middle of the shin.'
		Inanimate	Bollen sparkar jag mitt upp i krysset. 'The ball I kick right up in the top corner.'
	Experiencer	Animate	Killen glömmer jag sent på kvällen. 'The guy I forget late at night.'
		Inanimate	Bollen glömmer jag mitt på fotbollsplanen. 'The ball I forget in the middle of the soccer field.'

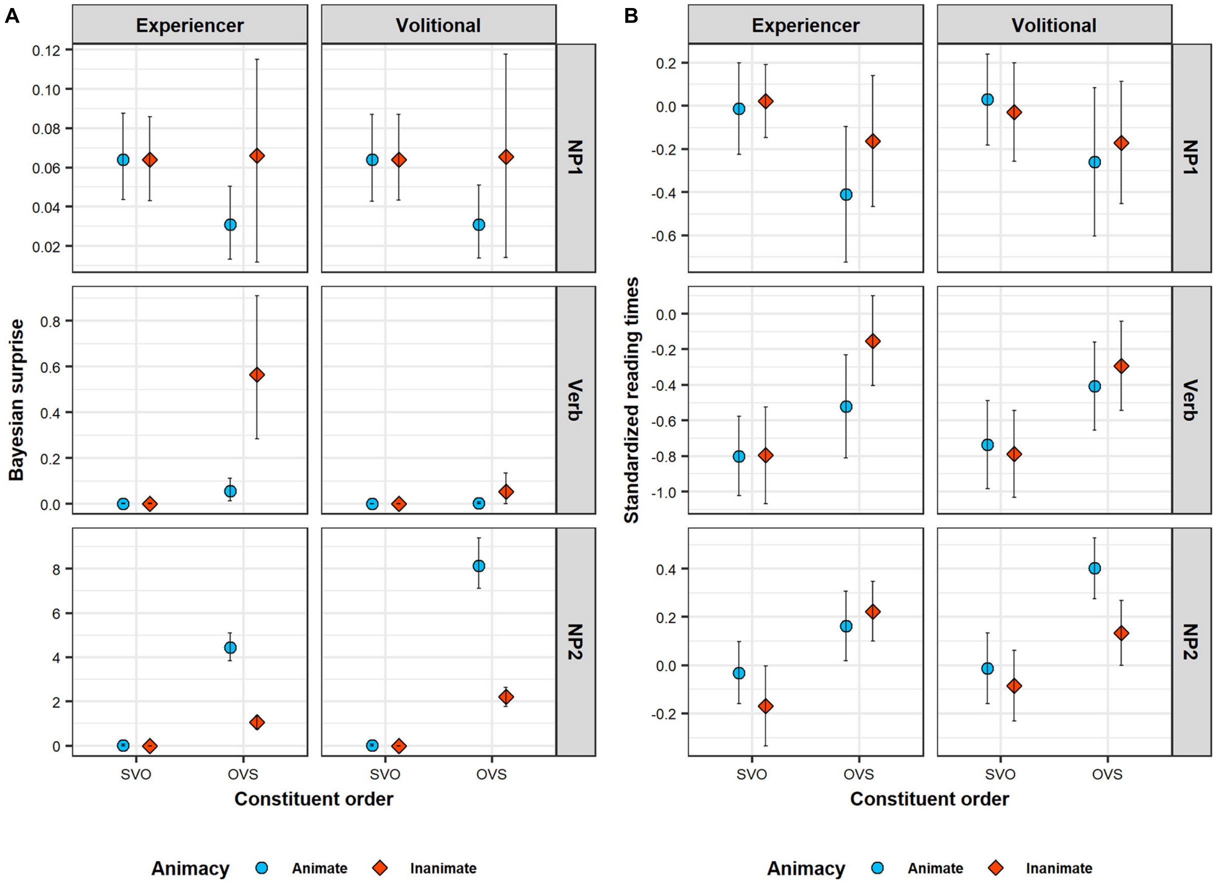


FIGURE 4 | Predicted Bayesian surprise (**Panel A**) compared to length-corrected reading times (**Panel B**) across sentence regions of the critical sentences (rows). Bayesian surprise is derived from the rational model described in Section "A Rational Model of Incremental Argument Interpretation." Length-corrected reading times are grand averages within each design condition. Note that the range of the y-axis differs between sentence regions. Error bars in **(Panel A)** illustrate 89% across-item average HPDIs of predicted Bayesian surprise, calculated on the basis of the posterior predictions of the underlying GLMMs. Error bars in **(Panel B)** are 89% confidence intervals, calculated on the basis of bootstrapping.

inanimate objects. Further, it also predicts slower verb RTs in OVS sentences with experiencer verbs when the object is inanimate rather than animate.

At NP2, Bayesian surprise is substantially higher in OVS than in SVO sentences in general, reflecting an increase in the probability of OVS order due to the disambiguating sentence-final subject pronoun (Hörberg et al., 2013). Importantly, however, this increase is strongly mediated by animacy and verb class. Overall, the effect is weaker when the initial object is inanimate. This is because the inanimate NP co-occurring with the verb has already provided some support for the object-initial interpretation, rendering the OVS interpretation more probable. However, the effect of animacy on the probability of OVS order is much more pronounced in sentences with volitional verbs. In experiencer verb sentences, the combination of a 3rd person NP and an experiencer verb has already provided additional support for the OVS interpretation independently of the object's animacy. The rational model thus predicts slower NP2 RTs in OVS than in SVO sentences. This effect should further be mediated by animacy and verb class in terms of even slower NP2 RTs in OVS sentences with volitional verbs and animate objects.

Materials and Methods

Participants

The self-paced reading experiment was conducted at the Department of Linguistics at Stockholm University. Participants were informed about the experimental procedure and that they could stop at any time without giving reason. They provided written informed consent. A total of 45 participants (15 male) performed the experiment. Their mean age was 28.4 years (*SD* = 9.93), and most of them were students at Stockholm University. Participants received a cinema voucher as reimbursement for their participation.

Materials

All sentences consisted of a one-word NP, a single verb, another one-word NP, and a sentence-final prepositional phrase between three to six words long. The stimulus material consists of 32 items, each of which formed an 8-tuple, representing the 2 × 2 × 2 design (as exemplified in **Table 2**) created from an animate and an inanimate noun, a 1st or 2nd person personal pronoun, and a volitional and an experiencer verb (see **Supplementary Table 9** for a full list of these lexical items).

As evident from **Table 2**, our design implies that a critical item starting with a lexical NP has OVS order. Since there is evidence that readers sometimes learn such experiment-specific statistical contingencies (e.g., Kaschak and Glenberg, 2004; Farmer et al., 2011; Fine et al., 2013; Fraundorf and Jaeger, 2016), we also included three types of SVO filler sentences with lexical subject NPs (see top three rows of **Table 3**). These filler sentences ensure that sentence-initial nouns occur both as subjects as well as objects, thereby avoiding that sentence-initial nouns become an unambiguous cue to OVS order within the context of the experiment. They consisted of 32 three-tuples of SVO filler sentences in which the lexical objects of the critical sentences instead function as sentence-initial subjects, and post-verbal objects consist of 1st or 2nd person pronouns (with object case-marking). For the animate lexical NPs, we used the same volitional and experiencer verbs as in the critical items. For the inanimate lexical NPs, we had to choose different verbs compatible with inanimate subjects. Additionally, we constructed 32 SVO fillers sentences with 1st and 2nd person pronominal NPs. An example stimulus is shown in the final row of **Table 3**. A full list of all stimuli is provided in the **Supplementary Table 9**.

All verbs and noun-verb co-occurrences were attested in the 13 billion word Korp collection (Borin et al., 2012). Within each item, different sentence-final prepositional phrases often had to be used in order for the sentences to make sense. Crucially, however, the two initial words of the phrases that directly follow the second NP were held as constant as possible within each item, always consisting of 2–4 letter function words or adverbs that in most cases were identical across sentences within items.

Each experimental sentence was matched with a comprehension question that probed the event described by the corresponding sentence (i.e., *Sparkar han bollen mitt upp i krysset?*—'Does he kick the ball right up into the top corner?' for the first example sentence in **Table 2**). Half of the comprehension questions were correctly answered with a yes, and the other half were to be answered with a no. In some of the "no"-questions the noun, verb, or the sentence-final prepositional phrase of the corresponding experimental sentence was replaced by another noun, verb, or prepositional phrase. In others, the subject and the object of the sentence were exchanged with each other. Each type of "no"-question occurred equally often.

Materials were arranged into four lists, resulting from a repeated Latin square design based on the design. Each participant read one list. First, a repeated Latin-square design was used to distribute the eight critical sentence conditions of each

TABLE 3 | Example sentence stimuli of the filler sentences used in the self-paced reading experiment.

Constituent order	Verb	Subject animacy	Example
SVO	Volitional	Animate	Killen sparkar mig mitt på smalbenet. 'The guy kicks me in the middle of the shin.'
	Experiencer	Animate	Killen glömmer mig sent på kvällen. 'The guy forgets me late at night.'
	Inanimate subject verbs	Inanimate	Bollen träffar mig mitt i pannan. 'The ball hits me in the middle of the forehead.'

item across four lists so that each list contained two instances of each item. These two instances were chosen such that they did not contain the same nouns or verbs, so that participants did not experience these stimuli as repeated items. This was possible because half of the conditions of each item contained a volitional verb and the other half contained an experiencer verb, and this manipulation was crossed with the animacy of the object. From the perspective of the participant, the two conditions of the items thus appeared unrelated. Across items, we further balanced the number of 1st and 2nd person pronouns in each list.

In order to ensure that participants saw the same sentence-initial nouns in both the subject and object functions, the three SVO filler sentences constructed from each critical item always occurred in a list with a critical OVS sentence from the same item. Within lists, filler sentences with volitional or experiencer verbs always co-occurred with critical sentences with the same verbs. Similarly, filler sentences with inanimate subjects were distributed across lists in a manner that ensured that each inanimate noun both occurred in the subject as well as in the object function. Each list also contained the identical set of 32 SVO filler sentences with 1st and 2nd person pronominal NPs. Each list therefore contained a total of 128 sentences (64 critical sentences, 32 filler sentences varying across lists, and 32 filler sentences that were the same in all lists).

Across participants, each of the four lists were presented in 8 different stimulus orders. Specifically, each list was divided in sequences of eight blocks with 16 sentences each, with item sets, conditions, question types as well as nouns, verbs and pronouns evenly distributed across blocks. Each noun and verb only occurred once within each block. Sentences within a block were presented in a pseudo-randomized fashion so that sentences of the same condition never were presented consecutively. Block order was counterbalanced across participants exposed to each respective list using a Latin square design, ensuring that each block occurred equally often in each of the eight possible list positions. This was done so as to avoid confounding of the conditions of interest with presentation order, since reading times are known to be affected by previous exposure to similar structures (e.g., Fine et al., 2010, 2013; Tooley et al., 2014; Tooley and Traxler, 2018; Yan and Jaeger, 2020).

During data collection, an error in the experimental setup resulted in the first 22 participants being assigned to one of the four lists created from the design factors (order was approximately balanced across those participants). When this error was detected, subsequent participants were exposed to three other lists in a counterbalanced fashion (with 8 participants each, 1 each for each order). Imbalanced data of this type does not violate the assumptions of the analysis approach we employ, and additional statistical analyses not reported here failed to find any significant differences between lists.

Procedure

The experiment was performed on a standard personal computer. Before the experimental trials started, written instructions were presented, and participants performed a practice session of 12 practice trials during which they received feedback on their performance.

Each trial consisted of a visual presentation of the sentence using a self-paced moving window paradigm (Just and Carpenter, 1980; Aaronson et al., 1984). First, a fixation cross appeared on the left-hand side of the screen for 800 ms, followed by a 400 ms blank screen. Then, the full sentence was shown with all non-space characters replaced by a hash symbol (#). Participants revealed each consecutive word of the sentence by pressing the space bar with their preferred hand. At each button press, the currently shown word reverted back to hash symbols as the next word was converted to letters, and button press durations were recorded.

After the presentation of the final word, the screen turned blank for 800 ms, and then the comprehension question was shown. The question remained visible until the participant answered it by pressing "y" for 'yes' or "n" for 'no.' A final blank screen then appeared for 1000 ms before the next trial started. Each experimental block was preceded by a screen that informed that the next block (showing the block number) was about to begin, and the block was started by a space bar press.

Data Exclusion and Correction for Word Length

All participants answered the comprehension questions with an accuracy of 80% or higher. Data from all participants was included in the analysis. Following Jegerski (2014), raw RTs below 100 ms or above 4000 ms (0.3% of the data) as well as RTs from incorrectly answered trials (5% of the data) were excluded from further analysis. Following common procedure, RTs were corrected for word length using linear mixed-effects regression: raw RTs were regressed against word length, while controlling for individual variation in RTs and sensitivity to word length across participants, using a by-participants random intercept and slope for word length (e.g., Fine et al., 2013). The residuals of this model are RTs for which the effect of word length and the individual variation and sensitivity to word length has been regressed out. Length-corrected RTs outside of three standard deviations from the participant's mean were excluded from further analysis (Jegerski, 2014). Taken together, our exclusion criteria removed 7.1% of all RTs from the analysis, leaving 8160 word RTs across the three sentence regions of critical stimuli.[6]

Results

We present three sets of analyses. We start by assessing the effect of Bayesian surprise on reading times in each of the three sentence regions (NP1, verb, NP1). This analysis tests whether the prediction error caused by changes in expectations—under a Bayesian surprise linking hypothesis—predicts variation in reading times. For comparison to previous work, our second set of analyses assesses the effect of linguistic cues—constituent order (OVS vs. SVO), animacy (inanimate vs. animate), verb class

[6] Additional analyses requested by a reviewer used a different approach to word length correction. Following the reviewer's suggestion, we analyzed log-transformed RTs, instead of length-corrected RTs, and included word length as an additional predictor in the main analysis rather than first regressing it out of the RTs. These alternative analyses largely yield the same results as reported here, except that neither Bayesian surprise, nor linguistic cues any longer had significant effects on NP1 RTs. Since these alternative analyses also did not address the convergence issues described in text footnote 7, we do not present them in further detail.

(volitional vs. experiencer), and their interactions—on reading times. These analyses parallel previous work that has investigated effects of linguistic cues on sentence processing (e.g., Ferreira and Clifton, 1986; Trueswell et al., 1994; Gennari and MacDonald, 2008; Wu et al., 2010). This second set of analyses also allows us to assess whether the effects of linguistic cues *qualitatively* follow the prediction of the rational model (whereas our first set of analysis focus on the quantitative fit). Third, we ask whether the effects of linguistic cues on reading times are fully accounted for by Bayesian surprise—the prediction error resulting from *expectations* based on those cues. Additional analyses reported in the **Supplementary Section 7**, show that the effects of Bayesian surprise can*not* be reduced to word-level surprisal—a measure that can be seen as approximating the Bayesian surprise across *all* levels of linguistic processing (Levy, 2008), and that has been found to be a good predictor of reading times (e.g., Frank and Bod, 2011; Smith and Levy, 2013; Brothers and Kuperberg, 2021).

All analyses employed Bayesian mixed-effects linear regression (LMM), again using the package *brms* (Bürkner, 2017, 2018) in *R* (R Core Team, 2020). The use of Bayesian, rather than frequentist, data analysis facilitates convergence under the full random effect structure (for an overview of additional advantages, see Wagenmakers, 2007). We used the standard weakly regularizing priors as recommended in the literature (e.g., Gelman, 2006; Gelman et al., 2008; Stan Development Team, 2017). For fixed effect parameters, we use 3 degree of freedom Student *t* priors with a mean of zero and a standard deviation of 2.5 units (following Gelman et al., 2008). For random effect standard deviations, we use a Cauchy prior with location 0 and scale 2. For random effect correlations, we use an LKJ-Correlation prior with the shape parameter set to 1 (Lewandowski et al., 2009), describing a uniform prior over correlation matrices. All analyses were fit using 12 chains with 1,000 warmup-samples and 4000 post-warmup samples per chain, resulting in 48,000 posterior samples for each analysis. In the **Supplementary Section 6**, we report frequentist analyses paralleling those presented here.

Effects of Bayesian Surprise

In order to evaluate the quantitative relationships between RTs and Bayesian surprise, we conducted separate LMMs for the NP1, verb, and NP2 regions, marked in example (3). Whereas NP1 and verb RTs were RTs of individual words (i.e., the initial single-word NP and the verb), NP2 RTs consisted of the region-averaged RT of the one-word, post-verbal NP and the initial word of the upcoming adverbial. This decision was made prior to data analysis, following the common approach to spill-over effects to capture effects that affect button presses on immediately subsequent words (Mitchell, 1984, among many others). All analyses reported in the main text are based on length-corrected RTs that averaged over the sentence regions exemplified in example (3). For the sake of comparison, the result figures we present below also show region-averaged RTs for the subsequent "adverbial region", consisting of the subsequent two words of the adverbial, as well as RTs of the sentence-final word.

3. [Bollen $_{\text{NP1}}$] [sparkar $_{\text{verb}}$] [jag mitt $_{\text{NP2 region}}$]
   Ball.the kick I middle
   [upp i $_{\text{adverbial region}}$] [krysset $_{\text{final word}}$].
   up in top.corner.the
   'The ball I kick right up in the top corner.'

We used standardized Bayesian surprise as the only fixed-effect predictor in the LMMs. Only by-participant intercepts were included since more complex random effect structures did not converge.[7] Model summaries contain maximum a posteriori (MAP) parameter estimates, corresponding 89% highest posterior density intervals (HPDIs), and the posterior probability ($p_{posterior}$) of the parameter taking on values in the direction of the MAP parameter estimate. These were obtained with the *describe_posterior()* function in R package *BayestestR* (Makowski et al., 2019).

We find very clear evidence for a positive effect of Bayesian surprise for all three sentence regions (NP1: $\hat{\beta}_{MAP} = 14.41$, SE = 6.21, HDPI = [3.53, 23.50], $p_{posterior} = 0.996$; Verb: $\hat{\beta}_{MAP} = 2.35$, SE = 0.48, HDPI = [1.58, 3.12], $p_{posterior} = 1.000$; NP2: $\hat{\beta}_{MAP} = 0.143$, SE = 0.03, HDPI = [0.10; 0.19], $p_{posterior} = 1.000$). These relationships are illustrated in **Figure 5**.

Effects of Linguistic Cues

Next, we analyzed the qualitative effects of linguistic cues (animacy, verb semantics, and constituent order) on the same three sentence regions. This facilitates the comparison to previous work, and sheds further light on the qualitative relation between the reading time patterns associated with linguistics cues and the predictions of the rational model.

The LMM of the NP1 region contained fixed effects for object animacy (sum-coded: 0.5 = animate vs. −0.5 = inanimate), constituent order (sum-coded: 0.5 = SVO vs. −0.5 = OVS), and the animacy × order interaction. The LMMs of the verb and NP2 region contained fixed effects for object animacy (same coding as for NP1), constituent order (same coding as for NP1), and verb (sum-coded: 0.5 = experiencer vs. −0.5 = volitional), as well as the full factorial interactions.[8] All

[7] By-item random intercepts or slopes likely did not converge because lexical content varied as much *within* items (see **Table 2**) as it did across items. There is thus little systematic cross-item variance (see **Supplementary Figure 6**). Additionally, Bayesian surprise varied almost exclusively by condition and thus *within* but not across items: by design, 99.97% (NP1 region), 99.99% (verb), and 99.99% (NP2) of the total variance in Bayesian surprise was accounted for by the eight design conditions. The remaining 0.01–0.03% variance is due to the fact that items differed in whether they employed singular or plural lexical NPs. This is a property that the rational model predicts to have small effect on Bayesian surprise and thus on RTs (see **Figure 3**).

[8] The inclusion of SVO vs. OVS order, which is *not* a linguistic cue but rather the variable to be inferred during argument interpretation, might seem counterintuitive. Here constituent order and its interaction with animacy together encode what cues occur in the three sentence regions of our stimuli. The predictor we refer to as constituent order encodes whether the NP1 is a case-marked pronoun or a non-case marked lexical noun.

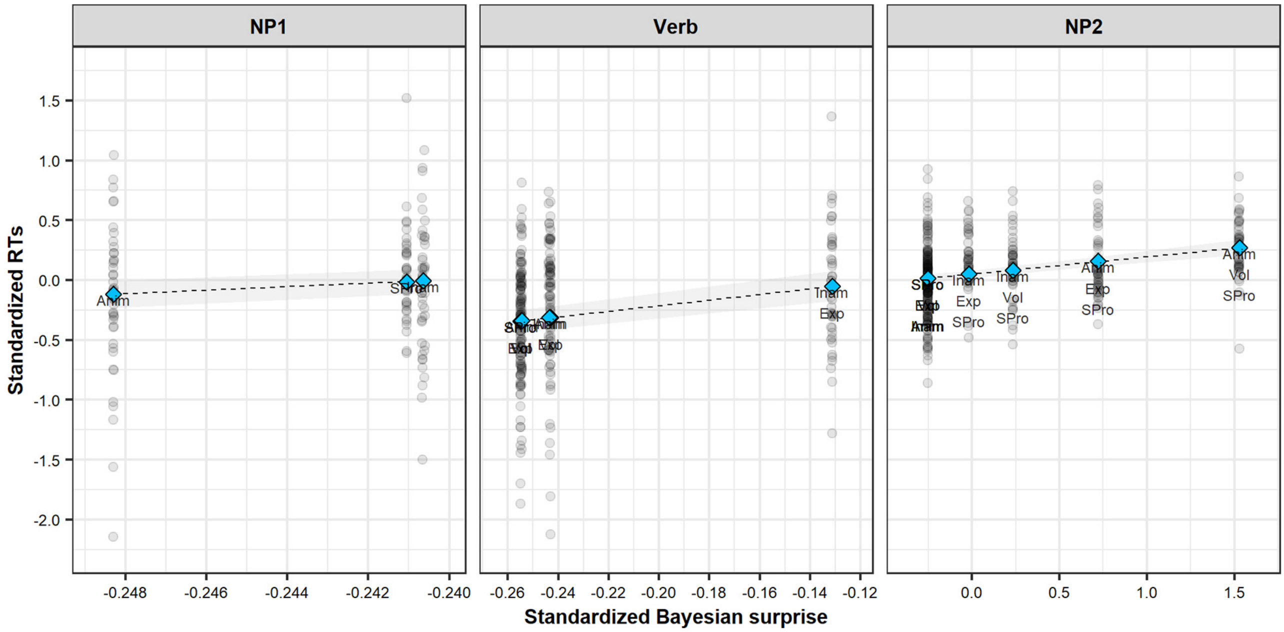


FIGURE 5 | Bayesian LMM MAP estimates of standardized RTs in each sentence region as function of standardized Bayesian surprise. Shaded areas illustrate 89% HPDIs. Shaded dots represent individual item-level RTs. Texts describe item-level sentence properties. Anim: animate; Inam: inanimate; Pro: pronominal; S: subject; O: object; Vol: volitional; Exp: experiencer.

LMMs also included the maximal random effect structure by-participants—i.e., by-participant random intercepts and slopes for all predictors in the analysis. No by-item random effects were included, since inclusion led to failure to converge (see text footnote 7).

The results are summarized in **Table 4**. **Figure 6** illustrates predicted RTs across sentence regions, as a function of linguistic cues.

For the NP1 region, we found a main effect of constituent order: length-corrected RTs were slower in SVO sentences (where NP1 is a case-marked pronoun) than in OVS sentences (where NP1 is a lexical noun). There was also evidence for an interaction between constituent order and object animacy, although this evidence did not reach the conventional frequentist threshold of significance. Simple effect analyses (see **Table 4**) showed that the effect of constituent order is primarily driven by the shorter RTs for animate object nouns in OVS sentences.

Of note is that the linguistic LMMs could—in theory—accommodate effects of animacy and constituent order in any direction and of any magnitude. Yet, this analysis finds that RTs on NP1 pattern in ways that closely resemble the qualitative predictions derived from the rational model of argument interpretation presented in Section "Testing the Predictions of the Rational Model Against Human Reading Times." **Figure 4** provides a direct comparison between patterns of predicted Bayesian surprise (Panel A) and average RTs (Panel B). In line with the predictions of the rational model, RTs are shorter for animate NP1s on OVS sentences, compared to all other conditions. Notably, these lexical NP1s in OVS sentences were read faster even than subject pronouns NP1s (in SVO sentences). This is the case despite the fact that subject pronouns are case-marked and thus morphologically unambiguous with respect to argument interpretation. Under the rational model, this makes sense: sentence-initial animate NPs do not provide much support in favor of either argument interpretation, leading to low Bayesian surprise. A subject pronoun, on the other hand, provides unequivocal support for an SVO interpretation. This support goes against the small but nevertheless existing expectation for OVS order, leading to comparatively larger Bayesian surprise (Similarly, an *in*animate lexical NP1 provides some additional support in favor of an OVS interpretation, violating the overall baseline expectation for SVO order, also leading to higher Bayesian surprise than the animate lexical NP1).

For the verb region, we again found a main effect of constituent order, but in the opposite direction than for the NP1 region: RTs were slower in OVS sentences (where the verb follows a non-case marked lexical noun) than in SVO sentences (where the verb follows a case-marked subject pronoun). In addition, evidence for an interaction of this effect with animacy reached the conventional frequentist threshold of significance. Simple effect analyses (see **Table 4**) found that object animacy affected verb RTs primarily for sentences with OVS order: verb RTs in OVS sentences were slower when the verb was preceded by an inanimate object noun than when it was preceded by an animate object noun. Simple effects analyses further showed that this effect of object animacy on verb RTs in OVS sentences was particularly pronounced for experiencer verbs.

For the verb region, too, the linguistic LMM thus returns effects that follow the qualitative predictions of the rational model (see **Figure 4**). The combination of an inanimate NP1 and either

TABLE 4 | Results of the Bayesian linear mixed-effects regressions (LMMs) of region-averaged length-corrected RTs investigating the effects of linguistic cues over the NP1, verb, and NP2 region.

Region	Predictor	$\hat{\beta}_{MAP}$	S.E. $(\hat{\beta})$	HDPI$_{lower}$	HDPI$_{upper}$	$p_{posterior}$
NP1	Intercept	−0.04	0.06	−0.14	0.06	0.740
	Constituent order (*OVS* vs. *SVO*)	0.14	0.06	0.05	0.25	0.991
	Object animacy (*anim.* vs. *inanim.*)	−0.04	0.03	−0.10	0.01	0.888
	Order × Animacy	0.09	0.07	−0.01	0.21	0.919
	SVO/Animacy	0.01	0.05	−0.07	0.08	0.553
	OVS/Animacy	−0.09	0.05	−0.17	−0.01	0.965
Verb	Intercept	−0.30	0.06	−0.40	−0.20	1.000
	Constituent order (*OVS* vs. *SVO*)	−0.25	0.04	−0.32	−0.19	1.000
	Object animacy (*anim.* vs. *inanim.*)	−0.06	0.04	−0.12	0.00	0.944
	Verb (*volitional* vs. *experiencer*)	−0.01	0.04	−0.07	0.05	0.594
	Order × Animacy	0.16	0.08	0.03	0.28	0.972
	Order × Verb	−0.01	0.08	−0.15	0.11	0.581
	Animacy × Verb	−0.10	0.08	−0.23	0.02	0.900
	Order × Animacy × Verb	0.11	0.16	−0.13	0.37	0.780
	SVO and Volitional/Animacy	0.04	0.08	−0.09	0.16	0.665
	SVO and Experiencer/Animacy	−0.01	0.08	−0.13	0.12	0.528
	OVS and Volitional/Animacy	−0.06	0.08	−0.18	0.06	0.776
	OVS and Experiencer/Animacy	−0.23	0.08	−0.35	−0.09	0.996
NP2	Intercept	0.07	0.02	0.05	0.11	1.000
	Constituent order (*OVS* vs. *SVO*)	−0.18	0.03	−0.23	−0.12	1.000
	Object animacy (*anim.* vs. *inanim.*)	0.06	0.03	0.01	0.11	0.971
	Verb (*volitional* vs. *experiencer*)	−0.04	0.03	−0.09	0.01	0.902
	Order × Animacy	0.00	0.06	−0.09	0.11	0.560
	Order × Verb	0.03	0.06	−0.08	0.13	0.654
	Animacy × Verb	−0.08	0.06	−0.18	0.02	0.893
	Order × Animacy × Verb	0.23	0.13	0.03	0.44	0.966
	SVO and Volitional/Animacy	0.05	0.06	−0.06	0.15	0.769
	SVO and Experiencer/Animacy	0.08	0.06	−0.02	0.19	0.908
	OVS and Volitional/Animacy	0.15	0.06	0.05	0.25	0.993
	OVS and Experiencer/Animacy	−0.04	0.07	−0.14	0.07	0.737

For each region, we show both the main LMM (top) and simple effects re-parameterization of the same LMM. The first number column provides the maximum a posteriori probability estimates for each coefficient ($\hat{\beta}_{MAP}$), the standard error of that estimate, the lower and upper bounds of the 89% highest posterior density interval (HPDI, following Kruschke, 2014), and the posterior probability that the effect has the sign of the MAP estimate. Effects that meet conventional frequentist significance criteria are highlighted by shading ($p_{posterior}$ > 0.95).

a volitional or experiencer verb provides some support for an OVS interpretation, over and above what is provided by the inanimate NP by itself. In contrast to what is observed for NP1 RTs, the rational model thus predicts verb RTs to be slower in OVS with inanimate objects. Further, because experiencer verbs frequently occur with 1st or 2nd person subjects (Dahl, 2000), the co-occurrence of a 3rd-person initial NP and an experiencer verb provides additional support for OVS order. Verb RTs are therefore predicted to be particularly slow in OVS sentences with an experiencer verb and an inanimate NP1.

Finally, for the NP2 region, we found a main effect of constituent order in the same direction as on the verb: NP2 RTs were slower in OVS sentences (where NP2 is a subject pronoun) than in SVO sentences (where NP2 consists of an object noun). There was also a main effect of animacy, showing that NP2 RTs overall are slower when the object noun is animate, irrespective of the position of the object. These effects need to be interpreted in light of the three-way interaction between constituent order, object animacy, and verb class. Simple effect analyses (see **Table 4**) found that NP2 RTs in OVS sentences are slowed down when the sentence-initial noun is animate—but only in sentences with volitional verbs. In OVS sentences with experiencer verbs, this animacy-induced slow-down instead already occurred on the verb.

Again, this RT pattern is qualitatively in line with the predictions of the rational model of argument interpretation (see **Figure 4**), and can be explained in terms of changes in the expectation for OVS word order. The sentence-final subject pronoun in OVS sentences disambiguates the sentence interpretation toward OVS. The slowdown on the NP2 for OVS sentences in comparison to SVO sentences is a predicted consequence of this change in expectations. The magnitude of this change depends on the extent to which NP1 animacy and verb class provides support for an OVS interpretation before NP2 has been encountered. In particular, an animate NP1 combined with a volitional verb provides no additional support for OVS

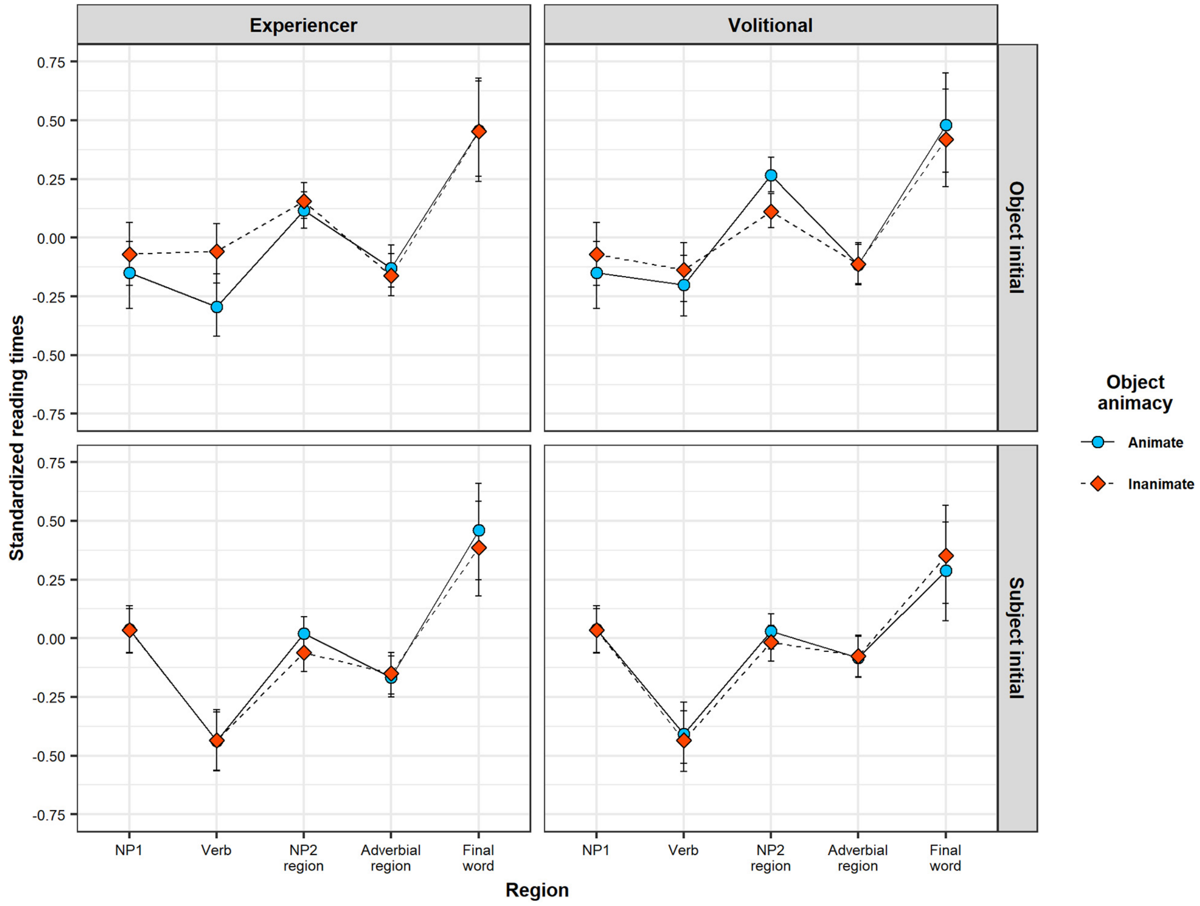


FIGURE 6 | Bayesian LMM model MAP estimates of standardized RTs in each sentence region as function of animacy, separated by verb class (experiencer vs. volitional) and constituent order (OVS vs. SVO). Error bars illustrate 89% HPDIs.

word order prior to the presentation of NP2. The sentence-final subject pronoun is therefore highly unexpected in such sentences, resulting in particularly slow RTs.

Can Bayesian Surprise Capture the Effects of Linguistic Cues on Reading Times?

In order to evaluate how much of the effects of linguistic cues Bayesian surprise can account for, we performed separate model comparisons for each of the three sentence regions. For each region, we refit the separate analyses of (i) Bayesian surprise and (ii) linguistic cues presented above but while including the full random effect structure from *both* analyses. Following recommendation for model comparison, the linguistic LMM and the Bayesian surprise LMM thus only differ in terms of their fixed effects.

We compare LMMs in terms of their out-of-sample predictive accuracy—the LMM's leave-one-out cross-validation information criterion (LOOIC—see Watanabe, 2013; Gelman et al., 2014; Vehtari et al., 2017). This LOOIC is related to an LMM's leave-one-out cross-validated log predictive density or $elpd_{LOO}$ (LOOIC = $-2 \times \text{elpd}_{LOO}$) in the same way that an LMM's deviance is related to its log-likelihood (deviance = $-2 \times$ log-likelihood). Smaller LOOICs indicate better predictive accuracy, similar to traditional deviance measures of model fit (e.g., the AIC or BIC). Unlike measures based on the log-likelihood, the elpd measures how well the LMM generalizes to held-out data. This takes into account the models' functional flexibility (which can lead to good fit on the observed sample but poor generalization to novel data). Additional analyses presented in the **Supplementary Table 11**, report model comparisons based on likelihood ratios, which captures the model's fit against the finite sample the researcher analyses. Unlike model comparison based on likelihood ratios, the elpd is not limited to comparison of nested models. This allows us to directly compare the linguistic and Bayesian surprise LMMs without comparing them indirectly through pairwise comparison to a superset LMM with the predictors from both LMMs.

The goal of the model comparison we conduct here is to assess to what extent reading time predictions based on linguistic cues are accounted for by Bayesian surprise with a single degree of freedom (DF). This conclusion would be supported if the Bayesian surprise LMM outperforms the linguistic LMM, or if the Bayesian surprise and linguistic LMMs do *not* differ in terms of their elpd. The latter outcome would indicate that the two LMMs achieve the same predictive accuracy but the Bayesian surprise LMM would do so with fewer DFs: each Bayesian surprise LMM only has a single DF in predicting RTs (all other DFs are fixed based on the corpus data, as described in Section "A Rational Model of Incremental Argument Interpretation"); the linguistic LMMs, however, have up to 7 DFs (resulting from the 2 × 2 × 2 design). If, however, the linguistic LMM outperforms the Bayesian surprise LMM, this would argue that the linguistic model—with its additional flexibility—can capture important predictive information about reading times that are not captured by the rational model that links changes in expectations to reading times.

We report differences in the LOOIC (ΔLOOIC). Following Bushong (2020), we consider a difference in LOOIC of more than 2.5 times of its estimated standard error (i.e., estimated differences outside the 99% error interval of the difference) as evidence for a difference in predictive accuracy between the models. **Table 5** summarizes the results. The Bayesian surprise LMM has a numerically better LOOIC than the linguistic LMM for both the NP1 and NP2 region, and vice versa for the verb region. However, all of these numerical differences fall well within the 99% interval. We thus do not have evidence that the two LMMs differ in their predictive accuracy at any of the three sentence regions. This suggests that Bayesian surprise largely captures the same predictive information about RTs as a model including the individual linguistic cues.

GENERAL DISCUSSION

Previous research has shown that the incremental interpretation of arguments is based on an interplay between form-based morpho-syntactic, meaning-based semantic and discourse-pragmatic NP properties, and verb-semantic cues (e.g., MacWhinney and Bates, 1989; MacDonald et al., 1994; Bornkessel and Schlesewsky, 2006; Bornkessel-Schlesewsky and Schlesewsky, 2009). On *linguistic accounts,* some of these cues—e.g., the prominence properties of arguments—are assumed to have a privileged role in language comprehension (Bornkessel and Schlesewsky, 2006; Kuperberg, 2007; Alday et al., 2014; see also Nakano et al., 2010; Szewczyk and Schriefers, 2011). For example, these cues might be assumed to be processed first, prior to other cues (Bornkessel and Schlesewsky, 2006), or to be processed by a separate mechanism (Kuperberg, 2007: 37).

In contrast, *linguistic accounts* attribute the effects of linguistic cues to implicit expectations based on the joint distribution of cues and argument assignments in previously experienced language input (e.g., MacDonald et al., 1994; Trueswell et al., 1994; McRae et al., 1998; Narayanan and Jurafsky, 1998; Kempe and MacWhinney, 1999; Vosse and Kempen, 2000, 2009; Tily, 2010; MacDonald, 2013; Bornkessel-Schlesewsky and Schlesewsky, 2019; Rabovsky, 2020).

The present study compared linguistic accounts of incremental argument interpretation, in which cues to argument interpretation have a direct effect, to expectation-based accounts, in which the cues are mediated through expectations. To this end, we developed a rational expectation-based model of incremental argument interpretation in simple transitive clauses in Swedish and then tested this model against reading time data from a self-paced reading experiment. The rational model predicts processing costs at different sentence regions for different constituent orders, morpho-syntactic and prominence properties of the NP arguments, and semantic properties of the verb. It estimates the incremental change in expectations about argument interpretation as a function of the cues provided by the subsequent sentence constituents (i.e., NP1, verb, and NP2), quantified in terms of Bayesian surprise—a shift from a prior to a posterior probability for a particular argument assignment.

We tested some of the most prominent predictions of this rational model against processing times in a moving window self-paced reading experiment of transitive sentences in Swedish. The model predicts that the processing difficulty associated with argument interpretation in locally ambiguous sentences depends on an interplay between prominence properties of the initial NP and the semantic class of the verb (see **Figure 3**). In particular, processing difficulty is predicted to vary as a function of the animacy of NP1 and whether the sentence verb is volitional or experiencer. We therefore used locally ambiguous OVS sentences and unambiguous SVO sentences with lexical objects and case-marked subject pronouns that varied with respect to the animacy of the object and whether the verb was volitional or experiencer. The results of the experiment confirmed most of the predictions of the rational model both quantitatively—Bayesian surprise is a significant predictor of within-region RTs (**Figure 5**)—and qualitatively—the effects of linguistic cues on RTs pattern similarly to their effect on Bayesian surprise (**Figure 4**). In all regions, higher Bayesian surprise predicted higher reading times, and the observed patterns of effects could be explained in terms of changes in the expectation for OVS order (see Section "Effects of Linguistic Cues"). This pattern of results is predicted under the hypothesis that listeners incrementally update their expectations about argument interpretations, with larger changes in expectations requiring more processing time.

In order to more directly compare the linguistic account of argument interpretation to the expectation-based account, we further investigated whether Bayesian surprise can predict RTs just as well as a model in which linguistic cues can have arbitrary direct effects on RTs. We found no evidence that direct effects of linguistic cues (as predicted by the linguistic account of argument interpretation) are required to predict RTs beyond the effects mediated through Bayesian surprise (as predicted by the rational expectation-based account). Thus, with only a single degree of freedom, Bayesian surprise derived from our rational model seems to achieve predictive accuracy for reading times that is about equally high as for the functionally much more flexible linguistic account.

At first blush, this finding might be surprising given that some previous studies have concluded that frequency information is insufficient to explain the interactions between different linguistics cues (Mitchell, 1987; Gibson et al., 1996; Pickering et al., 2000; Kennison, 2001; Van Gompel and Pickering, 2001; Bornkessel et al., 2002; McKoon and Ratcliff, 2003). For

TABLE 5 | Out-of-sample predictive accuracy of linguistic and Bayesian surprise LMMs for each sentence region.

Constituent	LOOIC		LOOIC differences (ΔLOOIC)			
	Linguistic	**Bayesian surprise**	**Estimate**	**S.E.**	**Lower**	**Upper**
NP1	6748.29	6746.13	2.16	2.98	−5.3	9.61
Verb	7725.25	7735.99	10.74	8.01	−9.29	30.76
NP2	6665.08	6664.97	0.12	5.92	−14.68	14.92

We compare models in terms of leave-one-out cross-validated (LOO) log predictive density (elpd$_{LOO}$), specifically the difference ΔLOOIC between LMMs in the LOO information criterion (LOOIC = −2 × elpd$_{LOO}$). Confidence intervals for the differences are 2.5 standard errors below and above each difference. A confidence interval excluding zero is considered as evidence for a difference in predictive accuracy between the models at hand.

example, Bornkessel et al. (2002) compared ERP responses associated with initial nominative-, accusative-, or dative-marked NPs in German complement clauses. In this sentence context, both accusative- and dative-marked NPs are infrequent, compared to nominate-marked NPs. Frequency-based accounts of argument interpretation, Bornkessel and colleagues argued, would thus predict increased processing costs—and hence enhanced amplitude of the N400 response—for both accusative and dative NPs, compared to nominate NPs. In contrast to this prediction, Bornkessel and colleagues observed increased N400 amplitudes only for accusative NPs. Critically though, this does not rule out expectation-based accounts of argument interpretation. As we have summarized here (but see also earlier works, e.g., McRae et al., 1998), the relevant theoretical construct in expectation-based accounts are the *contextual* expectations. These are based on the *conditional* probability distribution of argument assignments given the available cues (incl. the properties of the initial NP and the preceding context), not the overall frequency of different argument assignments. An interesting question for future work is thus to see whether results like those of Bornkessel et al. (2002) could be accounted for by a model like the one we presented here.

Taken together, these findings argue against accounts that attribute a privileged role to some types of cues (Bornkessel and Schlesewsky, 2006; Kuperberg, 2007; Alday et al., 2014; see also Nakano et al., 2010; Szewczyk and Schriefers, 2011). Instead, our findings provide further support for expectation-based accounts of incremental argument interpretation: the effects of morpho-syntactic, argument prominence and verb-semantic cues on argument interpretation seem to be indirect, mediated through implicit expectations that are based on the distribution of these cues in previous language input. Our results thus corroborate findings from earlier work on probabilistic sentence comprehension (MacDonald et al., 1994; Garnsey et al., 1997; Tabor et al., 1997; McRae et al., 1998; Spivey-Knowlton and Tanenhaus, 1998; Vosse and Kempen, 2000, 2009, among many others). In competition-based models, for example, the processing difficulty of argument interpretation is determined by the extent to which the cues introduced at the current sentence region disagree with the relative activation of competing argument assignments at the preceding sentence region. The present approach borrows from, and builds on, these previous works (see Levy, 2008 for a nuanced discussion of commonalities and differences between rational and competition accounts). Unlike earlier accounts, however, the rational model presented here does not contain any hidden parameters, thereby putting the expectation-based hypothesis to a stronger test. As far as we know, the present work is the first to directly pit the expectation-based account against a linguistic account, by directly comparing the rational model to a linguistic model with respect to their out-of-sample predictive accuracy.

A long line of research has entertained the idea that language comprehension is expectation-based and draws on statistical patterns in the input (for reviews, see MacDonald, 2013; Kuperberg and Jaeger, 2016). However, most of this work—in particular within the rational tradition—has focused on expectations for individual words, parts-of-speech, or syntactic parses (e.g., Hale, 2001; Demberg and Keller, 2008; Levy, 2008, 2011; Smith and Levy, 2013; Linzen and Jaeger, 2014; Frank et al., 2015; Brothers and Kuperberg, 2021). The present study is instead concerned with argument interpretation—the process by which the NP arguments are "assigned" or "linked" to the argument-slots required by the verb. Unlike models of word-level surprisal, the rational model introduced here transparently links linguistic cues to their effect on the probability of argument assignments. This, we hope, will facilitate transfer from, and comparison to, linguistic accounts, which have typically focused on the role of specific cues. For example, the rational model of argument interpretation allows us to quantify and predict the *magnitude of effects* associated with different types of linguistic cues (**Figure 3** above as well as **Supplementary Figures 4, 5** and **Supplementary Tables 6–8**). This makes apparent which cues are particularly important to argument interpretation, and how different cues interact. Additional analyses presented in the **Supplementary Section 7**, further found that Bayesian surprise over argument assignment captures different aspects of reading times than a model of word-level surprisal. This suggests that expectation-based models of argument interpretation might bridge the gap between expectation-based accounts of word-level surprisal and linguistic accounts of argument interpretation.

An obvious limitation of our *model*—as opposed to the general proposal to estimate Bayesian surprise over argument assignments—is that it only applies to Swedish transitive sentences presented in isolation. It thus implicitly assumes that the comprehender knows—or strongly expects— that all sentences have a subject and an object whose relative ordering is to be determined, and that the baseline probability of the two competing orders are always the same. Although these assumptions are likely to be warranted in the context of our experiment—where unrelated transitive sentences are presented in isolation— it is clearly violated for argument interpretation in natural discourse contexts. Although there is more uncertainty about, for instance, the number and types of NP arguments in sentences in natural language, there is also additional information about NP argument functions, as word order variations primarily are motivated by discourse-pragmatic relations (such as topic and contrast, see Hörberg, 2016, 2018). The theoretical proposal made here predicts that such discourse-pragmatic information plays an important role in argument interpretation in the processing of natural language, although the simple model we test here would not be able to account for them.

With that being said, the rational model tested here makes predictions for a wide variety of transitive clauses with different syntactic configurations (i.e., NP- versus adverbial-initial, with or without auxiliary verbs, with or without sentential adverbials, and with NP arguments of any length), and draws upon many different properties (nine NP properties, four verb-semantic classes, and two syntactic properties; see **Table 1**). The present experiment tested only a small subset of the predictions even this simple model makes. Future work could thus use the same model to derive predictions for further experiments, contrasting other sentence types and/or other linguistic cues that the model includes. Other experimental paradigms and/or more high-powered experimental designs should be able to detect more subtle effects that the model predicts.

SUMMARY

Incremental argument interpretation draws on an interplay between form-, meaning- and discourse-based argument properties, and verb-semantic information, that function as cues to argument assignment during incremental sentence comprehension. We have provided evidence for the hypothesis that the effects of these cues to argument interpretation are mediated through expectations, based on their joint distribution over NP arguments in previously experienced language input. Based on the distribution of these cues in a corpus of transitive sentences in Swedish, we develop a rational model of incremental argument interpretation. This model predicts the processing difficulty experienced at each sentence constituent (i.e., NP1, verb, and NP2) as a function of the Bayesian surprise associated with changes in expectations over possible argument interpretations. The predictions of the rational model were found confirmed by reading times from a self-paced reading experiment of Swedish transitive sentences, both quantitatively, by directly predicting reading times, and qualitatively, in terms of showing similar patterns with respect to linguistic cues.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in the Open Science Framework (OSF) repository: https://osf.io/rw5nf/.

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 participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Both authors wrote the manuscript and conceived of the general ideas underlying the rational model, with TFJ suggesting the theoretical framework behind it. With input from TFJ, TH collected the corpus data, designed the implementation of the rational model, designed and performed the self-paced reading experiment, and performed the statistical analyses.

FUNDING

This work was partially supported by Swedish Foundation for Humanities and Social Sciences (P17-0536:1) to TH and NSF CAREER 1150028 to TFJ.

ACKNOWLEDGMENTS

Earlier versions of this work were presented at *Rational Approaches in Language Sciences* (RAILS) 2019, Saarland University, Saarbrücken, Germany, *Discourse Expectations: Theoretical, Experimental, and Computational perspectives* (DETEC) 2019, Leibniz-Centre General Linguistics (ZAS), Berlin, Germany, *the Lund Symposium on Cognition, Communication, and Learning*, Lund University, Lund, *the 12th Conference of the Association for Linguistic Typology* (ALT), Australian National University, Canberra as well as in the first author's thesis (Hörberg, 2016). We are particularly grateful to Murathan Kurfalı for his help with word surprisal estimation in GPT-2. We also thank Maria Koptjevskaja Tamm, Mikael Roll, Östen Dahl, Robert Östling, Johan Sjons, and members of the Human Language Processing lab, for help with various aspects of this study, and Shawn Cummings for feedback on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.674202/full#supplementary-material

REFERENCES

Aaronson, D., Ferres, S., Kieras, D. E., and Just, M. A. (1984). "The Word-by-Word Reading paradigm: An Experimental and Theoretical Approach," in *New Methods in Reading Comprehension Research*, eds D. E. Kieras and M. A. Just (London: Lawrence Erlbaum Associates, Inc), 31–68. doi: 10.4324/9780429505379-3

Acuña-Fariña, C., Fraga, I., García-Orza, J., and Piñeiro, A. (2009). Animacy in the adjunction of Spanish RCs to complex NPs. *Eur. J. Cogn. Psychol.* 21, 1137–1165. doi: 10.1080/09541440802622824

Alday, P. M., Schlesewsky, M., and Bornkessel-Schlesewsky, I. (2014). Towards a Computational Model of Actor-Based Language Comprehension. *Neuroinformatics* 12, 143–179. doi: 10.1007/s12021-013-9198-x

Bernard, J.-B., and Castet, E. (2019). The optimal use of non-optimal letter information in foveal and parafoveal word recognition. *Vision Res.* 155, 44–61. doi: 10.1016/j.visres.2018.12.006

Bickel, B. (2010). "Grammatical Relations Typology," in *The Oxford Handbook of Linguistic Typology*, ed. J. J. Song (Oxford: Oxford University Press), 399–444. doi: 10.1093/oxfordhb/9780199281251.013.0020

Bicknell, K., and Levy, R. (2012). "Why long words take longer to read: the role of uncertainty about word length," in *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, (Montréal: Association for Computational Linguistics), 21–30.

Bock, K. J., and Irwin, D. E. (1980). Syntactic effects of information availability in sentence production. *J. Verbal Learn. Verbal Behav.* 19, 467–484. doi: 10.1016/S0022-5371(80)90321-7

Bock, K. J., and Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formation. *Cognition* 21, 47–67. doi: 10.1016/0010-0277(85)90023-X

Borin, L., Forsberg, M., and Roxendal, J. (2012). "Korp - the corpus infrastructure of Språkbanken," in *Proceedings of LREC 2012*, (Istanbul: LREC), 474–478.

Bornkessel, I., and Schlesewsky, M. (2006). The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages. *Psychol. Rev.* 113, 787–821. doi: 10.1037/0033-295X.113.4.787

Bornkessel, I., Schlesewsky, M., and Friederici, A. D. (2002). Grammar overrides frequency: Evidence from the online processing of flexible word order. *Cognition* 85, B21–B30. doi: 10.1016/S0010-0277(02)00076-8

Bornkessel-Schlesewsky, I., and Schlesewsky, M. (2009). Minimality as vacuous distinctness: Evidence from cross-linguistic sentence comprehension. *Lingua* 119, 1541–1559. doi: 10.1016/j.lingua.2008.03.005

Bornkessel-Schlesewsky, I., and Schlesewsky, M. (2019). Toward a Neurobiologically Plausible Model of Language-Related, Negative Event-Related Potentials. *Front. Psychol.* 10:298. doi: 10.3389/fpsyg.2019.00298

Bornkessel-Schlesewsky, I., Kretzschmar, F., Tune, S., Wang, L., Genç, S., Philipp, M., et al. (2011). Think globally: Cross-linguistic variation in electrophysiological activity during sentence comprehension. *Brain Lang.* 117, 133–152. doi: 10.1016/j.bandl.2010.09.010

Boston, M. F., Hale, J. T., Vasishth, S., and Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Lang. Cogn. Process.* 26, 301–349. doi: 10.1080/01690965.2010.492228

Bouma, G. J. (2008). *Starting a Sentence in Dutch*. Ph. D. thesis. Groningen: University of Groningen.

Brothers, T., and Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *J. Mem. Lang.* 116:104174. doi: 10.1016/j.jml.2020.104174

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *J. Stat. Softw.* 80, 1–28. doi: 10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *R J.* 10, 395–411. doi: 10.32614/RJ-2018-017

Bushong, W. (2020). *Maintenance of Subcategorical Information in Spoken Word Recognition*. Ph. D. thesis. Rochester: University of Rochester.

Crocker, M. W., and Brants, T. (2000). Wide-Coverage Probabilistic Sentence Processing. *J. Psycholinguist. Res.* 29, 647–669. doi: 10.1023/A:1026560822390

Czypionka, A., Spalek, K., Wartenburger, I., and Krifka, M. (2017). On the interplay of object animacy and verb type during sentence comprehension in German: ERP evidence from the processing of transitive dative and accusative constructions. *Linguistics* 55:0031. doi: 10.1515/ling-2017-0031

Dahl, Ö (2000). Egophoricity in discourse and syntax. *Funct. Lang.* 7, 37–77. doi: 10.1075/fol.7.1.03dah

Dahl, Ö, and Fraurud, K. (1996). "Animacy in grammar and discourse," in *Reference and referent accessibility, Pragmatics & Beyond New Series*, eds T. Fretheim and J. K. Gundel (Philadelphia: John Benjamins Publishing Company), 47–87. doi: 10.1075/pbns.38.04dah

Demberg, V., and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109, 193–210. doi: 10.1016/j.cognition.2008.07.008

Desmet, T., Brysbaert, M., and De Baecke, C. (2002). The correspondence between sentence production and corpus frequencies in modifier attachment. *Q. J. Exp. Psychol. Sect. A* 55, 879–896. doi: 10.1080/02724980143000604

Desmet, T., De Baecke, C., Drieghe, D., Brysbaert, M., and Vonk, W. (2006). Relative clause attachment in Dutch: On-line comprehension corresponds to corpus frequencies when lexical variables are taken into account. *Lang. Cogn. Process.* 21, 453–485. doi: 10.1080/01690960400023485

Dowty, D. (1991). Thematic Protoroles-Roles and Argument Selection. *Language* 67, 547–619. doi: 10.1353/lan.1991.0021

Du Bois, J. (2003). *Preferred Argument Structure: Grammar as Architecture for Function*. Amsterdam: John Benjamins. doi: 10.1075/sidag.14

Farmer, T. A., Monaghan, P., Misyak, J. B., and Christiansen, M. H. (2011). Phonological typicality influences sentence processing in predictive contexts: Reply to Staub, Grant, Clifton, and Rayner (2009). *J. Exp. Psychol. Learn. Mem. Cogn.* 37, 1318–1325. doi: 10.1037/a0023063

Feleki, E., and Branigan, H. (1999). "Conceptual accessibility and serial order in Greek speech production," in *Proceedings of the 21st Annual Conference of the Cognitive Science Society*, eds M. Hahn and S. C. Stones (Mahaw, NJ: Lawrence Erlbaum Associates), 96–101. doi: 10.4324/9781410603494-22

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cogn. Psych.* 47, 164–203. doi: 10.1016/S0010-0285(03)00005-7

Ferreira, F., and Clifton, C. (1986). The independence of syntactic processing. *J. Mem. Lang.* 25, 348–368. doi: 10.1016/0749-596X(86)90006-9

Ferreira, V. S., and Yoshita, H. (2003). Given-New Ordering Effects on the Production of Scrambled Sentences in Japanese. *J. Psycholinguist. Res.* 32, 669–692. doi: 10.1023/A:1026146332132

Fine, A. B., Jaeger, T. F., Farmer, T. A., and Qian, T. (2013). Rapid Expectation Adaptation During Syntactic Comprehension. *PLoS One* 8:e77661. doi: 10.1371/journal.pone.0077661

Fine, A. B., Qian, T., Jaeger, T. F., and Jacobs, R. A. (2010). "Is There Syntactic Adaptation in Language Comprehension?," in *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics CMCL '10*, (Stroudsburg,PA: Association for Computational Linguistics), 18–26.

Frank, S. L., and Bod, R. (2011). Insensitivity of the Human Sentence-Processing System to Hierarchical Structure. *Psychol. Sci.* 22, 829–834. doi: 10.1177/0956797611409589

Frank, S. L., and Haselager, W. (2006). "Robust semantic systematicity and distributed representations in a connectionist model of sentence comprehension," in *Proceedings of the 28th annual conference of the Cognitive Science Society*, eds R. Sun and N. Miyake (Mahwaw, NJ: Lawrence Erlbaum Associates), 226–231.

Frank, S. L., and Yang, J. (2018). Lexical representation explains cortical entrainment during speech comprehension. *PLoS One* 13:e0197304. doi: 10.1371/journal.pone.0197304

Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain Lang.* 140, 1–11. doi: 10.1016/j.bandl.2014.10.006

Fraundorf, S. H., and Jaeger, T. F. (2016). Readers generalize adaptation to newly-encountered dialectal structures to other unfamiliar structures. *J. Mem. Lang.* 91, 28–58. doi: 10.1016/j.jml.2016.05.006

Frenzel, S., Schlesewsky, M., and Bornkessel-Schlesewsky, I. (2011). Conflicts in language processing: A new perspective on the N400–P600 distinction. *Neuropsychologia* 49, 574–579. doi: 10.1016/j.neuropsychologia.2010.12.003

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Garnsey, S. M., Pearlmutter, N. J., Myers, E., and Lotocky, M. A. (1997). The Contributions of Verb Bias and Plausibility to the Comprehension of

Temporarily Ambiguous Sentences. *J. Mem. Lang.* 37, 58–93. doi: 10.1006/jmla.1997.2512

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* 1, 515–534. doi: 10.1214/06-BA117A

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Stat. Comput.* 24, 997–1016. doi: 10.1007/s11222-013-9416-2

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* 2, 1360–1383. doi: 10.1214/08-AOAS191

Gennari, S. P., and MacDonald, M. C. (2008). Semantic indeterminacy in object relative clauses. *J. Mem. Lang.* 58, 161–187. doi: 10.1016/j.jml.2007.07.004

Gibson, E., Schütze, C. T., and Salomon, A. (1996). The relationship between the frequency and the processing complexity of linguistic structure. *J. Psycholinguist. Res.* 25, 59–92. doi: 10.1007/BF01708420

Hale, J. (2001). "A Probabilistic Earley Parser as a Psycholinguistic Model," in *Proceedings of NAACL*, (Pittsburgh, PA: NAACL), 159–166. doi: 10.3115/1073336.1073357

Harrington Stack, C. M., James, A. N., and Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Mem. Cognit.* 46, 864–877. doi: 10.3758/s13421-018-0808-6

Hörberg, T. (2016). *Probabilistic and Prominence-driven Incremental Argument Interpretation in Swedish*. Ph. D. thesis. Stockholm: Stockholm University.

Hörberg, T. (2018). Functional motivations behind direct object fronting in written Swedish: A corpus-distributional account. *Glossa* 3:81. doi: 10.5334/gjgl.502

Hörberg, T., Koptjevskaja-Tamm, M., and Kallioinen, P. (2013). The neurophysiological correlate to grammatical function reanalysis in Swedish. *Lang. Cogn. Process.* 28, 388–416. doi: 10.1080/01690965.2011.651345

Hsiao, Y., and MacDonald, M. C. (2016). Production predicts comprehension: Animacy effects in Mandarin relative clause processing. *J. Mem. Lang.* 89, 87–109. doi: 10.1016/j.jml.2015.11.006

Itti, L., and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Res.* 49, 1295–1306. doi: 10.1016/j.visres.2008.09.007

Jaeger, T. F., and Norcliffe, E. J. (2009). The Cross-linguistic Study of Sentence Production. *Lang. Linguist. Compass* 3, 866–887. doi: 10.1111/j.1749-818X.2009.00147.x

Jegerski, J. (2014). "Self-paced reading," in *Research methods in second language psycholinguistics*, eds J. Jegerski and B. van Patten (New York, NY: Routledge), 20–49. doi: 10.4324/9780203123430

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cogn. Sci.* 20, 137–194. doi: 10.1207/s15516709cog2002_1

Just, M. A., and Carpenter, P. A. (1980). A Theory of Reading: From Eye Fixations to Comprehension. *Psychol. Rev.* 87, 329–354. doi: 10.1037/0033-295X.87.4.329

Just, M. A., and Carpenter, P. A. (1992). A Capacity Theory of Comprehension: Individual Differences in Working Memory. *Psychol. Rev.* 99, 122–149. doi: 10.1037/0033-295X.99.1.122

Kaiser, E., and Trueswell, J. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition* 94, 113–147.

Kamide, Y., Altmann, G. T. M., and Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *J. Mem. Lang.* 49, 133–156. doi: 10.1016/S0749-596X(03)00023-8

Kaschak, M. P., and Glenberg, A. M. (2004). This construction needs learned. *J. Exp. Psychol. Gen.* 133, 450–467. doi: 10.1037/0096-3445.133.3.450

Kempe, V., and MacWhinney, B. (1999). Processing of Morphological and Semantic Cues in Russian and German. *Lang. Cogn. Process.* 14, 129–171. doi: 10.1080/016909699386329

Kempen, G., and Harbusch, K. (2004). A corpus study into word order variation in German subordinate clauses: Animacy affects linearization independently of grammatical function assignment. *TRENDS Linguist. Stud. Monogr.* 157, 173–182. doi: 10.1515/9783110894028.173

Kennison, S. M. (2001). Limitations on the use of verb information during sentence comprehension. *Psychon. Bull. Rev.* 8, 132–138. doi: 10.3758/BF03196149

Kim, A., and Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *J. Mem. Lang.* 52, 205–225. doi: 10.1016/j.jml.2004.10.002

Kliegl, R., Hohenstein, S., Yan, M., and McDonald, S. A. (2013). How preview space/time translates into preview cost/benefit for fixation durations during reading. *Q. J. Exp. Psychol.* 66, 581–600. doi: 10.1080/17470218.2012.658073

Kretzschmar, F., Bornkessel-Schlesewsky, I., Staub, A., Roehm, D., and Schlesewsky, M. (2012). "Prominence Facilitates Ambiguity Resolution: On the Interaction Between Referentiality, Thematic Roles and Word Order in Syntactic Reanalysis," in *Case, Word Order and Prominence*, eds M. Lamers and P. de Swart (Dordrecht: Springer Netherlands), 239–271. doi: 10.1007/978-94-007-1463-2_11

Kruschke, J. K. (2014). *Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan*. Florida, FL: Academic Press. doi: 10.1016/B978-0-12-405888-0.00008-8

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Res.* 1146, 23–49. doi: 10.1016/j.brainres.2006.12.063

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

Kuperberg, G. R., Caplan, D., Sitnikova, T., Eddy, M., and Holcomb, P. J. (2006). Neural correlates of processing syntactic, semantic, and thematic relationships in sentences. *Lang. Cogn. Process.* 21, 489–530. doi: 10.1080/01690960500094279

Kuperberg, G. R., Kreher, D. A., Sitnikova, T., Caplan, D. N., and Holcomb, P. J. (2007). The role of animacy and thematic relationships in processing active English sentences: Evidence from event-related potentials. *Brain Lang.* 100, 223–237. doi: 10.1016/j.bandl.2005.12.006

Kuperberg, G. R., Sitnikova, T., Caplan, D., and Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cogn. Brain Res.* 17, 117–129. doi: 10.1016/S0926-6410(03)00086-7

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

Levy, R. (2011). "Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, (Pennsylvania, PA: Association for Computational Linguistics), 1055–1065.

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

Linzen, T., and Jaeger, T. F. (2014). "Investigating the role of entropy in sentence processing," in *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics*, (Pennsylvania, PA: Association for Computational Linguistics), 10–18. doi: 10.3115/v1/W14-2002

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Front. Psychol.* 4:00226. doi: 10.3389/fpsyg.2013.00226

MacDonald, M. C., and Seidenberg, M. S. (2006). "Constraint Satisfaction Accounts of Lexical and Sentence Comprehension," in *Handbook of Psycholinguistics*, eds M. J. Traxler and M. A. Gernsbacher (Amsterdam: Elsevier), 581–611. doi: 10.1016/B978-012369374-7/50016-X

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). The Lexical Nature of Syntactic Ambiguity Resolution. *Psychol. Rev.* 10, 676–703. doi: 10.1037/0033-295X.101.4.676

MacWhinney, B., and Bates, E. (1989). *The crosslinguistic study of sentence processing*. New York, NY: Cambridge University Press.

MacWhinney, B., Bates, E., and Kliegl, R. (1984). Cue Validity and Sentence Interpretation in English, German, and Italian. *J. Verbal Learn. Verbal Behav.* 23, 127–150. doi: 10.1016/S0022-5371(84)90093-8

Mak, W. M., Vonk, W., and Schriefers, H. (2006). Animacy in processing relative clauses: The hikers that rocks crush. *J. Mem. Lang.* 54, 466–490. doi: 10.1016/j.jml.2006.01.001

Mak, W. M., Vonk, W., and Schriefers, H. (2008). Discourse structure and relative clause processing. *Mem. Cognit.* 36, 170–181. doi: 10.3758/MC.36.1.170

Makowski, D., Ben-Shachar, M., and Lüdecke, D. (2019). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *J. Open Source Softw.* 4:1541. doi: 10.21105/joss.01541

McKoon, G., and Ratcliff, R. (2003). Meaning through syntax: Language comprehension and the reduced relative clause construction. *Psychol. Rev.* 110, 490–525. doi: 10.1037/0033-295X.110.3.490

McRae, K., Spivey-Knowlton, M. J., and Tanenhaus, M. K. (1998). Modeling the Influence of Thematic Fit (and Other Constraints) in On-Line Sentence Comprehension. *J. Mem. Lang.* 1998, 283–312. doi: 10.1006/jmla.1997.2543

Mitchell, D. (1987). "Lexical guidance in human parsing: Locus and processing characteristics," in *Attention and performance 12: The psychology of reading*, ed. M. Coltheart (London: Lawrence Erlbaum Associates, Inc), 601–618.

Mitchell, D. C. (1984). "An Evaluation of Subject-Paced Reading Tasks and Other Methods for Investigating Immediate Processes in Reading," in *New Methods in Reading Comprehension Research*, eds D. E. Kieras and M. A. Just (London: Lawrence Erlbaum Associates, Inc), 69–89. doi: 10.4324/9780429505379-4

Muralikrishnan, R., Schlesewsky, M., and Bornkessel-Schlesewsky, I. (2015). Animacy-based predictions in language comprehension are robust: Contextual cues modulate but do not nullify them. *Brain Res.* 1608, 108–137. doi: 10.1016/j.brainres.2014.11.046

Nakano, H., Saron, C., and Swaab, T. Y. (2010). Speech and span: Working memory capacity impacts the use of animacy but not of world knowledge during spoken sentence comprehension. *J. Cogn. Neurosci.* 22, 2886–2898. doi: 10.1162/jocn.2009.21400

Narayanan, S., and Jurafsky, D. (1998). "Bayesian Models of Human Sentence Processing," in *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, eds M. A. Gernsbacher and S. J. Derry (Mahaw, NJ: Lawrence Erlbaum Associates, Inc), 752–757.

Nice, K. Y., and Dietrich, R. (2003). Task sensitivity of animacy effects: evidence from German picture descriptions. *Linguistics* 41:027. doi: 10.1515/ling.2003.027

Nivre, J., and Megyesi, B. (2007). "Bootstrapping a Swedish Treebank Using Cross-Corpus Harmonization and Annotation Projection," in *Proceedings of the 6th International Workshop on Treebanks and Linguistic Theories*, eds K. de Smedt, J. Hajič, and S. Kübler (Pennsylvania, PA: Association for Computational Linguistics), 97–102.

Øvrelid, L. (2004). "Disambiguation of syntactic functions in Norwegian: modeling variation in word order interpretations conditioned by animacy and definiteness," in *Proceedings of the 20th Scandinavian Conference of Linguistics*, ed. F. Karlsson (Helsinki: University of Helsinki), 1–17.

Paczynski, M., and Kuperberg, G. R. (2011). Electrophysiological evidence for use of the animacy hierarchy, but not thematic role assignment, during verb-argument processing. *Lang. Cogn. Process.* 26, 1402–1456. doi: 10.1080/01690965.2011.580143

Paczynski, M., and Kuperberg, G. R. (2012). Multiple influences of semantic memory on sentence processing: Distinct effects of semantic relatedness on violations of real-world event/state knowledge and animacy selection restrictions. *J. Mem. Lang.* 67, 426–448. doi: 10.1016/j.jml.2012.07.003

Philipp, M., Bornkessel-Schlesewsky, I., Bisang, W., and Schlesewsky, M. (2008). The role of animacy in the real time comprehension of Mandarin Chinese: Evidence from auditory event-related brain potentials. *Brain Lang.* 105, 112–133. doi: 10.1016/j.bandl.2007.09.005

Philipp, M., Graf, T., Kretzschmar, F., and Primus, B. (2017). Beyond Verb Meaning: Experimental Evidence for Incremental Processing of Semantic Roles and Event Structure. *Front. Psychol.* 8:1806. doi: 10.3389/fpsyg.2017.01806

Pickering, M. J., Traxler, M. J., and Crocker, M. W. (2000). Ambiguity Resolution in Sentence Processing: Evidence against Frequency-Based Accounts. *J. Mem. Lang.* 43, 447–475. doi: 10.1006/jmla.2000.2708

Primus, B. (2006). "Mismatches in semantic-role hierarchies and the dimensions of role semantics," in *Semantic Role Universals and Argument Linking: Theoretical, Typological and Psycholinguistic Perspectives*, eds I. Bornkessel, M. Schlesewsky, B. Comrie, and A. D. Friederici (Berlin: Mouton de Gruyter), 53–89.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rabovsky, M. (2020). Change in a probabilistic representation of meaning can account for N400 effects on articles: A neural network model. *Neuropsychologia* 143:107466. doi: 10.1016/j.neuropsychologia.2020.107466

Rabovsky, M., Hansen, S. S., and McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nat. Hum. Behav.* 2, 693–705. doi: 10.1038/s41562-018-0406-4

Rahkonen, M. (2006). Some aspects of topicalization in Swedish declaratives. *Linguistics* 44, 23–55. doi: 10.1515/LING.2006.002

Ranganath, C., and Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nat. Rev. Neurosci.* 4, 193–202. doi: 10.1038/nrn1052

Roark, B., Bachrach, A., Cardenas, C., and Pallier, C. (2009). "Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, (Singapore: Association for Computational Linguistics), 324–333. doi: 10.3115/1699510.1699553

Roehm, D., Bornkessel-Schlesewsky, I., Rösler, F., and Schlesewsky, M. (2004). Fractionating language comprehension via frequency characteristics of the human EEG. *J. Cogn. Neurosci.* 15, 409–412. doi: 10.1097/00001756-200403010-00005

Sauppe, S. (2017). Symmetrical and asymmetrical voice systems and processing load: Pupillometric evidence from sentence production in Tagalog and German. *Language* 93, 288–313. doi: 10.1353/lan.2017.0015

Smith, N. J., and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition* 128, 302–319. doi: 10.1016/j.cognition.2013.02.013

Spivey-Knowlton, M. J., and Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *J. Exp. Psychol. Learn. Mem. Cogn.* 24, 1521–1543. doi: 10.1037/0278-7393.24.6.1521

Stan Development Team (2017). *Stan Modeling Language: User's Guide and Reference Manual*. Columbia: Columbia University.

Szewczyk, J. M., and Schriefers, H. (2011). Is animacy special? *Brain Res.* 1368, 208–221. doi: 10.1016/j.brainres.2010.10.070

Szewczyk, J. M., and Schriefers, H. (2013). Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. *J. Mem. Lang.* 68, 297–314. doi: 10.1016/j.jml.2012.12.002

Tabor, W., Juliano, C., and Tanenhaus, M. K. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Lang. Cogn. Process.* 12, 211–271. doi: 10.1080/016909697386853

Tanaka, M. N., Branigan, H. P., McLean, J. F., and Pickering, M. J. (2011). Conceptual influences on word order and voice in sentence production: Evidence from Japanese. *J. Mem. Lang.* 65, 318–330. doi: 10.1016/j.jml.2011.04.009

Tily, H. (2010). *The Role of Processing Complexity in Word Order Variation and Change*. Ph. D. thesis. Stanford, CA: Stanford University.

Tooley, K. M., and Traxler, M. J. (2018). Implicit learning of structure occurs in parallel with lexically-mediated syntactic priming effects in sentence comprehension. *J. Mem. Lang.* 98, 59–76. doi: 10.1016/j.jml.2017.09.004

Tooley, K. M., Swaab, T. Y., Boudewyn, M. A., Zirnstein, M., and Traxler, M. J. (2014). Evidence for priming across intervening sentences during on-line sentence comprehension. *Lang. Cogn. Neurosci.* 29, 289–311. doi: 10.1080/01690965.2013.770892

Traxler, M. J., Williams, R. S., Blozis, S. A., and Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *J. Mem. Lang.* 53, 204–224. doi: 10.1016/j.jml.2005.02.010

Trueswell, J. C., Tanenhaus, M. K., and Garnsey, S. M. (1994). Semantic Influences on Parsing: Use of Thematic Role Information in Syntactic Ambiguity Resolution. *J. Mem. Lang.* 1994, 285–318. doi: 10.1006/jmla.1994.1014

Van Gompel, R. P. G., and Pickering, M. J. (2001). Lexical guidance in sentence processing: A note on Adams, Clifton, and Mitchell (1998). *Psychon. Bull. Rev.* 8, 851–857. doi: 10.3758/BF03196228

van Herten, M., Chwilla, D. J., and Kolk, H. H. J. (2006). When Heuristics Clash with Parsing Routines: ERP Evidence for Conflict Monitoring in Sentence Perception. *J. Cogn. Neurosci.* 18, 1181–1197. doi: 10.1162/jocn.2006.18.7.1181

van Herten, M., Kolk, H. H. J., and Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cogn. Brain Res.* 22, 241–255. doi: 10.1016/j.cogbrainres.2004.09.002

Van Valin, R. D. J. (2005). *Exploring the Syntax-Semantics Interface*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511610578

Van Valin, R. D. (2006). "Semantic macroroles and language processing," in *Semantic Role Universals and Argument Linking: Theoretical, Typological, and Psycholinguistic Perspectives,* eds. I. Bornkessel, M. Schlesewsky, B. Comrie, and A. D. Friederici (Berlin: Mouton de Gruyter), 263–301.

Van Valin, R. D. J., and LaPolla, R. J. (1997). *Syntax: Structure, meaning and function*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139166799

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* 27, 1413–1432. doi: 10.1007/s11222-016-9696-4

Vosse, T., and Kempen, G. (2000). Syntactic structure assembly in human parsing: a computational model based on competitive inhibition and a lexicalist grammar. *Cognition* 75, 105–143. doi: 10.1016/S0010-0277(00)00063-9

Vosse, T., and Kempen, G. (2009). The Unification Space implemented as a localist neural net: predictions and error-tolerance in a constraint-based parser. *Cogn. Neurodyn.* 3, 331–346. doi: 10.1007/s11571-009-9094-0

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p*-values. *Psychon. Bull. Rev.* 14, 779–804. doi: 10.3758/BF03194105

Wang, L., Wlotko, E., Alexander, E., Schoot, L., Kim, M., Warnke, L., et al. (2020). Neural Evidence for the Prediction of Animacy Features during Language Comprehension: Evidence from MEG and EEG Representational Similarity Analysis. *J. Neurosci.* 40, 3278–3291. doi: 10.1523/JNEUROSCI.1733-19.2020

Warren, T., and Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition* 85, 79–112. doi: 10.1016/S0010-0277(02)00087-2

Watanabe, S. (2013). A Widely Applicable Bayesian Information Criterion. *J. Mach. Learn Res.* 14, 867–897.

Weckerly, J., and Kutas, M. (1999). An electrophysiological analysis of animacy effects in the processing of object relative sentences. *Psychophysiology* 36, 559–570. doi: 10.1111/1469-8986.3650559

Weissbart, H., Kandylaki, K. D., and Reichenbach, T. (2020). Cortical Tracking of Surprisal during Continuous Speech Comprehension. *J. Cogn. Neurosci.* 32, 155–166. doi: 10.1162/jocn_a_01467

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., and van den Bosch, A. (2016). Prediction During Natural Language Comprehension. *Cereb. Cortex* 26, 2506–2516. doi: 10.1093/cercor/bhv075

Wu, F., Kaiser, E., and Andersen, E. (2010). "Subject Preference, Head Animacy and Lexical Cues: A Corpus Study of Relative Clauses in Chinese," in *Processing and Producing Head-final Structures Studies in Theoretical Psycholinguistics*, eds H. Yamashita, Y. Hirose, and J. L. Packard (Dordrecht: Springer Netherlands), 173–193. doi: 10.1007/978-90-481-9213-7_9

Yan, S., and Jaeger, T. F. (2020). Expectation adaptation during natural reading. *Lang. Cogn. Neurosci.* 35, 1394–1422. doi: 10.1080/23273798.2020.1784447

Yan, S., Kuperberg, G. R., and Jaeger, T. F. (2017). Prediction (Or Not) During Language Processing. A Commentary On Nieuwland et al. (2017) And Delong et al. (2005). *bioRxiv* 2017:143750. doi: 10.1101/143750

Author Disclaimer: The views expressed here are not necessarily those of the funding agencies.

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 Communication

ORIGINAL RESEARCH
published: 15 October 2021
doi: 10.3389/fcomm.2021.674126

Pronominalization and Expectations for Re-Mention: Modeling Coreference in Contexts With Three Referents

Jet Hoek [1], *Andrew Kehler* [2] *and Hannah Rohde* [3]*

[1] Department of Language and Communication, Centre for Language Studies, Radboud University, Nijmegen, Netherlands, [2] Department of Linguistics, University of California San Diego, San Diego, CA, United States, [3] Department of Linguistics and English Language, University of Edinburgh, Edinburgh, United Kingdom

OPEN ACCESS

Edited by:
Juhani Järvikivi,
University of Alberta, Canada

Reviewed by:
Torgrim Solstad,
Bielefeld University, Germany
Jorrig Vogels,
University of Groningen, Netherlands

***Correspondence:**
Hannah Rohde
hannah.rohde@ed.ac.uk

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

Received: *28 February 2021*
Accepted: *01 September 2021*
Published: *15 October 2021*

Citation:
Hoek J, Kehler A and Rohde H (2021) Pronominalization and Expectations for Re-Mention: Modeling Coreference in Contexts With Three Referents.
Front. Commun. 6:674126.
doi: 10.3389/fcomm.2021.674126

The relationship between pronoun production and pronoun interpretation has been proposed to follow Bayesian principles, combining a comprehender's expectation about which referent will be mentioned next and their estimate of how likely it is that a potential referent will be re-mentioned using a pronoun. The Bayesian Model has received support from studies in several languages (English, Mandarin Chinese, Catalan, German), but tested contexts have been limited to two event participants, whereas natural language discourse often involves contexts with more than two event participants. In this study, we conducted three story continuation experiments to assess how the Bayesian Model performs in more complex contexts. Our results show that even in contexts with three event participants, comprehenders can behave rationally when interpreting pronouns, but that they appear to require sufficient context to build up a coherent representation of the situation to do so. In addition to testing the basic claim of the Bayesian Model (Weak Bayes), we test the central prediction of the Strong form of the hypothesis: that the two components of the model (next-mention expectations and choice of referring expression) are influenced by dissociated sets of factors. In a model comparison, Experiments 2 and 3 confirm the closest fit from the Bayesian Model, which supports Weak Bayes, and none of our experiments find evidence that the predictability of a referent affects pronominalization rates, which corroborates Strong Bayes. Finally, we test whether the rate of pronominalization is sensitive to factors related to ambiguity and argument/adjunct status of referents; we find that participants vary their production of pronouns most strongly based on the grammatical role of the antecedent (subject or not), with a smaller effect from the presence/absence of a gender-matched competitor and no effect from the syntactic position of this competing referent.

Keywords: coreference, pronoun production, pronoun interpretation, benefactives, ambiguity, bayesian coreference

1 INTRODUCTION

Reduced reference to previously mentioned entities–such as that achieved via pronominalization–is a hallmark of coherent discourse. Yet a speaker's decision to employ a reduced form poses an interpretation problem to the hearer, who needs to recover the speaker's intended referent.[1] A commonly held view is that speakers and hearers coordinate on the reference problem through a notion of entity salience: Speakers consult a set of factors that contribute to salience in deciding to use a pronoun, and hearers consult those same factors when interpreting it. Much of the literature has engaged with the question of what these factors are–including, for example, order of mention, grammatical role, thematic role, parallelism, information structure, and world knowledge–and how they are weighed with respect to one another.

There is also evidence, however, to suggest that the factors that condition pronoun production and interpretation are to some degree dissociated. In a context like (1), hearers are more likely to interpret a subsequent pronoun *she* as in (1-a) as referring to Jill than speakers are to produce a pronoun when referring to Jill in a subsequent sentence as in (1-b); likewise, speakers are more likely to use a pronoun to refer to Sue in (1-b) even though Sue will not be the preferred referent for the hearer in (1-a) (e.g., Stevenson et al., 1994; Kehler et al., 2008; Kehler and Rohde 2013).

1) a. Sue fired Jill. She ________
b. Sue fired Jill. ________

This asymmetry between the production of pronouns and their interpretation is posited to reflect a separation between the factors that guide choice of referring expression and the factors that guide expectations of next mention (both of which in turn influence interpretation).

Kehler et al. (2008) (see also Kehler and Rohde 2013; Rohde and Kehler 2014; Kehler and Rohde 2019) propose a rational Bayesian approach that is capable of capturing this asymmetry, according to which a hearer combines their expectation about which referent will be mentioned next and their estimate of how likely a speaker is to use a pronoun when re-mentioning a potential referent. The model produces quantitative estimates of interpretation biases that can be compared directly against actual biases collected in passage completion studies, and also allows for the factors that contribute to production and interpretation to be evaluated separately. Thus far, studies on English (Rohde and Kehler 2014; Kehler and Rohde 2019; Cheng and Almor 2019), Mandarin Chinese (Zhan et al., 2020), Catalan (Mayol, 2018), and German personal and demonstrative pronouns[2] have provided support for the model (but see Lam and Hwang 2021).

These studies have all focused on contexts with two event participants as potential referents for a pronoun. But natural language use, of course, often involves discourse contexts with more than two event participants. In light of the demands that a rational interpretation process might place on a hearer's cognitive apparatus (e.g., working memory, attention, probability estimation), an open question is how the model performs in more complex contexts: How well can hearers behave rationally when interpreting pronouns when there is a greater number of event participants to keep track of?

To address this question, we will employ contexts using the benefactive construction, exemplified in (2).

2) Adam scolded Russell for Diana.

Benefactive sentences describe situations in which an Agent engages in an action that affects a Patient for the benefit of a Beneficiary; these event participants appear as the grammatical subject, direct object, and object of a prepositional phrase adjunct, respectively.

In addition to its ability to introduce three event participants into the discourse, the benefactive construction allows us to address two other questions that currently exist in the literature. The first bears on the distinction between referents introduced from argument and adjunct positions and the rate at which they are pronominalized. Previous work that has compared two types of transfer-of-possession contexts–Source-Goal and Goal-Source constructions–has found a limited effect of thematic role on pronoun production favoring the Goal (Arnold 2001; Rosa and Arnold 2017; but see Rohde 2008 Expt VIII). In contrast, studies that have compared two types of implicit causality contexts–subject-biased and object-biased–have not (Rohde, 2008; Fukumura and Van Gompel, 2010; Rohde and Kehler, 2014). As we explain in further detail in Section 1.2, it has been suggested (Fukumura and Van Gompel, 2010; Rohde and Kehler, 2014) that the argument-adjunct distinction may have confounded results using transfer verbs, since the Goal occurs in an obligatory argument position in both types, whereas the Source is an argument in the Source-Goal construction but an adjunct in the Goal-Source construction. Benefactives provide a novel way to examine this question, since the subject and object appear in argument positions but the beneficiary occurs within a prepositional phrase adjunct. By utilizing benefactive contexts in three different configurations in which reference is two-ways ambiguous, we can run controlled studies that shed new light on this question.

A second question bears on comparing pronominalization rates for entities in referentially ambiguous and unambiguous contexts, a question for which different perspectives on the role of pronominalization make different predictions. On the one hand, on the view that referential form selection is intimately connected to audience design and ambiguity avoidance, we might expect to witness a big difference in the two scenarios: Speakers might be expected to pronominalize whenever possible in referentially-unambiguous contexts since referential success is not at stake, whereas they would presumably pronominalize less when the resulting expression would risk being ambiguous. On the other hand, on the view that pronominalization is driven primarily by the topicality of the referent (Grosz et al., 1995; Rohde and Kehler, 2014), little or no difference might be expected. Past work

[1] We use 'speaker' to generally refer to language producers and 'hearers' to generally refer to language comprehenders.

[2] Patterson, C., Schumacher, P. B., Nicenboim, B., Hagen, J., and Kehler, A. (2021). A Bayesian Approach to German Personal and Demonstrative Pronouns. Submitted for publication.

(Arnold and Griffin, 2007; Rohde, 2008; Rosa and Arnold, 2017) has shown a mixed picture: Ambiguity does affect rate of pronominalization, counter to a pure topicality account, but not to the extent one would expect if ambiguity avoidance was the only concern. This work has drawn comparisons across different contexts, however. Again, by utilizing benefactive contexts in three different configurations–each of which having three event participants, two that participate in referential ambiguity and one that does not–we can analyze this question across three different grammatical role pairings in the context of a single experiment.

This paper reports on three experiments designed to examine these issues, using discourse contexts that employ the benefactive construction. We focus on four central questions:

1. How well does the Bayesian Model predict the actual biases that hearers bring to the interpretation of pronouns in benefactive contexts, as measured in passage completion experiments?
2. Are the factors that influence pronoun production the same as those that influence predictability, or is there a dissociation between them?
3. Do pronoun production rates vary depending on whether the referent is introduced in an argument or adjunct position?
4. Do pronoun production rates vary depending on whether pronominal reference is ambiguous, keeping all else equal?

We elaborate on these questions in the sections that follow.

1.1 Three Models of Pronoun Interpretation

Bayesian Model

The Bayesian Model posits that a comprehender, upon encountering a pronoun, interprets it by reverse-engineering the speaker's intended referent following Bayesian principles (Kehler et al., 2008; Kehler and Rohde, 2013; Rohde and Kehler, 2014). The relationship between interpretation and production is captured by the model via the straightforward application of Bayes' Rule shown in **Eq. 1**.

$$P(\text{referent}|\text{pronoun}) = \frac{P(\text{pronoun}|\text{referent})\,P(\text{referent})}{\sum\limits_{\text{referent}\in\text{possible referents}} P(\text{pronoun}|\text{referent})\,P(\text{referent})} \tag{1}$$

The posterior term *P*(referent|pronoun) represents the comprehender's INTERPRETATION bias: upon encountering a pronoun, the probability that the comprehender will interpret it as referring to a particular referent. On the other hand, the likelihood term *P*(pronoun|referent) represents the PRODUCTION bias: the comprehender's estimate of the probability that the speaker would use a pronoun to refer to the potential referent under consideration. Finally, the prior term *P*(referent) denotes the comprehender's NEXT-MENTION bias: the hearer's estimate of the probability that the speaker would mention a specific referent at that point in the discourse, without regard for the form of referring expression that is chosen. On this model, therefore, pronoun interpretation biases result from comprehenders integrating their 'top-down' predictions about the content of the ensuing message (particularly, who will be mentioned next) with the 'bottom-up' linguistic evidence (particularly, the fact that the speaker opted to use a pronoun).

Strong vs Weak Bayes

Kehler et al. offer two varieties of the Bayesian Model. As it stands, **Eq. 1** says only that the relationship between pronoun interpretation and pronoun production follows Bayesian principles, without further specifying the types of contextual factors that affect the likelihood and prior terms. This claim is the sole prediction of the WEAK form of the hypothesis. That is, the weak hypothesis says that, given independent estimates of the prior, likelihood, and posterior probabilities, **Eq. 1** will approximately hold.

Whereas this is the central claim of the Bayesian Hypothesis, Kehler et al. also cited evidence that the two terms in the numerator of **Eq. 1** are conditioned by different types of contextual factors. On the one hand, they noted that the results of previous studies suggested that the factors that condition the next-mention bias *P*(referent) are primarily driven by meaning: semantic factors such as the verbs used in the context sentences and the eventualities they describe, and certain types of pragmatic inferences, including the coherence relations established between the clauses. On the other hand, the factors that condition the production bias *P*(pronoun|referent) appear to be grammatical and/or information structural in nature, for instance, based on grammatical role obliqueness or topichood respectively, both of which amount to a preference for pronouns when a sentential subject is re-mentioned. The resulting prediction, therefore, is that a speaker's decision about whether or not to pronominalize a reference will be insensitive to a set of semantic and pragmatic contextual factors that the comprehender will nonetheless bring to bear in interpretation. This is the central prediction of the STRONG form of the hypothesis.

The empirical status of the strong hypothesis remains under debate; while it is supported by for instance Rohde's (2008) (see also Rohde and Kehler (2014)) and Fukumura and Van Gompel's (2010) studies using implicit causality contexts, Rosa and Arnold (2017) report an effect of referent predictability on pronominalization in transfer-of-possession contexts. One consistent finding, however, is that insofar as semantic factors influence production at all, they do not affect production biases to the same extent that they do interpretation. We will examine the predictions of the strong model in the current experiments as well.

The Mirror Model

In order to provide benchmarks against which to evaluate the performance of the Bayesian Model, we will compare its quantitative predictions against those of two other models, each of which represent particular operationalizations of ideas drawn from the literature. The first such model we call the Mirror Model, which is designed to capture the idea that there is a single notion of entity prominence that the speaker and comprehender jointly use to mediate pronoun production and interpretation (posited by accounts of coreference put forward by, for instance, Ariel 1990, Givón 1983, and Gundel et al., 1993). On this conception–under which the comprehender is using the same cues to referential prominence that the speaker is–the ultimate interpretation bias toward a referent on the comprehension side should be proportional to the likelihood of the referent being pronominalized by the speaker, as reflected in **Eq. 2**.

$$P(\text{referent}|\text{pronoun}) \leftarrow \frac{P(\text{pronoun}|\text{referent})}{\sum_{\text{referent}\in\text{referents}} P(\text{pronoun}|\text{referent})} \tag{2}$$

Here we use the assignment operator to capture the fact that this model, unlike (1), does not follow the standard laws of probability theory. This model captures the idea that comprehenders will assign pronouns based on their consideration of what entities the speaker is most likely to refer to using a pronoun instead of a competing referential form, which is cached out by taking the comprehenders' estimate of the probability that a speaker will produce a pronoun for a particular referent, normalized by the sum of the probabilities for all suitably prominent referents that are consistent with any constraints imposed by the pronominal form (gender, number, etc).

The Expectancy Model

The second competing model we refer to as the Expectancy Model, which represents a particular way of operationalizing of an insight from Arnold (1998) regarding the role of predictive processing. According to Arnold's Expectancy Hypothesis, "listeners focus their attention on discourse entities in proportion to their estimation of the likelihood that the entity will be mentioned" (Arnold, 2008, p. 505). Comprehenders use referential expectations as a proxy for their estimates of speaker's focus of attention (p. 506); the higher this level of attention for a particular entity, the higher the likelihood that the speaker, when uttering a pronoun, is using it to refer to that entity. Here we operationalize this idea using next-mention bias *P*(referent) in **Eq. 3**, normalized by the next-mention probability of all referents that are compatible with the constraints (gender, number) imposed by the pronominal form.

$$P(\text{referent}|\text{pronoun}) \leftarrow \frac{P(\text{referent})}{\sum_{\text{referent}\in\text{referents}} P(\text{referent})} \tag{3}$$

We again use the ← assignment operator to emphasize the fact that the equality of the terms on the left and right hand sides does not follow from the laws of probability theory. On this model, therefore, the influence of context is mostly 'top-down', creating expectations about who will be mentioned next, with pronoun interpretation biases following these expectations.

1.2 Thematic Roles and Pronoun Production

The primary evidence for the impact of thematic role on pronoun production comes from work by Arnold and colleagues. First, Arnold (2001) found an effect that favored the pronominalization of Goal antecedents over Source antecedents when comparing two types of transfer-of-possession contexts: Goal-Source frames (*The butler got some ice from the chef*) and Source-Goal frames (*The chef gave some ice to the butler*). However, the effect was relatively small, and only found when the antecedent was a non-subject. More recently, Rosa and Arnold (2017) ran three follow-up experiments using the same types of frames, one which used an event-retelling task with more situated contexts (Exp 1) and two standard story-continuation tasks (Exps 2-3). Effects were found in Exps. 1 and 3, but much more strongly for subject antecedents than non-subject antecedents in the same-gender condition of Exp. 1 and only for non-subject antecedents in Exp. 3.[3] We will return to these findings in the General Discussion, after presenting our results using benefactive contexts.

There is an additional complication that arises when it comes to disentangling the effects of thematic role and grammatical role in transfer-of-possession frames. As expected, across Rosa and Arnold's experiments there was a large effect of grammatical role whereby referents introduced in subject position are re-mentioned with pronouns at higher rates than those introduced in object position. The thematic role effect arises when comparing Goal and Source subjects and likewise Goal and Source non-subjects. However, there is a relevant asymmetry here: whereas the Goal in a Source-Goal frame is mentioned from an obligatory argument position (*Sue handed the book *(to Mary)*), the Source in a Goal-Source frame is mentioned from within an optional adjunct (*Mary received the book (from Sue)*). The reason this is relevant is that according to some theories of information structure (Lambrecht, 1994, inter alia), the potential for topicality of a constituent decreases as one moves down the obliqueness hierarchy (subjects > objects > other arguments > adjuncts). On a theory in which pronominalization biases are driven by topicality (Grosz et al., 1995; Rohde and Kehler, 2014), it follows that the increased pronominalization rates for Goals in subject position could be attributed to the fact that it competes for topicality with an adjunct, whereas Source subjects compete with another argument. Similar logic applies for non-subjects: as arguments, Goals may be more topical than adjunct Sources. To shed new light on this question, we use the benefactive construction in contexts with three event participants but where only two of them match the gender of the pronoun in the pronoun-prompt condition. By running all three possible configurations–where NP1 and NP2 compete, NP1 and NP3 compete, and NP2 and NP3 compete–we can hold constant the status of a given referent and analyze its pronominalization rate when it competes with a gender-matched referent in an argument or adjunct position.

1.3 Ambiguity Avoidance

Hearer-oriented models of pronoun production make the assumption that speakers take into account the hearer's discourse model when producing referring expressions. Many studies suggest that speakers avoid producing ambiguous referring expressions to make sure they are understood correctly by their audience (e.g., Horton and Keysar 1996; Nadig and Sedivy 2002; Matthews et al., 2006; Hendriks et al., 2014). Under this assumption, speakers are less

[3] Rosa and Arnold report reliable effects for their Exp 2, but it is clear from their descriptive statistics that no effect exists for the condition of interest for evaluating the predictions of the strong Bayesian Model, in which the two event participants are of the same gender and hence reference is ambiguous. Here the pronominalization rates for subjects were identical (69% for both Goal and Source antecedents), and only negligibly different for non-subjects and in the wrong direction (18% for Goals and 19% for Sources).

likely to produce a pronoun for a referent when there is another referent in the immediate discourse that matches the intended referent in features relevant to the pronoun (e.g., gender, number, or animacy in English).

Evidence for a role of ambiguity avoidance in pronoun production, however, is not undisputed. Fukumura and van Gompel (2012), for instance, find that speakers produce pronouns to refer to referents in the preceding discourse, regardless of whether their addressee has knowledge of the preceding discourse. In addition, Arnold and Griffin (2007) show that an additional potential referent in the discourse leads to a decrease in the proportion of pronouns produced, even if the pronoun would nonetheless be unambiguous. In explaining this effect, Arnold and Griffin take a speaker-oriented approach by arguing that additional referents influence pronoun production by competing for attention in the speaker's representation of the discourse (and that similarity between referents, for instance in terms of gender, increases this effect; see also Fukumura et al., 2011). Offering a similar speaker-based explanation for Arnold and Griffin's findings, Rohde and Kehler (2014) propose that more referents entering the discourse decreases the chance that a referent is the topic, which in turn reduces the pronominalization rate. The question thus remains whether speakers strive to avoid ambiguity when producing referring expressions.

2 EXPERIMENT 1

In a story continuation experiment, we tested participants' pronoun interpretations, re-mention preferences, and pronominalization rates in contexts containing sentence frames with three event participants: an Agent (NP1), a Patient (NP2), and a Benefactive (NP3), as in for instance *Ben followed Sophia for David*. We varied prompt type (pronoun vs full-stop) and the position of the pair of gender-matched referents (NP1&NP2 vs NP1&NP3 vs NP2&NP3).

Crucial to determining whether different factors influence the prior and the likelihood (i.e., Strong Bayes), we expect that in these sentence frames, like in the implicit-causality and transfer-of-possession constructions commonly used in previous research on pronoun production and interpretation, the topicality and predictability of the referents do not coincide. Regarding topicality, if we assume that the grammatical subject position is the default position for topics in English, then the Agent in these benefactive constructions is the most topical referent. On the other hand, the predictability for re-mention does not necessarily favor the subject. For coherence-driven reasons, the Benefactive may be preferred if the next sentence provides an explanation of the event and one assumes that the initiative for the event is attributed to the Benefactive (i.e., *why did David want Sophia followed and why didn't he do this himself*?). Alternatively, the Patient may be preferred if the next sentence describes what happened next and the Patient is the referent most closely associated with the end state of the event. The point is that these benefactive sentences are posited to disfavor the subject referent for re-mention, a scenario that allows us to test the effectiveness of coreference models in contexts in which next-mention and pronominalization biases are dissociated.

2.1 Method

2.1.1 Participants

Participants were recruited through Amazon Mechanical Turk. 143 monolingual speakers of English completed the experiment and wrote correct continuations for the catch trials (see Materials) (mean age 37.2, age range 18–66, 65 women). Monolingual status was defined as an answer of 'no' to a question of whether any other language was spoken at home before the age of 6. All participants were paid for their participation ($5.25).

2.1.2 Materials

Stimuli consisted of 30 target prompts that featured three referents (subject, direct object, benefactive) and varied in prompt type (full stop vs pronoun), as in (3). Proper names were used to manipulate which potential referents were gender-matched: NP1&NP2, as in (3), NP1&NP3, or NP2&NP3.[4]

3) a. Adam$_{\mathrm{NP1}}$ scolded Russell$_{\mathrm{NP2}}$ for Diana$_{\mathrm{NP3}}$. ________
[full-stop prompt]
b. Adam$_{\mathrm{NP1}}$ scolded Russell$_{\mathrm{NP2}}$ for Diana$_{\mathrm{NP3}}$. He ________
[pronoun prompt]

The target items were distributed over six lists, with each item occurring only once per list, in one of the six conditions. The target items were interspersed with 32 fillers, including two 'catch' items that had an obvious correct continuation (e.g., *Caleb's favorite TV series is Game of [Thrones]*); these two items were used to filter out any participants who were not taking the task seriously. The other fillers varied in the number of (human) arguments they contained and whether they ended in a full stop or after the first word of a second sentence (similar to the pronoun prompt items).

2.1.3 Procedure

Continuations were collected via a web-based interface embedded in the Amazon Mechanical Turk environment. After reading a short instruction, signing a consent form, and supplying some demographic information, participants were asked to write a natural continuation for the prompts in the supplied text box. Each item was displayed on a separate page.

2.1.4 Annotation

For all target items in all three experiments, we annotated which referent was the subject of the continuation (next-mention: NP1, NP2, NP3) and how that referent was re-mentioned (form of referring expression: full NP vs pronoun). To ensure reliable coding, we (first author and a trained linguistics undergraduate student) double-coded data

[4] All materials and analysis scripts can be found at https://tinyurl.com/BenefactivesFrontiers.

from approximately 85 participants for all three experiments (approx. 60% for Experiment 1, 100% for Experiment 2, and 55% for Experiment 3). Inter-annotator agreement was very high on both next-mention (Experiment 1: 93%, $\kappa = 0.90$, Experiment 2: 94%, $\kappa = 0.90$, Experiment 3: 93%, $\kappa = 0.91$) and form of referring expression (Experiment 1: 99.5%, Experiment 2: 100%, Experiment 3: 99.3%). In all three experiments, the majority of disagreements on next-mention were due to one coder making a decision, while the other indicated they were not completely sure who was being referred to. All disagreements on form of referring expression were due to coding errors (5 in Experiment 1 and 7 in Experiment 3). After considering all disagreements, one coder (first author) finished annotation of the data from Experiments 1 and 3.

2.1.5 Data Analysis

We analyze the data in R (R Core Team, 2019). We compare the predictability and pronominalization rates of the referents using generalized linear mixed-effect regression (GLMM: Jaeger 2008) using the lme4 package (Bates et al., 2015).

For our questions regarding the efficacy of the Bayesian Model in benefactive contexts and the separation of referent predictability from pronominalization, we consider participants' next-mention and pronoun production behavior in the full-stop condition. To compare the predictability of the referents, we model the binary value of next-mention (yes vs no) in the full stop prompt subset of the data, with fixed effects of Referent (three levels: NP1/NP2/NP3) and Ambiguous Pair (three levels: NP1&NP2, NP1&NP3, NP2&NP3), as well as the interaction between Referent and Ambiguous Pair. To compare the pronominalization rates, we model the binary value of form of referring expression (pronoun or not) with Referent, Ambiguous Pair, and their interaction as fixed effects. Finally, we compare whether the pronoun prompts resulted in more NP1 continuations than the full stop prompts by modeling the binary value of NP1 continuation (yes vs no) on the entire dataset, with Prompt Type as fixed effect.

For our questions regarding the effect of argument/adjunct status and referential ambiguity on pronominalization, we compare pronominalization rates for ambiguous and unambiguous referents across the three ambiguous pair conditions in which a referent's gender-matched competitor is either an argument or adjunct. We model the binary value of form of referring expression (pronoun or not), with referent (three-level) and ambiguity (yes or no), as well as their interaction as fixed effects.

All models contained by-participant and by-item random effects. For each model, we started with a maximal random effects structure, only simplifying the model in case of non-convergence (cf. Barr et al., 2013). All categorical predictor variables in all analyses were deviation coded. The significance of fixed effects was determined by performing likelihood ratio tests to compare the fit of the model to that of a model with the same random effects structure that did not include the fixed effect. In case of significant three-level categorical predictor variables, we obtained pair-wise comparisons using a subset of the data that only contained the relevant conditions with re-centered predictor variables.

TABLE 1 | Proportion of next-mention in Experiment 1, per referent, per prompt type.

	Full stop	Pronoun
NP1	0.24	0.51
NP2	0.30	0.24
NP3	0.46	0.25

For a comparison between the three models of pronoun interpretation, we follow Rohde and Kehler (2014). We use the free prompt continuations to calculate Bayes-derived estimates of $p(referent|pronoun)$ via the prior $p(referent)$ and likelihood $p(pronoun|referent)$, as well as estimates for the Expectancy Model (normalized prior) and the Mirror Model (normalized likelihood). We then compare the model estimates with the pronoun interpretations measured in the pronoun prompt condition. We calculate the correlation between the model estimates and the observed pronoun interpretations. For these estimates, we only consider the subset of continuations in a given Ambiguous Pair condition that mention the referent who the ambiguous pronoun could refer to. While Rohde and Kehler (2014) calculate observed pronoun interpretations and model estimate both by-participant and by-item, we only compare the by-item model estimates to the by-item observed pronoun interpretation rates. A crucial difference between our experiments and Rohde and Kehler (2014) study is that we have to take into account which two out of three referents compete with each other for coreference. Obtaining, per participant, a number of observations per ambiguous pair (NP1&NP2, NP1&NP3, NP2&NP3) similar to the number of observations on which the Rohde and Kehler (2014) calculations are based requires triple the number of target items. This would make the experiments infeasibly long and very likely diminish the quality of participants' output. Since by-item and by-participant analyses in previous studies yielded similar results, we opt to only compare the model estimates to the observed pronoun interpretations *by item*. While this creates a similar data sparsity issue as the by-participant analyses, we compensate for this by increasing the number of participants.

2.2 Results

First, we replicate the well-established finding that pronoun prompts yield more NP1 continuations than full stop prompts ($\beta = 0.36$, $SE = 0.07$, $z = 4.85$, $p < 0.001$); see **Table 1** for the means collapsed across condition or **Table 2** for the same data broken down by condition. When it comes to the predictability of the referents (measured in the full stop prompts), there is a main effect of Referent (reflecting the bias away from NP1 towards NP2 and NP3; $p < 0.001$), no main effect of Ambiguous Pair ($p = 0.81$) and a Referent × Ambiguous Pair interaction ($p < 0.001$), whereby the re-mention rates of NP2 and NP3 generally differ more across the ambiguous pair conditions than does the re-mention rate of NP1. Follow-up analyses confirm that there is a main effect of ambiguous pair in the NP2 and NP3 subsets of the

TABLE 2 | Proportion of next-mention in Experiment 1, per referent, per prompt type, per ambiguous pair. The vertical columns sum to one (e.g., the re-mention rates in the NP1&NP2 condition are distributed 0.24/.25/.51 across the three referents).

	Full stop			Pronoun		
	NP1&NP2	NP1&NP3	NP2&NP3	NP1&NP2	NP1&NP3	NP2&NP3
NP1	0.24	0.23	0.28	0.82	0.70	x
NP2	0.25	0.32	0.31	0.18	x	0.55
NP3	0.51	0.45	0.41	x	0.30	0.45

TABLE 3 | Proportion of pronominalization by ambiguous vs unambiguous referents in Experiment 1 in the full stop prompt condition.

	Ambiguous	Unambiguous
NP1	0.77	0.79
NP2	0.26	0.33
NP3	0.26	0.31

TABLE 4 | Proportion of pronominalization of ambiguous referents in the full stop prompt items in Experiment 1, per referent, per ambiguous pair.

	NP1&NP2	NP1&NP3	NP2&NP3
NP1	0.82	0.72	x
NP2	0.23	x	0.27
NP3	x	0.25	0.25

TABLE 5 | Correlations between observed data and model predictions in Experiment 1, by items. *indicates significance at or below 0.001.

		Bayes	Expectancy	Mirror
by-item	R^2	0.346*	0	0.455*

data ($p < 0.01$ for both), but no main effect in the NP1 subset of the data ($p = 0.16$).

In keeping with the strong Bayes account in which factors that influence the predictability of re-mention are distinct from those that influence pronominalization, the pronominalization rates of the referents (measured in the full stop prompts) did not differ across the ambiguous pair conditions ($p = 0.98$) and the interaction between Referent and Ambiguous Pair was also not significant ($p = 0.84$). There was, however, a main effect of Referent influencing pronominalization ($p < 0.001$): The subject referent NP1 is more often re-mentioned with a pronoun than NP2 ($\beta = 49.23$, $SE = 15.60$, $z = -3.16$, $p < 0.001$) or NP3 ($\beta = 78.61$, $SE = 10.83$, $z = 7.26$, $p < 0.001$). There is no difference between NP2 and NP3 ($\beta = 4.28$, $SE = 9.53$, $z = 0.45$, $p = 0.68$); see **Table 3** for the pronominalization rates broken down by ambiguity of referent or **Table 4** for those rates broken down by referent and by condition.

We also test the effect of referent ambiguity on pronoun production. We find that unambiguous referents were more often pronominalized than ambiguous referents ($\beta = 0.50$, $SE = 0.18$, $z = 2.74$, $p < 0.01$). The interaction between Ambiguity and Referent was not significant at $p = 0.64$; see **Table 3**.

Finally, we are interested in which model yields the best correlations with the observed pronoun interpretation behavior. As in earlier work, the Bayesian Model's correlation with observed pronoun interpretation is stronger than that of the Expectancy Model; see **Table 5**. In contrast, however, the Mirror Model provided the best fit to the observed data. **Figure 1** visualizes the estimates for all three models compared to the observed pronoun interpretations for each referent in each ambiguous pair. This first of all reveals that the Expectancy Model consistently overestimates the influence of predictability: In the NP1&NP2 and NP1&NP3 ambiguous pairs, for instance, the pronoun is often interpreted as NP1 by participants, but since it is not the preferred referent for re-mention, the Expectancy Model underestimates coreference to NP1. Similarly, the Bayesian also appears to place too much importance on predictability (the prior), though not to the same extent as the Expectancy Model.

2.3 Discussion

The first question we ask is how well the Bayesian Model predicts the interpretation biases witnessed in the passage completions the participants provided, as compared to the other two models. Whereas the Bayesian Model outperformed the Expectancy Model, its predictions were not as accurate as those made by the Mirror Model. The difference between the two models is that the Bayesian Model incorporates the next-mention biases witnessed in the free-prompt data, which favored NP3 over the other two event participants. This overlaid effect of the prior was not witnessed as strongly in the interpretation biases estimated in the pronoun prompt condition, resulting in the Mirror Model being the most empirically adequate.

The second question we ask is whether there is evidence for the independence between factors that determine predictablity and pronominalization, as predicted by the strong form of the Bayesian Model. The answer here is affirmative. The most predictable referent (NP3) is not the one most often pronominalized, while the least predictable referent (NP1) is. Furthermore, comparing the re-mention rates in **Table 1** and the pronominalization rates in **Table 3**, the overall re-mention rates of NP1 and NP2 are similar (0.24 vs 0.30), but their pronominalization rates are not (0.77 versus 0.26). Conversely, the re-mention rates of NP2 and NP3 differ (0.30 versus 0.46), but their pronominalization rates do not (0.26 for both). We thus find no evidence of a dependence between predictability and pronominalization.

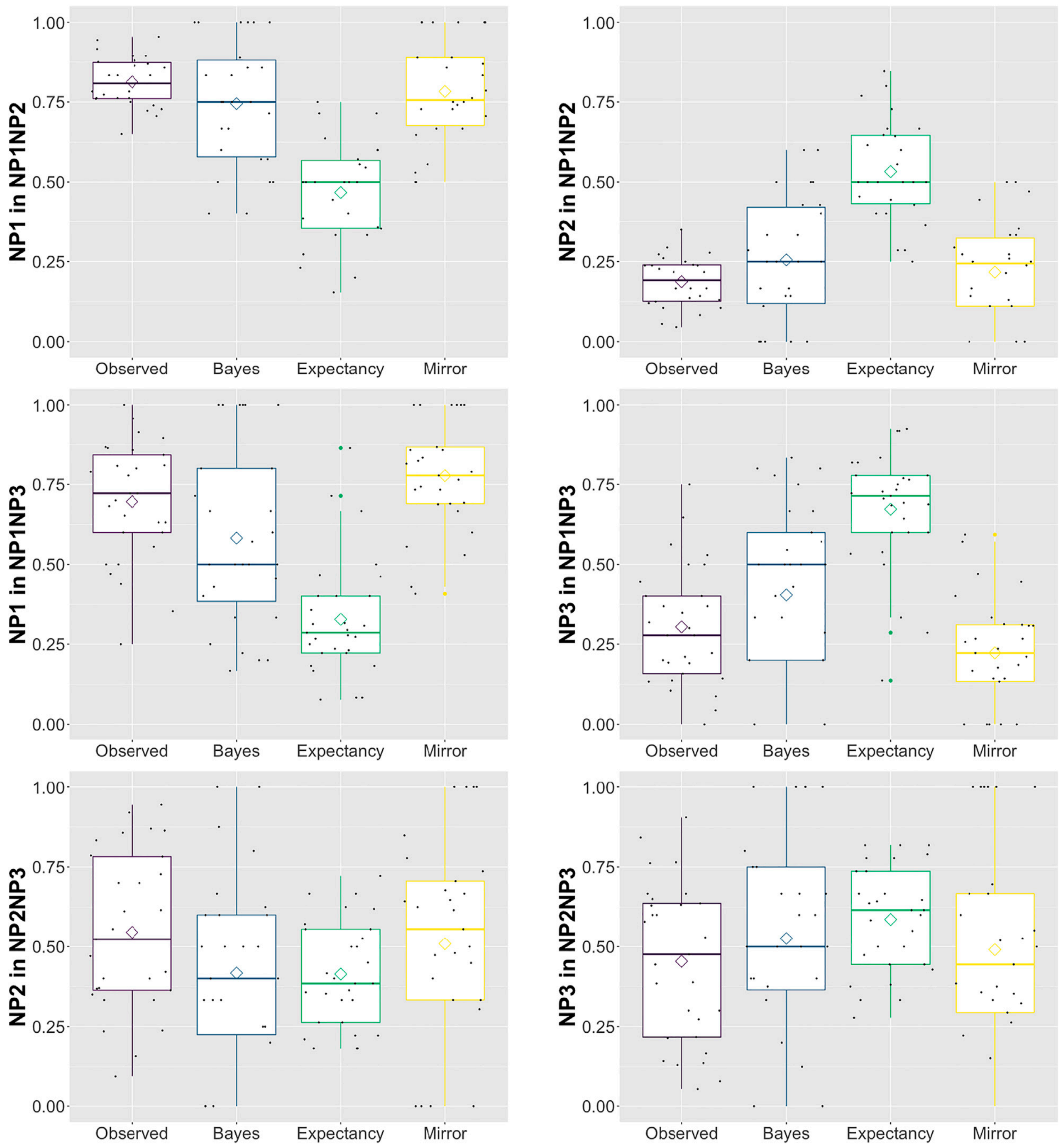


FIGURE 1 | By-item observed values and model estimates for each referent in each ambiguous pair in Experiment 1. Tilted squares indicate the means. Random jitter has been added to the individual data points to avoid overplotting.

The results, somewhat curiously, therefore appear to support the added predictions of the strong form of the Bayesian Model, but ultimately not the basic claims of the weak form, a result not seen in previous work. Comparing this study to previous ones that have found the Bayesian Model to make the best predictions, we see that our materials differ in two ways: We used a different construction in our context sentences than previous work, and also increased the number of event participants introduced in those sentences. We attempt to tease apart these two possible sources in Experiment 2 by keeping the benefactive sentence frame while reducing the number of human event participants it introduces by employing a non-human Patient. If the results witnessed in Experiment 1 are due to particular properties associated with benefactive contexts, we expect the Mirror Model to continue to outperform the Bayesian Model. On the other hand, if the issue bears on the cognitive load imposed by having to track three event participants who are introduced by name out of the blue in the context sentence, the Bayesian Model might do better in contexts where only two human event participants need to be tracked.

Our third question asks whether pronoun production is sensitive to argument/adjunct status. The results from Experiment 1 do not support the hypothesis that pronominalization rates vary systematically with argument/adjunct status beyond the well-known effects of subjecthood. If argument/adjunct status played a role in pronominalization, we would have expected variation by Ambiguous Pair such that the pronominalization rate of, for example, NP1 varied depending whether its gender-matched competitor was NP2 (an argument of the verb) or NP3 (an adjunct). Contra an account in which pronominalization rates of referents are consistently higher when their competing referent is an adjunct or consistently lower when the referent itself occupies an adjunct position (such as the account proposed to explain Rosa and Arnold 2017 thematic role effects), NP1 and NP2 show divergent behavior. For NP1, there is no increase in the pronominalization rate between the condition where the competing gender-matched referent is an argument (the NP1&NP2 condition) and that where the competing referent is an adjunct (the NP1&NP3 condition); rather there is a numeric decrease. This pattern is reversed for NP2, where the pronominalization rate does increase from the condition with an argument competitor (NP1&NP2) to the condition with an adjunct competitor (NP2&NP3). However, these numeric patterns were not sufficient to give rise to a main effect of Ambiguous Pair on pronominalization. There is thus no evidence of a consistent pattern which would support the proposed alternative explanation of the previously reported effects of thematic role on pronominalization.

Finally, the fourth question asks whether pronominalization rate is sensitive to the potential ambiguity of a pronoun. The results indicate that presence of other referents that make pronominal reference ambiguous does reduce the rate of pronominalization. Since this effect was the same across all three referents, the effect does not seem to have been influenced by the referents' topicality or predictability. Looking at **Table 3**, however, the effect of ambiguity on pronominalization appears to be modest. If ambiguity avoidance is the primary concern, one might expect this effect to be larger; as is, it is not on a par with the larger main effect of grammatical role.

3 EXPERIMENT 2

In order to ease the cognitive load of tracking three human, discourse-new referents, we replicate the setup for Experiment 1, except that we modify the stimuli so as to employ a non-human event participant in the NP2 position.

3.1 Method

3.1.1 Participants

Participants were recruited through Amazon Mechanical Turk. 85 monolingual speakers of English completed the experiment and wrote correct continuations to the catch trials (see Materials) (mean age 36.9, age range 21–71, 47 women, 2 participants preferred not to supply their gender identity). All participants were paid in exchange for their participation ($5.25).

TABLE 6 | Proportion of next-mention in Experiment 2, per referent, per prompt type.

	Full stop	Pronoun
NP1	0.24	0.67
NP3	0.76	0.33

TABLE 7 | Pronominalization rates in Experiment 2, per referent. All pronominal references are ambiguous.

	Full stop
NP1	0.86
NP3	0.18

3.1.2 Materials

Stimuli consisted of 28 target prompts that featured three arguments. Unlike in Experiment 1, however, the second argument was a non-human, usually inanimate, event participant, as in (4). The two human event participants in this experiment were of the same gender, as signalled by the default gender associated with their names. Since we are only interested in whether and how the two human potential referents are picked up, the items in this experiment correspond only to the NP1&NP3 conditions from Experiment 1.

4) a. Jacob$_{NP1}$ called the hospital for Max$_{NP3}$. ________
[full stop prompt]
b. Jacob$_{NP1}$ called the hospital for Max$_{NP3}$. He ________
[pronoun prompt]

The prompts were adapted from the items from Experiment 1 as much as possible, but not all verbs were compatible with a non-human second argument. In total, half the prompts used a verb that was also included in Experiment 1.[5]

The target items were distributed over two lists, with each item occurring only once per list, in one of the two conditions. The target items were interspersed with 32 fillers, including the same two 'catch' items that were used in Experiment 1. The other fillers varied in the number of (human) arguments they contained and whether they ended in a full stop or after the first word of a second sentence (similar to the pronoun prompt items).

3.1.3 Procedure and Annotation

The task setup and the subsequent annotation followed that of Experiment 1.

[5] All materials and analysis scripts can be found at https://tinyurl.com/BenefactivesFrontiers.

TABLE 8 | Correlations between observed data and model predictions in Experiment 2, by items. * indicates significance at or below 0.001.

		Bayes	Expectancy	Mirror
by-item	R^2	0.727*	0.300*	0.719*

3.1.4 Data Analysis

The analysis followed that of Experiment 1, except that the fixed effect of Referent was binary (NP1/NP3) and there was no fixed effect of Ambiguous Pair.

3.2 Results

As in Experiment 1, there were more NP1 re-mentions in the pronoun prompt condition than in the full stop condition (β = 2.88, *SE* = 0.25, *z* = 11.29, *p* < 0.001), as shown in **Table 6**. In addition, we again find no evidence that predictability influences pronominalization rates: While NP3 is more predictable than NP1 (β = 3.52, *SE* = 0.60, *z* = 5.87, *p* < 0.001), as shown in **Table 6**, NP1 is pronominalized more often than NP3 (β = 5.89, *SE* = 0.82*z* = 7.20, *p* < 0.001), as shown in **Table 7**.

Unlike in Experiment 1, however, the Bayesian Model yields the best correlations with the observed pronoun interpretations, as shown in **Table 8**. As can be seen from **Figure 2**, the Expectancy Model again overestimates the importance of predictability: The pronoun is more often interpreted as NP1 by participants than would be expected on the basis of the next-mention rates. In contrast, by not taking into account the predictability of the referents at all, the Mirror Model overestimates how often participants interpret the pronoun as referring to NP1 in this experiment.

Regarding our research questions about the status of competing referents and the role of ambiguity, Experiment 2 does not provide data to speak to these since the items contain only two human referents.

3.3 Discussion

Unlike in Experiment 1, the Bayesian estimates derived from the Experiment 2 data match the observed pronoun interpretation data more closely than the other two models. Also, Experiment 2 again finds no evidence in favor of a dependence between predictability and pronominalization, lending support for the strong form of the Bayesian Model.

These results suggest that the Bayesian Model's poor fit for the observed pronoun interpretation data in Experiment 1 was likely not due to properties intrinsic to the benefactive construction, but rather to the number of human event participants in the prompts. Since Bayesian reasoning relies heavily on expectations about the upcoming discourse, it could be the case that the prompts in Experiment 1 were too complex–due to their introduction of three discourse-new event participants with no other supporting context–to enable participants to create a sufficiently rich mental representation to allow for fully rational reasoning processes to take hold. The Mirror Model might then function as a sort of 'default' pronoun interpretation strategy: If participants are unable to appropriately track priors and default to the uniform distribution over the three human event participants, the Bayesian and Mirror Models make the same predictions.

In fact, previous authors have worried about the limits of the passage completion paradigm with single-sentence contexts in this respect. For instance, in their analysis of predictability on pronoun production rates, Rosa and Arnold (2017) argued that a more richly contextualized paradigm than that offered by a simple passage completion task might facilitate the development of a richer discourse representation on the part of the participant. Whereas we opted not to adopt the type of continued-story task they used in their Experiment 1 (we return to this point in the General Discussion), we agree that contexts that support richer discourse representations might better approximate natural language understanding scenarios, particularly when constructions as syntactically and semantically complex as benefactives are involved. To test this potential explanation, in Experiment 3 we return to employing benefactive prompts with three human event participants, but provide more context to facilitate the building of a mental representation of the discourse.

4 EXPERIMENT 3

In Experiment 3, like in Experiment 1, we use benefactive prompts with three human event participants, but use

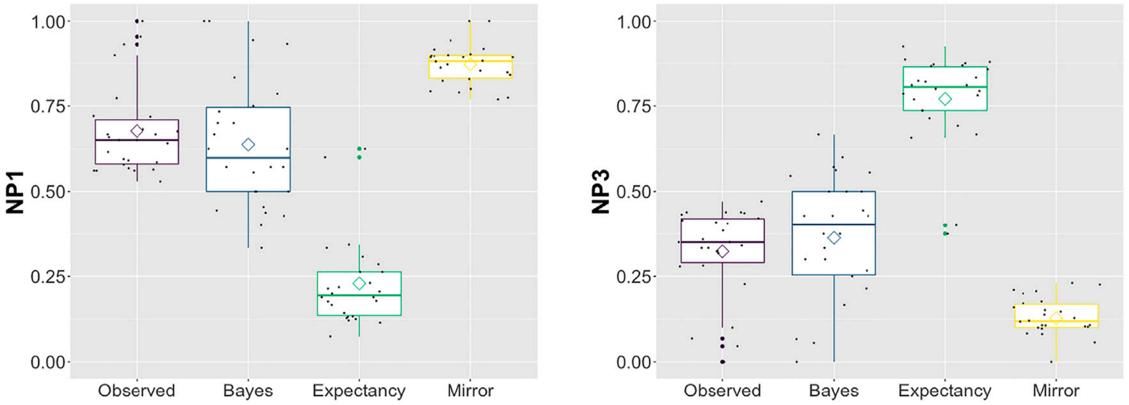


FIGURE 2 | By-item observed values and model estimates for each referent in Experiment 2. Tilted squares indicate the means.

There was an incident at the hospital with a nurse, a doctor, and a patient.
It happened late last night.
The nurse called the doctor for the patient.

FIGURE 3 | Sample item from Experiment 3 in the NP1&NP2 full stop condition.

descriptive NPs instead of proper names. In addition, we add both a verbal and visual context to help participants build a mental representation of the situation.

4.1 Method

4.1.1 Participants

Participants were recruited through Amazon Mechanical Turk. 157 monolingual speakers of English completed the experiment and wrote correct continuations for the catch trials (see Materials) (mean age 38.5, age range 20–71, 67 women, 2 participants preferred not to supply their gender identity). All participants were paid in exchange for their participation ($10).

4.1.2 Materials

Similar to Experiment 1, the stimuli consisted of a target sentence featuring three human event participants: a subject, a direct object, and a benefactive. This time, however, the referents were referred to using descriptive NPs (instead of proper names) and the target sentences followed a two-sentence context; the first sentence introduced the three event participants and the second provided a scene-setting transition that didn't mention any event participants (see **Figure 3**). In the first sentence, the referents were introduced as conjoined NPs and thus had the same grammatical and thematic role. This was done to avoid effects of the linguistic context on next-mention biases as much as possible. Right above the sentences, images of the referents were displayed, along with the corresponding descriptive NPs.[6] The order of the images corresponded to the surface order of the referents in both the context and the target sentence.

TABLE 9 | Proportion of next-mention in Experiment 3, per referent, per prompt type.

	Full stop	Pronoun
NP1	0.23	0.54
NP2	0.42	0.26
NP3	0.35	0.20

As in Experiment 1, we manipulated which two referents were gender-matched (NP1&NP2, NP1&NP3, NP2&NP3) and whether the prompt ended in a full stop or a pronoun. Pronoun prompts were ambiguous between two of the three referents (*she* in the sample item in **Figure 3**). The stimuli were distributed over 6 lists, interspersed with 30 fillers that were similar in length and composition to the target fillers and the 2 catch fillers used in Experiments 1 and 2, adapted to match the other experimental items.[7]

4.1.3 Procedure and Annotation

The task setup and the subsequent annotation followed that of Experiments 1 and 2.

4.1.4 Data Analysis

The analysis followed that of Experiment 1, which also had three referents and a manipulation of Ambiguous Pair.

4.2 Results

As in Experiments 1 and 2, there are more NP1 continuations following pronoun prompts than following full stop prompts (β = 1.63, *SE* = 0.18, *z* = 9.17, *p* < 0.001), as shown in **Table 9**. When it comes to the predictability of the referents (measured in the full stop prompts), the results follow those of Experiment 1: There is again a main effect of Referent (reflecting the bias away from NP1 towards NP2 and NP3; *p* < 0.01), no main effect of Ambiguous Pair (*p* = 0.12) and a Referent × Ambiguous Pair interaction (*p* < 0.01), whereby the re-mention rates of NP2 and NP3 generally differ more across the ambiguous pair conditions than does that of NP1, see **Table 10**. Unlike in Experiment 1, the follow-up analyses show no main effect of Ambiguous Pair in any of the Referent subsets (NP1 *p* = 0.94, NP2 *p* = 0.17, NP3 *p* = 0.32), indicating that the interaction is only apparent at the level of the whole dataset.

As in Experiment 1, the pronominalization rates of the referents do not differ between ambiguous pairs (*p* = 0.13), and the interaction between Ambiguous Pair and Referent is also not significant (*p* = 0.20). What does influence the rates of pronominalization is the grammatical role of the referent (*p* < 0.001), as shown in **Tables 11** and **12**: NP1 is pronominalized more than NP2 (β = 3.67, *SE* = 0.44, *z* = 8.27, *p* < 0.001) and NP3 (β = 4.37, *SE* = 0.68, *z* = 6.42, *p* < 0.001). There is no difference in pronominalization rate between NP2 and NP3 (β = 1.16, *SE* = 0.95, *z* = 1.23, *p* = 0.20). Since differences in re-mention rates are

[6] The images were adapted from images from the open source illustration website https://undraw.co.

[7] All materials and analysis scripts can be found at https://tinyurl.com/BenefactivesFrontiers.

TABLE 10 | Proportion of next-mention in Experiment 3, per referent, per prompt type, per ambiguous pair. The values in each column sum to one (e.g., the re-mention rates in the full-stop NP1&NP2 condition are distributed 0.22/.39/.39 across the three referents).

	Full stop			Pronoun		
	NP1&NP2	NP1&NP3	NP2&NP3	NP1&NP2	NP1&NP3	NP2&NP3
NP1	0.22	0.23	0.24	0.79	0.80	x
NP2	0.39	0.45	0.43	0.21	x	0.57
NP3	0.39	0.32	0.33	x	0.20	0.43

TABLE 11 | Proportion of pronominalization overall and for ambiguous vs unambiguous referents in Experiment 3 in the full stop prompt condition.

	Ambiguous	Unambiguous
NP1	0.63	0.68
NP2	0.11	0.13
NP3	0.12	0.15

TABLE 12 | Proportion of pronominalization of ambiguous referents in the full stop prompt items in Experiment 3, per referent, per ambiguous pair.

	NP1&NP2	NP1&NP3	NP2&NP3
NP1	0.65	0.60	x
NP2	0.12	x	0.10
NP3	x	0.08	0.15

TABLE 13 | Correlations between observed data and model predictions in Experiment 3, by items. * indicates significance at or below 0.001.

		Bayes	Expectancy	Mirror
by-item	R^2	0.385*	0	0.355*

not matched by differences in pronominalization, we again find no evidence of predictability influencing choice of referring expression.

For the effect of referent ambiguity on pronoun production, as in Experiment 1, we find that the unambiguous referents were more often pronominalized than ambiguous referents ($\beta = 0.54$, $SE = 0.22$, $z = 2.40$, $p < 0.05$). This effect is significant alongside a significant main effect of referent ($p < 0.001$) whereby NP1 is pronominalized more than the other two referents. The interaction between Ambiguity and Referent was not significant; see **Table 11**.

Looking at the correlations between the model estimates and the observed pronoun interpretations, we find that in this experiment, like in Experiment 2, the Bayesian Model makes the best predictions; see **Table 13** and **Figure 4**.

4.3 Discussion

The results from Experiment 3, like the results from Experiments 1 and 2, indicate that the prior and the likelihood are driven by different factors, as captured by the strong form of the Bayesian Model. Unlike in Experiment 1, the Bayesian Model is indeed the best fit for the observed pronoun interpretation data in Experiment 3. The crucial difference between Experiments 1 and 3 was how much contextual information was offered alongside the prompts participants were asked to continue. Whereas in Experiment 1, participants were asked to continue prompts in isolation featuring three human event participants introduced by proper names, in Experiment 3 the referents were introduced using descriptive role nouns and embedded in a longer passage with more discourse context. In addition, the prompts were accompanied by both a verbal and a visual context. The fact that the Bayesian Model outperformed the Mirror Model in this experiment suggests that predictability played a bigger role in interpreting the ambiguous pronouns in Experiment 3 than in Experiment 1.

As in Experiments 1 and 2, participants again showed a bias away from NP1 in their next mention preferences, a feature of the benefactive contexts that is useful for testing the competing models because they make different predictions in such cases. We note that Experiments 1 and 3 differ in their bias to NP2 versus NP3. This difference likely reflects the fact that the materials for the two experiments are quite different: They use different verbs and Experiment 3 contains short preceding discourse contexts, descriptive role nouns, and visual context.

Regarding the rates of pronominalization across argument/adjunct positions, the results from Experiment 3 follow Experiment 1 in providing no support for the proposed alternative explanation of the previously reported effects of thematic role on pronominalization. For both NP1 and NP2, there is no increase in the pronominalization rate between the condition where the competing gender-matched referent is an argument (the NP1&NP2 condition) and the condition where the competing referent is an adjunct (the NP1&NP3 condition for NP1 and the NP2&NP3 condition for NP2); rather there is a numeric decrease.

Regarding ambiguity, Experiment 3, like Experiment 1, shows that ambiguity appears to play a role in pronoun production. Again, participants produced more unambiguous than ambiguous pronouns, an effect that did not differ between the different referents. As in Experiment 1, however, the effect of pronoun ambiguity was small; see **Table 11**.

5 GENERAL DISCUSSION

Three experiments were conducted to evaluate the predictions of the Bayesian Model of pronoun use against those of two competing models: the Mirror Model, which derives an interpretation bias from the hearer's estimates of which entities the speaker is most

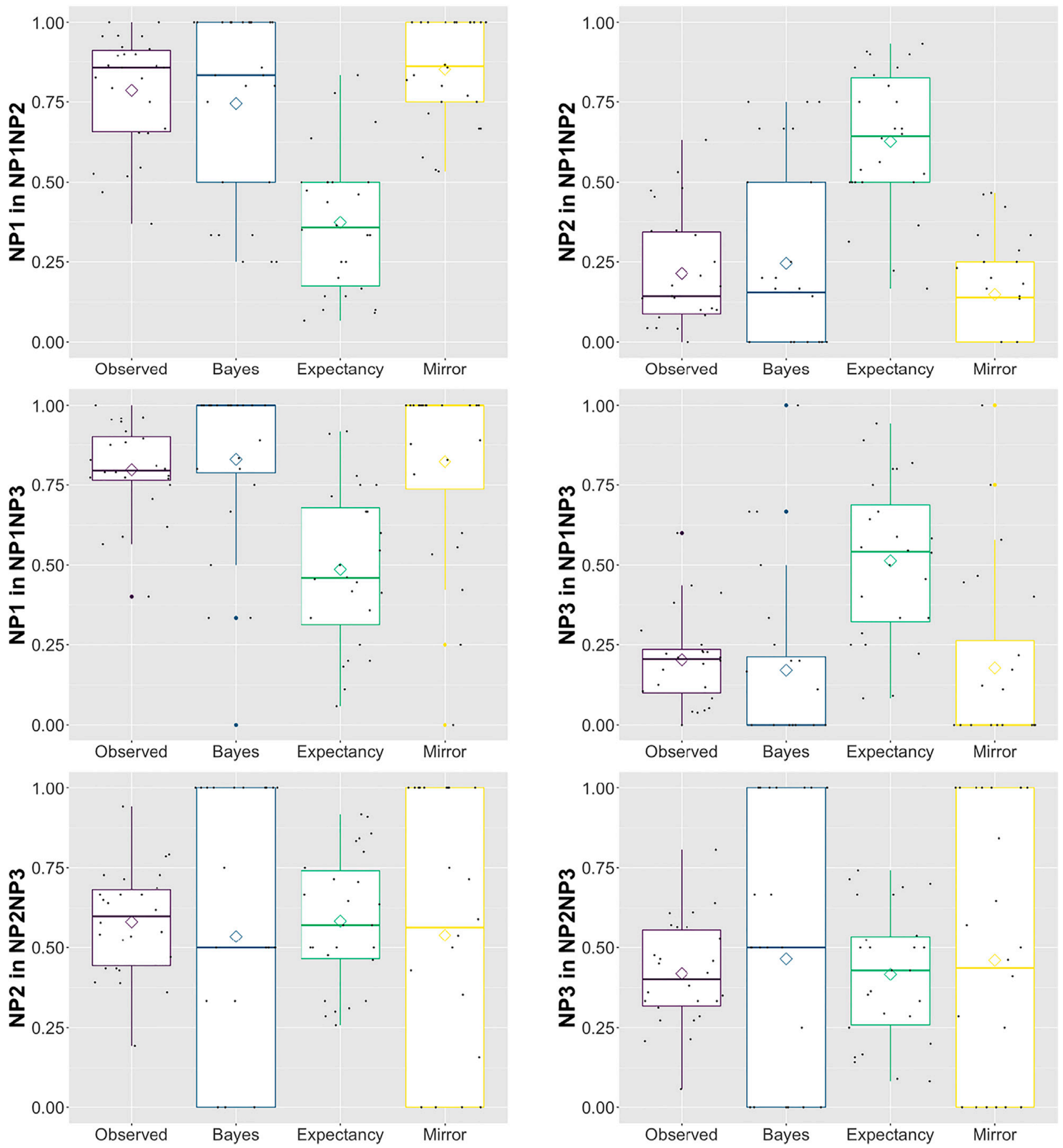


FIGURE 4 | By-item observed values and model estimates for each referent in each ambiguous pair in Experiment 3. Tilted squares indicate the means. Note: the medians at the extremes (0 and 1) in the graphs for the NP1&NP3 ambiguous pair arise in part due to the relatively low number of observations on which the model estimates are based for the NP1 and NP3 referents. In the NP1&NP3 condition, the majority of continuations were about the NP2 referent, who does not figure into the calculations here.

likely to refer to with a pronoun, and the Expectancy Model, which derives an interpretation bias from a hearer's predictions about what entities the speaker is most likely to mention next. Previous work has supported the predictions of the weak Bayesian Model in passage completions with implicit causality contexts (Rohde and Kehler, 2014; Kehler and Rohde, 2019), whereby Bayes-derived estimates of pronoun interpretation behavior yielded the best fit to participants' observed behavior when compared to those of the competing models. The current work extends the range of context types evaluated to include benefactive contexts, which mention three event participants rather than two. Interestingly, the Mirror Model yielded the best fit in Experiment 1, raising the question of what property of the target passages overrode the good fit achieved by the Bayesian Model in prior work: Was it the increased complexity from having three characters involved in the bias estimation process, or was it something about the benefactive construction itself? Experiment 2 used the benefactive construction again but reduced the number of event participants that are compatible with gendered personal pronouns *he/she* to two. The results revealed that the Bayesian Model has the best fit, suggesting that it was not the benefactive construction that derailed the Bayesian Model in Experiment 1. Experiment 3 then tested benefactive passages with three human event participants again, this time with enriched contexts including characters described with role nouns, a longer verbal context, and a visual context with images that corresponded to each event participant. In this more situated task, the Bayesian Model yielded the best fit.

These studies also lend support to a prediction of the strong Bayesian Model, revealing that the factors that influence referent re-mention are different from those that influence referent pronominalization. This pattern was evident in all three experiments, whereby re-mention biases consistently favored non-subjects and pronominalization biases consistently favored subjects. An example of this dissociation is provided by the results of Experiment 1, where there was no evidence of dependence between predictability and pronominalization: The re-mention rate of NP3 is higher than NP2 but the pronominalization rates do not differ between them, and conversely the re-mention rates of NP1 and NP2 do not differ but their pronominalization rates do. These findings uphold the strong Bayesian hypothesis.

The two experiments whose setups are most similar are Experiments 1 and 3, but they show several differences in the coreference behavior they give rise to. Overall the rate of pronominalization was higher in Experiment 1 than 3, perhaps reflecting the difficulty of tracking too many unanchored proper names in Experiment 1. Moreover, the referent who was favored for re-mention also differed. Whereas Experiment 1 favored NP3, the Beneficiary, Experiment 3 favored NP2, the Patient. This divergence may be due to the different verbs used across experiments or simply the cognitive availability of the referents for re-mention. Experiment 1 favored re-mention of NP3, often as part of an explanation of the event (e.g., *Why did Ben follow David for Sophia?* Because *Sophia wanted to know what David has been doing*), whereas Experiment 3 favored re-mention of NP2, possibly because the role nouns in the passages made the NP3 referent more peripheral to the situation (e.g., *The security guard followed the alleged shoplifter for the store manager*, with continuations about what happened to the two main characters involved in the scene: *The shoplifter tried to run but the guard tackled him* or *The security guard stopped the shoplifter in the parking lot before she could get into her car*). This comparison demonstrates how a variety of contextual factors–some of which might at first blush appear subtle or even inert–can have strong semantic and pragmatic effects on expectations about what event participants are most likely to be mentioned next. These effects on the prior in turn affect biases with respect to pronoun interpretation, as predicted by the Bayesian Model.

As we have discussed, the model fits likewise differed between Experiments 1 and 3, with the best fits being achieved by the Mirror and Bayesian Models respectively. A possible explanation for this difference is that Bayesian reasoning requires participants to have a sufficiently fine-grained mental model of the situation in order to engage in the estimation of both referent predictability and pronoun production likelihood, so as to combine them when interpreting a pronoun. Of these two components, there can be little doubt that the estimation of the prior is the more complex, as any of a number of factors that draw on semantics, pragmatics, world knowledge, and inference will come into play in predicting what the ensuing message is likely to be. The production bias, in being primarily governed by grammatical (e.g., subjecthood) and information structural (e.g., topichood) factors, does not similarly require an exploration of the (virtually infinite) ways in which a discourse might be continued in terms of content. With the more complex demands associated with making predictions from the short, one-sentence contexts in Experiment 1 that nonetheless introduced three new discourse participants with no additional information to ground them, it could be that participants proceeded with poor estimates of the priors, or even fell back on the uniform distribution. When the prior is uninformative, the Bayesian Model makes the same predictions as the Mirror Model. However, while the Bayesian Model achieved the best fit for the observed data in Experiment 3, it is clear from both the correlations (see **Table 13**) and the graphs from **Figure 4** that the Mirror Model was a close competitor. If the poor fit of the Bayesian Model in Experiment 1 was indeed due to participants being unable to estimate a reliable prior, the enriched contexts in Experiment 3 still seem to have been fairly limited in helping them do so. Compared to natural language use, even the context provided by our more situated prompts is, of course, fairly insubstantial. The hypothesis that Bayesian reasoning requires enough context for language users to build a sufficient mental representation of the situation, especially when situations get more complex (for instance with more than two referents to keep track of), should be further tested in future work.

In addition to testing the predictions of the Bayesian Model, the data from Experiments 1 and 3 also provided an opportunity to consider the role of referential ambiguity in a speaker's choice about whether to use a pronoun. Recall that the contexts in both experiments provided three potential referents, one of which could be referred to with a gender-unambiguous pronoun in the free prompt condition (for instance, NP3 in the NP1&NP2 condition) and two that would require a gender-ambiguous pronoun (NP1 and NP2 in the NP1&NP2 condition). Our comparison of the pronominalization rates of referents when they were and were not part of the pair sharing the same gender showed that ambiguity does indeed have an effect. That having been said, on an account in which likelihood of pronominalization is dependent on referential ambiguity (Hendriks et al., 2014; Horton and Keysar 1996; Matthews et al., 2006; Nadig and Sedivy 2002, though cf.; Fukumura and van Gompel 2012; Arnold and Griffin 2007), one might expect to see higher rates of pronominalization in contexts in which the referent can be referred to unambiguously, since referential success in such contexts is not at stake. What we see instead, however, is a remarkable similarity in pronominalization rates across the unambiguous and ambiguous cases. If ambiguity avoidance is as influential as grammatical role, for instance, one might expect to see an effect of similar magnitude. Instead, the effect of ambiguity, while significant, does not rival grammatical role in effect size. Such results raise the question of why ambiguity effects emerge but are far smaller than what an ambiguity avoidance account might predict.

Finally, we note that research on the Bayesian Model has primarily focused on two context types, Source-Goal transfer-of-possession verbs and object-biased implicit causality verbs. This is for good reason: Whereas in most contexts the next-mention and pronominalization biases are likely to both favor the subject, the next-mention biases for these two constructions point away from the subject, thereby providing an opportunity to study divergences between production biases that favor the subject and next-mention biases that favor a non-subject. A result of the current study is the identification of benefactives as a third construction type of this sort, whereby the re-mention rate of NP3 was consistently higher than that of NP1 (albeit lower than NP2 in Experiment 3). We see two potential explanations for the high next-mention bias to NP3. The first bears on the meaning of the benefactive construction and its role in generating discourse expectations. In the case of object-biased implicit causality verbs, the hypothesis is that these verbs have the ability both to generate an expectation for an ensuing explanation and to impute causality to the direct object, thereby creating an expectation that the object will be mentioned next. In the case of Source-Goal transfer-of-possession verbs, the bias plausibly results from an expectation that the speaker will next describe what the recipient did with the object-of-transfer they just received. It could be that benefactives generate a high next-mention bias to NP3 for similar reasons, e.g., by creating an expectation that the speaker will next describe why the beneficiary would want the event to be performed or how the beneficiary reacted to the event that was performed on their behalf.

As pointed out by a reviewer, however, another possible explanation stems from the fact that the NP3 argument is optional in the benefactive construction. Arnold (2001) previously compared next-mention biases within Source–Goal and Goal–Source transfer-of-possession contexts, and found that non-subject (Source) referents were re-mentioned unexpectedly often in Goal–Source cases. Unlike Source–Goal sentences, in which all three thematic roles are presented in obligatory arguments, the Source is optional in Goal–Source sentences (e.g., *Mary received the book from Sue* and *Mary received the book* are both acceptable). Arnold hypothesized that participants may have felt the need to re-mention the Source in the continuation in order to justify its inclusion in the story. In a study that compared active and passive IC contexts, Rohde and Kehler (2014) similarly found that the re-mention rate of the logical subject in their free-prompt condition was higher in passive contexts–where it is mentioned from within an optional *by*-adjunct–than in the active condition, and followed Arnold in speculating that the optionality of including the *by*-adjunct was the reason for the effect. As such, it is possible that the bias toward NP3 in benefactives is due to the same reason. The results presented here do not inform the question of which explanation is correct, but whichever one proves to be, benefactives can be added to the list of context types capable of evaluating claims concerning the dissociation between pronoun production and interpretation biases.

Our results using benefactive contexts largely revealed that the types of semantic factors that affect next-mention biases do not similarly affect production biases, in line with recent work using IC contexts, but in contrast with Rosa and Arnold (2017) results on transfer-of-possession. One of our goals was to evaluate a hypothesis expressed in previous work (Fukumura and Van Gompel, 2010; Rohde and Kehler, 2014) that the effects found for transfer contexts may be due to the imbalance between the argument status of the Goal in Source-Goal frames and adjunct status of the Source in Goal-Source frames. Whereas we investigated this question in the context of benefactive instead of transfer contexts, our results do not support that explanation of Rosa and Arnold's results: Whereas in Experiment 1 the pronominalization rate of NP2 went up slightly when the competing referent was an adjunct compared to an argument, the effect wasn't significant, and in the cases of NP1 in Experiment 1 and both NP1 and NP2 in Experiment 3, the differences went numerically in the wrong direction.

This leads us to wonder about other explanations for the effects found by Rosa and Arnold. One obvious possibility is that the results are sound, and that the strong form of the Bayesian analysis is, well, too strong. This conclusion would of course be welcome if it captures the reality of the facts, and would not itself provide any evidence against the weak form of the hypothesis. It should nonetheless give us pause in light of our current state of knowledge, however, since effects of predictability have been not been found in IC contexts nor (now) benefactive contexts. The most obvious explanation for why predictability would affect pronominalization is the rationale behind the common wisdom outlined in the introduction, whereby the speaker and hearer are coordinating via a singular notion of entity salience when producing and interpreting a pronoun respectively. The recent data however, when considered as an ensemble, provides little evidence to support that view: no effect of predictability has been found for IC and benefactive contexts, and the effects reported for transfer-of-possessive contexts are smaller and more varied than this explanation would predict. We are thus left with the question of what type of model would predict this mixed pattern of effects.

Further commentary must necessarily remain speculative. The primary support for an effect of thematic role on pronominalization comes from Rosa and Arnold's first experiment, where an effect for both grammatical roles was found, albeit much more strongly for subjects.[8] Their Experiment 1 utilized a paradigm in which the stimuli were presented as a continuous story, which carries with it complications that one does not find in the standard passage completion paradigm. In particular, while the continuous story paradigm clearly does not affect theoretical predictions regarding the effect of grammatical role on pronominalization, it is much less clear that the same is true for theories that tie pronominalization

[8] As mentioned earlier, Rosa and Arnold's Exp 2 yielded no apparent effect for gender-ambiguous contexts like those studied here and in previous work, and Experiment 3 revealed a small effect for non-subjects only. There is a potential worry concerning the results of both Exps. 2 and 3, however, in that role nouns were used without clip art to disambiguate gender, as used in their Experiment 1 and the studies presented here. This means that one cannot be sure what contexts were viewed by participants to be gender ambiguous vs unambiguous. This worry receives support from the fact that Rosa and Arnold saw cases of this based on the nature of the continuations that participants provided, which led them to recategorize the gender of two of their characters post-hoc.

rates to topicality, since inferences about the relative topicality of referents can be affected by any of a number of factors as the mental models of the interlocutors evolve throughout a discourse. The fact that participants themselves produced half of the utterances that comprised each discourse means that each discourse was unique, and hence the topicality status of potential referents at different points in the discourse would be expected to vary as well. This worry receives support from the fact that Rosa and Arnold found a significant effect of stimulus order: Two lists were employed, and which list a participant saw reliably affected their pronominalization rates, despite the fact that the individual prompts were the same. In contrast, while the design of our Experiment 3 followed Rosa and Arnold in using more extended contexts, care was taken to control for topicality: The three event participants were introduced from a coordinate noun phrase that offered no topicality advantage for any potential referent, with an intervening scene-setting clause that did not mention any of them. Using these carefully constructed discourses that were nonetheless richer than the single-sentence contexts used in our Experiment 1, the expected effects of semantic factors on next-mention biases were found, but no effects of these factors were found on production biases. An obvious next step for future work is to examine transfer contexts with extended, albeit more carefully controlled, stimuli.[9]

In sum, the results presented here demonstrate that the Bayesian Model scales well from its previous applications to a new domain: benefactive constructions with two or three human event participants. However, this was only true in the three event participant case when a verbal and visual context was present to (by hypothesis) allow participants to track the available referents and build an adequate mental representation of the situation being described. This hypothesis, of course, immediately evokes questions for future work: For instance, do all language users use Bayesian reasoning when faced with ambiguous pronouns, regardless of mental capacity or task demands? Indeed, there is evidence that not everyone can always engage in predictive processing [e.g., children, non-native speakers, and non-student and older adults (Huettig, 2015; Pickering and Gambi, 2018; Grüter et al., 2012)]. For example, non-native speakers don't make the same coreference predictions that native speakers do in contexts with transfer-of-possession verbs, a finding that has been attributed to the increased difficulty of real-time next-mention computations during second language processing (Grüter and Rohde, 2021). Further research can thus shed light on whether our hypothesis regarding the differences witnessed in Experiments 1 and 3 is on the right track.

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://tinyurl.com/BenefactivesFrontiers.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by School of Philosophy, Psychology, and Languages Sciences ethics panel, University of Edinburgh. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JH contributed to conceptualisation, methodology, analysis, and writing—original draft. HR and AK contributed to design, methodology, and writing—review and editing.

ACKNOWLEDGMENTS

We thank Joseph Cohen and Emily Williamson for help in annotation. This work was supported by a Leverhulme Trust Prize in Languages and Literatures to HR.

[9] Indeed, there are other aspects of Rosa and Arnold's stimuli that could potentially be cause for concern. For one, an examination of their stimuli suggests that in some context sentences, the event participants were introduced with different referential forms, varying among proper names, definite lexical NPs, and indefinite NPs (e.g., *The maid handed a piece of cake to Sir Barnes; Sir Barnes bought earrings from a sales clerk*). Information structure theorists have posited that form of reference, like grammatical role, influences the likelihood of an entity being the topic, with pronominalized antecedents being the strongest indicator, followed by other definites (proper names; *the*-NPs), and finally with indefinites being the poorest prospects (Lambrecht, 1994, p. 165, inter alia). Thus, mixing these forms across potential referents in a single context sentence potentially creates a confound. There are also other irregularities of smaller scope. First, included are transfer verbs that appear in the double object construction (*The maid gave Lady Mannerly a glass of champagne; Lady Mannerly handed the maid a duster and a broom*). The hypothesis that topicality conditions pronominalization rates does not treat Goals introduced as indirect objects to be on a par with those introduced as the object of a PP (with indirect objects being more topical, by virtue of their being higher on the obliqueness hierarchy), and no double object construction is available for the corresponding Goal-Source transfer verbs (* *Lady Mannerly received the maid a glass of champagne*). Second, some sentences we understand as being part of the stimuli are not transfer-of-possession verbs at all (ex 3b, *Lady Mannerly gave a backrub to Sir Barnes*)—such cases do not create an expectation that the next sentence will describe what the Goal did next with the object of transfer–and others only involve transfer-of-possession in an abstract, metaphorical sense (e.g., *The chauffeur taught shooting techniques to the butler*). Third, at least one stimulus–*Sir Barnes received a painting of the two of them from Lady Mannerly*, given in **Figure 2** of the paper–contains a pronoun that refers to both participants, one anaphorically and one cataphorically. This should be avoided, since additional mentions can influence the salience and topicality status of event participants beyond the mentions that fill the grammatical and thematic roles under scrutiny. Finally, some verbs occurred multiple times in the same stimulus set (e.g., *handed* occurs seven times by our count), and verb re-use is not balanced between the Source-Goal and Goal-Source contexts. We of course cannot say for sure that any or all of these factors influenced the effects found, but do nonetheless suggest that a follow-on study that remedies these issues is in order.

REFERENCES

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. London: Routledge.

Arnold, J. E., and Griffin, Z. M. (2007). The Effect of Additional Characters on Choice of Referring Expression: Everyone Counts. *J. Mem. Lang.* 56, 521–536. doi:10.1016/j.jml.2006.09.007

Arnold, J. E. (1998). Reference Form and Discourse Patterns. Ph.D. thesis. Stanford, CA: Stanford University.

Arnold, J. E. (2008). Reference Production: Production-Internal and Addressee-Oriented Processes. *Lang. Cogn. Process.* 23, 495–527. doi:10.1080/01690960801920099

Arnold, J. E. (2001). The Effect of Thematic Roles on Pronoun Use and Frequency of Reference Continuation. *Discourse Process.* 31, 137–162. doi:10.1207/s15326950dp3102_02

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, 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

Cheng, W., and Almor, A. (2019). A Bayesian Approach to Establishing Coreference in Second Language Discourse: Evidence from Implicit Causality and Consequentiality Verbs. *Bilingualism* 22, 456–475. doi:10.1017/s136672891800055x

Fukumura, K., and Van Gompel, R. P. G. (2010). Choosing Anaphoric Expressions: Do People Take into Account Likelihood of Reference?. *J. Mem. Lang.* 62, 52–66. doi:10.1016/j.jml.2009.09.001

Fukumura, K., Van Gompel, R. P. G., Harley, T., and Pickering, M. J. (2011). How Does Similarity-Based Interference Affect the Choice of Referring Expression?. *J. Mem. Lang.* 65, 331–344. doi:10.1016/j.jml.2011.06.001

Fukumura, K., and van Gompel, R. P. G. (2012). Producing Pronouns and Definite Noun Phrases: Do Speakers Use the Addressee's Discourse Model?. *Cogn. Sci.* 36, 1289–1311. doi:10.1111/j.1551-6709.2012.01255.x

Givón, T. (1983). *Topic Continuity in Discourse: A Quantitative Cross-Language Study*, Vol. 3. Amsterdam: John Benjamins Publishing.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A Framework for Modelling the Local Coherence of Discourse. *Comput. Linguistics* 21, 203–225. doi:10.21236/ada324949

Grüter, T., Lew-Williams, C., and Fernald, A. (2012). Grammatical Gender in L2: A Production or a Real-Time Processing Problem?. *Second Lang. Res.* 28, 191–215. doi:10.1177/0267658312437990

Grüter, T., and Rohde, H. (2021). Limits on Expectation-Based Processing: Use of Grammatical Aspect for Co-Reference in L2. *Appl. Psycholinguistics* 42, 51–75. doi:10.1017/s0142716420000582

Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive Status and the Form of Referring Expressions in Discourse. *Language* 69, 274–307. doi:10.2307/416535

Hendriks, P., Koster, C., and Hoeks, J. C. J. (2014). Referential Choice Across the Lifespan: Why Children and Elderly Adults Produce Ambiguous Pronouns. *Lang. Cogn. Neurosci.* 29, 391–407. doi:10.1080/01690965.2013.766356

Horton, W. S., and Keysar, B. (1996). When Do Speakers Take into Account Common Ground?. *Cognition* 59, 91–117. doi:10.1016/0010-0277(96)81418-1

Huettig, F. (2015). Four central Questions about Prediction in Language Processing. *Brain Res.* 1626, 118–135. doi:10.1016/j.brainres.2015.02.014

Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAs (Transformation or Not) and Towards Logit Mixed Models. *J. Mem. Lang.* 59, 434–446. doi:10.1016/j.jml.2007.11.007

Kehler, A., Kertz, L., Rohde, H., and Elman, J. L. (2008). Coherence and Coreference Revisited. *J. Semant* 25, 1–44. doi:10.1093/jos/ffm018

Kehler, A., and Rohde, H. (2013). A Probabilistic Reconciliation of Coherence-Driven and Centering-Driven Theories of Pronoun Interpretation. *Theor. Linguistics* 39, 1–37. doi:10.1515/tl-2013-0001

Kehler, A., and Rohde, H. (2019). Prominence and Coherence in a Bayesian Theory of Pronoun Interpretation. *J. Pragmatics* 154, 63–78. doi:10.1016/j.pragma.2018.04.006

Lam, S. Y., and Hwang, H. (2021). "Interpretation of Null Pronouns in Mandarin Chinese Does Not Follow a Bayesian Model," in Paper presented at the 34th Annual CUNY Conference on Human Sentence Processing. Philadelphia, United States.

Lambrecht, K. (1994). *Information Structure and Sentence Form*. Cambridge: Cambridge University Press.

Matthews, D., Lieven, E., Theakston, A., and Tomasello, M. (2006). The Effect of Perceptual Availability and Prior Discourse on Young Children's Use of Referring Expressions. *Appl. Psycholinguistics* 27, 403–422. doi:10.1017/s0142716406060334

Mayol, L. (2018). Asymmetries between Interpretation and Production in Catalan Pronouns. *Dialogue & Discourse* 9 (2), 1–34. doi:10.5087/dad.2018.201

Nadig, A. S., and Sedivy, J. C. (2002). Evidence of Perspective-Taking Constraints in Children's On-Line Reference Resolution. *Psychol. Sci.* 13, 329–336. doi:10.1111/j.0956-7976.2002.00460.x

Pickering, M. J., and Gambi, C. (2018). Predicting while Comprehending Language: A Theory and Review. *Psychol. Bull.* 144, 1002–1044. doi:10.1037/bul0000158

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rohde, H. (2008). Coherence-Driven Effects in Sentence and Discourse Processing. Ph.D. thesis. UC San Diego.

Rohde, H., and Kehler, A. (2014). Grammatical and Information-Structural Influences on Pronoun Production. *Lang. Cogn. Neurosci.* 29, 912–927. doi:10.1080/01690965.2013.854918

Rosa, E. C., and Arnold, J. E. (2017). Predictability Affects Production: Thematic Roles Can Affect Reference Form Selection. *J. Mem. Lang.* 94, 43–60. doi:10.1016/j.jml.2016.07.007

Stevenson, R. J., Crawley, R. A., and Kleinman, D. (1994). Thematic Roles, Focus and the Representation of Events. *Lang. Cogn. Process.* 9, 519–548. doi:10.1080/01690969408402130

Zhan, M., Levy, R., and Kehler, A. (2020). Pronoun Interpretation in Mandarin Chinese Follows Principles of Bayesian Inference. *PLOS ONE* 15 (8), e0237012. doi:10.1371/journal.pone.0237012

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.

Using Rational Models to Interpret the Results of Experiments on Accent Adaptation

Maryann Tan[1,2*†], *Xin Xie*[2,3†] and *T. Florian Jaeger*[2,4]

[1] Centre for Research on Bilingualism, Department of Swedish Language & Multilingualism, Stockholm University, Stockholm, Sweden, [2] Brain & Cognitive Sciences, University of Rochester, Rochester, NY, United States, [3] Department of Language Science, University of California, Irvine, Irvine, CA, United States, [4] Computer Science, University of Rochester, Rochester, NY, United States

Exposure to unfamiliar non-native speech tends to improve comprehension. One hypothesis holds that listeners adapt to non-native-accented speech through distributional learning—by inferring the statistics of the talker's phonetic cues. Models based on this hypothesis provide a good fit to incremental changes after exposure to atypical *native* speech. These models have, however, not previously been applied to non-native accents, which typically differ from native speech in many dimensions. Motivated by a seeming failure to replicate a well-replicated finding from accent adaptation, we use ideal observers to test whether our results can be understood solely based on the statistics of the relevant cue distributions in the native- and non-native-accented speech. The simple computational model we use for this purpose can be used predictively by other researchers working on similar questions. All code and data are shared.

Keywords: non-native speech, L2 speech, adaptation, distributional learning, ideal observer, computational modeling, rational cognition

OPEN ACCESS

Edited by:
Rory Turnbull,
Newcastle University, United Kingdom

Reviewed by:
Michelle R. Kapolowicz,
University of California, Irvine,
United States
Stefania Degaetano-Ortlieb,
Saarland University, Germany

****Correspondence:***
Maryann Tan
maryann.tan@biling.su.se

†These authors share first authorship

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

Received: *04 March 2021*
Accepted: *14 September 2021*
Published: *05 November 2021*

Citation:
Tan M, Xie X and Jaeger TF (2021)
Using Rational Models to Interpret the
Results of Experiments on Accent
Adaptation.
Front. Psychol. 12:676271.
doi: 10.3389/fpsyg.2021.676271

1. INTRODUCTION

Understanding strongly non-native-accented speech can be challenging: native listeners unfamiliar with a non-native accent tend to process it more slowly and with decreased accuracy (Munro and Derwing, 1995; Witteman et al., 2013). There is now ample evidence that this initial processing disadvantage can decrease with exposure to the accented talker (e.g., Weil, 2001; Bradlow and Bent, 2008; Adank et al., 2009), with some improvements emerging within mere minutes of exposure (Clarke and Garrett, 2004; Xie et al., 2018b). What has remained less well-understood are the mechanisms underlying these changes in speed and accuracy of processing.

Two broad classes of (mutually compatible) hypotheses have emerged. One holds that changes in native listeners' processing of non-native-accented speech arise from a general relaxation of decision criteria for phonological categorization (e.g., "general expansion", Schmale et al., 2012). The other hypothesis holds that listeners learn talker- or even accent-specific characteristics, including information about specific segmental features and super-segmental properties of the accented speech (e.g., Bradlow and Bent, 2008; Sidaras et al., 2009). This latter hypothesis has received further elaboration: that adaptation to non-native accents is at least in part achieved through distributional learning (Wade et al., 2007; Idemaru and Holt, 2011; Schertz et al., 2015; Kartushina et al., 2016) of the type assumed in exemplar (Pierrehumbert, 2001) or Bayesian theories of speech perception (Kleinschmidt and Jaeger, 2015).

Distributional learning models have been found to provide a good qualitative and quantitative explanation of certain adaptive changes listeners exhibit in response to shifted or otherwise atypical pronunciations by native talkers (Clayards et al., 2008; Bejjanki et al., 2011; Kleinschmidt and Jaeger, 2015, 2016; Theodore and Monto, 2019). This includes changes in categorization boundaries observed in perceptual recalibration (e.g., Norris et al., 2003; Eisner and McQueen, 2005; Kraljic and Samuel, 2006; Drouin et al., 2016) or unsupervised learning paradigms (Clayards et al., 2008; Nixon et al., 2016). However, tests of distributional learning models have almost exclusively been limited to comparatively small deviations from the expected means or variances of two phonological categories along a single phonetic dimension (for examples with two phonetic dimensions, see Hitczenko and Feldman, 2016; Xie et al., 2021a). Whether distributional learning can explain adaptation to the types of more complex deviations from expected pronunciations that are observed in unfamiliar non-native accents is an open question. Specifically, non-native accents differ from the expected native pronunciation along many acoustic and linguistic dimensions, including both supra-segmental and segmental differences. Non-native speech might, for example, realize segmental or supra-segmental categories with means that are shifted relative to native means (Best, 1995; Flege, 1995) and with expanded or reduced variance (Smith et al., 2019; Vaughn et al., 2019; Xie and Jaeger, 2020), including deviation in terms of the relative reliance on different cues to signal the same phonological contrast (Flege et al., 1992; Xie et al., 2017). In short, adaptation to a talker with an unfamiliar non-native accent constitutes a more complex problem than adjustments in response to more limited differences between native talkers, and it is possible that these challenges require a different set of mechanisms (for related discussion see Goslin et al., 2012; Porretta et al., 2017).

We take a hugely simplified step toward addressing this question. Our approach is *post-hoc* and confirmatory (although future work might employ the same approach *pre*dictively prior to data collection). We ask whether a simple model of speech perception (an ideal observer, Clayards et al., 2008; Norris and McQueen, 2008; Kleinschmidt and Jaeger, 2015) can be employed to make informative predictions as to whether exposure to a specific set of non-native-accented speech stimuli is expected to result in detectable adaptation (see also Hitczenko and Feldman, 2016). To demonstrate the potential value of such an approach, we ask whether an ideal observer sheds light on what appeared to be, at first blush, a failure to replicate previous findings from accent adaptation (Eisner et al., 2013; Xie et al., 2017), despite very similar design and procedure.

We emphasize that our goal here is not to convincingly argue that distributional learning is the best explanation for the data at hand. Rather, we aim to demonstrate *how* one can use a simple normative model of speech perception to derive predictions for the perception of, and adaptation to, non-native-accented speech. By comparing the responses of human listeners to the predictions of this computational model, researchers can achieve a clearer sense of which results (null or not) should be treated as surprising (see also Massaro and Friedman, 1990, on the value of normative models for speech perception). While models of speech perception suitable for this purpose now exist (Clayards et al., 2008; Kleinschmidt and Jaeger, 2015), they are rarely employed in the interpretation of experimental results (but see e.g., Lancia and Winter, 2013; Kleinschmidt et al., 2015; Hitczenko and Feldman, 2016; Theodore and Monto, 2019; Xie et al., 2021a). Research in experimental psychology often remains focused on *effects* with less discussion of whether these effects are *predicted by existing theories or models* (see discussion in Jaeger et al., 2019). When models are evoked, it is not uncommon that predictions are attributed to them without verifying that a computational model would actually make those predictions. These practices are arguably particularly problematic when applied to human behavior that is known to be affected by previously experienced input (as is the case for speech perception in general and accent adaptation in particular). Even for theories of speech perception that are conceptually simple, the predictions of these models tend to depend on the statistics of previously experienced speech in non-trivial ways. This is precisely the type of situation in which computational studies can provide a deeper understanding of experimental findings, and prevent misunderstandings of existing theory.

The present report aims to demonstrate how even the *post-hoc* application of computational models to experimental data can aid interpretation. It also holds the potential to reduce the "file drawer" problem (Rosenthal, 1979)—the bias to not publish null results—as well as to pre-empt the "over-interpretation" of null results. As we illustrate below, not every null result is a Type II error; null results can be precisely what a model predicts given the specific stimuli of an experiment. We thus hope this report can serve as a helpful guide for researchers, encouraging experimenters to interpret results relative to theoretical models that are sufficiently specified to make predictions. To this end, this report is accompanied by detailed **Supplementary Material** written as executable, richly documented, R markdown (Allaire et al., 2021) and compiled into an interactive HTML. These **Supplementary Material**, along with all data, are shared via the Open Science Framework (https://osf.io/72fkx/). The main text aims to provide a high-level overview of the approach and results.

2. THE 'PUZZLE'

The two perception experiments we aim to understand share the same exposure-test design and procedure (**Figure 1**), but differ in the L1-L2 pair investigated. Both experiments investigate adaptation to non-native-accented speech of a single unfamiliar talker (see also Clarke and Garrett, 2004; Eisner et al., 2013, a.o.). Both experiments focus on the realization of the same phonological category—syllable-final stop voicing of /d/, and its contrast to /t/—present in the L2s, but absent in L2 talkers' L1s.

The first experiment exposed native speakers of American English to Mandarin-accented English speech (Xie et al., 2017) while the second exposed native speakers of Swedish to Flemish-accented Swedish. Both Mandarin-accented English and Flemish-accented Swedish are known to differ from native English and Swedish, respectively, in the realization of final

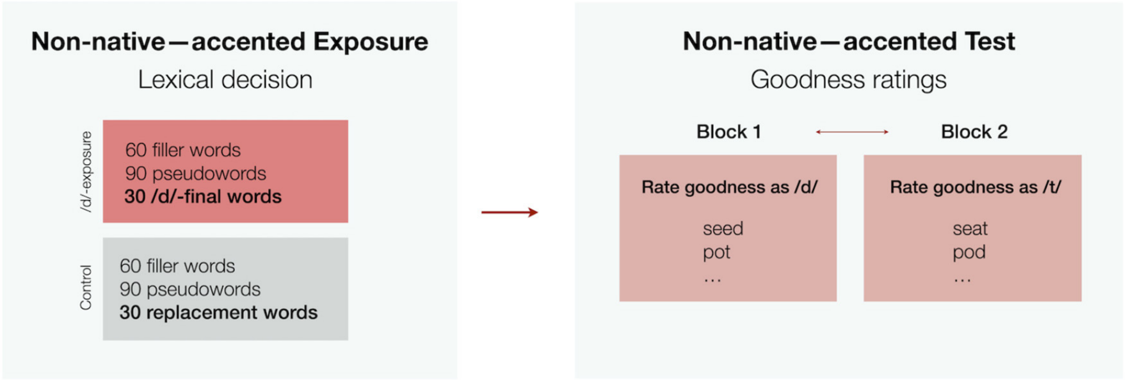


FIGURE 1 | Design of English and Swedish experiment analyzed here. The order of /d/- and /t/-goodness test blocks was counter-balanced across participants.

stop-voicing (Tan et al., 2019; Xie and Jaeger, 2020). As the Swedish study was designed as a replication of the English study, we describe the English study first.

Unlike English, Mandarin does not have stops in syllable-final position. As would be expected from theories of L2 learning (e.g., Flege, 1995), the realization of final stop-voicing differs between native English and Mandarin-accented English (Flege et al., 1992; Xie and Jaeger, 2020). This was also confirmed specifically for the non-native-accented speech materials used in the experiment (Xie et al., 2017).

Exposure was manipulated between participants. Both groups heard 90 words and 90 pseudowords while conducting a lexical decision task. For the *\/d/-exposure* group, this included 30 words containing a syllable-final /d/ (e.g., *lemonade*). These exposure words were chosen to not have minimal pair neighbors with syllable-final /t/, allowing lexical guidance on the non-native talker's /d/ productions. Participants in the *control group* heard no words with syllable-final /d/ (for details about the materials, see **Supplementary Material**). Neither groups heard syllable-final /t/ productions during exposure.

During test, participants in both groups heard the same minimal pair words with syllable-final /d/ or /t/ (e.g., a recording of *seed* or *seat*). Participants had to rate how "good" the word sounded as an instance of /d/ (one block) or /t/ (another block, with the order of blocks counter-balanced across participants). Words within the same minimal pair did not appear in the same block (see **Figure 1**).

Goodness ratings have been used to analyze listeners' representations of the internal structure of phonological categories (e.g., Samuel, 1982; Volaitis and Miller, 1992; Allen and Miller, 2001), including after exposure to shifted native categories in perceptual recalibration (e.g., Drouin et al., 2016). Xie et al. (2017) found that /d/-exposure led to improved goodness ratings for the non-native-accented /d/- and /t/-final words during test, compared to the control group. We refer to this as the English data. Xie and colleagues replicated the effect of /d/-exposure in three additional experiments using the same recordings and similar exposure-test paradigms but different tasks and participants (Xie and Myers, 2017; Xie et al., 2017, 2018a). The same effect has also been found in experiments with similar designs on syllable-final /d/ in Dutch-accented English, which tends to devoice final stops (Eisner et al., 2013).

In a recent experiment however, we failed to find the effect of /d/-exposure for another L1-L2 pair, Flemish-accented Swedish. Unlike Swedish, Flemish (a dialect of Dutch) devoices voiced stops in syllable-final position (Booij, 1999; Verhoeven, 2005). This type of phonological rule is well-documented to transfer from a talker's first language to their second language and was confirmed in the L2-accented speech materials used in the Swedish experiment (Tan et al., 2019). Like with Dutch- and Mandarin-accented English, we thus expected exposure to Flemish-accented Swedish syllable-final /d/ to affect ratings during test. Both the English and Swedish experiments used lexically-guided exposure with the same task. Both experiments manipulated exposure to the non-native-accented sound (syllable-final /d/) in the same two between-participant conditions, including the same amount of exposure. Both experiments used /d/ and /t/ goodness ratings of /d/-/t/-final minimal pair words during test. Unlike Xie et al. (2017), however, the Swedish data did *not* yield an effect of /d/-exposure on ratings during test. In fact, the effect of exposure went numerically in the opposite direction in the Swedish data.

Figure 2 (top) shows the rating results from both experiments. Linear mixed-effects regression presented in the **Supplementary Material** (section 3.2.2) confirmed that the effects of exposure differed significantly between the two experiments (coefficient-based t-test, $p < 0.002$): whereas /d/-exposure resulted in significant facilitation for English ($\hat{\beta} = 0.04, p < 0.0001$), it did not for Swedish—in fact, trending in the opposite direction ($\hat{\beta}$ =-0.03, $p > 0.1$).

At first blush, the Swedish data seem to constitute a failure to replicate the English experiment. In particular, since the effect found in the English data has been replicated a number of times, it would be tempting to consider the Swedish result a Type II error (rather than the English result a Type I error). Further, adding to this interpretation, the Swedish experiment collected substantially less data: while the English data consists of 120 ratings each from 48 participants, the Swedish data

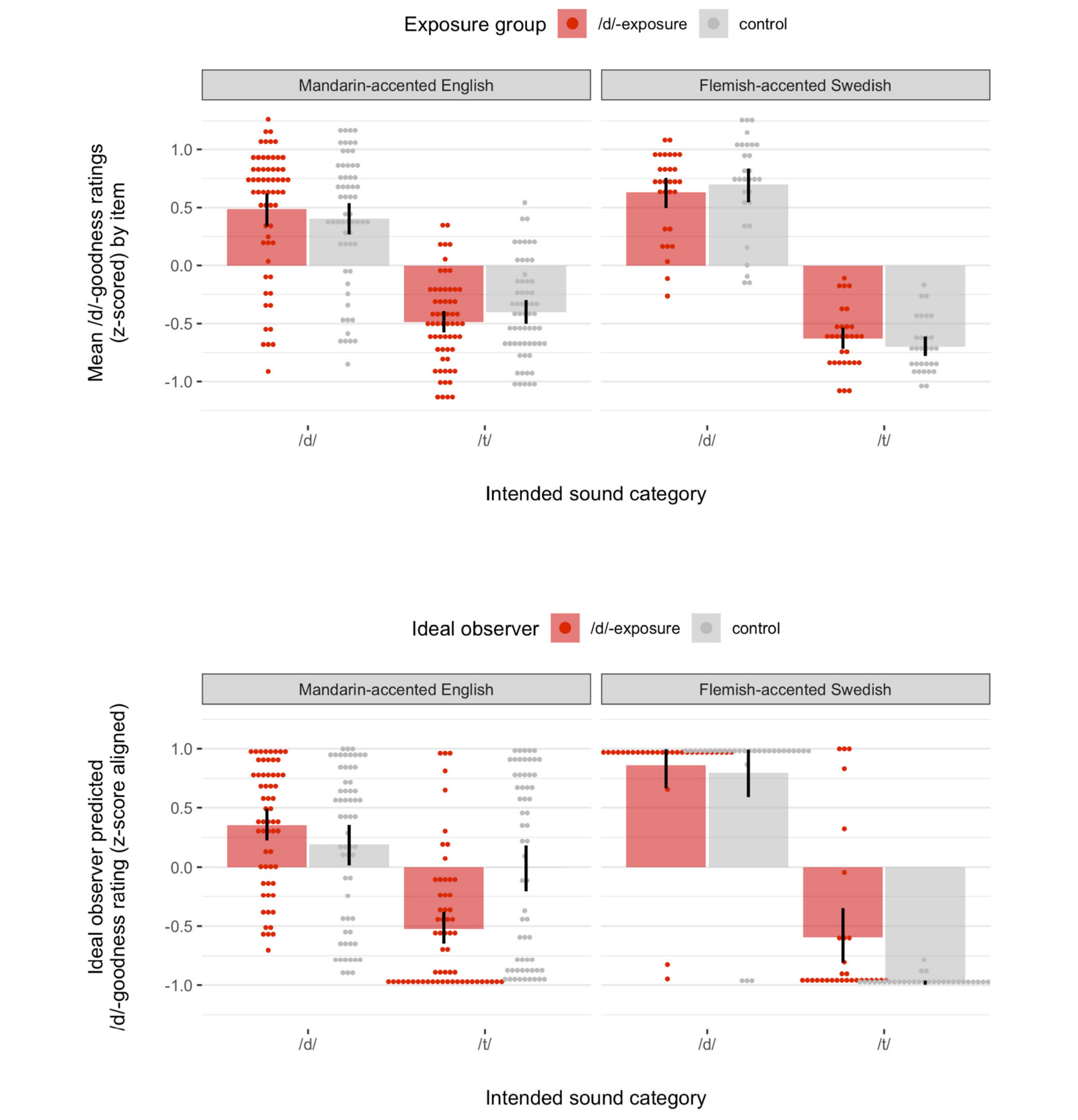


FIGURE 2 | (Top) Results of behavioral experiment on native listeners' perception of syllable-final /d/ and /t/ in Mandarin-accented English (left) and Flemish-accented Swedish (right). Points show by-item means of z-scored /d/-goodness ratings (standardized within each participant) for non-native productions of syllable-final /d/ and /t/ during test, depending on the whether participants received exposure to the relevant non-native realization of syllable-final /d/ (/d/-exposure) or not (control). Bars show means and 95% bootstrapped confidence intervals of the by-item means. **(Bottom)** Ideal observer-predicted /d/-goodness ratings described in section 3.2.

consist of 60 ratings each from 23 participants—about a fourth of the English data. This would seem to suggest lack of statistical power as a straightforward explanation for the null effect in the Swedish experiment. However, even when the English data was down-sampled to the size and structure of the Swedish data, the difference between the two data sets remained significant 57.6% of the time (out of 1,000 hierarchical bootstrap samples, **Supplementary Material**, section 3.2.5). For

English, the simple effect of /d/-exposure went in the predicted direction 89.6% of the time, reaching significance in 44.4% of all bootstrap samples (vs. 0.6% significant effects in the opposite direction). For Swedish, the simple effect went in the predicted direction 29.2% of the time, and was significant in 7.6% of all samples (vs. 40.2% significant effects in the opposite direction).

Overall, this suggests that power differences alone are unlikely to fully explain the difference between the English and Swedish results. Indeed, the same hierarchical bootstrap analyses found that the Swedish results are very unlikely to result if the English experiment is taken as the "ground truth": only 12 out of 1000 (1.2%) random resamples of the English experiment resulted in *t*-values as small or smaller than the one observed in the Swedish experiment.

What then caused the difference in results? And do the Swedish data really constitute a Type II error? The **Supplementary Material** (section 2) discusses a comprehensive list of differences in methodology between the experiments. This comparison revealed that the recordings for two experiments had been obtained in different ways. The Flemish-accented Swedish materials were elicited by first playing a native-accented recording of the word, whereas the Mandarin-accented English materials were elicited without such assistance (**Supplementary Material**, section 2.2). This raised the possibility that the Flemish-accented Swedish recordings

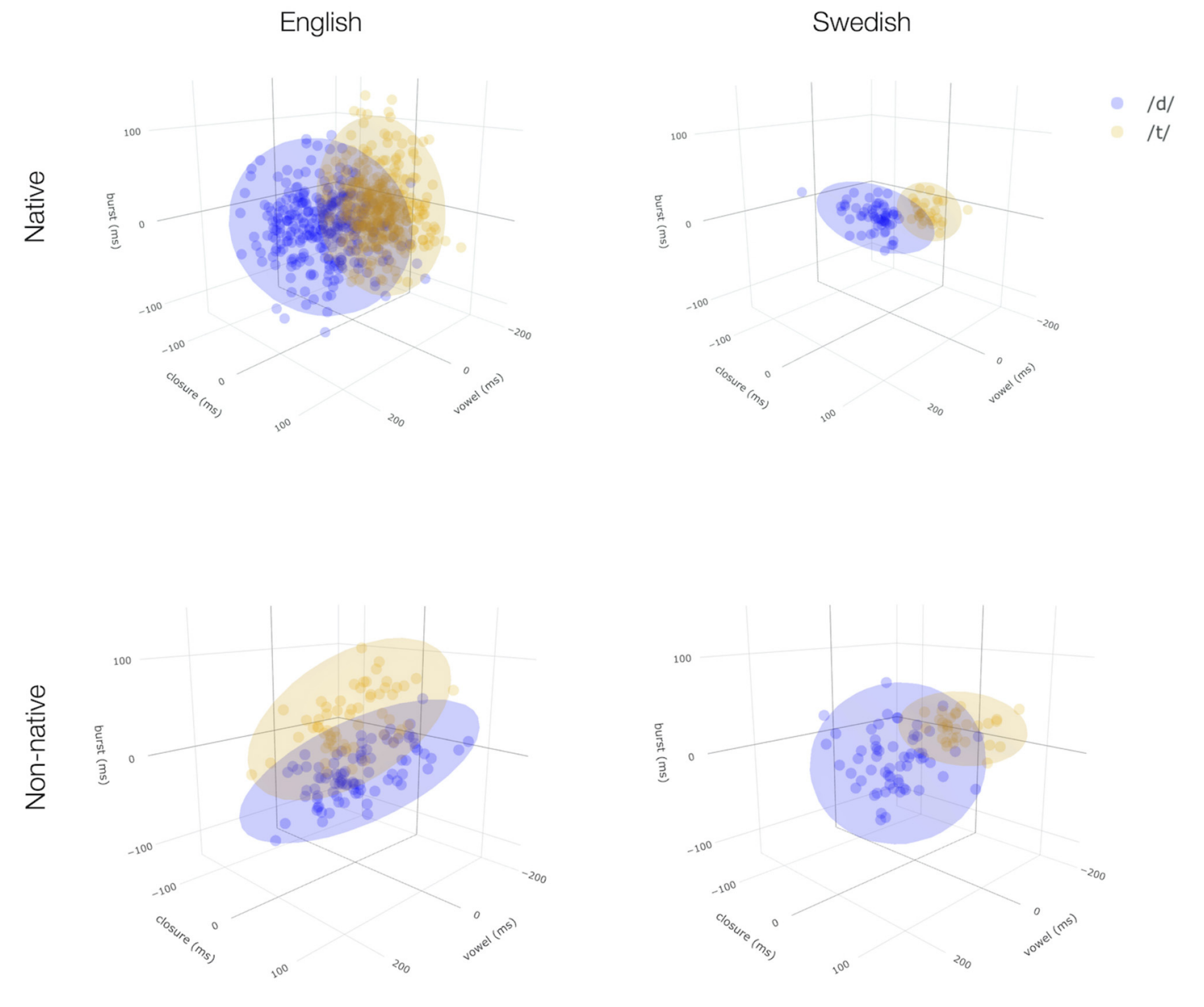


FIGURE 3 | Comparison of native- **(top)** and non-native-accented **(bottom)** syllable-final /d,t/ for both the English (left) and the Swedish (right) data. Productions combine information from multiple databases and are corrected for phonotactic context effects (see **Supplementary Material**, section 4.3) and are shown in the 3D space defined by three important cues (duration of vowel, closure, burst) to syllable-final voicing. Ellipses contain 95% of the probability mass centered around the mean under the assumption that categories form multivariate Gaussian distributions. To facilitate comparison, axis limits are held constant across panels. See **Supplementary Material** (section 4.3.2) for interactive visualization, 1D density plots, and 2D pair-wise correlation plots.

deviated less from native Swedish than the Mandarin-accented English recordings deviated from native English, which would reduce the perceptual benefit of /d/-exposure.

An initial comparison of the non-native-accented /d,t/ productions during test to productions of the same test words by a Swedish native speaker (not included in the experiment, but recorded using a similar procedure) lends credence to this hypothesis. **Figure 3** shows native- and non-native-accented syllable-final /d,t/ productions of all test items for both English and Swedish. Native productions were obtained from one or more gender-matched speakers similar in age to the non-native speakers employed in the experiments (for details, see **Supplementary Material**, section 2.2.1). We annotated native- and non-native-accented production for three cues known cross-linguistically to signal syllable-final stop voicing: the duration of the preceding vowel, the duration of the closure interval, and the duration of the burst release (for details on the annotation procedure, see **Supplementary Material**, section 4.1). The computational studies we present below confirm that these three cues were indeed highly informative about stop voicing in both English and Swedish, though it is possible, if not likely, that listeners employ different (related) or additional cues. Syllable-final stop voicing in Mandarin-accented English is known to differ in the use of these three cues, compared to native-accented English (Xie and Jaeger, 2020), as also clearly visible in the left panels of **Figure 3** (replicating Xie et al., 2017). At least at first blush, the Flemish-accented recordings seem to deviate less strongly from the native Swedish productions (right panels) than the Mandarin-accented recordings deviate from native English productions (left panels).

In line with this initial impression, the Flemish-accented Swedish recordings were substantially easier to process for the Swedish participants compared to the Mandarin-accented English recordings for the English participants: lexical decision accuracy during exposure was substantially higher for the Swedish data (Swedish, d-exposure: 96%, control: 97%) than for the English data (/d/-exposure: 78%, control: 74%). This included accuracy on the critical exposure words with syllable-final /d/ (English, /d/-exposure: 78%, SD = 9%; Swedish, /d/-exposure: 94%, SD = 6%; for further detail, see **Supplementary Material**, section 3.1)[1].

We thus decided to estimate the predicted consequences for the benefit of /d/-exposure for each experiment given the specific distributional properties of (1) the non-native-accented /d/ in the /d/-exposure group in that experiment, (2) the "typical" native-accented /d/ and /t/ in that language, and (3) the non-native-accented /d,t/-final minimal pair words during test. From this point on—having ruled out a number of alternative explanations for the seemingly diverging results—our approach is confirmatory: our goal is not to rule out alternative mechanisms for accent adaptation but rather to explore how a simple but fully specified computational model of distributional learning can aid data interpretation. This, we hope, may be informative for researchers who find themselves in a situation similar to the one described here: trying to understand (or even predict) the results of an experiment—specifically, the expected results based on the distributional properties of the speech stimuli employed in the experiment.

3. MODELING THE EFFECT OF EXPOSURE

We approach this question using ideal observers, specifically ideal categorizers, though we note that exemplar models would make similar predictions for the present purpose (for demonstration, see Shi et al., 2010). We use ideal observers because they provide an analytic framework to derive how an ideal/rational listener should respond to input given a certain set of assumptions (for early discussion of the value of this approach, see Massaro and Friedman, 1990). Like exemplar models, ideal observers link distributional patterns in the speech input—which listeners are assumed to have successfully learned, or at least approximated, through exposure (e.g., McClelland and Elman, 1986; Luce and Pisoni, 1998; Norris and McQueen, 2008; for reviews, see MacDonald, 2013; Kuperberg and Jaeger, 2016)—to the categorization decision listeners make during speech perception. Specifically, the posterior probability of recognizing an input as category c is a function of both the category's prior probability, $p(c)$, and the probability of observing the input under the hypothesis that the speaker intended to produce category c (the "likelihood"), $p(cues|c)$. These two pieces of information are assumed to be integrated optimally, as described by Bayes' theorem:

$$p(c|cues) = \frac{p(cues|c) * p(c)}{\Sigma_i p(cues|c_i) * p(c_i)} \tag{1}$$

Just as listeners are assumed to acquire the distributional parameters in Equation (1) from the speech input, researchers can estimate the resulting implicit knowledge of a typical listener from databases of speech production. Of appeal is that this approach makes predictions about *perception* based on only data from *production*, with zero computational degrees of freedom: the likelihood and prior distributions in Equation (1) are fully determined by the production data (unlike in, for example, exemplar models). This makes it noteworthy that ideal observers have been found to provide a good explanation for a variety of phenomena in speech perception and spoken word recognition (e.g., Luce and Pisoni, 1998; Clayards et al., 2008; Norris and McQueen, 2008; Feldman et al., 2009; Bejjanki et al., 2011; Kleinschmidt and Jaeger, 2015; Kronrod et al., 2016).

Here we use ideal observers as a methodological tool to estimate how an idealized participant who has adapted to the phonetic distributions in the input during exposure would

[1] The difference in exposure accuracy could also be explained if the Swedish participants were more familiar with accents that involve syllable-final devoicing than the American participants. For example, exposure to German-accented Swedish is common in Stockholm (as our Swedish colleagues were eager to point out). Post-experiment surveys found that none of the Swedish participants was able to guess the L1 of the non-native accent, and only one (4.3%) of the participants guessed another L1 that leads to syllable-final devoicing (German). It is possible, however, that participants nevertheless had subconscious familiarity with syllable-final devoicing. This would explain the lack of an effect of exposure. It would not, however, explain the differences in the degree of accentedness in the *productions*, shown in **Figure 3**. We also note that analyses presented in the **Supplementary Material** (section 5.4) suggest that, if anything, prior familiarity with the L2 accent was higher amongst the participants in the English experiment, compared to participants in the Swedish experiment.

respond to the test items. The lack of additional computational degrees of freedom is of particular appeal for this purpose, since fewer degrees of freedom reduce the risk of over-fitting the model to the data. In the same spirit, the models we present in the main text make a number of simplifying assumptions—many of them known to be wrong, but none of them trivially explaining the predictions we derive. These assumptions are summarized in **Supplementary Table 1**. Here we emphasize only the assumptions that make the models ideal*ized* rather than *ideal* (for the same distinction, see also Qian et al., 2016): rather than model ideal incremental adaptation to the exposure stimuli (Kleinschmidt and Jaeger, 2015), we model listeners that (1) have *completely* adapted by the end of exposure, and (2) do not adapt further during the test phase or at least not much. While (2) is plausible (inputs during test are not lexically labeled since they are minimal pair words; and adaptation seems to proceed most quickly upon initial exposure to talkers, Kraljic and Samuel, 2007), assumption (1) is likely wrong. Indeed, ideal adaptation should weight and integrate the observed input from a talker with prior expectations, so that only partial adaptation is expected after exposure to 30 critical words—partial in the sense that listeners' representations are not a replica of the statistics of the non-native speech, but rather somewhere between the native and non-native speech (Kleinschmidt and Jaeger, 2015).

3.1. Methods

We developed four ideal observer models, matching the four combinations of experimental conditions: 2 experiment (Swedish vs. English) X 2 exposure group (/d/-exposure vs. control). Our goal was to approximate the effects of exposure in these four conditions. All models encode listeners' beliefs about /d/ and /t/ as multivariate Gaussian distributions in the 3D space defined by vowel, closure, and burst duration. Category priors, $p(c)$ in Equation (1), were assumed to be uniform, with each category having a prior probability of 0.5 in all models. This is not meant to entail that syllable final /t/ and /d/ are equally probable in English (they are likely not), but rather that participants expect the two sounds to be equally probable in the context of the experiments (in which they repeatedly observe minimal pair words during test).

To approximate the effect of /d/-exposure, we estimated the mean and covariance of the /d/ category from the 30 non-native-accented recordings of the syllable-final /d/ employed during the experiments' exposure phase. To approximate the effect of control exposure, we estimated the mean and covariance of the /d/ category from recordings of the same 30 exposure words by a gender- and age-matched native speaker. Since by design, neither /d/- nor control exposure contained similarly lexically-labeled instances of syllable-final /t/, we made the simplifying assumption that both idealized listeners would have native /t/ categories. This ignores that listeners might adapt their expectations about /t/ based on exposure to the talker's /d/ or other categories whose realization is correlated with that of the /t/ category (see, e.g., Chodroff and Wilson, 2017). The **Supplementary Material** describes the databases (section 4.1) and annotation procedure (section 4.2) we employed to estimate the means and covariances of the native /t/ and non-native /t/ and /d/ categories.

While test words formed minimal pairs, holding phonotactic context constant across productions of /d/ and /t/, this was not the case between exposure and test productions. We thus use multiple linear regression to correct cue values for effects of segmental, supra-segmental and talker context (for details, including interactive plots illustrating the consequence of the correction procedure, see **Supplementary Material**, section 4.3). This approach closely follows the influential C-CuRE model of cue normalization (McMurray and Jongman, 2011), extending it to the contrast between native and non-native speech. C-CuRE has been found to provide a good fit against human categorization responses, including influences of coarticulation due to phonotactic context (Apfelbaum and McMurray, 2015). All ideal observers were fitted to and evaluated on these context-corrected cue values (**Supplementary Material**, section 4.5).

Both the control and d-exposure ideal observers were then applied to the *non-native-accented* minimal pair words from the test phase of the experiments (**Supplementary Material**, section 4.6). For each test token, we calculated the ideal observer's posterior probability of /d/ (and /t/), using Bayes theorem. In order to relate the posterior probabilities of /d/ and /t/ to participants' goodness ratings, it is necessary to specify a linking hypothesis. Conveniently, human categorization responses for the same stimuli and the same exposure conditions as analyzed here are available from a separate experiment in Xie et al.. Paralleling Xie and colleagues' rating experiment, the categorization experiment found the predicted shift in the /d/-/t/ category boundary following /d/-exposure, compared to control exposure (Xie et al., 2017). This allowed us to investigate the relation between human goodness ratings and proportions of categorization responses, using generalized additive mixed models (GAMMs, Hastie, 2017). These analyses (presented in the **Supplementary Material**, section 4.6.4) revealed a clearly linear relation between proportion /d/-responses in categorization and /d/-goodness ratings (and, vice versa, for /t/), at least for the type of stimuli analyzed here. For our analyses, we thus assume a simple identity link between the ideal observers' predicted posterior probability of a category and listeners' goodness ratings for that category. For visualizations (e.g., **Figure 2**, bottom), we facilitate comparison of ideal observers' prediction to human ratings by scaling the ideal observer-predicted posterior probabilities (range = 0–1) to have the same range as human rating responses across the combined English and Swedish data (range = −1 to 1). In those visualizations, we refer to the resulting predictions as posterior ratings. This scaling does not affect correlations between the ideal observers' predictions and human rating responses.

3.2. Results: Goodness Ratings Predicted by Ideal Observer

Figure 4 (bottom row) shows the results for the control and /d/-exposure ideal observers and both exposure conditions. Paralleling participants' goodness ratings for Mandarin-accented English in **Figure 2**, posterior ratings were improved under the

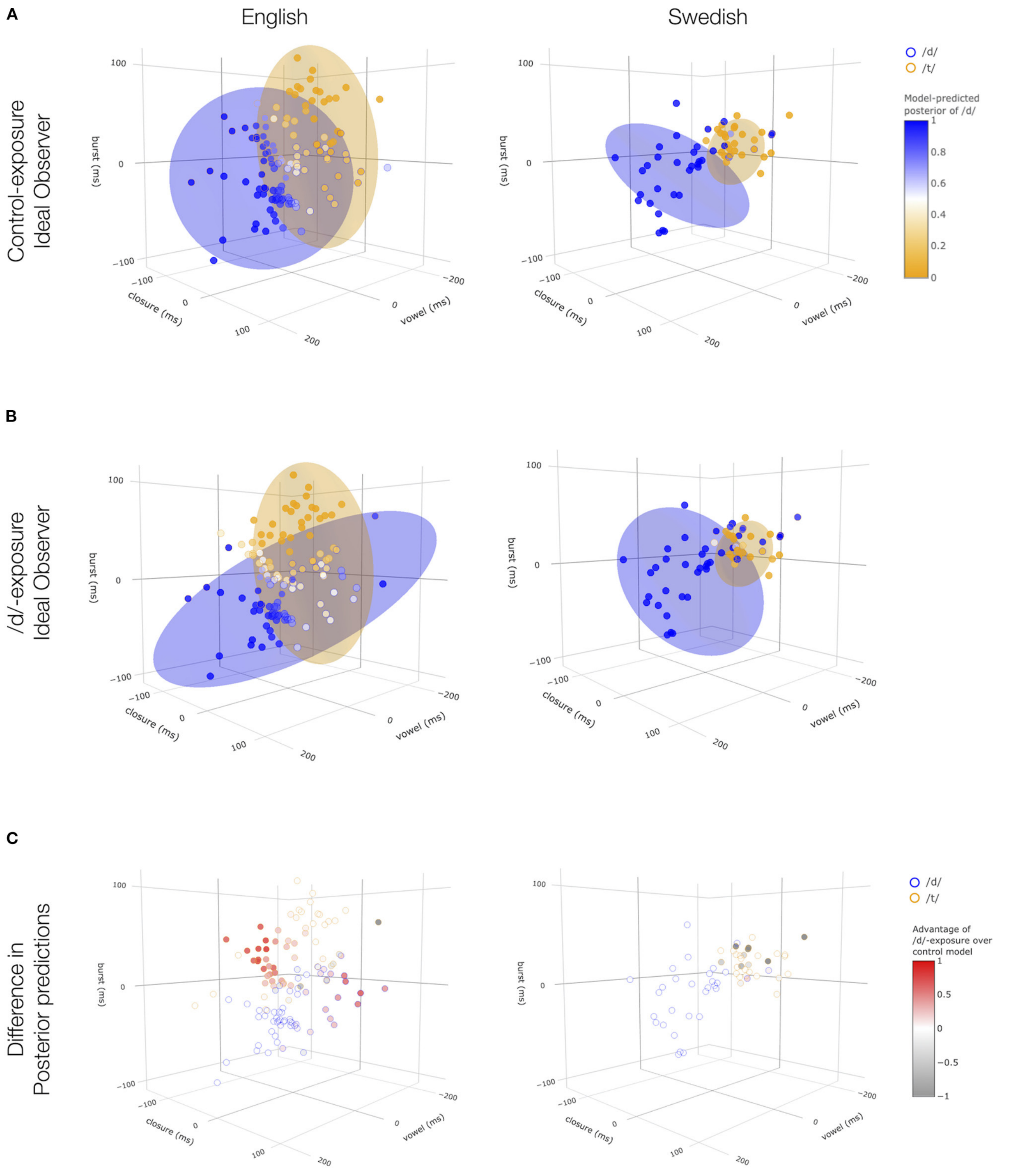


FIGURE 4 | Predicted ratings of control **(A)** and /d/-exposure ideal observers **(B)**, as well as their difference **(C)** shown for all test tokens. These models were constructed from the distributions shown in **Supplementary Material** (section 4.3.2) so as to simulate a learner who has either been exposed to, and perfectly learned, the statistics of the non-native /d/ (/d/-exposure model) or has not been exposed to non-native /d/ and thus assumes native /d/ statistics (control model). Ellipses visualize the category likelihoods assumed by the respective ideal observers (specifically 95% of the probability density). Across all panels, the outline color of points indicate their intended category. In **(A,B)**, the degree of color match between a point's outline color and fill color indicates a more accurate prediction matching the intended category. In **(C)**, the color fill of points indicate the difference in posterior predictions between the /d/-exposure and control exposure models: redness indicates better performance in the /d/-exposure model (relative to control) and grayness indicates the opposite pattern. See **Supplementary Material** (section 5.3) for interactive visualization.

non-native English model compared to the native English model. And, paralleling participants' goodness ratings for Flemish-accented Swedish, no such improvement of posterior ratings was observed under the non-native Swedish model compared to the native Swedish model. Further analysis presented in the **Supplementary Material** (section 5.2), confirmed that these results held across randomly sampled subsets of the data (training and test folds).

The ideal observers thus predict effects of exposure condition on goodness ratings that *qualitatively* resemble the results of both the English and the Swedish data. In particular, had we applied the ideal observers to the exposure and test stimuli from both experiments *prior to collecting data*, we would have correctly predicted an effect for the English experiment and a null effect for the Swedish experiment. In this sense then, the Swedish experiment would *not* constitute a Type II error. The quality of fit was also confirmed by trial-level linear mixed-effects regressions reported in the **Supplementary Material** (section 5.3). These analyses found that the posterior probability of the /d/ category was a significant predictor of listeners' /d/-goodness ratings ($\hat{\beta} = 1.22$, $p < 0.001$). This effect remained significant when the experiment (English vs. Swedish), exposure group (/d/-exposure vs. control), and their interaction were included in the analysis ($\hat{\beta} = 0.17$, $p < 0.02$; for additional details, see **Supplementary Material**, section 5.3).

To further elucidate the reason for the differences in the ideal observers' predictions for the two experiments, **Figure 4** shows the ideal observers' predictions for each of the items participants heard during test, shown in a 3D cue space. A distributional learning framework predicts failure to observe evidence for adaptation if (a) the non-native exposure stimuli provide misleading information about the non-native stimuli during test or (b) if the distributions of cues in the non-native exposure stimuli do not differ much from native distributions. From the first two rows of **Figure 4**, it is apparent that the predicted null effect for the Swedish experiment is an example of case b): rather than the /d/-exposure model performing badly on the test items, both the control and the /d/-exposure model perform well on the test items. The reason for this is also obvious: the realization of native and non-native /d/ did not differ much for the Swedish recordings (see also **Figure 3**). For the English recordings, on the other hand, the cue distributions for the Mandarin-accented /d/ stimuli differed starkly from those of the native-accented /d/ stimuli. Deviating from native pronunciations, the Mandarin-accented talker showed no distinction between /d/ and /t/ in vowel and closure duration but clear separation along the burst dimension (**Figure 3**, bottom left). This gave listeners in the /d/-exposure group a clear learning advantage over the control exposure group.

4. DISCUSSION

Critical reviews of standard practices in the psychological sciences have called out the tendency to dismiss null results as uninformative (Franco et al., 2014). A welcome consequence of this is that it is now easier to publish null results, often as failures to replicate. This reduces the "file drawer" problem (Rosenthal, 1979). The present work can be seen as building on this idea, aiming to understand *why* a null effect is observed. Specifically, the motivation for the present report grew out of an attempt to extend a previously replicated result of accent adaptation to a new L1-L2 pair, Flemish-accented Swedish. Apart from the language, test talker, and lexical materials, this experiment closely followed the design and procedure of previous work, specifically an experiment on Mandarin-accented English (Xie et al., 2017). Beyond the rating results from Xie and colleagues, several other studies with similar design had previously found the predicted effect of /d/-exposure, indexed either by increased auditory priming effects (Eisner et al., 2013; Xie and Myers, 2017; Xie et al., 2017) or improved segment identification (Xie et al., 2017). We thus expected that the experiment on Swedish would find positive evidence of adaptation, yet it seemingly failed to do so. After having ruled out differences in statistical power as a likely cause for the difference in results, we turned to computational models of speech perception to understand whether differences in the statistical properties of the exposure and test stimuli can explain the difference in results.

We found that ideal observers predict both the positive evidence for an effect for Mandarin-accented English in Xie et al. (2017) and the lack thereof in our experiment on Flemish-accented Swedish. This suggests that the original results were not a Type I error, nor are the Swedish results a Type II error. Rather, our ideal observer analyses suggest that the Swedish experiment would not find an effect even if repeated as a large-scale replication, at least as long as the same exposure and test stimuli are used. Indeed, even a much longer exposure phase that repeatedly presents the same non-native /d/ pronunciation as in our experiment on Swedish would not be expected to yield significant changes in participants' goodness ratings. The reason for this is clear from **Figure 4**: while the Flemish-accented talker differs from native speakers of Swedish in her realization of Swedish syllable-final /d/, these differences are small compared to the non-nativeness observed in the Mandarin-accented speech employed in the experiment on English.

At least qualitatively, ideal observer models provide a good fit against listeners' rating responses. This is noteworthy since the modeling approach employed here does not include *any* degrees of freedom to mediate the effect of input statistics on perception. The only parameters of ideal observers describe the statistics of categories' cue distributions in the speech input. These parameters are thus not fitted to participants' responses during the perception experiment but rather are fixed by data from speech *production*—specifically, speech data that is assumed to have formed listeners' prior expectations based on native speech input and speech data that listeners observe during exposure in the experiment. Based on these speech data, ideal observers make predictions about listeners' *perception* during a subsequent test phase (here goodness ratings). In this sense, ideal observers offer a particularly parsimonious explanation for the differences in results between the two experiments.

The present findings thus are compatible with the hypothesis that adaptation to non-native accents involves similar mechanisms as adaptation to talker-specific differences between native talkers (see also Eisner et al., 2013; Reinisch and Holt, 2014), and that these mechanisms include some form of distributional learning (see also Wade et al., 2007; Xie and Myers, 2017; Xie et al., 2017). Notably, recent work might be seen as calling into question the existence of such shared mechanisms (Zheng and Samuel, 2020). Zheng and Samuel report a failure to find a correlation between individuals' changes after exposure to shifted native speech (perceptual recalibration) and exposure to non-native accented speech (in a paradigm not unlike the present one). However, unlike the present work, the analyses presented by Zheng and Samuel do not assess whether such a correlation would actually be predicted by theories of distributional learning for the particular exposure and test recordings of their study. And, while Zheng and Samuel present power analyses, these analyses are based on arbitrarily selected effect sizes rather than effect sizes expected under theories of distributional learning. This and similar studies are thus an interesting venue for future applications of the modeling approach presented here, allowing researchers to shed light on the informativeness of null findings.

The present study also contributes to efforts to facilitate the theoretical interpretation of perception experiments through computational modeling (e.g., Clayards et al., 2008; Feldman et al., 2009; Toscano and McMurray, 2010; McMurray and Jongman, 2011; Kleinschmidt and Jaeger, 2015; Kronrod et al., 2016; Chodroff and Wilson, 2018). In particular, an emerging body of work has used ideal observers and ideal adaptors to quantify how changes in the distributional statistics of phonetic cues affect listeners' categorization decisions (e.g., Clayards et al., 2008; Kleinschmidt and Jaeger, 2011, 2016; Kleinschmidt et al., 2012, 2015; Theodore and Monto, 2019). When listeners are exposed to speech in which categories' cue distributions deviate from those of typical talkers—e.g., in terms of changes in categories' means or variances—this affects how listeners perceive and categorize subsequent input from the same talker. This manifests in changes in the location (Kleinschmidt and Jaeger, 2011, 2015; Kleinschmidt et al., 2012) or the steepness of listeners' categorization functions (Clayards et al., 2008; Theodore and Monto, 2019) that are well-described by ideal observer and adaptor models. More recent work has begun to go one step further, using exposure-induced changes in categorization behavior from multiple exposure conditions to probe the structure of listeners' prior expectations about cross-talker variability (Kleinschmidt and Jaeger, 2016; Kleinschmidt, 2020).

Previous work in speech perception has employed ideal observers mostly for 2AFC or n-AFC tasks (ideal categorizers, e.g., Clayards et al., 2008; Hitczenko and Feldman, 2016; Xie et al., 2021a). However, with suitable link functions, ideal observers can be applied to other types of tasks and dependent variables. Ideal observers have, for example, been used to model perceptual discrimination (Feldman et al., 2009; Kronrod et al., 2016) and sentence transcription (Xie et al., 2021b, **Supplementary Material**). Here, we have extended them to model category goodness ratings from 7-point Likert scales (see **Supplementary Material**, section 4.6.4).

In the present study, we used one case study to demonstrate how computational modeling aids the interpretation of experimental results that run counter to expectations. But computational models can provide substantial gain even when the result of experiments seemingly conform to expectations. A case in point that is directly relevant to the present study comes from recent work by Hitczenko and Feldman (2016). Like the present work, Hitczenko and Feldman employed computational models *post-hoc* to inform the theoretical interpretation of a previously reported finding from an experiment on adaptation to a synthesized accent (Maye et al., 2008). Maye and colleagues exposed listeners to synthesized American English in which all front vowels were simulated to have undergone phonological lowering (e.g., [i] became [ɪ] and [ɪ] became [ɛ], etc.). Listeners subsequently completed a lexical decision task of previously unheard words by the same synthesized voice with front vowels either lowered or raised. Based on the specific pattern of results, Maye and colleagues concluded that listeners adapted to the synthesized accent by shifting the means of their category representations, rather than merely becoming more accepting of *any* type of input. This finding and its interpretation has been influential, with almost 300 citations since 2008. Hitczenko and Feldman (2016) revisit these results, comparing them to the predictions of different types of ideal distributional learners (ideal adaptors, an extension to the simpler ideal observers employed here Kleinschmidt and Jaeger, 2015). Based on these computational comparisons, Hitczenko and Feldman conclude that shifted category representations are *not* the only way, or even the best, way to explain the specific changes in listeners' perception after exposure to the synthesized accent.

4.1. Limitations and Future Directions

These studies and the present work serve as examples of how computational models can inform the theoretical interpretation of empirical findings. One strength of the computational approach is that it compels deeper introspection about the assumptions that are necessary to derive predictions from a theory, and to make those assumptions explicit. We refer the reader to Table 4.2 in the **Supplementary Material**, which aims to list all assumptions we made in the present study. In the remainder, we discuss some of these assumptions, their limitations, and how future work might go about relaxing and revising them.

First, we made simplifying assumptions about what sources of noise contribute to listeners' estimates of the relevant cue distributions. Acoustic noise in the environment and neural noise in listeners' perceptual systems distort the speech signal produced by talkers beyond whatever variability results from noise during the planning and execution of speech articulation. By estimating distributions from speech recordings, our ideal observers ignore whatever acoustic noise our participants experienced beyond those in the recordings, as well as any noise within listeners' perceptual systems[2]. This might explain why the responses

[2] At the same time, our ideal observers' estimates of all relevant cue distributions are likely perturbed by measurement errors due to the annotation procedure we used.

predicted by the ideal observers are more categorical than the actual responses made by human listeners: adding perceptual noise to our ideal observers would increase the variance of cue distributions, leading to more shallow categorization functions, and thus less categorical predicted rating responses. Previous work has demonstrated that noise effects can be quantitatively estimated from separate perceptual data and integrated into ideal observers (Feldman et al., 2009; Kronrod et al., 2016). It would be informative to see whether the inclusion of perceptual noise improves the fit between the ideal observers' predictions and human perceptual decisions.

Second, we applied normalization procedures on the acoustic cues to correct for phonotactic context effects. We made the simple assumption that—for native listeners, whose perception we were aiming to model—such correction is based on previous experience with native speech, rather than being shaped by the exposure to non-native speech in the experiment. That is, neither the control, nor the /d/-exposure model assumed learning of non-native phonotactic regularities. On the one hand, this would seem to be in the spirit of C-CuRE and related normalization approaches (Lobanov, 1971; Nearey, 1978; McMurray and Jongman, 2011). For example, C-CuRE computes acoustic cues relative to expectations about the mean of cues in a particular phonotactic or talker context. Critically, the C-CuRE model presented in McMurray and Jongman (2011) assumes that these adjustments are made independent of each other—i.e., this normalization procedure corrects for talker-specific differences in cue distributions and for phonotactics, but not for talker-specific phonotactics. On the other hand, there is evidence that non-native speech deviates from native speech in not only the overall realization of categories, but also in how specific phonotactic contexts affect pronunciation (as found in, e.g., Flege and Wang, 1989; Lahiri and Marslen-Wilson, 1991; Xie and Jaeger, 2020). Whether listeners in the accent adaptation experiments learn these non-native phonotactics in addition to changes in category-to-cue distributions is an open question. Future work could therefore compare models like ours without learning talker- or accent-specific phonotactic patterns against models that also learn this information.

Third, we constructed the /d/-exposure and the control models directly from the input statistics in each accent (non-native vs. native). These models assumed complete learning whereby listeners are assumed to have fully converged toward exposure statistics. In reality, rational listeners are expected to be guided by prior beliefs based on their native experience. While such priors facilitate adaptation to talker-specific statistics that meet prior expectations (Kleinschmidt and Jaeger, 2015), the same priors slow-down and constrain learning of unexpected non-native statistics (Kleinschmidt and Jaeger, 2016; Kleinschmidt, 2020). Learners are thus not expected to fully converge against the statistics experienced during exposure. Future work might consider the same type of incremental Bayesian belief updating applied in previous work on the perception of native speech (Kleinschmidt and Jaeger, 2011; Theodore and Monto, 2019) or synthesized speech (Hitczenko and Feldman, 2016) to investigate adaptation to the perception of non-native speech.

Fourth, and related to the third point, we adopted an assumption commonly made in research on accent adaptation—that participants were unfamiliar with the non-native accents in the experiments. This assumption is almost always questionable. We followed previous work (Reinisch and Holt, 2014), and asked participants to guess the native language of the talker. Based on this measure, participants in either experiment did not seem to be familiar with the accent prior to the experiment. However, explicit identification of accents is likely an unreliable measure of participants' previous experience with an accent (McCullough, 2015; McKenzie, 2015; Gnevsheva, 2018). It is thus possible that some of the results we discussed here are due to participants prior familiarity with the L2 accent in the experiment. Indeed, additional analyses reported in the **Supplementary Material** (section 5.4) found that participants in the English experiment might have had prior familiarity with Mandarin-accented English or similar L2 accents. The effects observed by Xie et al. (2017) thus do not necessarily reflect the same adaptation as listeners that are completely unfamiliar with Mandarin-accented English or similar L2 accents: on the one hand, prior familiarity might lead to faster adaptation; on the other hand, prior familiarity likely would reduce the difference between the two exposure conditions, since it means that both groups of participants have exposure to Mandarin-accented /d/. As pointed out by a reviewer, it is further possible that Swedish listeners were more familiar with Flemish-accented Swedish (or similar accents) than L1 English listeners are familiar with Mandarin-accented English (or similar accents). This would provide an alternative explanation for the null results in the Swedish experiment. The additional analyses in the **Supplementary Material** (section 5.4) did not, however, reveal support for this possibility. If anything, these analyses argued against this possibility though we note that the lack of a significant exposure effect makes it difficult to rule it out entirely (see discussion in the **Supplementary Material**).

Beyond the aforementioned specifics of the models, there are limitations to the specific way in which the present study employed ideal observers: our approach has been both *post-hoc* and confirmatory. With regard to the latter, future work could follow in the footsteps of Hitczenko and Feldman (2016), and compare the ideal observers developed here against alternative hypotheses. For example, instead of distributional learning, the effects of different exposure on listeners' rating responses during test might reflect changes in response biases (Clarke-Davidson et al., 2008) or a general relaxation of response criteria (Hitczenko and Feldman, 2016). Similarly, future work might employ the same methods we have used here *post-hoc*, but do so *predictively* prior to conducting the experiment. As we have illustrated here, the distributional statistics of the specific input—and more specifically the way in which such statistics differ between native and non-native speech—can be linked to predicted changes in subsequent perception. Future work could, for example, use ideal observer-predicted categorization or rating responses in power analyses to inform experimental designs prior to the experiment (for similar approaches in other domains, see Jaeger et al., 2019; Bicknell et al., in revision[3]).

[3] Bicknell, K., Bushong, W., Tanenhaus, M. K., and Jaeger, T. F. (in revision). Listeners can maintain and rationally update uncertainty about prior words.

Finally, it is important to recall that the reliability and generalizability of the results presented here is limited by two types of data sparsity. First, we used phonetically annotated databases to estimate the implicit distributional knowledge that listeners are hypothesized to have learned from previous speech input. Even for well-studied languages like English, these databases tend to be small. For Swedish, we had access to only one talker. While efforts were taken to record a 'typical' talker of Swedish, with the hope that the phonetic distributions of this talker would be representative of what native listeners might have come to expect through a lifetime of exposure, the results reported here might change once a larger database with more Swedish talkers is considered. In short, the fact the we obtained a decent fit against human performance for both experiments does *not* show that the amount of data we used to develop the ideal observers was sufficient. Additional analyses presented in the **Supplementary Material** (section 5.2) address this question. By subsetting both the training and test data for the ideal observers into multiple separate folds, we find that the qualitative match between model predictions and human ratings seems to be surprisingly robust even for the small data sets we had access to. We do, however, also find that the results are considerably more robust for English (trained on 6 native talkers) than for Swedish (trained on 1 native talker). Overall, the results of these additional analyses suggests (1) that 15 training tokens per category and 15 test tokens per category *can* be sufficient for the type of analysis conducted here, but that (2) having access to data from multiple talkers is important for the estimation of listeners' prior (in this case native) knowledge. The second way in which data sparsity limits the conclusions we can draw from the present study is likely more severe. It is also shared with the majority of work on talker-specific accent adaptation: both experiments analyzed here employed a single non-native accented talker. There is now evidence that the results of such experiments can depend on the specific talker (for evidence and discussion, see Xie et al., 2021b). Moving forward, the same models employed here for talker-specific adaptation can be used to understand adaptive changes in listeners' perception and categorization following exposure to multiple talkers, or listeners' ability to generalize previously experienced input to unfamiliar talkers (for discussion and model development, see Kleinschmidt and Jaeger, 2015, Part II.

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://osf.io/72fkx/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Connecticut Institutional Review Board (for the English version). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XX and MT conducted the behavioral experiments in English and Swedish, respectively. TJ and XX led in the statistical analyses. MT contributed to the statistical analyses. All authors contributed to the writing of the manuscript.

FUNDING

This study was supported by a travel grant from the Knut & Alice Wallenberg Foundation (2018) awarded to MT.

ACKNOWLEDGMENTS

We are grateful to Shawn Cummings for feedback on earlier versions of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://osf.io/72fkx/

REFERENCES

Adank, P., Evans, B. G., Stuart-Smith, J., and Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *J. Exp. Psychol.* 35:520. doi: 10.1037/a0013552

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., et al. (2021). *rmarkdown: Dynamic Documents for R. R Package Version 2.7.*

Allen, J. S., and Miller, J. L. (2001). Contextual influences on the internal structure of phonetic categories: a distinction between lexical status and speaking rate. *Percept. Psychophys.* 63, 798–810. doi: 10.3758/BF03194439

Apfelbaum, K. S., and McMurray, B. (2015). Relative cue encoding in the context of sophisticated models of categorization: separating information from categorization. *Psychon. Bull. Rev.* 22, 916–943. doi: 10.3758/s13423-014-0783-2

Bejjanki, V. R., Clayards, M., Knill, D. C., and Aslin, R. N. (2011). Cue integration in categorical tasks: insights from audio-visual speech perception. *PLoS ONE* 6:e19812. doi: 10.1371/journal.pone.0019812

Best, C. T. (1995). "A direct realist view of cross-language speech perception," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, ed W. Strange (Timonium, MD: York Press), 171–206.

Booij, G. *The Phonology of Dutch*. Oxford: Oxford University Press (1999).

Bradlow, A. R. and Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition* 106, 707–729. doi: 10.1016/j.cognition.2007.04.005

Chodroff, E., and Wilson, C. (2017). Structure in talker-specific phonetic realization: covariation of stop consonant VOT in American English. *J. Phonet.* 61, 30–47. doi: 10.1016/j.wocn.2017.01.001

Chodroff, E., and Wilson, C. (2018). Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice. *Linguist. Vanguard* 4:s2. doi: 10.1515/lingvan-2017-0047

Clarke, C. M., and Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *J. Acoust. Soc. Am.* 116, 3647–3658. doi: 10.1121/1.1815131

Clarke-Davidson, C. M., Luce, P. A., and Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Percept. Psychophys*. 70, 604–618. doi: 10.3758/PP.70.4.604

Clayards, M., Tanenhaus, M. K., Aslin, R. N., and Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition* 108, 804–809. doi: 10.1016/j.cognition.2008.04.004

Drouin, J. R., Theodore, R. M., and Myers, E. B. (2016). Lexically guided perceptual tuning of internal phonetic category structure. *J. Acoust. Soc. Am*. 140, EL307–EL313. doi: 10.1121/1.4964468

Eisner, F., and McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Percept. Psychophys*. 67, 224–238. doi: 10.3758/BF03206487

Eisner, F., Melinger, A., and Weber, A. (2013). Constraints on the transfer of perceptual learning in accented speech. *Front. Psychol*. 4:148. doi: 10.3389/fpsyg.2013.00148

Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychol. Rev*. 116:752. doi: 10.1037/a0017196

Flege, J. E. (1995). Second language speech learning: theory, findings, and problems. *Speech Percept. Linguist. Exp*. 92, 233–277.

Flege, J. E., Munro, M. J., and Skelton, L. (1992). Production of the word-final english/t/-/d/contrast by native speakers of English, Mandarin, and Spanish. *J. Acoust. Soc. Am*. 92, 128–143. doi: 10.1121/1.404278

Flege, J. E., and Wang, C. (1989). Native-language phonotactic constraints affect how well Chinese subjects perceive the word-final English /t/-/d/contrast. *J. Phonet*. 17, 299–315. doi: 10.1016/S0095-4470(19)30446-2

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication bias in the social sciences: unlocking the file drawer. *Science* 345, 1502–1505. doi: 10.1126/science.1255484

Gnevsheva, K. (2018). Variation in foreign accent identification. *J. Multiling. Multicult. Dev*. 39, 688–702. doi: 10.1080/01434632.2018.1427756

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

Hastie, T. J. (2017). "Generalized additive models," in *Statistical Models in S*, eds J. M. Chambers and T. J. Hastie (Boca Raton, FL: Routledge), 249–307. doi: 10.1201/9780203738535-7

Hitczenko, K., and Feldman, N. H. (2016). "Modeling adaptation to a novel accent," in *Proceedings of the Annual Conference of the Cognitive Science Society*, 1367–1372.

Idemaru, K., and Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *J. Exp. Psychol*. 37:1939. doi: 10.1037/a0025641

Jaeger, T., Burchill, Z., and Bushong, W. (2019). *Strong evidence for expectation adaptation during language understanding, not a replication failure. A reply to Harrington Stack, James, and Watson* (2018) (New York, NY: University of Rochester). Retrieved from: https://wbushong.github.io/publications/pub_files/Jaeger_etal_response.pdf (accessed on August 31, 2021).

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., and Golestani, N. (2016). Mutual influences between native and non-native vowels in production: evidence from short-term visual articulatory feedback training. *J. Phonet*. 57, 21–39. doi: 10.1016/j.wocn.2016.05.001

Kleinschmidt, D. (2020). *What Constrains Distributional Learning in Adults?* University of Rochester; Rutgers University. Available online at: https://doi.org/10.31234/osf.io/6yhbe (accessed on August 31, 2021).

Kleinschmidt, D., and Jaeger, T. F. (2011). "A Bayesian belief updating model of phonetic recalibration and selective adaptation," in *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 10–19.

Kleinschmidt, D., Raizada, R., and Jaeger, T. F. (2015). "Supervised and unsupervised learning in phonetic adaptation," in *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci15)* (Austin, TX: Cognitive Science Society).

Kleinschmidt, D. F., Fine, A. B., and Jaeger, T. F. (2012). "A belief-updating model of adaptation and cue combination in syntactic comprehension," in *Proceedings of the Annual Meeting of the Cognitive Science Society, Vol. 34*.

Kleinschmidt, D. F., and Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychol. Rev*. 122:148. doi: 10.1037/a0038695

Kleinschmidt, D. F., and Jaeger, T. F. (2016). "What do you expect from an unfamiliar talker?," in *Conference: The 38th Annual Meeting of the Cognitive Science Society*.

Kraljic, T., and Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychon. Bull. Rev*. 13, 262–268. doi: 10.3758/BF03193841

Kraljic, T., and Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *J. Mem. Lang*. 56, 1–15. doi: 10.1016/j.jml.2006.07.010

Kronrod, Y., Coppess, E., and Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychon. Bull. Rev*. 23, 1681–1712. doi: 10.3758/s13423-016-1049-y

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

Lahiri, A., and Marslen-Wilson, W. (1991). The mental representation of lexical form: a phonological approach to the recognition lexicon. *Cognition* 38, 245–294. doi: 10.1016/0010-0277(91)90008-R

Lancia, L., and Winter, B. (2013). The interaction between competition, learning, and habituation dynamics in speech perception. *Lab. Phonol*. 4, 221–257. doi: 10.1515/lp-2013-0009

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *J. Acoust. Soc. Am*. 49, 606–608. doi: 10.1121/1.1912396

Luce, P. A., and Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear Hear*. 19:1. doi: 10.1097/00003446-199802000-00001

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

Massaro, D. W., and Friedman, D. (1990). Models of integration given multiple sources of information. *Psychol. Rev*. 97:225. doi: 10.1037/0033-295X.97.2.225

Maye, J., Aslin, R. N., and Tanenhaus, M. K. (2008). The Weckud Wetch of the Wast: lexical adaptation to a novel accent. *Cogn. Sci*. 32, 543–562. doi: 10.1080/03640210802035357

McClelland, J. L., and Elman, J. L. (1986). The trace model of speech perception. *Cogn. Psychol*. 18, 1–86. doi: 10.1016/0010-0285(86)90015-0

McCullough, E. A. (2015). "Open-set identification of non-native talkers' language backgrounds," in *ICPhS*.

McKenzie, R. M. (2015). The sociolinguistics of variety identification and categorisation: free classification of varieties of spoken English amongst non-linguist listeners. *Lang. Awareness* 24, 150–168. doi: 10.1080/09658416.2014.998232

McMurray, B., and Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychol. Rev*. 118:219. doi: 10.1037/a0022325

Munro, M. J., and Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Lang. Speech* 38, 289–306. doi: 10.1177/002383099503800305

Nearey, T. (1978). *Phonetic Feature Systems for Vowels*. Bloomington, IN: Indiana University Linguistics Club.

Nixon, J. S., van Rij, J., Mok, P., Baayen, R. H., and Chen, Y. (2016). The temporal dynamics of perceptual uncertainty: eye movement evidence from cantonese segment and tone perception. *J. Mem. Lang*. 90, 103–125. doi: 10.1016/j.jml.2016.03.005

Norris, D., and McQueen, J. M. (2008). Shortlist b: a Bayesian model of continuous speech recognition. *Psychol. Rev*. 115:357. doi: 10.1037/0033-295X.115.2.357

Norris, D., McQueen, J. M., and Cutler, A. (2003). Perceptual learning in speech. *Cogn. Psychol*. 47, 204–238. doi: 10.1016/S0010-0285(03)00006-9

Pierrehumbert, J. B. (2001). Exemplar dynamics: word frequency, lenition and contrast. *Typol. Stud. Lang*. 45, 137–158. doi: 10.1075/tsl.45.08pie

Porretta, V., Tremblay, A., and Bolger, P. (2017). Got experience? PMN amplitudes to foreign-accented speech modulated by listener experience. *J. Neurolinguist*. 44, 54–67. doi: 10.1016/j.jneuroling.2017.03.002

Qian, T., Jaeger, T. F., and Aslin, R. N. (2016). Incremental implicit learning of bundles of statistical patterns. *Cognition* 157, 156–173. doi: 10.1016/j.cognition.2016.09.002

Reinisch, E., and Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *J. Exp. Psychol*. 40:539. doi: 10.1037/a0034409

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychol. Bull*. 86:638. doi: 10.1037/0033-2909.86.3.638

Samuel, A. G. (1982). Phonetic prototypes. *Percept. Psychophys*. 31, 307–314. doi: 10.3758/BF03202653

Schertz, J., Cho, T., Lotto, A., and Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *J. Phonet*. 52, 183–204. doi: 10.1016/j.wocn.2015.07.003

Schmale, R., Cristia, A., and Seidl, A. (2012). Toddlers recognize words in an unfamiliar accent after brief exposure. *Dev. Sci*. 15, 732–738. doi: 10.1111/j.1467-7687.2012.01175.x

Shi, L., Griffiths, T. L., Feldman, N. H., and Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychon. Bull. Rev*. 17, 443–464. doi: 10.3758/PBR.17.4.443

Sidaras, S. K., Alexander, J. E., and Nygaard, L. C. (2009). Perceptual learning of systematic variation in Spanish-accented speech. *J. Acoust. Soc. Am*. 125, 3306–3316. doi: 10.1121/1.3101452

Smith, B. L., Johnson, E., and Hayes-Harb, R. (2019). ESL learners' intra-speaker variability in producing American English tense and lax vowels. *J. Second Lang. Pronunc*. 5, 139–164. doi: 10.1075/jslp.15050.smi

Tan, M. S. L., Xie, X., and Jaeger, T. F. (2019). "Analysing L2 Swedish word-final stops," in *10th Tutorial and Research Workshop on Experimental Linguistics*, 193–196. doi: 10.36505/ExLing-2019/10/0049/000411

Theodore, R. M., and Monto, N. R. (2019). Distributional learning for speech reflects cumulative exposure to a talker's phonetic distributions. *Psychon. Bull. Rev*. 26, 985–992. doi: 10.3758/s13423-018-1551-5

Toscano, J. C., and McMurray, B. (2010). Cue integration with categories: weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cogn. Sci*. 34, 434–464. doi: 10.1111/j.1551-6709.2009.01077.x

Vaughn, C., Baese-Berk, M., and Idemaru, K. (2019). Re-examining phonetic variability in native and non-native speech. *Phonetica* 76, 327–358. doi: 10.1159/000487269

Verhoeven, J. (2005). Belgian standard Dutch. *J. Int. Phonet. Assoc*. 35, 243–247. doi: 10.1017/S0025100305002173

Volaitis, L. E., and Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *J. Acoust. Soc. Am*. 92, 723–735. doi: 10.1121/1.403997

Wade, T., Jongman, A., and Sereno, J. (2007). Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. *Phonetica* 64, 122–144. doi: 10.1159/000107913

Weil, S. (2001). Foreign accented speech: encoding and generalization. *J. Acoust. Soc. Am*. 109:2473. doi: 10.1121/1.4744779

Witteman, M. J., Weber, A., and McQueen, J. M. (2013). Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation. *Attent. Percept. Psychophys*. 75, 537–556. doi: 10.3758/s13414-012-0404-y

Xie, X., Buxó-Lugo, A., and Kurumada, C. (2021a). Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition* 211:104619. doi: 10.1016/j.cognition.2021.104619

Xie, X., Earle, F. S., and Myers, E. B. (2018a). Sleep facilitates generalisation of accent adaptation to a new talker. *Lang. Cogn. Neurosci*. 33, 196–210. doi: 10.1080/23273798.2017.1369551

Xie, X., and Jaeger, T. F. (2020). Comparing non-native and native speech: are L2 productions more variable? *J. Acoust. Soc. Am*. 147, 3322–3347. doi: 10.1121/10.0001141

Xie, X., Liu, L., and Jaeger, T. F. (2021b). Cross-talker generalization in the perception of nonnative speech: a large-scale replication. *J. Exp. Psychol*. doi: 10.1037/xge0001039. [Epub ahead of print].

Xie, X., and Myers, E. B. (2017). Learning a talker or learning an accent: acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *J. Mem. Lang*. 97, 30–46. doi: 10.1016/j.jml.2017.07.005

Xie, X., Theodore, R. M., and Myers, E. B. (2017). More than a boundary shift: perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories. *J. Exp. Psychol*. 43:206. doi: 10.1037/xhp0000285

Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., and Jaeger, T. F. (2018b). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *J. Acoust. Soc. Am*. 143, 2013–2031. doi: 10.1121/1.5027410

Zheng, Y., and Samuel, A. G. (2020). The relationship between phonemic category boundary changes and perceptual adjustments to natural accents. *J. Exp. Psychol*. 46:1270. doi: 10.1037/xlm0000788

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 reviewer MK declared a shared affiliation with one of the authors, XX to the handling editor at the time of the review.

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 Psychology

ORIGINAL RESEARCH
published: 11 November 2021
doi: 10.3389/fpsyg.2021.661785

Fast but Not Furious. When Sped Up Bit Rate of Information Drives Rule Induction

Silvia Radulescu[1]*, *Areti Kotsolakou*[1], *Frank Wijnen*[1], *Sergey Avrutin*[1] *and Ileana Grama*[2]

[1] *Utrecht Institute of Linguistics-OTS, Utrecht University, Utrecht, Netherlands,* [2] *Amsterdam Centre for Language and Communication, Faculty of Humanities, University of Amsterdam, Amsterdam, Netherlands*

The language abilities of young and adult learners range from memorizing specific items to finding statistical regularities between them (*item-bound generalization*) and generalizing rules to novel instances (*category-based generalization*). Both external factors, such as input variability, and internal factors, such as cognitive limitations, have been shown to drive these abilities. However, the exact dynamics between these factors and circumstances under which rule induction emerges remain largely underspecified. Here, we extend our information-theoretic model (Radulescu et al., 2019), based on Shannon's noisy-channel coding theory, which adds into the "formula" for rule induction the crucial dimension of *time*: the rate of encoding information by a time-sensitive mechanism. The goal of this study is to test the *channel capacity*-based hypothesis of our model: if the *input entropy per second* is higher than the maximum rate of information transmission (bits/second), which is determined by the *channel capacity,* the encoding method moves gradually from *item-bound generalization* to a more efficient *category-based generalization*, so as to avoid exceeding the *channel capacity*. We ran two artificial grammar experiments with adults, in which we sped up the bit rate of information transmission, crucially not by an arbitrary amount but by a factor calculated using the *channel capacity* formula on previous data. We found that increased bit rate of information transmission in a repetition-based XXY grammar drove the tendency of learners toward *category-based generalization*, as predicted by our model. Conversely, we found that increased bit rate of information transmission in complex non-adjacent dependency *aXb* grammar impeded the *item-bound generalization* of the specific *a_b* frames, and led to poorer learning, at least judging by our accuracy assessment method. This finding could show that, since increasing the bit rate of information precipitates a change from *item-bound* to *category-based generalization,* it impedes the *item-bound generalization* of the specific *a_b* frames, and that it facilitates *category-based generalization* both for the intervening *Xs* and possibly for a/b categories. Thus, sped up bit rate does not mean that an unrestrainedly increasing bit rate drives rule induction in any context, or grammar. Rather, it is the specific dynamics between the *input entropy* and the maximum *rate of information transmission*.

Keywords: rule induction, entropy, channel capacity (information rate), generalization (psychology), category formation, bit rate

OPEN ACCESS

Edited by:
Matthew W. Crocker,
Saarland University, Germany

Reviewed by:
Nuria Sagarra,
Rutgers, The State University of New Jersey, United States
Ivan Yuen,
Macquarie University, Australia

****Correspondence:***
Silvia Radulescu
S.Radulescu@uu.nl;
sil.radulescu@gmail.com

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

Received: *31 January 2021*
Accepted: *06 October 2021*
Published: *11 November 2021*

Citation:
Radulescu S, Kotsolakou A, Wijnen F, Avrutin S and Grama I (2021) Fast but Not Furious. When Sped Up Bit Rate of Information Drives Rule Induction. Front. Psychol. 12:661785. doi: 10.3389/fpsyg.2021.661785

INTRODUCTION

Both young and adult learners possess a domain-general distributional learning mechanism for finding statistical patterns in the input (Saffran et al., 1996; Thiessen and Saffran, 2007), and a learning mechanism that allows for category (rule) learning (Marcus et al., 1999; Wonnacott and Newport, 2005; Smith and Wonnacott, 2010; Wonnacott, 2011). While previously cognitive psychology theories claimed that there are two qualitatively different mechanisms, with rule learning relying on encoding linguistic items as abstract categories (Marcus et al., 1999), as opposed to learning statistical regularities between specific items (Saffran et al., 1996), recent views converge on the hypothesis that one mechanism, *statistical learning*, underlies both item-bound learning and rule induction (Aslin and Newport, 2012, 2014; Frost and Monaghan, 2016; Radulescu et al., 2019). Rule induction (generalization or regularization) has often been explained as resulting from processing input variability (quantifiable amount of statistical variation), both in young and adult language learners (Gerken, 2006; Hudson Kam and Chang, 2009; Hudson Kam and Newport, 2009; Reeder et al., 2013).

This study looks into the factors that drive the inductive step from encoding specific items and statistical regularities to inferring abstract rules. While supporting the *single-mechanism hypothesis* and a *gradient of generalization* proposed previously (Aslin and Newport, 2012, 2014), in Radulescu et al. (2019), we took a step further in understanding the two qualitatively different representations discussed in previous research, which we dubbed, in accordance with previous suggestions (Gómez and Gerken, 2000), *item-bound generalizations* and *category-based generalizations*. While *item-bound generalizations* describe relations between specific physical items (e.g., a relation based on physical identity, like "*ba* always follows *ba*" or "*ke* always predicts *mi*"), *category-based generalizations* are operations beyond specific items that describe relationships between categories (variables), e.g., "Y always follows X," where Y and X are variables taking different values. In order to explain *how* and *why* a single mechanism outputs these two qualitatively different forms of encoding, Radulescu et al. (2019) proposed an information-theoretic model of rule induction as an encoding mechanism. In this model, based on Shannon's communication theory (1948), we put together both the statistical properties of the input, i.e., *input entropy*, and the finite capacity of the brain to encode the input. In information-theoretic terms at the computational level, in the sense of Marr (1982), we define encoding capacity as *channel capacity*, that is, the finite rate of information transmission (entropy per unit of time, bits/s), which might be supported by certain cognitive capacities, e.g., memory capacity, at the algorithmic level.

Indeed, previous research hinted at cognitive constraints, i.e., memory limitations, on rule learning: the *Less-is-More* hypothesis (Newport, 1990, 2016) proposed that differences in tendency to generalize between young and adult learners stem from maturational differences in memory development: limited memory capacity leads to difficulties in storing and retrieving low-frequency items, which prompts the overuse of more frequent forms leading to overgeneralization. A few studies investigating the nature of these cognitive constraints showed that, while there is some evidence for the *Less-is-More* hypothesis (Hudson Kam and Newport, 2005, 2009; Hudson Kam and Chang, 2009; Wonnacott, 2011), it is not yet clear under *what* specific circumstances and *why* memory constraints should drive rule learning (Perfors, 2012; Hudson Kam, 2019). Cognitive constraints on regularization were also found in nonlinguistic domains (Kareev et al., 1997; Ferdinand et al., 2019), while constrained regularization tendencies were found to be similar across language domains, morphology vs. word order (Saldana et al., 2017).

Nevertheless, the exact cognitive load and mechanisms at stake in rule induction have yet to be thoroughly specified. To this end, Radulescu et al. (2019) offer an extended and more refined information-theoretic approach to the *Less-is-More* hypothesis, by proposing an entropy model for rule induction, which quantifies the specific pattern of statistical variability in the input (i.e., *input entropy*, measured in bits) to which the brain is sensitive, and hypothesizes that rule induction is driven by the interaction between the input entropy and the finite encoding capacity of the brain (i.e., *channel capacity*). Crucially, the model proposes that rule induction is an automatic process that moves *gradually – bit by bit* – from a high-fidelity item-specific encoding (*item-bound generalization*) to a more general abstract encoding (*category-based generalization*), as a result of the input entropy being higher than the *channel capacity*, i.e., the maximum rate of information encoding (bits/s). The model is based on Shannon's *entropy* and noisy-channel coding theory (Shannon, 1948), which says that in a communication system, a message (or information) can be transmitted reliably (i.e., with the least loss in bits of information), if, and only if, encoded using an encoding method that is efficient enough so that the rate of information transmission (i.e., per unit of time), including noise, is below the capacity of the channel. If the rate of information transmission (bit rate) is higher than the *channel capacity*, then another more efficient encoding method can be found, but the *channel capacity* cannot be exceeded.

Based on these concepts, our entropy model for rule induction posits that the change in encoding method, i.e., from *item-bound* to *category-based generalization*, is driven by a kind of a regulatory mechanism, which moves from an inefficient encoding method (with loss of information), to a more efficient encoding method, which allows for higher input entropy to be encoded reliably (with the least loss possible) per second, but crucially below the capacity of the channel. The reliability of encoding should be understood intuitively as given by the least loss of information (caused by noise interference) against the sent message. Thus, this model adds into the rule induction "formula" the crucial dimension of time, i.e., the rate of encoding information by a time-sensitive encoding mechanism, and, consequently, the decrease in loss of information by moving to a more efficient encoding.

A few studies used different (not information-theoretic) methods of quantifying and manipulating a time-dependent variable to investigate the role it plays in category learning (exposure time, Endress and Bonatti, 2007; Reeder et al., 2013), in nonadjacent dependency learning (speech rate, Wang et al., 2016, 2019) and in auditory statistical learning (inter-stimulus temporal distance, Emberson et al., 2011). Although these studies used different designs, stimulus materials, and forms of operationalization to the temporal variable, nevertheless, a clear pattern stands out: generally, a shorter time is beneficial to auditory rule (category) learning. However, the exact amount of time, and the mechanism and reasons for it having a positive effect on rule learning are still to be fully investigated and understood.

In order to address these gaps, this study further extends the entropy model we proposed in Radulescu et al. (2019), and puts forth an innovative information-theoretic quantification of the time-dependent variable, that is not an arbitrary manipulation of inter-stimulus temporal distance or exposure time, but the information-theoretic concept of *channel capacity* and Shannon's noisy-channel coding theory.

AN ENTROPY AND CHANNEL CAPACITY MODEL FOR RULE INDUCTION

Among other studies that used entropy measures to look into regularization patterns (Perfors, 2012, 2016; Ferdinand, 2015; Saldana et al., 2017; Samara et al., 2017; Ferdinand et al., 2019), Radulescu et al. (2019) and this study take a step further and propose an information-theoretic model that captures the dynamics of the interaction between the *input entropy* and the encoding capacity (*channel capacity*). This model specifies a quantitative measure for the likelihood of transitioning from encoding specific probability distributions to category formation. Specifically, our model hypothesizes that the *gradient of generalization* (Aslin and Newport, 2012) results from a *bit by bit* increase in *input entropy* per unit of time, which gradually adds up to the maximum rate of information transmission (bits/s), i.e., *channel capacity* of the learning system.

Given a random variable X, with n values $\{x_1, x_2 \ldots x_n\}$, Shannon's entropy (Shannon, 1948), denoted by $H(X)$, is defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)\,[1];$$

where $p(x_i)$ is the occurrence probability of x_i. This quantity (H) measures the information per symbol produced by a source of input, i.e., it is a measure of the average uncertainty (or surprise) carried by a symbol produced by a source, relative to all the possible symbols (values) contained by the set (Shannon, 1948).

In Radulescu et al. (2019), in two artificial grammar experiments, we exposed adults to a three-syllable XXY artificial grammar. We designed six experimental conditions with increasing input entropy (2.8, 3.5, 4, 4.2, 4.58, and 4.8 bits). The results showed that an increase in input entropy gradually shaped *item-bound generalization* into *category-based generalization* (Radulescu et al., 2019). Thus, we obtained a precise measure of the sensitivity of a learner to the input entropy: the information load of a learner (=surprise) of the XXY structure decreases logarithmically as the input entropy increases. These findings bring strong evidence for the *gradient of generalization* depending on the probabilistic properties of the input, as proposed by Aslin and Newport (2014).

While in Radulescu et al. (2019) we probed the effect of the first factor (*input entropy*), in this study we further develop and test the model by probing the effect of the second factor – *channel capacity* – on rule induction.

Channel Capacity in Information-Theoretic Terms

This section elaborates on the other factor of our entropy model, namely *channel capacity,* which is another information-theoretic concept in Shannon's *noisy-channel coding theory* of a communication system. Shannon (1948) defines a communication system as having five main components: an information source (which produces a message), a transmitter (which encodes the message into a signal), a channel (the medium used to transmit the signal), a receiver (which does the inverse operation of the transmitter, that is, decodes the signal to reconstruct the message), and a destination (the person or thing for which the message is intended). In short, an information source produces a message, which is encoded by a transmitter into a signal that is suitable for transmission over a channel to a destination. The main factor under investigation here is the medium used for the transmission of information, i.e., the *channel,* and its capacity for information transmission. It follows, and it must be specified that the process of information transmission encompasses all processes starting with the transmission of information from the source to the destination, that is, all the transmission and encoding-decoding processes.

In order to define *channel capacity,* we first have to define the two main factors that are relevant for *channel capacity*: the source rate of information transmission and noise. Since the process of information transmission occurs in time, Shannon defined *source rate of information transmission* as the amount of information that a source transmits per unit of time. Information is measured using *entropy*, so *source rate of information transmission* (H') is the amount of entropy that the source produces per unit of time (bits/s), or the source rate of information production.

The ideal case of a noiseless transmission is nearly impossible under normal real-life conditions; thus, transmission is affected by another variable, *noise*. *Noise* is defined as any random perturbations that interfere with the signal, thus rendering a *noisy channel*. The noise might perturb the signal during transmission through the channel or at either terminal end, i.e., transmitter and receiver's end. As a result, there are missing bits of information because of a noisy transmission. Shannon (1948) defined this loss of information as *rate of equivocation (E).*

The actual rate of information transmission (R) *via* a *noisy channel* is obtained by subtracting the *rate of equivocation* (E) from the *source rate of information transmission*, H' (Shannon, 1948):

$$R = H' - E.$$

Note that *actual rate of information transmission* (R) is different from *source rate of information transmission* (H'), since it takes

[1] *Log* should be read as log to the base 2 here and throughout the paper.

into account the loss of information due to noise (E), which occurs in the transmission of information from the source to the destination. *Source rate of information transmission* (H′) is the rate at which the source produces and transmits information, while *actual rate of information transmission* (R) is quantified at the other terminal end, i.e., the receiver, after the noise had caused a loss in information (E).

Shannon (1948) demonstrated mathematically that the capacity of *noisy channel* should be the maximum possible rate of information transmission (R), which can be obtained only if the encoding method is adequate and efficient:

$$C = Max\ (R) = Max\ (H' - E).$$

The formula above means that the maximum rate of information transmission, i.e., *channel capacity,* can be achieved by an adequate and efficient encoding method. The efficiency of the encoding method means that the rate of equivocation (E) is kept at a minimum, in order for the actual information transmission to be as close as possible to the source rate of production. That means the received signal matches closely the sent signal, and, consequently, the message is received with the least loss of information.

According to Theorem 11 by Shannon (1948), given a certain source with a rate of information production H′ (entropy per unit of time), if $H' \leq C$, information can be sent through a noisy channel at the rate *C* with an arbitrarily small frequency of errors using a proper encoding method. If $H' > C$, it is possible to find an encoding method to transmit the signal over the channel, such that the rate of equivocation is minimum, as specified by Shannon, less than $H' - C + e$ (*e* stands for errors*)*, but the rate of transmission can never exceed C. If there is an attempt to transmit a message at a higher rate than C, using the same encoding method, then there will be an equivocation rate at least equal to the excess rate of transmission. In other words, a message can only be communicated reliably if it is encoded in such a way, i.e., using an efficient encoding method, so that the rate of information transmission, including noise, is below the capacity of the channel. In this study, we will focus on the first factor in the *channel capacity* formula, namely *source rate of information transmission.*

Main Hypotheses of the Model About the Effect of Channel Capacity on Rule Induction

(1) *Item-bound generalization* and *category-based generalization* are outcomes of the same information encoding mechanism that *gradually* goes from a high-specificity form of encoding (*item-bound generalization)* to a more general abstract encoding (*category-based generalization)*, as triggered by the interaction between *input entropy* and the finite encoding capacity of the learning system. The encoding mechanism moves from *item-bound* to *category-based generalization* as *input entropy per unit of time* increases and becomes higher than the maximum rate of information transmission, i.e., *channel capacity,* as follows:

(a) If the source rate of information transmission (H′–input entropy per second) is below or matches *channel capacity*, then the information can be encoded using an encoding method that matches the statistical structure of the input (the probability distribution of the specific items). Thus, if $H' \leq C$, the information about specific items with their uniquely identifying (acoustic, phonological, phonotactic, prosodic, distributional, etc.) features and probability distribution (i.e., input entropy) can be encoded with a high-fidelity item specificity, and transmitted through the channel, with little loss of information, at the channel rate, the maximum rate of information transmission, and encoded by *item-bound generalization.* If $H' > C$, *item-bound generalization* is impeded.

(b) If an attempt is made to exceed the finite *channel capacity* of the encoding system, that is, the source rate of information transmission (H′–input entropy per second) does not match *channel capacity*, but it is higher than *channel capacity*, it is possible to find a proper method that encodes more information (entropy), but the rate of information transmission cannot exceed the available *channel capacity*. According to Theorem 11 (Shannon, 1948), if there is an attempt to transmit information at a rate higher than C, using the same encoding method, then there will be an equivocation rate at least equal to the excess rate of transmission. In other words, the increased source rate of information ($H' > C$) brings higher inflow of *noise*, which interferes with the signal and causes an increased equivocation rate or information loss (as explained above). Thus, we hypothesize that it is precisely the *finite channel capacity* that drives the restructuring of the information, in order to find another more efficient encoding method. A more efficient encoding allows for higher input entropy per second to be encoded reliably (with the least information loss possible). As we argued in Radulescu et al. (2019), information is re-structured by (unconsciously) re-observing the item-specific features and structural properties of the input. Noise introduces random perturbations that interfere with the signal and feature configuration. This leads to instability, which unbinds features and sets them free to interact and bind into new structures. Then, similarities (shared features) that have higher significance (i.e., are "stronger" because of their higher probability) are kept in the new encoding, while differences between items (unshared features), which are insignificant features (e.g., low-probability "noisy" features) are erased or "forgotten." This leads to a compression of the signal by reducing the number of unshared "noisy" features encoded with individual items (i.e., bits of information) and grouping them in "buckets" (categories). As a result, a new form of encoding is created, which allows for higher *input entropy* to be encoded using the available *channel capacity*, thus yielding a more general (less specific) *category-based* encoding method. Thus, *finite channel capacity* is designed to drive the re-structuring of the information for the purpose of adapting to noisier (=increasingly entropic) environments, by the principle of self-organization in line with Dynamic Systems Theory invoked in studies on other cognitive mechanisms, e.g., Stephen et al. (2009).

(2) *Channel capacity* is used here as an information-theoretic measure of the encoding capacity used in linguistic rule induction (at the computational level, in the sense of Marr (1982))[2]. In order to identify psychological correlates (at the algorithmic level), we follow experimental evidence from the *Less-is-More* hypothesis line of research, which suggests that memory constraints drive linguistic rule induction (Hudson Kam and Newport, 2005, 2009), and we embed this in classical and recent models of memory capacity and attention (Miller, 1956; Cowan, 2005; Oberauer and Hein, 2012; Baddeley et al., 2015). Hence, we hypothesize that the cognitive capacity that underlies *channel capacity*, specifically in linguistic rule induction (and, implicitly, in category formation), is the attentional capacity focused on activated representations in long-term memory, in other words working-memory capacity (WM), as defined in Cowan (2005). Rule induction can be argued to rely on the storage and online time-dependent processing capacities that support the ability to maintain active goal-relevant information (the rule), while concurrent processing (of other possible hypotheses and of noise) takes place (which is what defines WM as well, Conway et al., 2002). Corroborating evidence comes from positive correlations found between WM and domain-general categorization tasks (Lewandowsky, 2011).

Thus, while we generally deem linguistic rule induction to be supported by a domain-general WM capacity, rather than language-specific algebraic rule learning as proposed by early prominent research (Marcus et al., 1999), in this study, we are exploring specific possible memory components and WM-correlated abilities that are directly involved in linguistic rule induction (besides more general storage and retrieval components tested in previous studies under the *Less-is-More* hypothesis, Hudson Kam and Chang, 2009; Perfors, 2012). Hence, we specifically predict that one of the components underlying *channel capacity* in linguistic rule induction is a domain-general pattern recognition capacity, given that a rule induction task can be intuitively envisaged as a task of finding patterns/rules in the input.

A possible candidate test of domain-general pattern recognition is the RAVENS test (Raven et al., 2000), which was shown to be based on rule induction (Carpenter et al., 1990; Little et al., 2012) and to rely on similar storage and online time-dependent processing capacities to maintain active goal-relevant information (the rule) while concurrent processing takes place (Conway et al., 2002). Although this pattern recognition test and WM capacity are not identical (Conway et al., 2003), and apparently WM is not a causal factor for pattern recognition either (Burgoyne et al., 2019), high positive correlations were found between measures of WM capacity and tests for this domain-general pattern-recognition capacity (such as RAVENS, e.g., Conway et al., 2002; Little et al., 2014; Dehn, 2017).

TESTING THE PREDICTION OF SPEEDING UP THE SOURCE BIT RATE OF INFORMATION TRANSMISSION

The goal of this study is to probe the effect of the time-dependent variable of the second main factor of our entropy model, *channel capacity*, on rule induction, by directly increasing *source rate of transmission* (H′), in order to attempt to exceed *channel capacity*. Theoretically, following the definition of *channel capacity* and Shannon's Theorem 11 (Shannon, 1948), this can be achieved in two ways: either by increasing the amount of entropy (bits) at a constant rate or by speeding up the rate of feeding information (at constant bit value) into the channel. It follows that, practically, there are two methods to attempt to exceed *channel capacity*:

(1) Add stimulus-unrelated entropy (*noise*) in the input to render a noisier channel, while keeping the time variable constant. This method aims at exceeding *channel capacity* by specifically modulating the *noise* variable of *channel capacity*.
(2) Increase the source rate of information production to directly modulate the time-dependent variable of *channel capacity*. This method reduces the time that the same amount of entropy is sent through the channel, i.e., speeds up the bit rate of information transmission.

We employed the first method in another study (Radulescu et al., 2020 unpublished data), and we found that added stimulus-irrelevant entropy (*noise*) drove a higher tendency toward *category-based generalization*. In this study, we employed the second method: we increased the source rate of information transmission (*input entropy per second*) in order to directly modulate the time-dependent variable of *channel capacity*. According to our entropy model, speeding up the source rate of transmission (i.e., to a higher rate than *channel capacity*) leads to a change in encoding method, so as to avoid increased equivocation rate. Why? Because increased rate of equivocation is in fact information loss. Thus, the encoding method transitions to another encoding method in order to achieve more efficient transmission of information: that is, faster encoding rate with least information loss. Specifically, we hypothesize that increasing the source rate of information transmission leads to

[2] Although with different definitions and applications, *channel capacity* has previously been used in an early study on capacity in memory studies on psychology (Miller, 1956) and in more recent mathematical modeling for inferring workload capacity using response time hazard functions (Townsend and Ashby, 1978; Townsend and Eidels, 2011).

higher tendency to move from *item-bound* to *category-based generalization* for the purpose of achieving a more efficient encoding, with the least loss of information possible.

We tested the effect of speeding up the source rate of information transmission on both the repetition-based XXY grammar from the study of Radulescu et al. (2019) and a more complex grammar, non-adjacent-dependency grammar (aXb). The learning of a repetition-based XXY grammar requires learners to abstract away from specific items of the X and Y categories, and to move from *item-bound* to *category-based generalization,* that is, to learn a *same-same-different* rule between categories, regardless of their specific items. A source rate of transmission higher than *channel capacity* is hypothesized to boost this transition and, thus, have a positive effect on learning an XXY grammar. However, learning a non-adjacent dependency grammar is a more complex process: it entails learning item-bound dependencies between specific *a* and *b* elements and *category-based generalization* of the rich category of intervening *Xs* (Gómez, 2002; Onnis et al., 2004; Frost and Monaghan, 2016; Grama et al., 2016; Wang et al., 2019). This type of artificial grammar learning models the mechanisms needed in language acquisition to acquire rules such as *is* go-*ing, is* learn-*ing.* Thus, the learning of this type of *aXb* grammar requires learners to move from *item-bound* to *category-based generalization* for the *X* category of middle elements, while, crucially, sticking to *item-bound generalization* for specific *a_b* dependencies. If increased source rate of information transmission drives *category-based generalization* for the *X* category, it follows that it should impede *item-bound generalization* for the specific *a_b* dependencies of such an *aXb* grammar. So how does the model perform when tested on such a complex type of grammar?

Given an entropy (H) of a source and an average number of symbols produced by the source per second (m), we can calculate the source rate of information transmission, $H' = mH$ (Shannon, 1948). Using this formula, we estimated a source rate of transmission of information in experiments carried out by Radulescu et al. (2019). Then, we specifically predicted that, if we keep the same information content (input entropy) of the lowest entropy grammar from Radulescu et al. (2019), where there was no evidence of *category-based generalization*, but we increase the source rate of transmission up to the source rate of transmission of the highest entropy condition from the same study, where that study found high tendency toward *category-based generalization*, then we should see a higher tendency toward *category-based generalizations*, even though the statistical properties (entropy) of the input are the same.

Specifically, let us denote the source rate of information transmission in the highest entropy grammar from Radulescu et al. (2019) as $\mathrm{H'_H = m_1 H_H}$, and the source rate of information transmission in the lowest entropy version as $\mathrm{H'_L = m_1 H_L}$. Note that the average rate of symbols per second ($\mathrm{m_1}$) was the same in both versions. For the purpose of the manipulation we are aiming for, we would like to obtain $\mathrm{H'_H = H'_L}$ but by keeping $\mathrm{H_L}$ constant and increasing the average rate of symbols/s to obtain $\mathrm{m_2}$ such that $\mathrm{m_2 > m_1}$. Thus, in the three-syllable XXY grammar from Radulescu et al. (2019), for a constant m_1 *(symbols/s)*:

$\mathrm{H_L = 2.8b/symbol}$: $H'_\mathrm{L} = m_1\ H_\mathrm{L}$

$\mathrm{H_H = 4.8b/symbol}$: $H'_\mathrm{H} = m_1\ H_\mathrm{H}$.

For the purpose of increasing the source rate of transmission up to H_H*'* while keeping entropy constant ($\mathrm{H_L}$), and by increasing the average rate of symbols/s, we calculated the necessary $\mathrm{m_2}$ as follows:

$m_2\ H_\mathrm{L} = H'_\mathrm{H}$

$m_2\ H_\mathrm{L} = m_1\ H_\mathrm{H}$

$m_2/\ m_1 = H_\mathrm{H}\ /H_\mathrm{L}$

$m_2 = (4.8/2.8)\ m_1$

$m_2 = 1.71\ m_1$

Thus, we obtained $m_2 = 1.71m_1$, and translated it into duration of syllables and within- and between-string pauses, such that we sped up all syllables and pauses proportionally by a coefficient of 1.71. As a result, we created a faster source rate of information transmission, i.e., entropy per second ($\mathrm{H'_L = H'_H}$), but we kept the entropy per symbol constant $\mathrm{H_L = 2.8b/symbol}$.

Next, for the aXb grammar, we created two versions of the grammar with different levels of entropy ($\mathrm{H_L}$; $\mathrm{H_H}$), but the same average rate of symbols/s (m_3):

$\mathrm{H_L = 3.52b/symbol}$: $H'_\mathrm{L} = m_3\ H_\mathrm{L}$

$\mathrm{H_H = 4.71b/symbol}$: $H'_\mathrm{H} = m_3\ H_\mathrm{H}$.

For the purpose of increasing the source rate of information transmission up to H'_H while keeping entropy constant ($\mathrm{H_L}$), and by increasing the average rate of symbols/s, we calculated the necessary $\mathrm{m_4}$ as follows:

$m_4\ H_\mathrm{L} = H'_\mathrm{H}$

$m_4\ H_\mathrm{L} = m_3\ H_\mathrm{H}$

$m_4/\ m_3 = H_\mathrm{H}\ /H_\mathrm{L}$

$m_4 = (4.71/3.52)\ m_3$

$m_4 = 1.34\ m_3$.

Thus, we obtained $m_4 = 1.34m_3$, and translated it into duration of syllables and within- and between-string pauses, such that we sped up all elements (syllables and pauses) proportionally by a coefficient of 1.34. As a result, we created a faster source rate of information transmission, i.e., entropy per second ($\mathrm{H'_L = H'_H}$), but we kept the entropy per symbol constant $\mathrm{H_L = 3.52b/symbol}$.

Besides probing the direct effect of the time variable of *channel capacity*, as presented above, this study also looked into the effect of individual differences in cognitive capacities on rule induction, to explore the cognitive capacities that underlie *channel capacity*: short-term memory capacity and a domain-general pattern-recognition capacity, as a component that reflects the working memory capacity we deem relevant for rule induction. To this end, we tested each participant on three independent tests: Forward Digit Span, as a measure of explicit short-term memory (Baddeley et al., 2015), an incidental memorization task, which measures implicit memory capacity, i.e., the ability to memorize information without being explicitly instructed to do so (Baddeley et al., 2015), and RAVENS Standard Progressive Matrices (Raven et al., 2000), which is a standardized test based on visual pattern-recognition (Carpenter et al., 1990; Little et al., 2014).

We ran two experiments to test the effect of increased rate of information on rule induction in an XXY grammar and in an aXb non-adjacent dependency grammar. Importantly, we tested the same participants in both experiments, which were conducted in two separate sessions, on two different days (at least 3 days between sessions). For practical reasons, all the participants took part first in the aXb grammar experiment (Experiment 2) and then in the XXY grammar experiment (Experiment 1). For theoretical presentation reasons, which have to do with the logic and theoretical development of the entropy model and its hypotheses, here we present the XXY experiment first, followed by the aXb experiment.

To the best of our knowledge, these are the first language learning experiments that investigate the effect of the time-dependent variable of *channel capacity* in rule induction by specifically testing information-theoretic predictions made by an entropy model.

EXPERIMENT 1

In Experiment 1, the participants carried out three tasks. The first task presented the three-syllable XXY grammar in two different conditions: a slow source rate of information transmission (Slow Rate condition) and a fast source rate of information transmission (Fast Rate condition). In the Slow Rate condition, we used the exact stimuli and source rate of information transmission (H'_L) as in the lowest entropy condition from Radulescu et al. (2019), 2.8 bits. In the Fast Rate condition, the same stimuli were used ($H_L = 2.8$), but the source rate of information transmission was increased by a factor of 1.71 (see section "Testing the Prediction of Speeding up the Source Bit Rate of Information Transmission"). In the test phases, the participants heard four different types of test strings (from Radulescu et al., 2019), as presented below. The participants answered a yes/no question to indicate whether the test strings could be possible in familiarization language.

Familiar-syllable XXY (XXY structure with familiar X-syllables and Y-syllables), correct answer: accept. This type of test strings probed the learning of familiar strings. Both groups were expected to accept these strings as grammatical because they were encoded as either *item-bound generalizations* (Slow Rate condition) or *category-based generalizations* (Fast Rate condition).

New-syllable XXY (XXY structure with new X-syllables and Y-syllables), correct answer: accept. This type tested whether learners moved from *item-bound* to *category-based generalization,* which enables them to accept XXY strings with new syllables. We expected that the Fast Rate group was more likely to accept these strings, as compared with the Slow Rate group. However, the absolute mean acceptance rate of these strings does not represent direct evidence for *category-based generalization*. As we argued in Radulescu et al. (2019), this rate should be compared with the mean acceptance rate of Familiar-syllable XXY strings: if the difference of the mean acceptance rate between New-syllable XXY strings and Familiar-syllable XXY strings is significantly smaller in the Fast Rate as compared with the Slow Rate condition (i.e., effect size), this would suggest that the Fast-Rate learners were more likely to have formed *category-based generalization* than the Slow-Rate learners.

Familiar-syllable X_1X_2Y (X_1X_2Y structure with familiar syllables), correct answer: reject. The participants are expected to reject these strings because the input was encoded as either *item-bound generalizations* (Slow-Rate learners) or *category-based generalizations* (Fast-Rate learners). Slow-Rate learners are expected to reject this type of strings, as their memory trace of the Familiar-syllable XXY strings is expected to be strong enough to highlight a mismatch between these strings and the Familiar-syllable X_1X_2Y strings. Fast-Rate learners are expected to form *category-based generalizations*, thus they should reject the Familiar-syllable X_1X_2Y strings as deviant from the *same-same-different* rule. However, as argued in Radulescu et al. (2019), we expect both *item-bound* and *category-based generalization* to support accuracy scores on X1X2Y strings because of different reasons: if *item-bound generalization* is developed, as (per hypothesis) learners encoded the strings as frozen *item-bound generalization*, which highlight clear mismatches between familiar and noncompliant combinations of specific items. However, memory traces of familiar items (i.e., syllables) might prompt incorrect acceptance of familiar-syllable X1X2Y. On the other side, if *category-based generalization* is fully encoded, these strings will be much more frequently rejected as non-compliant with the *same-same-different* rule, regardless of any memory trace. Thus, the higher rejection rate of these strings suggests stronger category-based encoding.

New-syllable X_1X_2Y (X_1X_2Y structure with new syllables), correct answer: reject. The participants are expected to reject this type of strings, because the input was encoded as either *item-bound generalizations* (Slow Rate group) or *category-based generalizations* (Fast Rate group).

The second task was a Forward Digit Span (Baddeley et al., 2015), and the third task was an incidental memorization task (Baddeley et al., 2015). According to the hypotheses of our entropy model, we predicted a negative effect of the explicit/incidental memory capacities on the tendency of learners to move from *item-bound* to *category-based generalization*. The rote memorization capacity (Baddeley et al., 2015) is hypothesized to have a negative effect on the transition from *item-bound* to *category-based generalization*, since a strong memory capacity for specific items and their probability configuration would support a higher *input entropy* to be encoded per unit of time (i.e., a higher *channel capacity,* in computational terms).

Participants

Fifty-six adults, Dutch native speakers (10 males, age range 18–72, $M_{age} = 26.39$, $SD_{age} = 11.06$) participated. All the participants were naïve to the aim of the experiment, had no known language, reading, or hearing impairment or attention deficit, and received €5.

Materials

Task 1: XXY Grammar

Familiarization Stimuli

The participants in both the Slow Rate and the Fast Rate conditions listened to the same three-syllable XXY[3] artificial grammar used in the low entropy condition of Experiment 2 from Radulescu et al. (2019). Each string consisted of two identical

[3] Each letter stands for a set of syllables that do not overlap, that is the subset of X-syllables does not overlap with the subset of Y-syllables.

syllables (XX) followed by another different syllable (Y): e.g., *ke:ke:my, da:da:li*. All syllables consisted of a consonant followed by a long vowel, to resemble common Dutch syllable structure. Seven X-syllables and seven Y-syllables were used to generate seven strings (see **Supplementary Appendix A** for complete stimulus set). Each string was repeated four times in each of the three familiarization phases (7 strings x 4 repetitions = 28 strings in each familiarization phase). The same 28 strings were used in all three familiarization phases, such that the entropy was the same, 2.8 bits. The participants were randomly assigned to either the Slow Rate or the Fast Rate condition, in a between-subjects design, and the presentation order of strings was randomized per participant. For entropy calculations, we employed the same method as in Radulescu et al. (2019), which is a fine-tuned extension of a related entropy calculation method proposed by Pothos (2010) for finite state grammars (see **Table 1** for complete entropy calculations). In the Slow Rate condition, there was a pause of 50 ms between the syllables within strings, and a pause of 750 ms between the strings. In the Fast Rate condition, all X and Y syllables, as well as the within-and between-string pauses, were sped up separately by a factor of 1.71 using Praat (Boersma and Weenink, 2019).

Test Stimuli

There were three familiarization phases, interleaved with three intermediate test phases and a final (longer) test phase. Each intermediate test included four test strings, one of each type. The final test had eight test strings (two of each type): $4 + 4 + 4 + 8 = 20$ test strings in total (see **Supplementary Appendix A** for complete stimulus set). Accuracy scores were measured as correct acceptance of Familiar-syllable XXY and New-syllable XXY strings, and correct rejection of Familiar-syllable X_1X_2Y and New-syllable X_1X_2Y strings.

We recorded all the yes/no answers and coded them as correct/incorrect answers. From all the 20 correct/incorrect answers for each participant, we calculated a proportion of correct answers per each type of test item. We performed an empirical logarithmic transformation on the proportions, to analyze the data using a linear model.

Task 2: Forward Digit Span

The participants were explicitly told that this was a memory test, during which a series of digits would be presented aurally, and that they would have to recall them in the same order. To prevent the participants from creating a visual pattern on the keypad while listening to the digits, we modified the standard Forward Digit Span task such that no physical keyboard was made available to the participants; rather, a row with buttons for each digit was displayed in a line on the screen only in the moment when they were asked to enter the digits by clicking the buttons, and disappeared during the listening phases. We used the standard scoring method: we measured the highest span of each participant, and recorded it as one data point per participant.

TABLE 1 | Entropy value for Experiment 1, taken from Radulescu et al. (2019).

Low entropy
$H[bX] = H[7] = -\Sigma[0.143^*\log 0.143] = 2.8$
$H[XX] = H[7] = 2.8$
$H[XY] = H[7] = 2.8$
$H[Ye] = H[7] = 2.8$
$H[bXX] = H[7] = 2.8$
$H[XXY] = H[XYe] = H[7] = 2.8$
$H[bigram] = 2.8$
$H[trigram] = 2.8$
$H[total] = \frac{H[bigram]+H[trigram]}{2} = 2.8$

Task 3: Incidental Memorization Test

The participants listened to 30 bisyllabic nonsense words resembling Dutch phonology. Crucially, the participants were not told in advance that a memory test would be administered. They were only told that they were about to listen to words from another forgotten language. They were instructed to imagine what the word might have meant in the forgotten language and to pick a category (flower, animal, or tool) based on what the word sounded like to them. They had 3 s to choose a category for each word by pressing the button for flowers, animals, or tools.

After this phase, a message informed the participants that they would be given a memory test, which would check whether they remembered the words they categorized during the previous phase. They were instructed to press a yes/no button on the screen, depending on whether they have heard the word previously or not. In the memorization test, the participants gave answers on 13 targets and 13 foils. We recoded all the correct/incorrect answers into a *d'* value for each participant.

Procedure

The participants completed the tasks in the order presented above. For Task 1, they were told that they would listen to a "forgotten language" that would not resemble any language they might know, and that the language had its own rules and grammar. The participants were informed that the language had more words than what they heard in the familiarization phases. They were told that each intermediate test would be different from the other tests, and that the tests were meant to check what they had noticed about the language. They had to decide, by pressing a Yes or a No button, if the words they heard in the tests could be possible in the language. This task lasted around 5 min. For Task 2, they were explicitly instructed that it was a memory test. For Task 3, they were not told in advance about the memory test. The entire experiment lasted for about 20 min.

Results

Figure 1 presents the mean correct acceptance rate (proportion of correct acceptances per group) for Familiar-syllable XXY strings and New-syllable XXY strings, across the two conditions (Slow Rate, Fast Rate). The mean correct acceptance rate in the Slow Rate condition for Familiar-syllable XXY strings was $M = 0.96$ ($SD = 0.1$), and for New-syllable XXY strings it was $M = 0.75$ ($SD = 0.27$). The mean rate of correct acceptance in the Fast Rate condition for Familiar-syllable XXY strings was $M = 0.99$ ($SD = 0.04$), and for New-syllable XXY strings it was $M = 0.9$ ($SD = 0.18$).

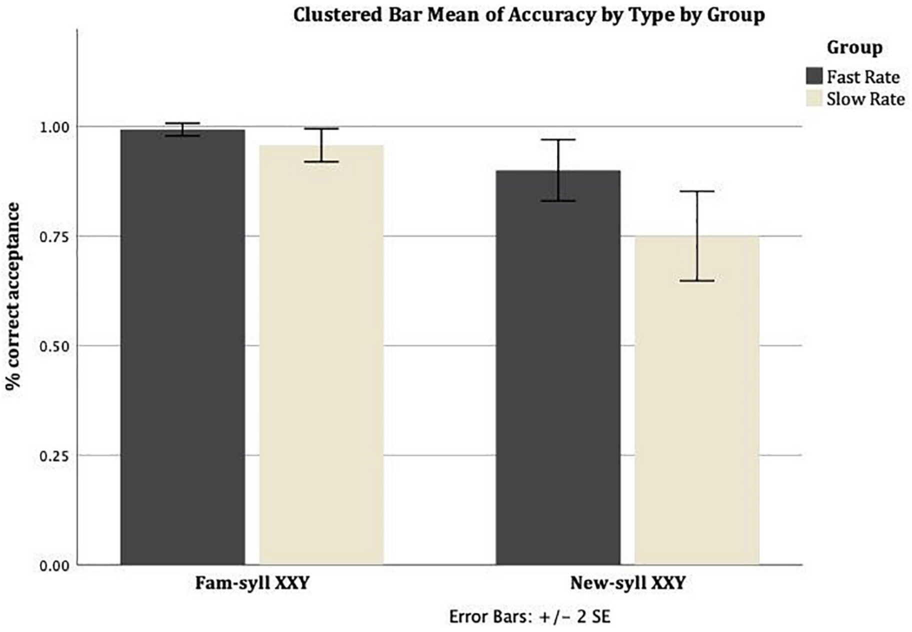


FIGURE 1 | Mean rate of correct acceptance for Familiar-syllable XXY and New-syllable XXY strings in both conditions: Fast Rate and Slow Rate. Error bars show standard error of the mean.

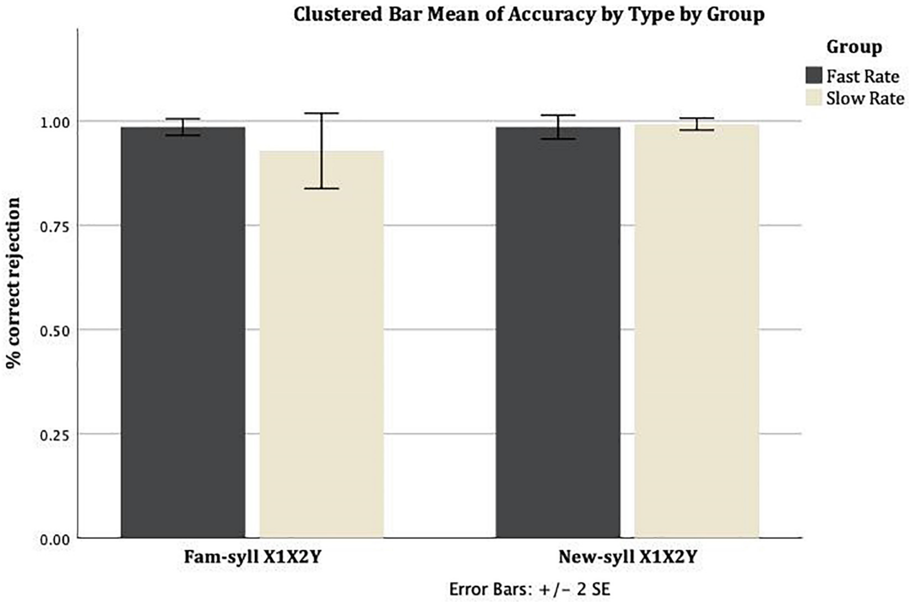


FIGURE 2 | Mean rate of correct rejection for Familiar-syllable X1X2Y and New-syllable X1X2Y strings in both conditions: Fast Rate and Slow Rate. Error bars show standard error of the mean.

Similarly, **Figure 2** shows the mean correct rejection rate (proportion of correct rejections per group) for Familiar-syllable X_1X_2Y strings and New-syllable X_1X_2Y strings, across the Slow Rate and Fast Rate conditions. In the Slow Rate condition, the mean correct rejection rate for Familiar-syllable X_1X_2Y strings was $M = 0.93$ ($SD = 0.24$), and for New-syllable X_1X_2Y strings it was $M = 0.99$ ($SD = 0.04$). In the Fast Rate condition, the mean correct rejection rate for Familiar-syllable X_1X_2Y strings was $M = 0.99$ ($SD = 0.05$), and for New-syllable X_1X_2Y strings it was $M = 0.99$ ($SD = 0.08$).

Figure 3 shows the distribution of individual mean rates per test type in both conditions.

In order to probe the effect of *channel capacity* on rule induction, we used IBM SPSS 26 to compare the performance in the two conditions (Slow Rate and Fast Rate groups) in a general linear mixed effects analysis of the relationship between Accuracy (correct acceptance of the grammatical test items and correct rejection of the ungrammatical ones) and the Rate of Transmission (Slow Rate, Fast Rate) as well as the Type of Test Strings (Familiar-syllable XXY, New-Syllable XXY, Familiar-syllable X_1X_2Y, New-Syllable X_1X_2Y). As a dependent variable, we entered Accuracy score into the model. As fixed effects, we entered Rate of Transmission, Type of Test Strings, and Rate of Transmission x Type of Test Strings interaction. As a random effect we had intercepts for subjects. The scores for Forward Digit Span, Incidental Memorization Task, and RAVENS tests[4] were entered one by one as covariates in the model. An alpha level of.05 was used for all the statistical tests. We started fitting the data from the intercept-only model and added the random and fixed factors one by one. The model reported here is the best fitting model, both in terms of the accuracy of the model in predicting the observed data, and in terms of Akaike Information Criterion.

We found a significant main effect of Type of test strings [$F(3,213) = 5.742$, $p = 0.001$], a Rate of Transmission × Type interaction that did not reach significance [F(4,213) = 2.039, $p = 0.09$], a non-significant Forward Digit Span effect [$F(1,213) = 0.069$, $p = 0.793$], a non-significant Incidental Memorization Task effect [$F(1,213) = 0.880$, $p = 0.349$], and a non-significant RAVENS effect [$F(1,213) = 2.326$, $p = 0.129$].[5]

Pairwise comparisons of the Estimated Marginal Means (adjusted to the mean values of the covariates in the model, i.e. Forward Digit Span = 6.68, Incidental Memorization Task = 1.968, RAVENS = 71.54) revealed a significant difference between the Rate of Transmission conditions (Fast Rate and Slow Rate groups) for the New-syllable XXY [$M = 0.101$, $SE = 0.045$, $F(1,213) = 4.936$, $p = 0.027$], and a nearly significant difference for the Familiar-syllable X1X2Y [$M = 0.085$, $SE = 0.045$, $F(1,213) = 3.522$, $p = 0.062$]. For the other two Types of test, pairwise comparisons of the Estimated Marginal Means adjusted for the same level of the covariates revealed a non-significant difference between the Rate of Transmission conditions (Fast Rate and Slow Rate groups): Familiar-syllable XXY [$M = 0.01$, $SE = 0.045$, $F(1,213) = 0.051$, $p = 0.822$] and New-syllable X1X2Y [$M = 0.012$, $SE = 0.045$, $F(1,213) = 0.069$, $p = 0.793$].

Cohen's effect size value (d) and the effect-size correlation (r) for the difference in acceptance between Familiar-syllable XXY and New-syllable XXY were higher in the Slow Rate condition ($d = 1.03$, $r = 0.45$; large effect size), than in the Fast Rate condition ($d = 0.69$, $r = 0.32$; medium effect size).

Discussion

The results of Experiment 1 show that the mean acceptance of new XXY strings as grammatical in the familiarization language was higher in the Fast Rate condition than in the Slow Rate condition, as predicted by our model. Moreover, there was a difference between the rates of acceptance of new XXY strings vs. familiar XXY strings depending on the rate of transmission: there was a smaller difference between the mean acceptance of the new XXY strings vs. familiar XXY strings in the Fast Rate condition compared with the Slow Rate condition. This shows differences between groups in terms of how they encoded the input: if learners do not make a clear distinction between a new XXY string and a familiar XXY string, we conclude that they encoded the input as *category-based generalization,* which

[4] RAVENS scores were obtained for the participants during the second experiment presented in this paper, since the same participants participated in both experiments (see section "Experiment 2" below).

[5] We also checked the main effect of Rate of Transmission, and since it was non-significant [$F(1,213) = 2.558$, $p = 0.111$], it did not improve the model, and it created effects of an overfitted model, we excluded it from the final model presented here.

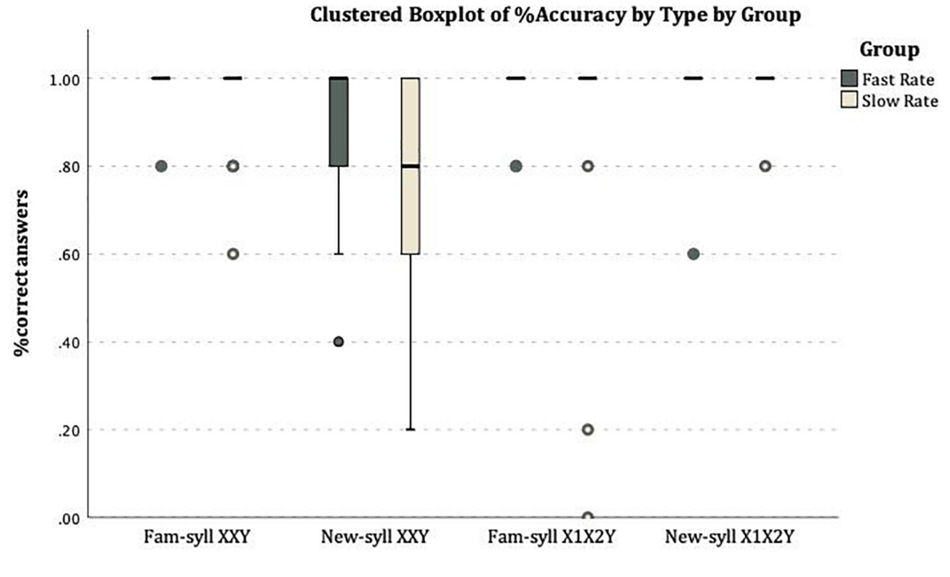


FIGURE 3 | On the *X*-axis the four types of test items: Familiar-Syllable XXY, New-syllable XXY, Familiar-syllable X1X2Y, New-syllable X1X2Y. On the *Y*-axis the mean rate of correct answers: correct acceptance for XXY strings (with familiar or new syllables) and correct rejections of X1X2Y (with familiar or new syllables).

allows them to accept any XXY string based on the *same-same-different* rule regardless of new or familiar syllables. Hence, a smaller difference between the means of acceptance of these test types in the Fast Rate condition shows a higher tendency toward *category-based generalization* than in the Slow Rate condition. Also the rate of correct rejection of X1X2Y strings with familiar syllables was higher in the Fast Rate condition than in the Slow Rate condition, which supports the same hypothesis of our model: when speeding up the source rate of transmission, learners formed *category-based generalizations,* which helped them reject strings that violated the *same-same-different* rule, regardless of their familiar syllables. Thus, these results, together, show that there was a higher tendency toward *category-based generalization* when the source rate of transmission was increased to a rate higher than *channel capacity*, even though the input entropy was the same in both conditions, which supports the predictions of our entropy model regarding the effect of the time-dependent variable of *channel capacity* on rule induction.

We did not find a significant main effect of any of the individual differences in explicit/implicit memory capacity or RAVENS, but they improved the model as covariates. A logical possible explanation under the hypotheses of our model could be that the effect of the source rate of information was increased to such a high extent (shown by the almost at ceiling overall performance in the Fast Rate condition) that individual cognitive abilities do not make any difference. Alternatively, these particular cognitive differences do not underlie the *channel capacity* relevant for linguistic rule induction.

These results show that, even with a low input of entropy (Radulescu et al., 2019), increasing the source rate of information transmission, while controlling for individual differences in explicit/implicit memory capacity and RAVENS, drives a change in the encoding method toward a more efficient encoding. As hypothesized, the same transition to a more efficient encoding method, from *item-bound* to *category-based generalization*, was obtained by either increasing the *input entropy* (H) in Radulescu et al. (2019) or reducing the time that the same input entropy is fed into the channel, i.e., by speeding up the source bit rate of information transmission.

EXPERIMENT 2

In Experiment 2, the participants carried out three tasks. In Task 1, the adults were exposed to an *aXb* language (Gómez, 2002; Grama et al., 2016) where they had to learn item-bound dependencies between *a* and *b* (*item-bound generalization*), while also generalizing *a_b* dependencies over a category of *X* words (*category-based generalization*). For example, they had to learn the item-bound dependency *tɛp_jɪk* and generalize it over new *X* elements (like *nilbo, perxɔn*): *tɛp_nilbo_jɪk*, *tɛp_perxɔn_jɪk*, etc.

We designed two experimental conditions: a slow source rate of information transmission (Slow Rate condition) and a fast source rate of information transmission (Fast Rate condition). As presented in section "Testing the Prediction of Speeding up the Source Bit Rate of Information Transmission," we first created two entropy versions of the grammar, with the same average rate of symbols/s (m_3), then we increased the average rate of symbols/s (m_4), in order to reach the same source rate of information transmission of the high entropy version while, crucially, keeping the input entropy low.

Unlike Gómez (2002), we kept *X* set size constantly high (18 *Xs*) and manipulated entropy by combining each of the three *a_b* frames with different subsets of 6 *Xs* (3 *a_b** 6 *Xs*), which generated a rather low entropy grammar version ($H_L = 3.52$ bits/symbol). For the high entropy condition, the *aXb* grammar combined exhaustively each of the three *a_b* frames with all the 18 *Xs* (three *a_b** 18 *Xs*), which resulted in a rather high entropy ($H_H = 4.7$ bits/symbol). Since such evaluations of low/high entropy could be seen as relative, depending on the grammar/language, we took into account previous studies on nonadjacent dependency learning (Gómez, 2002; Grama et al., 2016; Radulescu and Grama, 2020 unpublished data) in order to estimate the set size and variability necessary to achieve a low and a high entropy version. For entropy calculations, we used the same method as in Radulescu et al. (2019), see **Table 2** for complete entropy calculations.

In the Slow Rate condition, we used the low entropy version as presented above $H_L = 3.52$b/symbol. In the Fast Rate condition, the same stimuli were used ($H_L = 3.52$b/symbol), but the source rate of information was sped up by a factor of ($H_H/H_L = 4.71/3.52 =$) 1.34 (as per calculations in section "Testing the Prediction of Speeding up the Source Bit Rate of Information Transmission").

In the test phase, the participants were asked to give grammaticality judgments on aXb strings with either correct (familiar) or incorrect (unfamiliar) *a_b* frames. Whereas familiar *a_b* frames where the same as presented during familiarization ($a_i\_b_i$, where a_i predicted b_i with 100% probability), unfamiliar *a_b* frames consisted of combinations between familiar a and b elements that were mismatched ($a_i\_b_j$, where *a* predicted another *b*). Importantly, all test strings (correct and incorrect) included new X elements that were not present in the familiarization, since we aimed at testing for generalization of non-adjacencies to new intervening elements.

Recall that, according to our entropy model, rule induction is a *phased* mechanism that moves from the first phase of *item-bound generalization* to the next-level phase of *category-based generalization* as a function of the interaction between the input entropy and *channel capacity*. Learning *aXb* strings requires both

TABLE 2 | Entropy values for the two entropy versions of the *aXb* grammar.

Low entropy	**High entropy**
H[begin-*a*] = H[3] = −Σ[0.333*log0.333] = 1.58	H[begin-*a*] = H[3] = −Σ[0.333*log0.333] = 1.58
H[aX] = H[18] = 4.17	H[aX] = H[54] = 5.75
H[Xb] = H[18] = 4.17	H[Xb] = H[54] = 5.75
H[*b*-end] = H[3] = 1.58	H[*b*-end] = H[3] = 1.58
H[begin-aX] = H[18] = 4.17	H[begin-aX] = H[54] = 5.75
H[aXb] = H[Xb-end] = H[18] = 4.17	H[aXb] = H[Xb-end] = H[54] = 5.75
H[bigram] = 2.86	H[bigram] = 3.67
H[trigram] = 4.17	H[trigram] = 5.75
H[total] = $\frac{H[\text{bigram}]+H[\text{trigram}]}{2}$ = 3.52	H[total] = $\frac{H[\text{bigram}]+H[\text{trigram}]}{2}$ = 4.71

item-bound generalization of the *a_b* frames simultaneously with *category-based generalization* of these frames over a category of *X* elements. In this case, if the sped-up source rate of information transmission drives faster transition to *category-based generalization*, the item-bound encoding mechanism for the specific *a_b* dependencies might be phased out, and the encoding method might move to *category-based generalization* for the *a/b* elements as well, not only for the *X* category. Specifically, learners might encode the *a/b* elements as categories, which do not restrict to a specific $a_i\_b_i$ dependency. That is, learners might not encode an $a_i\_b_i$ relationship, but a relationship between a category of *a* elements and a category of *b* elements, which also allows for an $a_i\_b_j$ dependency to be legit ("class-words," Endress and Bonatti, 2007). To sum up, the predictions for this task could be opposite for the two types of relationships encoded in such an *aXb* grammar: increasing the source rate of information transmission impedes *item-bound generalization* (of the specific $a_i\_b_i$ relationship), but it facilitates *category-based generalization* (i.e., generalizing a relationship between a/*b* categories over a category of *Xs*).

The second task that the participants had to complete was RAVENS Standard Progressive Matrices (Raven et al., 2000). According to the hypotheses of our entropy model, we predicted a positive effect of RAVENS on the tendency to move from *item-bound* to *category-based generalization*.

In the third task, the participants completed a word-recall task, designed to test item memorization, i.e., detailed phonological representations of the *a*, *b* and *X* elements, in order to test for a correlation between learners' representations of specific items and their accuracy scores. We expected accurate memorization of the *a*/*b* elements to support better learning of the *a_b* dependencies and, thus, better accuracy scores. Conversely, failing to recall *X*s would indicate better generalization of the *X* category, hence better scores.

Participants

The same 56 participants from Experiment 1 participated in Experiment 2. We tested one more participant in Experiment 2 (as Experiment 2 was conducted before Experiment 1, one participant did not return to participate in Experiment 1). Therefore, in total, 57 adults participated in Experiment 2 (10 males, age range 18–72, $M_{age} = 26.28$, $SD_{age} = 11$) and received €10.

Materials

Task 1: aXb Grammar Learning

Familiarization Stimuli

All the *a* and *b* elements were monosyllabic nonsense words (e.g., *tɛp, jɪk*), while all the *X* elements were bisyllabic nonsense words (e.g., *naspu, dyfo:*), based on Grama et al. (2016). Each *a_b* pair was combined with a different, non-overlapping set of six *X* elements (see **Supplementary Appendix B** for the complete stimulus set). In both Slow Rate and Fast Rate conditions, two versions of the *aXb* language were used: Language 1 (L1) and Language 2 (L2). The only difference between L1 and L2 was the specific legit combination of the three *a* and *b* elements into pairs: *tɛp _lœt, sɔt_ jɪk,* and *rak_tuf* (L1), and *tɛp _ jɪk, sɔt_tuf,* and *rak_lœt* (L2). Therefore, every a_i $\_b_i$ pair in L1 was ungrammatical ($a_i\_b_j$) in L2, and vice versa. We used two different versions to prevent an effect of idiosyncrasies of particular *a_b* combinations (L1 or L2). Therefore, each version of the *aXb* grammar (L1 and L2) consisted of (3 $a_i\_b_i$ * 6 X_i =) 18 different $a_iX_ib_i$ strings. Each participant listened to only one version of the *aXb* grammar (either L1 or L2), and to only one source rate of transmission condition (either Slow Rate or Fast Rate).

The 18 different $a_iX_ib_i$ strings were presented 12 times, resulting in a total of 216 strings, in a randomized order for each participant. In the Slow Rate condition, there was a 100-ms within-string pause, and a 750-ms between-string pause. In the Fast Rate condition, all the *a, b,* and *X* elements, as well as the within-string and between-string pauses for each *aXb* string, were sped up by a factor of 1.34 (see section "Testing the Prediction of Speeding up the Source Bit Rate of Information Transmission") using Praat (Boersma and Weenink, 2019). The duration of each *a, b,* and *X* word was shortened separately by the 1.34 factor, and then the elements were spliced into the specific *aXb* strings.

Test Stimuli

Each *a_b* frame of each language (L1 and L2) was combined with two novel *X* elements to yield (6 *a_b* * 2 *X*=) 12 new test items (see **Supplementary Appendix B**). Each participant listened to 12 new *aXb* strings: six grammatical and six ungrammatical. The six new *aXb* strings that contained the L1 *a_b* pairs were counted as ungrammatical for the L2 learners, while the six new *aXb* strings with the L2 *a_b* pairs were ungrammatical for the L1 learners. Accuracy scores for learning the *aXb* grammar were calculated as correct acceptances of the grammatical strings and correct rejections of the ungrammatical strings.

Task 2: RAVENS

The second task was Raven's Standard Progressive Matrices (Raven et al., 2000), for which the participants had to solve 60 matrices by identifying which pattern is missing in a multiple choice task. Each matrix consists of a set of nine patterns, of which one is missing, arranged in a particular order according to some underlying rules. The standard RAVENS allows 50 min for completion, but after a pilot, we allowed the participants only 35 min, to avoid a time-consuming and exhausting experiment session. We used the standard scoring method: we counted all correct answers, and then we used the standard tables to transform them into age-corrected percentiles.

Task 3: Word Recall Task

The Word Recall task had two tests. In the first test, the participants were presented visually with 12 familiar two-syllable *X* words from the *aXb* language, and 12 new bisyllabic foils, similar to the familiar ones, which overlapped in one syllable with the target words. The second test presented the participants visually with six monosyllabic familiar *a* or *b* elements of the *aXb* language, and six new nonsense word foils, which differed from the target words only by one letter (see **Supplementary Appendix C** for stimulus set). The participants had to indicate for each word whether they heard it during the first task. Accuracy scores were measured as correct acceptances of the familiar items and correct rejections of the foils.

Procedure

Before the familiarization phase of Task 1, the participants were instructed that they would listen to an "alien language" that does not resemble any language that they might be familiar with, and that the language has its own rules and grammar. To avoid any motivation to explicitly look for patterns in the stimuli, the participants were not informed of the subsequent test phase until after the end of the familiarization phase. Before the test phase, the participants were instructed that they would listen to new sentences in the same "alien language," and that none would be identical to the sentences they had heard before. They were then asked to decide for each sentence whether it was correct or not, according to the grammar of the language they had just heard, by clicking on "Yes" or "No." They were instructed to answer quickly and intuitively. Afterward, the other tasks were administered in the order stated above. Experiment 2 lasted approximately 1 h.

Results

Table 3 shows the means and standard deviations of accuracy scores (proportion correct responses) for both conditions (Slow Rate vs. Fast Rate).

Figure 4 shows a bimodal distribution of individual accuracy scores in the Slow Rate condition: this shows that most of the participants either performed around chance level or achieved a very high accuracy score. **Figure 5** shows most of the participants in the Fast Rate condition performed between 40 and 60%.

Because the data were not normally distributed, a nonparametric statistical test, a two-tailed one-sample Wilcoxon signed-rank test, was conducted to assess whether response rates were significantly different from chance. The accuracy score of Fast-Rate learners ($M = 0.55$, $SD = 0.5$) was significantly different from chance at the 0.05 level of significance, with a moderate effect size ($p = 0.017$, 95% CI for mean difference 0.5 to 0.63, $r = 0.45$). The accuracy score of Slow-Rate learners ($M = 0.69$, $SD = 0.46$) was significantly different from chance at the 0.05 level of significance, with a large effect size ($p < 0.001$, 95% CI for mean difference 0.67 to 0.83, $r = 0.73$).

To compare performance across the two conditions, we used R (R Core Team, 2017) and the lmerTest package (Kuznetsova et al., 2017) to perform a general linear mixed effects analysis of the relationship between Accuracy (correct acceptance of grammatical test strings and correct rejection of ungrammatical test strings) and Rate of Transmission (Slow Rate, Fast Rate). As a dependent variable, we entered Accuracy in the model, and as fixed effects we entered Rate of Transmission (Slow Rate, Fast Rate) and Language (L1, L2), without interaction term. As random effects we had intercepts for subjects[6]. An alpha level of 0.05 was used for all the statistical tests. We started fitting the data from the intercept-only model and added the random and fixed factors one by one. The model reported here is the best fitting model, both in terms of model accuracy in predicting the observed data and Akaike Information Criterion. Likelihood Ratio Tests were performed separately as a means to attain p-value for the effect of each predictor (Rate of Transmission, Language).

TABLE 3 | Descriptive statistics of mean correct score in two conditions of exposure. Experiment 2.

Condition	*M*	*SD*	*n*	*SE*	*95% CI for Mean Difference*
Slow rate	0.69	0.46	29	0.09	0.51, 0.87
Fast rate	0.55	0.50	28	0.09	0.37, 0.74

A significant main effect of Rate of Transmission [$\chi^2(1) = 8.43$, $p = 0.003$, conditional $R^2 = 0.1$] on Accuracy was found, indicating that the participants in the Fast Rate condition had significantly lower Accuracy scores as compared with the participants in the Slow Rate condition. Language was not a significant predictor [$\chi^2(1) = 3.2$, $p = 0.07$, conditional $R^2 = 0.09$]. Finally, we ran an additional model that included the interaction between Rate of Transmission and Language (although this was not the best fitting model, we wanted to verify that our specific stimuli did not prompt different performance). No significant interaction effect was found between Rate of Transmission and Language [$\chi^2(1) = 0.14$, $p = 0.7$, conditional $R^2 = 0.1$]. The scores of individual differences tests (Forward Digit Span, Incidental Memorization Test, Raven's Progressive Matrices, Word Recall Test) were added to this model as fixed factors, one by one. However, the only one that improved the model was the accuracy score in the Word Recall Test for *a*/*b* (but not *X*) elements of the *aXb* grammar, and it also had a significant positive effect on the Accuracy scores [$\chi^2(1) = 3.8$, $p = 0.05$, conditional $R^2 = 0.1$].

Discussion

In Experiment 2, we tested the effect of speeding up the source rate of transmission on learning a complex *aXb* grammar, which required both *item-bound generalization* of the specific *a_b* dependencies and *category-based generalization* in order to generalize those dependencies over a category of intervening *X* elements. According to our entropy model, our predictions for this experiment were opposite for the two types of relationships encoded in an a_iXb_i grammar: increasing the source rate of information transmission impedes *item-bound generalization* (of the specific $a_i\_b_i$ relationship), but it facilitates *category-based generalization* (i.e., generalizing a relationship between *a* and *b* categories over a category of *Xs*). The results showed that there was indeed a significant effect of increasing the source rate of transmission on learning the *aXb* grammar, such that the Fast Rate group scored lower than the Slow Rate group. This shows that increasing the source rate of transmission by a factor of 1.34 in this particular a_iXb_i grammar with an entropy of 3.52 bits/symbol makes learning of the specific

[6] Due to convergence issues, random intercepts for items were excluded due to convergence issues [their estimated variance was zero, they did not improve the model and their effect was insignificant – $\chi^2(1) = 0$, $p = 1$)].

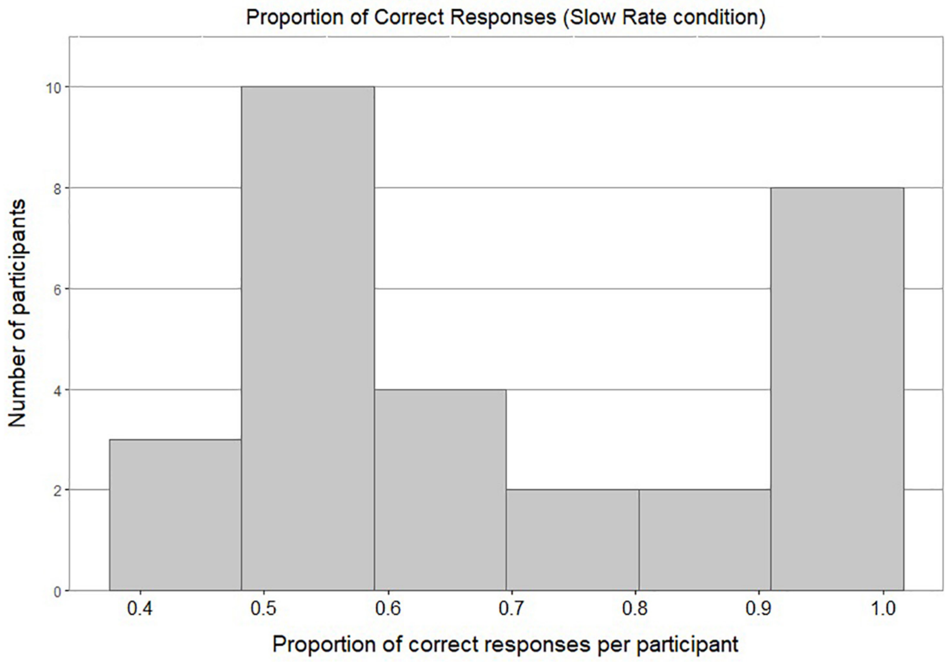


FIGURE 4 | Histogram of proportion of correct responses per participant in Slow Rate Condition.

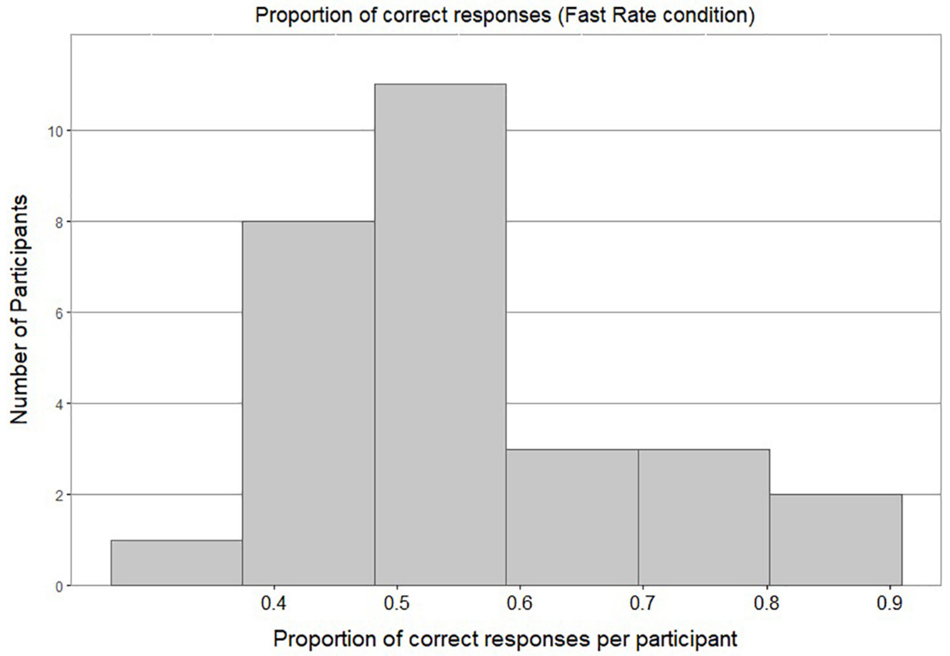


FIGURE 5 | Histogram of proportion of correct responses per participant in Fast Rate Condition.

$a_i\_b_i$ frames and generalizing them over novel intervening X elements more difficult than a slower rate of transmission. Moreover, participants who recalled the *a/b* elements better across conditions learned the specific $a_i\_b_i$ frames better. Thus, the learning of a_iXb_i grammar is correlated with item-specific encoding of the *a/b* elements. All these results, taken together, support the predictions of our entropy model, namely, that an increased source rate of information transmission impedes *item-bound generalization* (of the specific $a_i\_b_i$ relationship).

As we argued above, if learners correctly accept new *aXb* strings with the specific familiar $a_i\_b_i$ dependencies and new *X* elements, it shows they were both able to encode *item-bound generalizations* ($a_i\_b_i$ frames), and to generalize them over a category of *X* elements, i.e., *category-based generalization*. This is what happened both in the Slow Rate and Fast Rate conditions. However, the Fast Rate group had a lower tendency to do so compared with the Slow Rate group. There could be several logical interpretations: Fast-Rate learners failed at *category-based generalization* of the *Xs,* they failed at *item-bound generalization* of the $a_i\_b_i$ frames, or they were simply confused. Therefore, we looked into the acceptance/rejection ratios. If the first case was true, rejection rates should be higher than acceptance rates, since all the test items had new *Xs*. This was not the case. Actually, the Fast-Rate learners show similarly high acceptance rates for both language-specific a_iXb_i strings (specific to the exposure language, e.g., L1) and language-deviant a_iXb_j strings (specific to the other language, e.g., L2), with a rather high acceptance rate for the language-deviant a_iXb_j strings (Median = 0.58) compared with the Slow-Rate learners (median = 0.33) (**Figures 6**, **7**). This points to the fact that the Fast-Rate learners failed to learn the specific $a_i\_b_i$ dependencies, that is, *item-bound generalization* was impaired in the Fast Rate group.

If this was the case, this result can be accounted for by our entropy model: as we argued in section "Experiment 2", a sped up source rate of information transmission precipitates the transition to *category-based generalization* faster, such that the item-bound encoding mechanism for the specific $a_i\_b_i$ frames might be phased out, and the encoding method moves to *category-based generalization* for the $a_i\_b_i$ frames as well. This would be a case of overgeneralization: categories of the *a*/*b* elements would be inferred (i.e., *category-based generalization*), not just the item-bound specific $a_i\_b_i$ frames, so any *a* could freely combine with any *b*, such that the $a_i\_b_j$ frames would also be accepted ("class-words"). Since all the test items show new combinations with *X* elements, the learner might find it highly probable that the *a/b* elements could yield new combinations, as long as they preserve the main *aXb* order and word characteristics (i.e., monosyllabic *a* followed by a bisyllabic *X* and then a monosyllabic *b*).

Following this logic, if the Fast-Rate learners actually overgeneralized, they must have started the test by accepting both language-specific and language-deviant *aXb* strings, and after the first acceptances they would question why all the test items seem to be acceptable, which might have led to an increased rate of rejections in the last part of the test. Alternatively, if the Fast-Rate learners were just confused, the acceptances should be randomly scattered over test trials.

An inspection of the acceptance rate of both language-specific and language-deviant *aXb* strings, in the Fast Rate condition, showed a higher tendency to accept all the test strings in the first three trials of the test [$t(11) = -1.951$, $p = 0.05$], regardless of exposure language, than in the last trials. These results might point to a case of overgeneralization in the Fast Rate condition.

Thus, it is possible that the source rate of information transmission was increased to an extent higher than required to actually learn the a_iXb_i grammar, and that it led to overgeneralization. Further research should specifically test the overgeneralization hypothesis, and look further into the effect of sped-up source rate of information transmission at a lower rate, i.e., a speeding up factor $m < 1.34$, to find the adequate source rate of transmission for learning this complex grammar.

GENERAL DISCUSSION AND CONCLUSION

This article contributes to the ongoing research on the underlying mechanisms and factors that drive both *item-bound generalization* and *category-based generalization* by extending further the entropy model for rule induction that we proposed in Radulescu et al. (2019). Our entropy model offers a more refined formal approach to the classical *Less-is-More* hypothesis (Newport, 1990) and takes a step further by bringing together two factors in one information-theoretic account based on Shannon's noisy-channel coding theory (Shannon, 1948). Specifically, our model hypothesizes that an increase in the source *input entropy per second* to a rate higher than the time-sensitive encoding capacity of our brain, *channel capacity,* drives the transition from *item-bound* to *category-based generalization*. In Radulescu et al. (2019), in two artificial grammar experiments, we found evidence that an increase in input entropy gradually shapes *item-bound generalization* into *category-based generalization*. Hence, our model specifically predicts that it is not high entropy in absolute terms that is the factor at stake in this mechanism. Rather, our finite entropy-processing *channel capacity,* places an upper bound on the amount of entropy per second, which drives the self-organization of information from an encoding method to another, in line with Dynamic Systems Theory (Stephen et al., 2009).

In two artificial grammar experiments, an XXY grammar and a more complex aXb grammar, we sped up the source rate of information transmission to tax *channel capacity*, which was hypothesized to drive the transition from *item-bound* to *category-based generalization*. Learning an XXY grammar requires abstracting away from specific items of the X and Y categories, to move from *item-bound* to *category-based generalization,* that is, to learn the *same-same-different* rule between categories, regardless of specific items. The results showed that this transition was driven by an increase in the source rate of information transmission, i.e., *input entropy per second*, while the statistical properties of the input, i.e., *input entropy per symbol,* remained constant at a low level, which did not support the generalization in Radulescu et al. (2019). Crucially, as hypothesized by our entropy model, moving from *item-bound* to *category-based generalization* was driven by either increasing the *input entropy* (H) in Radulescu et al. (2019) or increasing the time that the same input entropy enters the channel, thus, taxing the *channel capacity* in this study.

Learning an *aXb* grammar requires moving from *item-bound* to *category-based generalization* for the category of middle *Xs*, while, crucially, sticking to *item-bound generalization* for the specific *a_b* dependencies. If increased source rate of information transmission drives *category-based generalization* for

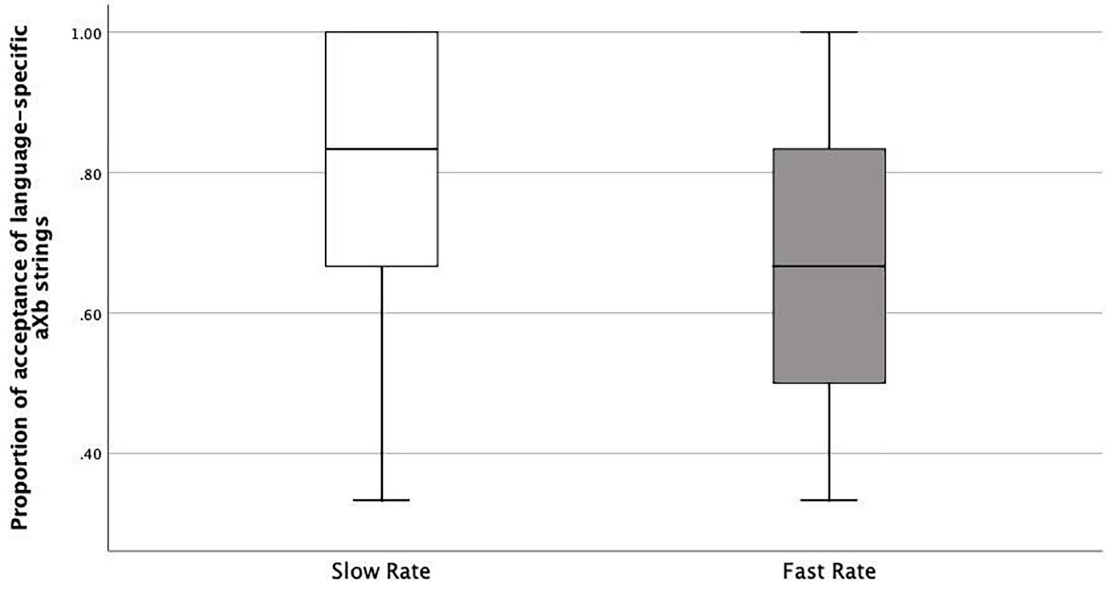


FIGURE 6 | Boxplots of proportions of acceptance of language-specific aXb strings in Slow Rate Condition as compared to Fast Rate Condition.

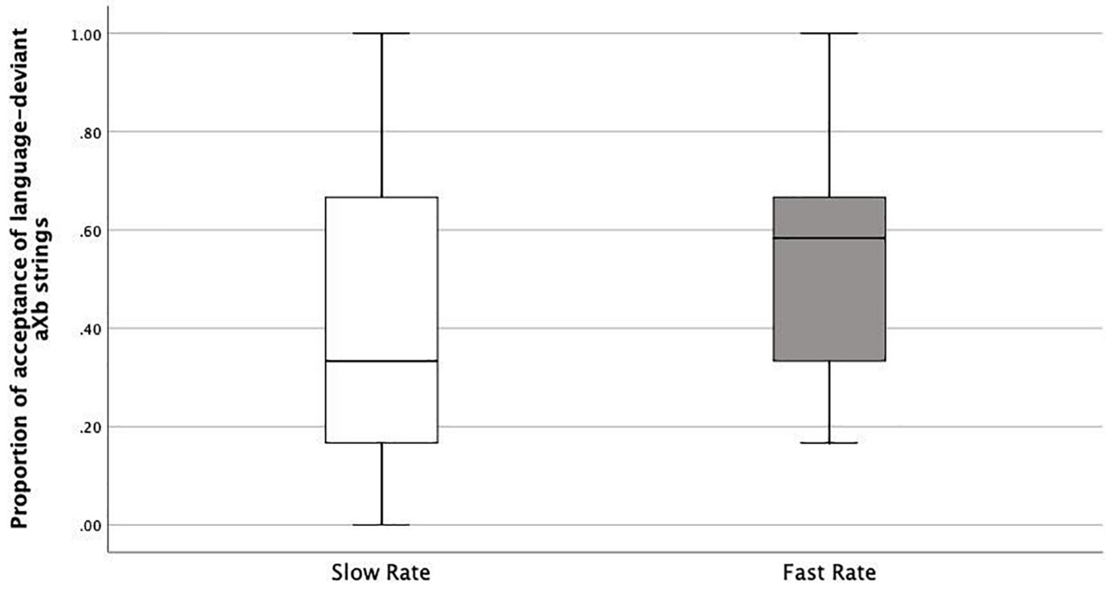


FIGURE 7 | Boxplots of proportions of acceptance of language-deviant aXb strings in Slow Rate Condition as compared to Fast Rate Condition.

the *X* category, it follows that it should phase out *item-bound generalization* for the specific *a_b* dependencies. Indeed, the results showed that faster source rate of information caused lower accuracy than slower source rate of information on this grammar. As per our model, one logical interpretation of these results would be that the source rate of transmission was too high for this type of grammar with the specific input entropy that we tested (3.52 bits), and that it precipitated the transition to *category-based generalization* for the specific *a_b* dependencies as well, not only for the *X* elements. This points to a possible overgeneralization, where learners might have learned an *AXB* grammar, where A and B stand for categories instead of *item-bound* relationships between specific *a*/*b* elements. Indeed, it is possible that for this type of grammar fast, but not furious, might yield better learning. Future research should look into a slower rate of transmission for an *aXb* grammar with this specific entropy (3.52 bits).

Altogether, these results show that, as hypothesized by our entropy model, rule induction is an encoding mechanism that moves from *item-bound* to *category-based generalization* driven by the interaction between the *input entropy* and the finite *channel capacity*. Future research should look into the exact mathematical relationship between input entropy and rate of transmission, by also considering the other variable of *channel capacity*, i.e., the rate of equivocation caused by noise interference, in order to calculate an estimation of the *channel capacity* for rule induction.

Although having used other methods than information-theoretic approaches to investigate the effect of a time-dependent variable on category learning (Reeder et al., 2009, 2013), on non-adjacent dependency learning (Endress and Bonatti, 2007; Wang et al., 2016, 2019) and on auditory statistical learning (Emberson et al., 2011), converging evidence from these studies highlights a clear pattern: generally a shorter time is beneficial to auditory rule (category) learning. This hypothesis is also supported by evidence from neural network research showing that reduced training time leads to lower generalization error (Hardt et al., 2016). Our study contributes to this research topic by taking a step further: it applies a purely information-theoretic measure directly derived from Shannon's noisy-channel coding theory and based on the quantified amount of input entropy per second.

Our model is compatible with another information-theoretic hypothesis derived from Shannon's noisy-channel coding theory: the hypothesis of Uniform Information Density (Jaeger, 2006, 2010; Levy and Jaeger, 2007). Although proposed in a different domain of application, this hypothesis proposes that in language production speakers prefer (intuitively) to encode their message by a uniform distribution of information across the signal, with a rate of information transfer close to the channel capacity, but without exceeding it. In other words, language production is inherently a mechanism designed for efficient communication, in that it balances the amount of information per time or signal (dubbed "information density"), such that the channel is never under- or overutilized (Jaeger, 2010). Underutilization means a waste of channel, while overutilization risks information loss, as per Shannon's noisy-channel coding theory, hence, as per the Uniform Information Density. By posing the noisy-channel capacity as an upper bound of the rate of information transmission for the purpose of efficient transmission without information loss, our model accounts for the Uniform Information Density hypothesis, and takes a step further by offering a more general domain of application (i.e., learning and generalization).

At the algorithmic level (in the sense of Marr, 1982), our entropy and channel capacity model for rule induction in artificial grammar is compatible with recent models of recognition memory (Cox and Shiffrin, 2017) and exemplar models applied to artificial grammar learning (Jamieson and Mewhort, 2010). Future research should look into the link between our entropy model and these formal approaches based on encoding instances as vectors of features, with generalization being triggered by vector similarity (Chubala and Jamieson, 2013). Indeed, as we argued in Radulescu et al. (2019), by refining the feature similarity approach to the category formation proposed by Aslin and Newport (2012, 2014), our entropy model suggests that information is re-structured from *item-bound* to *category-based generalization* by (unconsciously) re-observing the structural properties of the input and identifying similarities (shared features) and specific differences (unshared features) between items. Crucially, our model proposes *channel capacity* as the upper bound on the amount of similarities/differences encoded. The degree of specificity of the encoding (i.e., *item-bound* specificity) is given by the amount of differences encoded with specific items, which results from a lower or higher *input entropy* (measured in bits of information): the more differences are encoded (higher *input entropy*), the higher the degree of specificity of the encoding (i.e., *item-bound generalization*). Conversely, when the degree of specificity of the encoding reaches the upper bound placed by *channel capacity* on the number of bits encoded per second, a reduction or "gradual forgetting" of the encoded differences is triggered in order to avoid an inefficient, i.e., noisy, encoding (Radulescu et al., 2019). Hence, more and more similarities between items are highlighted, which drives an automatic gradual grouping of items under the same "bucket." Hence, the degree of specificity decreases and the degree of generality increases *gradually* with each bit of information. Thus, a gradient of specificity/generality on a continuum from *item-bound* to *category-based generalizations* can be envisaged in terms of number of bits of information encoded in the representation (analogous to the degree of stability/plasticity in terms of strength of memory pathways in neural networks, Abraham and Robins, 2005).

A follow-up topic would be to better define and specify *channel,* be it a communication channel between speakers or an abstract channel as we mostly hinted in this study: an abstract channel between an abstract source, a grammar, and a learner. However, we would briefly suggest a more in-depth and granular understanding of the abstract concept of *channel* as a system of channels: intuitively, and oversimplifying here, the acoustic signal from the environment enters the acoustic channel of a learner, which has a specific rate of information transmission, then the output of this channel becomes the input to the perception channel, whose output becomes the input to the cognitive channel. Estimates of the bit rate of information processing by applying information theory were proposed in some perception and cognitive domains, e.g., in visual attention (Verghese and Pelli, 1992), visual processing (Koch et al., 2006), unconscious vs. conscious processing (Dijksterhuis and Nordgren, 2006), and cognitive control (Wu et al., 2016). However, we suggest that the concept of *channel* should be first and foremost defined and specified in physical and biological terms (i.e., at the level of brain structure and neural networks), and further investigated in terms of its link to the cognitive capacities (at the algorithmic level). That would mean further investigating and applying Shannon's *channel* and *noisy-channel coding theory* to recent developments in neurobiology, where it was shown that artificially induced forgetting at the cellular level drives generalization (Migues et al., 2016). Moreover, since information is physical (Laughlin et al., 1998; Machta, 1999; Karnani et al., 2009), further research should look into the information-theoretic concept of *channel* and *rate of information transmission* at the level of neural networks. Neural networks are the physical/biological medium (i.e., channel) transmitting one form of information (acoustic energy) to the brain that is transcoded into another form of information (i.e., neuronal energy, patterns of electric activity at the neuronal level). Physical bioprocesses of energy transformation from acoustic information into electric signal and transmission through neural networks were proposed to underlie abstract memory representations (Varpula et al., 2013).

Before concluding, it is imperative to clarify one aspect. A model of *finite* and *noisy*-channel capacity might lead the reader to assume a kind of a cognitive limitation as in a flaw of the cognitive system, which is definitely not the case. We do not propose a model in which the emergence of rules and categories, i.e., structure, is merely the side effect of some constraints of a limited biological system. In accordance with innovative theories and findings in neurobiology (Frankland et al., 2013; Hardt et al., 2013; Migues et al., 2016; Richards and Frankland, 2017), we deem our *finite* and *noisy*-channel capacity to be a design feature of our biological system for adaptive purposes. More precisely, neurobiological evidence shows that our memory system is designed to encode memories not as in-detail representations of the past, but as simplified models better suited for future generalization in noisy environments (Richards and Frankland, 2017). The brain employs several strategies to undermine faithful in-detail representations to prevent overfitting to past events (in accordance with neural networks research (MacKay, 2003; Hawkins, 2004), which promotes better generalization (among which is *noise* injection, a neurobiological mechanism that increases random variability in the synaptic connections, Villarreal et al., 2002).

Fast but not furious, reads the title of this article. Speed up, but not wildly and in an unrestrained fashion. The channel capacity acts as a speedometer, and determines the maximum rate of information transmission with adequate encoding. In this study, we proposed an innovative method to increase the rate of information to tax *channel capacity*. We found that increasing the rate of transmission with a specific factor calculated by applying Shannon's formula to experimentally obtained data indeed has the hypothesized effect on rule learning: it drives *category-based generalization,* and it interferes with *item-bound generalization*. Thus, we deem it necessary to specify that by sped-up bit rate we do not mean that an unrestrained increased bit rate, in absolute terms, up to very high bit rates, drives rule induction in any context, or grammar. In other words, the very specific dynamics between the *input entropy* and the maximum *rate of information transmission* drive rule induction. Further research should investigate this sweet spot and find the mathematical relationship between these two factors.

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 Ethics Assessment Committee Linguistics (ETCL), Utrecht Institute of Linguistics OTS. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SR developed the entropy model and the idea for the study (with input from SA and FW). SR and IG designed the experiments. SR, IG, and AK created the materials. AK recruited and tested the participants. AK and SR analyzed the data. AK wrote a shorter preliminary draft as her internship report. SR wrote the publishable manuscript with input from IG, FW, SA, and AK. All the authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Netherlands Organization for Scientific Research (NWO) (322-75-009/1116).

ACKNOWLEDGMENTS

We would like to thank the editor and the reviewers for their valuable comments on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.661785/full#supplementary-material

REFERENCES

Abraham, W. C., and Robins, A. (2005). Memory retention–the synaptic stability versus plasticity dilemma. *Trends Neurosci.* 28, 73–78. doi: 10.1016/j.tins.2004.12.003

Aslin, R. N., and Newport, E. L. (2012). Statistical learning: From acquiring specific items to forming general rules. *Curr. Direct. Psychol. Sci.* 21, 170–176. doi: 10.1177/0963721412436806

Aslin, R. N., and Newport, E. L. (2014). Distributional Language Learning: Mechanisms and Models of ategory Formation. *Lang. Learn.* 64, 86–105. doi: 10.1111/lang.12074

Baddeley, A., Eysenck, M. W., and Anderson, M. C. (2015). *Memory*, 2nd Edn. United Kingdom: Psychology Press.

Boersma, P., and Weenink, D. (2019). *Praat: Doing Phonetics by Computer [Computer program]. Version 6.0.49*. Available online at: http://www.praat.org/,

Burgoyne, A. P., Hambrick, D. Z., and Altmann, E. M. (2019). Is working memory capacity a causal factor in fluid intelligence? *Psychon. Bull. Rev.* 26, 1333–1339. doi: 10.3758/s13423-019-01606-9

Carpenter, P. A., Just, M. A., and Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychol. Rev.* 97, 404–431. doi: 10.1037/0033-295X.97.3.404

Chubala, C. M., and Jamieson, R. K. (2013). Recoding and representation in artificial grammar learning. *Behav. Res. Methods* 45, 470–479. doi: 10.3758/s13428-012-0253-6

Conway, A. R., Kane, M. J., and Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends Cogn. Sci.* 7, 547–552. doi: 10.1016/j.tics.2003.10.005

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., and Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence* 30, 163–184. doi: 10.1016/S0160-2896(01)00096-4

Cowan, N. (2005). *Essays in Cognitive Psychology.Working Memory Capacity*. United Kingdom: Psychology Press, doi: 10.4324/9780203342398

Cox, G. E., and Shiffrin, R. M. (2017). A dynamic approach to recognition memory. *Psychol. Rev.* 124, 795–860. doi: 10.1037/rev0000076

Dehn, M. J. (2017). How working memory enables fluid reasoning. *Appl. Neuropsychol.* 6, 245–247. doi: 10.1080/21622965.2017.1317490

Dijksterhuis, A., and Nordgren, L. F. (2006). A Theory of Unconscious Thought. *Perspect. Psychol. Sci.* 1, 95–109. doi: 10.1111/j.1745-6916.2006.00007.x

Emberson, L. L., Conway, C. M., and Christiansen, M. H. (2011). Timing is everything: changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. *Q. J. Exp. Psychol.* 64, 1021–1040. doi: 10.1080/17470218.2010.538972

Endress, A. D., and Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition* 105, 247–299. doi: 10.1016/j.cognition.2006.09.010

Ferdinand, V. (2015). *Inductive Evolution: Cognition, Culture, and Regularity in Language*. Ph D thesis, Edinburgh: University of Edinburgh.

Ferdinand, V., Kirby, S., and Smith, K. (2019). The cognitive roots of regularization in language. *Cognition* 184, 53–68. doi: 10.1016/j.cognition.2018.12.002

Frankland, P. W., Köhler, S., and Josselyn, S. A. (2013). Hippocampal neurogenesis and forgetting. *Trends Neurosci.* 36, 497–503. doi: 10.1016/j.tins.2013.05.002

Frost, R. L., and Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition* 147, 70–74. doi: 10.1016/j.cognition.2015.11.010

Gerken, L. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition* 98, B67–B74. doi: 10.1016/j.cognition.2005.03.003

Gómez, R.L. (2002). Variability and detection of invariant structure. *Psychol. Sci.* 13, 431–436. doi: 10.1111/1467-9280.00476

Gómez, R. L., and Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends Cogn. Sci.* 4, 178–186. doi: 10.1016/s1364-6613(00)01467-4

Grama, I. C., Kerkhoff, A., and Wijnen, F. (2016). Gleaning structure from sound: The role of prosodic contrast in learning non-adjacent dependencies. *J. Psychol. Res.* 45, 1427–1449. doi: 10.1007/s10936-016-9412-8

Hardt, M., Recht, B., and Singer, Y. (2016). "Train faster, generalize better: Stability of stochastic gradient decent," in *Proceedings of the 33rd International Conference on Machine Learning*, (New York, NY: ICML).

Hardt, O., Nader, K., and Wang, Y. T. (2013). GluA2-dependent AMPA receptor endocytosis and the decay of early and late long-term potentiation: possible mechanisms for forgetting of short- and long-term memories. *Philosoph. Transact. Roy. Soc. London Ser B* 369:20130141. doi: 10.1098/rstb.2013.0141

Hawkins, D. M. (2004). The problem of overfitting. *J. Chem. Inf. Comput. Sci.* 44, 1–12.

Hudson Kam, C. (2019). Reconsidering retrieval effects on adult regularization of inconsistent variation in language. *Lang. Learn. Devel.* 15, 317–337. doi: 10.1080/15475441.2019.1634575

Hudson Kam, C. L., and Chang, A. (2009). Investigating the cause of language regularization in adults: Memory constraints or learning effects? *J. Exp. Psychol.* 35, 815–821. doi: 10.1037/a0015097

Hudson Kam, C. L., and Newport, E. L. (2009). Getting it right by getting it wrong: when learners change languages. *Cogn. Psychol.* 59, 30–66. doi: 10.1016/j.cogpsych.2009.01.001

Hudson Kam, C. L. H., and Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Lang. Learn. Devel.* 1, 151–195. doi: 10.1207/s15473341lld0102_3

Jaeger, T. F. (2006). *Redundancy and Syntactic Reduction*. Ph.D. thesis, California: Stanford University.

Jaeger, T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cogn. Psychol.* 61, 23–62. doi: 10.1016/j.cogpsych.2010.02.002

Jamieson, R. K., and Mewhort, D. J. (2010). Applying an exemplar model to the artificial-grammar task: String completion and performance on individual items. *Q. J. Exp. Psychol.* 63, 1014–1039. doi: 10.1080/17470210903267417

Kareev, Y., Lieberman, I., and Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *J. Exp. Psychol.* 126, 278–287. doi: 10.1037/0096-3445.126.3.278

Karnani, M., Pääkkönen, K., and Annila, A. (2009). The physical character of information. *Proc. R. Soc. A* 465, 2155–2175. doi: 10.1098/rspa.2009.0063

Koch, K., McLean, J., Segev, R., Freed, M. A., Berry, M. J. II, Balasubramanian, V., et al. (2006). How much the eye tells the brain. *CB* 16, 1428–1434. doi: 10.1016/j.cub.2006.05.056

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *J. Statist. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13

Laughlin, S. B., de Ruyter, van Steveninck, R. R., and Anderson, J. C. (1998). The metabolic cost of neural information. *Nat. Neurosci.* 1, 36–41. doi: 10.1038/236

Levy, R., and Jaeger, T. F. (2007). "Speakers optimize information density through syntactic reduction," in *Advances in Neural Information Processing Systems (NIPS)*, Vol. 19, eds B. Schlökopf, J. Platt, and T. Hoffman (Cambridge, MA: MIT Press), 849–856.

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *J. Exp. Psychol.* 37, 720–738. doi: 10.1037/a0022639

Little, D. R., Lewandowsky, S., and Craig, S. (2014). Working memory capacity and fluid abilities: The more difficult the item, the more more is better. *Front. Psychol.* 5:239. doi: 10.3389/fpsyg.2014.00239

Little, D. R., Lewandowsky, S., and Griffiths, T. L. (2012). A bayesian model of rule induction in raven's progressive matrices. *Proc. Ann. Meet. Cogn. Sci. Soc.* 34, 1918–1923. doi: 10.1.1.300.2871

Machta, J. (1999). Entropy, information, and computation. *Am. J. Phys.* 67:1074.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press.

Marcus, G. F., Vijayan, S., Bandi Rao, S., and Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science* 283, 77–80. doi: 10.1126/science.283.5398.77

Marr, D. (1982). *Vision: A Computational Approach*. San Francisco: Freeman and Co.

Migues, P. V., Liu, L., Archbold, G. E., Einarsson, E. Ö, Wong, J., Bonasia, K., et al. (2016). Blocking Synaptic Removal of GluA2-Containing AMPA Receptors Prevents the Natural Forgetting of Long-Term Memories. *J. Neurosci.* 36, 3481–3494. doi: 10.1523/JNEUROSCI.3333-15.2016

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63, 81–97. doi: 10.1037/h004315

Newport, E. L. (1990). Maturational constraints on language learning. *Cogn. Sci.* 14, 11–28. doi: 10.1207/s15516709cog1401_2

Newport, E. L. (2016). Statistical language learning: Computational, maturational, and linguistic constraints. *Lang. Cogn.* 8, 447–461. doi: 10.1017/langcog.2016.20

Oberauer, K., and Hein, L. (2012). Attention to information in working memory. *Curr. Direct. Psychol. Sci.* 21, 164–169. doi: 10.1177/0963721412444727

Onnis, L., Monaghan, P., Christiansen, M. H., and Chater, N. (2004). "Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies," in *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, (Mahwah, NJ: Lawrence Erlbaum).

Perfors, A. (2012). When do memory limitations lead to regularization? An experimental and computational investigation. *J. Mem. Lang.* 67, 486–506. doi: 10.1016/j.jml.2012.07.009

Perfors, A. (2016). Adult Regularization of Inconsistent Input Depends on Pragmatic Factors. *Lang. Learn. Devel.* 12, 138–155. doi: 10.1080/15475441.2015.1052449

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* Available online at: https://www.R-project.org/

Radulescu, S., Wijnen, F., and Avrutin, S. (2019). Patterns bit by bit. An entropy model for rule induction. *Lang. Learn. Devel.* 16, 109–140. doi: 10.1080/15475441.2019.1695620

Raven, J., Raven, J. C., and Court, J. H. (2000). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3: The Standard Progressive Matrices.* Oxford, UK: Oxford Psychologists Press.

Reeder, P. A., Newport, E. L., and Aslin, R. N. (2009). "The role of distributional information in linguistic category formation," in *Proceedings of the 31st Annual Conference of the Cognitive Science Society,* eds N. Taatgen and H. van Rijn (Austin, TX: Cognitive Science Society), 2564–2569.

Reeder, P. A., Newport, E. L., and Aslin, R. N. (2013). From shared contexts to syntactic categories: the role of distributional information in learning linguistic form-classes. *Cogn. Psychol.* 66, 30–54. doi: 10.1016/j.cogpsych.2012.09.001

Richards, B. A., and Frankland, P. W. (2017). The Persistence and Transience of Memory. *Neuron* 94, 1071–1084. doi: 10.1016/j.neuron.2017.04.037

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

Saldana, C., Smith, K., Kirby, S., and Culbertson, J. (2017). "Is the strength of regularisation behaviour uniform across linguistic levels?," in *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, eds G. Gunzelmann, A. Howes, T. Tenbrink, and E. J. Davelaar (New York, NY: Cognitive Science Society), 1023–1028.

Samara, A., Smith, K., Brown, H., and Wonnacott, E. (2017). Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. *Cogn. Psychol.* 94, 85–114. doi: 10.1016/j.cogpsych.2017.02.004

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Techn. J.* 27, 379–423. doi: 10.1002/bltj.1948.27.issue-3

Smith, K., and Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition* 116, 444–449. doi: 10.1016/j.cognition.2010.06.004

Stephen, D. G., Dixon, J. A., and Isenhower, R. W. (2009). Dynamics of representational change: Entropy, action, and cognition. *J. Exp. Psychol.* 35, 1811–1832. doi: 10.1037/a0014510

Thiessen, E. D., and Saffran, J. R. (2007). Learning to Learn: Infants' Acquisition of Stress-Based Strategies for Word Segmentation. *Lang. Learn. Devel.* 3, 73–100. doi: 10.1207/s15473341lld0301_3

Townsend, J. T., and Ashby, F. G. (1978). "Methods of modeling capacity in simple processing systems," in *Cognitive Theory*, Vol. III, eds J. Castellan and F. Restle (Hillsdale: Erlbaum), 200–239.

Townsend, J. T., and Eidels, A. (2011). Workload capacity spaces: a unified methodology for response time measures of efficiency as workload is varied. *Psychon. Bull. Rev.* 18, 659–681. doi: 10.3758/s13423-011-0106-9

Varpula, S., Annila, A., and Beck, C. (2013). Thoughts about thinking: cognition according to the second law of thermodynamics. *Adv. Stud. Biol.* 5, 135–149.

Verghese, P., and Pelli, D. G. (1992). The information capacity of visual attention. *Vis. Res.* 32, 983–995. doi: 10.1016/0042-6989(92)90040-P

Villarreal, D. M., Do, V., Haddad, E., and Derrick, B. E. (2002). NMDA receptor antagonists sustain LTP and spatial memory: active processes mediate LTP decay. *Nat. Neurosci.* 5, 48–52.

Wang, F. H., Zevin, J., and Mintz, T. H. (2016). "Learning Non-Adjacent Dependencies in Continuous Presentation of an Artificial Language," in *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, (Austin, TX: Cognitive Science Society).

Wang, F. H., Zevin, J., and Mintz, T. H. (2019). Successfully learning non-adjacent dependencies in a continuous artificial language stream. *Cogn. Psychol.* 113:101223. doi: 10.1016/j.cogpsych.2019.101223

Wonnacott, E. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *J. Mem. Lang.* 65, 1–14.

Wonnacott, E., and Newport, E. L. (2005). "Novelty and regularization: The effect of novel instances on rule formation," in *BUCLD 29: Proceedings of the 29th Annual Boston University Conference on Language Development*, eds A. Brugos, M. R. Clark-Cotton, and S. Ha (Somerville, MA: Cascadilla Press).

Wu, T., Dufford, A. J., Mackie, M. A., Egan, L. J., and Fan, J. (2016). The Capacity of Cognitive Control Estimated from a Perceptual Decision Making Task. *Scient. Rep.* 6:34025. doi: 10.1038/srep34025

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 Psychology

ORIGINAL RESEARCH
published: 26 November 2021
doi: 10.3389/fpsyg.2021.650969

Information Density and the Extraposition of German Relative Clauses

Sophia Voigtmann[1,2]* *and Augustin Speyer*[1,2]

[1] *Department of Modern German Linguistics, Saarland University, Saarbrücken, Germany,* [2] *Collaborative Research Center on Information Density and Linguistic Encoding, Saarland University, Saarbrücken, Germany*

OPEN ACCESS

Edited by:
Hannah Rohde,
University of Edinburgh,
United Kingdom

Reviewed by:
Ulrike Demske,
University of Potsdam, Germany
Kalle Müller,
University of Tübingen, Germany
Łukasz Jedrzejowski,
University of Cologne, Germany

****Correspondence:***
Sophia Voigtmann
sophia.voigtmann@uni-saarland.de

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

Received: *08 January 2021*
Accepted: *27 September 2021*
Published: *26 November 2021*

Citation:
Voigtmann S and Speyer A (2021)
Information Density and the
Extraposition of German Relative
Clauses. Front. Psychol. 12:650969.
doi: 10.3389/fpsyg.2021.650969

This paper aims to find a correlation between Information Density (ID) and extraposition of Relative Clauses (RC) in Early New High German. Since surprisal is connected to perceiving difficulties, the impact on the working memory is lower for frequent combinations with low surprisal-values than it is for rare combinations with higher surprisal-values. To improve text comprehension, producers therefore distribute information as evenly as possible across a discourse. Extraposed RC are expected to have a higher surprisal-value than embedded RC. We intend to find evidence for this idea in RC taken from scientific texts from the 17th to 19th century. We built a corpus of tokenized, lemmatized and normalized papers about medicine from the 17th and 19th century, manually determined the RC-variants and calculated a skipgram-Language Model to compute the 2-Skip-bigram surprisal of every word of the relevant sentences. A logistic regression over the summed up surprisal values shows a significant result, which indicates a correlation between surprisal values and extraposition. So, for these periods it can be said that RC are more likely to be extraposed when they have a high total surprisal value. The influence of surprisal values also seems to be stable across time. The comparison of the analyzed language periods shows no significant change.

Keywords: information density, Early New High German, relative clauses, extraposition, corpus linguistics

INTRODUCTION

Attributive Relative Clauses (RC) provide more information about their head noun in their matrix clause. The head noun is taken up by the relative pronoun, the first word of the RC itself. One characteristic of German RC is that they can be placed adjacent to (1) or separated from their head noun (2).

When RC is separated from their head nouns, they are mostly placed in the Postfield (PoF), a position that is not usually mandatory. So, the question arises why RC can be frequently found there.

Prefield	LSB	Middle field	RSB	Postfield[1] (PoF)
1) Peter	Hat	Maria das *Buch*, das sie dringend braucht,	gegeben.	
Peter	Has	Maria the *book* that she urgently needs	given.	
2) Peter	Hat	Maria *das Buch*	gegeben	das sie dringend braucht.
Peter	Has	Maria *the book*	given	that she urgently needs.
"Peter has given Maria a book that she needs urgently."				

Explanations for the phenomenon of RC extraposition[2] vary between the length of the RC, the distance between PoF and RC head noun, the RC type, which can be divided into restrictive and non-restrictive RC, and the phenomenon called information disentanglement ["Informationsentflechtung" Zifonun et al. (1997, p. 1,669)]. This study examines the correlation between information and the position of relative clauses. There has been much research [Vinckel-Roisin (2015), Poschmann and Wagner (2016), among others] which links RC extraposition with information status or focus [both are often understood according to Chafe (1976), Prince (1981), Krifka (2007), among others].

For the purpose of this study, we, however, use the term information in reference to the Information Density (ID) of Shannon (1948) and define it as the "amount of information per unit comprising the utterance" (Levy and Jaeger, 2007, p. 1). Information is the likelihood of the occurrence of a word given a context of *n* words in terms of ID. Words that are frequent in a certain context have lower surprisal values than words that rarely occur in that context. Surprisal values correlate with perceiving and production difficulties [Hale (2001), Jaeger (2010), among others]. So, the impact of words with a high surprisal value on the working memory is higher than the influence of words with a low surprisal value. Therefore, speakers tend to distribute information as evenly as possible across clauses and discourses. Aylett and Turk (2004) found these effects in spoken languages. Levy and Jaeger (2007) extended their hypothesis for more contexts, like syntax, and formulated this principle in their "Uniform Information Density Hypothesis (UID)."

Information Density is well-established for measuring cognitive load and has already been used to explain RC extraposition in English with experiments and corpus studies [Francis and Michaelis (2012, 2014, 2017), Levy et al. (2012), among others]. Nonetheless, it has been rarely used for other languages such as, for instance, German (e.g., Voigtmann and Speyer, forthcoming). It is possible to connect all explanations for RC extraposition in German to the establishment of successful communication and the prevention of perceived difficulties. Only a few studies, however, immediately correlate perceiving difficulties with RC extraposition [e.g., Hawkins (1994) or Gibson (1998), Uszkoreit et al. (1998) for modern German] or test this correlation using ID (Voigtmann and Speyer, forthcoming; Speyer and Lemke, 2017). In this study, we apply the principle of establishment of successful communication, the main goal pursued by Shannon (1948), to the explanation of RC extraposition. The principles of ID are considered to be universal and testable on corpus data and are thus applied to historical data where RC extraposition, in general, is still under-researched. We aim to fill that gap.

In this study, we pose and discuss two hypotheses. First, regarding the extraposition of RC, we claim that the variability of RC positions is connected to perceiving difficulties that are caused by high surprisal values. Following Hawkins (1994) and Gibson (1998), the RSB marks the end of a clause. Processing capacities are free again so that RC with higher surprisal values are placed there without causing information loss. If RC with high surprisal values would be placed adjacent to their head noun between the sentence brackets, their processing could strain the processing capacities too much and information loss would happen. So, our first hypothesis is the following:

(H1) Higher surprisal values in RC favor their extraposition.

Second, we take the diachronic perspective of our corpus into account. We conducted a corpus study for Early New High German (ENHG)[3] and early Modern German medical texts to test this hypothesis and provide information about an earlier stage of German. Due to few scientific texts in German in the seventeenth century, as scientific writing in Germany was done in Latin before that point, the effect might be different for the seventeenth than the nineteenth century. While the seventeenth century authors might only have a few scientific texts as a model, for nineteenth century authors, scientific articles written in their native language were already common. Developments in style and commonness of writing in native language of an individual instead of a Lingua Franca are taken into account by dividing the timespan into different parts. As the main goal

[1] This way of dividing a sentence follows the "Topologisches Feldermodell" (Drach, 1937; Wöllstein, 2014), a model describing the distribution of constituents and verbs across German clauses. The verbal parts built the framework of the clause *via* the left and right sentence bracket. In main clauses, the left sentence bracket (LSB) is filled with the finite verb, whereas the right sentence bracket (RSB) holds infinite verbal parts or verbal particles. In subordinate clauses, conjunctions fill the LSB and the verb the RSB. The position in front of the LSB (Prefield) can usually be filled with one constituent and must be present to signal that a clause is a main clause. It can only be dropped in polar questions and in German dependent clauses with a complementizer, while the field between the brackets (Middlefield) can—theoretically—be filled with an arbitrary number of constituents or remain empty. The Postfield (PoF) is, if present, mostly occupied with subordinate clauses—independent and dependent clauses as in the case of RC.

[2] We use the word extraposition only as an expression which describes the separation of head noun and antecedent without referring to any generative theory for RC placement.

[3] Early New High German is commonly understood as the period of the German language spoken from 1350 to 1650. Predecessors of ENGH were Old High German (500–1,050) and Middle High German (1,050–1,350). The New High German period begins about 1700 (Nübling et al., 2013). In this study, we use the term late ENHG until 1700, following Polenz (2010).

of all authors in each time span is, however, still to ensure successful communication by means of their texts, we propose in our second hypothesis:

(H2) The correlation between the extraposition of RC and their surprisal values is consistent over the centuries.

We divided this time span of 250 years into periods of 50 years to be able to account for a change. Note that the New High German period (from around 1650 to 1900) is not subdivided like former language periods. Research concentrating on German in the eighteenth or nineteenth century does not base the division of the timespan on intra-linguistic criteria but takes century borders. For our subperiods, we used a smaller time span of only 50 years and understand this as an exploratory approach.

In this study, we try to find evidence for both hypotheses. Furthermore, we check whether information density is a better predictor for extraposition than restrictiveness and length because all factors factor frequently mentioned in literature are connected to perceiving difficulties. For a complete picture, we first present a more detailed description of RC and RC extraposition along with that of the ID of Shannon (1948) and the principles mentioned above (Section Theoretical Background). Then, we describe our corpus and method (Section Corpus and Method) before presenting the results (Section Results: Information Density and Length). The study closes with a discussion about the results (Section Discussion) and a conclusion (Section Conclusion).

THEORETICAL BACKGROUND

In the following section, we present the kind of RC used for this investigation and reasons for extraposition which includes restrictiveness, length, and information management for German RC. As far as possible we include research about ENHG but as this period is highly underrepresented in research, synchronic research will be included as well as an approach to Modern German standards.

The second part of this section gives an overview of Information Density, its usage, and some of its advantages. We concentrate mostly on the study of Shannon (1948) itself and only include some more recent research where it is relevant for our hypotheses. The main goal is to show how information is defined and how ID correlates with processing difficulties. For more details and mathematical evidence for the way ID is calculated, please see Shannon (1948) or Levy (2008), among others.

The connection between RC extraposition and ID will be drawn in section "Methodological Considerations About RC Extraposition and ID" because it also concerns the predictors used in our model and is, therefore, more suitable there.

Relative Clauses

As mentioned in the introduction, RC are subordinate clauses. Besides bound or attributive RC (example 1, 2), on which this paper focuses, there are also free and continuous RC. They are excluded from this investigation either because they do not have an antecedent that is present in the sentence (free RC) or take the whole sentence as an antecedent and therefore can only be placed in the right periphery [continuous RC, for more information, refer to Gallmann (2005)].

Our definition of bound RC follows Hentschel and Weydt (2003), who define RC as clauses that apply attributively to their antecedent and which are introduced by the relative pronouns "der/die/das" or "welcher/welche/welches."[4] To take our diachronic approach into account we also included the relative particle "so" (Pfeifer, 1995) because it is more or less comparable to the English "that," and has no additional meaning and is not bound to specific nouns.

Having established the kind of RC we want to investigate, we want to come to the main point: the position of RC. They can be placed adjacent and extraposed to their antecedent without changing the proposition of the sentence. The head noun can, for example, stand in the middle field while the RSB separates it from the RC (Birkner, 2008, p. 50; for the classification we use here, see section Method). Lehmann (1984) describes the extraposition as the process in which the RC is moved to the end of the sentence while Fritsch (1990, p. 114) specifies it as a movement to the right but no further than to the end of the smallest clause in which its antecedent is found. The most frequent explanations for the varying positioning are RC type, RC length, informational aspects, and the distance between RC and antecedent which is or could theoretically be covered.

We start with the RC type as it requires additional information. RC can be divided into restrictive and non-restrictive RC. Restrictive RC restricts the possible references of the antecedent (Birkner, 2008) when the antecedent is not sufficiently determined (3).

(3)	Diejenigen	Studenten,	die	ihre
	Those	students	who	their

Hausaufgaben	machen,	bekommen	bessere	Noten.
homework	do	get	better	grades.

"Those students who do their homework get better grades."

In this example, the RC limits the number of students getting good grades. Restrictive RC often follow determiners or pronouns like "jeder" (everybody) or "derjenige" [the one; Lehmann (1984), Fritsch (1990), Eisenberg (1999), Birkner (2008), among others]. There are also non-restrictive RC which only illustrate their antecedent and give further information about the antecedent. A German non-restrictive relative clause can be identified by adding "ja" or "eben" to the RC without creating a marked sentence. The antecedent of a non-restrictive RC is already completely determined (4)[5].

[4] The definition of Hentschel and Weydt (2003) contrasts with the one presented in Eisenberg (1999) or Helbig and Buscha (2001). The latter two claim that subordinate conjunctions can also initiate relative clauses.
We do not follow their works because, unlike relative pronouns, subordinate conjunctions can only relativize clauses to certain nouns which have a similar meaning as the conjunction itself. A temporal conjunction can, for example, only follow a noun which denotes a point in time. Relative pronouns, on the other hand, can follow any noun.

[5] In some cases, it is not easily determinable whether a RC is restrictive or not when the context or the world knowledge does not disambiguate the RC type. That will be relevant especially for the historical data because modern readers, including us

(4) Sabrina, die (ja) Literatur studiert, sitzt in der Bibliothek.
Sabrina who [PRT] literature reads sits in the library.

"Sabrina who reads literature sits in the library."

Restrictive RC is said to be more separable from their head nouns than non-restrictive ones. This can be attributed to the fact that the RC is necessary to complete the sentence. The recipient knows that the head noun is still incomplete and can therefore keep cognitive capacities free, which can be filled by the RC even at the end of the matrix clause. Non-restrictive RC may be too surprising when separated from the head noun over a long distance because they are not needed for the referential identification of the head noun. Though several scholars state but do not test, the influence of restrictiveness [Lehmann (1984), Fritsch (1990), Zifonun et al. (1997), Poschmann and Wagner (2016), among others], this could not be shown in the study of Poschmann and Wagner (2016) on RC where restrictiveness could not be determined as a predictor for extraposition.

The second, most frequent explanation for RC extraposition is length. On the basis of avoiding a stain of memory capacities and enabling effective processing of a sentence (Uszkoreit et al., 1998), RC length is often correlated with extraposition. Uszkoreit et al. (1998) found in their corpus study on German newspaper articles that extraposed RC in the PoF are on average one to three words longer than adjacent RC. The maximum distance is up to nine words but mostly varies between one and four words. In the latter case, the length of the RC becomes a more relevant factor. The longer an RC is, the more likely is its extraposition even over short distances. They showed that distance is the most crucial factor, followed by length (Uszkoreit et al., 1998, p. 130). Zifonun et al. (1997) also saw the length of extraposed material as one of the most important factors besides distance. One possible reason for this is combined with information disentanglement. Clauses are understood incrementally [Levy (2008), among others]. Words are ordered in a way that allows fast processing. On the one hand, this would mean that an RC should be adjacent as no dependencies must be kept in mind while processing the rest of the sentence (Hawkins, 1994; Gibson, 1998). On the other hand, Hawkins (1994) describes a complex interaction between the advantages and disadvantages of adjacency when the embedded material such as the RC is so complex that a recipient is no longer able to remember or incorporate the part of the clause which occurred in front of the RC. This holds true especially for RC placed in the middle field where the prefield and the first part of the middle field itself must be incorporated in order to understand the whole sentence. The longer the RC, the more cognitively challenging the processing of the RC and the matrix clause will be [see Gibson (1998) "memory load"].

The last reason frequently used to explain extraposition is information management. It is highly connected to the formerly mentioned memory load. Poschmann and Wagner (2016), for instance, showed a connection between the length of German RC and information structure. They saw a correlation between length and focus, as new material is usually longer than given material (Poschmann and Wagner, 2016, p. 1,022). They referred to the concept that easily accessible (that is: given or inferable) information is usually presented early in the sentence while new information tends to follow later.

Furthermore, the integration cost is influenced by the number of intervening new referents (Bader, 2014). They use a production experiment to find out how to focus, word order, and RC-type effect extraposition. In a second step, participants had to rate the acceptability of RC extraposition under the manipulation of word order, focus, and RC-Type. Though corpus data suggests an even distribution of RC, extraposition was rated worse than adjacency. This might be caused by long distances between antecedent and RC. RC with a wide focus is more acceptable than those with a narrow focus and the interaction between extraposition and focus shows that extraposition is rated better "when the NP it modifies or the RC itself is in focus" (Poschmann and Wagner, 2016, p. 1,057). They link their findings to other research investigating the influence of predictability, namely, the one presented by Levy et al. (2012) on English RC. The more constituents intervene between antecedent and RC, the more unlikely it becomes to find an extraposed RC. Levy et al. (2012, p. 29) show in their reading time experiments that "[r]elative clauses extraposed from simple [determiner + noun] NPs across a verb are harder to process than their corresponding *in situ* variants. RC extraposed from a direct object NP across a PP are harder to process than *in situ* RC modifying either the direct object (but following the PP) or the PP-internal NP. Nevertheless, a preceding context (specifically, NP-internal premodifiers) that sets up a strong expectation for a RC modifying a given noun can strongly facilitate comprehension of an extraposed RC modifying that noun." They also assume that the scarcity of some collocations tested might also cause the shown difficulties of extraposed RC comprehension. We should keep in mind that one of their most precious findings is the influence of expectancy on reading times. It is interesting to mention that the PoF and, with it, RC extraposition has undergone a changeover centuries. In Old High German, information structural considerations, namely, focus, are the most important next to restrictiveness which is also explained by information structure (Coniglio and Schlachter, 2015). The importance of information structure slowly decreases over time (Speyer, 2016) along with the usage of the PoF.

Even throughout ENHG itself, we see changes when the sentence brackets are finally established. According to Schildt (1976), from 1470 to 1530, 68% of the sentences in the corpus had no filled PoF, whereas from 1670 to 1730, this number decreased by 81%. Together with the decreasing frequency of PoF filling, this position also becomes more permissive in information structural terms. Early ENHG allowed especially new material in this position, while late ENHG does not make a real distinction between new and given material there. This might be a result of

as researchers, might not be able to reconstruct the knowledge of the author and his audience. The fictional example (5) shall explain this:

5) The physician treats the patient with a bezoar that is taken from a cat.

A modern reader cannot say whether a bezoar—an ingredient for a remedy that is mentioned in two texts in the corpus (Purmann, 1680; Abel, 1699)—is always taken from cats or not and whether it was assumed to make a difference for the success of the treatment and is thus not able to determine the RC type.

the not yet fully established sentence bracket structure in this period (Sahel, 2015, p. 168). The structure however becomes more pronounced over time (1650 to 1800) which is indicated by an increasing number of RC found there (Sahel, 2015, p. 172).

For the early New High German period, research, especially on RC extraposition, is rare. The sentence frame, finally established in the eighteenth century (Admoni, 1990; Konopka, 1996; Takada, 1998), can already be found in Old High German (OHG) but it was not as necessary as it is today and has begun to be in the eighteenth century. The material was more often placed on the borders of the clause and for various reasons (Paul, 2007). Besides the decline of phrasal material in the PoF, clauses were frequently placed there. Konopka (1996, p. 178) gives three reasons for placement of material in the postfield: "A. die Gestaltung der Informationsperspektive, B. die Sicherung der Textkonnexion, C. die Entlastung des überfüllten Satzrahmens." (A. to shape the information perspective, B. to ensure text connectivity, C. to relieve strain on the sentence frame). Scholars from the seventeenth century agree with the latter: Though the sentence frame should be kept, the placement of clauses behind the RSB can ensure better processing of information [e.g., Schottelius (1641) in Takada (1998)].

In summary, bound RC can be placed adjacent or extraposed to their head noun. Reasons which have been proposed in the literature for the extraposition are the RC length, the distance to the head noun, restrictiveness, and information management. Closer examined, these reasons all refer to successful communication and information transmission.

Information Density

Explaining, characterizing, and measuring successful communication and information transmission is the key feature of Information Density. A change in the position of certain linguistic material aims to improve communication by improving the transmission of information. In most literature on RC, information disentanglement or avoiding sentence fields that are too long are given as reasons for extraposition (cf. Section Relative Clauses). However, such approaches lack measurability, and even research results that deal with focus or givenness [Coniglio and Schlachter (2015) for example] can only include certain freedoms in the position in their considerations. Therefore, it is necessary to use a method that makes processing effort objectively calculable and does not differentiate between the information content of certain word forms. Such a theory is offered by Information Density theory of Shannon (1948).

In short, ID describes information as the probability of occurrence of a word in its context. The idea behind this is as follows: the less expectable a word is in its context, the more information it contains. Likelihood and information value correlate negatively with each other. The significance of this approach is that it offers explanatory potential for intra-linguistic variations, which, however, have no influence on the proposition of the sentence. According to Shannon (1948), the aim is not to write better messages, but to encode messages more effectively.

In almost all languages, there is a wide range of variations in the area of coding. In spoken language, the length of phones can be varied. In the field of morphology, speakers and writers can use abbreviations. Lexicology offers the possibility of variation between semantically very similar terms to express the same facts. For pragmatics, different reference expressions can be used to obtain variation in expressing the same facts. Syntactically, certain liberties in word order (see example 6) are offered (Gibson et al., 2019).

6) **Yesterday**, I gave him the book. → I gave him the book **yesterday**.

Both sender and recipient can select and decode different codes from a set of codes during the transmission of the message. All possible choices from this set of codes are equally probable according to Shannon (1948)[6]. The logarithmic function on the basis 2 is used as a mathematical description for the selection process (see below). Bits are thus the unit for information in context.

The signals and the coding must be adapted to the kind of transmission without exceeding the limits of the channel through which the message is sent and its specific capacity. The goal is to transfer the message into a language. This language already gives guidelines for the structure and thus, defines a natural frequency of certain elements. Both the sender and the receiver are aware of these structures. This leads either to time saving in the transmission of the message or to a less heavy load on the channel if the message sequence has been correctly encoded into the signal sequence (Shannon, 1948, p. 384). So, the transmission of the symbols is both incremental and dependent on the previous symbol and the symbol itself. The system of selecting the subsequent symbols can therefore be described as a stochastic process and is thus subject to the conditions of probability theory (Shannon, 1948). This can be represented as follows: $p_i(j)$, which describes the probability that j follows i (Shannon, 1948, p. 384).

If only the element itself is considered in its frequency, it is called unigram frequency. This is the simplest way to approach the stochastic process of element selection. However, this simple approach does not even come close to existing languages. For this purpose, more context must be considered, which then can be called bigram, trigram... n-gram. The larger the context, the more the results converge to the actual language, even if no attention is paid to conveying a specific content. "A sufficiently complex stochastic process will give a satisfactory representation of a discrete source" (Shannon, 1948, p. 386).

The core question that Shannon (1948) pursues consists of describing and mathematically explaining the conditions for optimal message transmission through a *noisy* channel. The considerations presented so far in the present work refer to a channel in which no interference is present, a so-called "noiseless channel" (Shannon, 1948, p. 19). However, this is only the case in a few situations. Nevertheless, most conversations are successful even if the speaker says something different than the receiver understands and the input is no longer equal to the output (Shannon, 1948, p. 19). This is highly dependent on context.

Certain words are more expectable in their context than other words. Let us consider (7):

[6] They can still have different surprisal values or contain more or less information depending on the context these variations occur in.

7) You may now kiss the [bride].

This sentence might have been heard so often in wedding scenarios that the recipient has a strong expectation that the *bride* follows after a kiss and the definite article. So, the surprisal for the *bride* should be very low. Surprisal is usually calculated by the negative logarithm of the probability of an element given in a context: P(word) = $-\log_2$ (*word*|*context*). Again, a distinction must be made as to how far the context is defined. In the case of unigram-surprise values, only the frequency of the element is relevant. In the case of bigram-surprise values, the probability of occurrence of the element before the considered element is included. Lastly, in the case of trigram-surprise values, the two preceding elements are included, etc.

Due to the very narrow context, even small changes can be decisive for other surprisal values. If the predicate was changed to *lecture* in example 9, the *bride* would no longer be a word marked with a low surprisal value as the *lecture* is more likely to occur in an educational context. These examples may be very simplified and may not capture the whole problem of positional variants. However, they do allow the first impression of ID theory and touch on a problem that can rightly be identified, namely, the strong focus of classical surprisal calculation on single words. Recent research shows that extralinguistic contexts such as script or world knowledge have an influence on the likelihood of a word and the difficulty in processing it [Ostermann (2020) among others]. However, it is precisely for historical contexts that the strong intra-linguistic orientation of theory of Shannon (1948) is useful since world knowledge can only be reconstructed to a limited extent and the knowledge of individual writers, on the other hand, can hardly be traced. The orientation toward purely written sources facilitates the objective evaluation of data.

Furthermore, the relationship between the predictability of linguistic material and efficient communication exists at all linguistic levels (Gibson et al., 2019), and a relationship between processing effort, i.e., psycholinguistic reality, and information density could be shown as well [Levy (2008) and others].

According to Levy (2008, p. 1,127), there is a probabilistic and expectation-based theory of syntactic understanding. Some syntactic structures consume more resources or memory than others. At the same time, human resources are limited which is why processing problems can occur in structures that consume a lot of resources. Therefore, the channel is virtually overloaded, so information is lost. Theories of syntactic processing gain importance. Thus, the understanding of information is based on different sources: structural, lexical, pragmatic, and discourse-based (Levy, 2008, p. 1,128). This results in a competition of similar analyses since these sources are combined for understanding (Jurafsky, 2003). The processing effort thus corresponds to the surprisal of a word. It is the interface between the linguistic representation during the comprehension of the sentence and the processing difficulties which can be found for a particular word within a sentence (Levy, 2008, p. 1,128). The recipient thereby preserves the complete set of the different, probable, and partially processed constituents from the already seen or heard input. They assign to it a possible probability distribution over the complete structure to which the already received constituents can expand. Surprisal is thus seen as the difficulty of replacing an old distribution with a new one (Levy, 2008, p. 1,132).

To facilitate communication, an even distribution of information is important at all linguistic levels, not only at the phoneme and grapheme but also at the syntactic level. Speakers design their utterances in such a way that there are no strong fluctuations in the information profile (Levy and Jaeger, 2007). This is achieved by exploiting the freedom of expression offered by languages or by omitting optional material. To prove this for the syntactic design of utterances, Levy and Jaeger (2007) investigated syntactic reductions and found them to be "a phenomenon in which speakers have the choice of either marking a phrase with an optional word, or leaving it unmarked" (Levy and Jaeger, 2007, p. 2). Their research topic is optional *that* in English RC. In their corpus study, they find that *that* is inserted when the surprisal on the first word of the RC would otherwise be too high, thereby exceeding the assumed channel capacity and causing a loss of information. Thus, they found the first evidence for what is known as the "Uniform Information Density Hypothesis." It can be shown for both spoken and written English that speakers drop an optional relative pronoun, and this finding is also common across standard varieties (Jaeger, 2010, p. 163). This phenomenon and the UID can also be integrated into existing processing approaches and preferences. It can be compared both with "dependency processing accounts," which assume that preference is given to variants that have shorter dependency relationships. They also take up the "Gesetz der wachsenden Glieder" (law of increasing constituents) by Behagel (1932). Furthermore, it concerns "alignment accounts" which regard access to referents as a major factor for linguistic preferences. These accounts rely on the conceptual accessibility and pre-mentioning of referents and can be combined with "availability accounts," which focus more on the referent and claim that material that is cognitively available appears earlier in the sentence (Jaeger, 2010, p. 165). Incremental speech production is also related to this. What is available earlier can be expressed earlier, which in turn can be combined with the other approaches mentioned above.

While the language processing system works basically incrementally, at least for the hearer, there is still the need to keep the elements of a clause together in the working memory as syntactic dependencies must be reconstructed by the hearer and the verb valency has to be checked. Therefore, another factor in the calculation must be the sum of the surprisal values of the individual lexical items within a clause as they must be related to each other and thus, to some degree, processed together. It is reasonable to assume that a clause containing some words with high surprisal is a whole lot more difficult to process than a clause containing only words with low or medium surprisal values. To account for this fact, we use two measures that are derived from surprisal: the *cumulative surprisal* of a clause is the sum of all individual surprisal values of the words in the clause, and the *mean surprisal* is the arithmetic mean of the surprisal values in a clause, that is, cumulative surprisal divided by the number of words in the clause (cf. Section Methodological Considerations About RC Extraposition and ID).

In summary, ID according to Shannon (1948) determines the information content of a word in a certain context and links this information content to the likelihood of the word in the context. The surprisal value is calculated by the logarithmic function and expressed in bits. The aim of ID theory is to provide a descriptor for the optimal encoding of a message and thus, to be able to demonstrably describe how information loss can be prevented. In the classical method of calculation with n-grams, *all* words, namely, content and function words, are considered in the calculation of the surprisal values, whereas in classical information-structural studies often only content words are considered. Thus, no positional changes can already lead to visible effects. A description of why and how this concept is applied to RC follows in the methodology section.

CORPUS AND METHOD

This section presents the basis for our research. We will present the corpus we used and provide further reasons for our decision to work on early New High German. The second part of the section is concerned with our method. We present our annotation process and our language model. The section is closed by an explanation of the predictors we consider relevant for extraposition. A special goal is to show that while length might have already proven important for extraposition, it is not necessarily the best predictor for extraposition. Using ID as a predictor instead might lead to a different conclusion. We are aware of the rather exploratory character of the study.

Corpus

Our corpus is built on texts from the *Deutsches Textarchiv* (DTA). The DTA is a collection of texts from different genres and periods ranging from the seventeenth to the twentieth century. Balanced samples from newspapers, novels, literature for a specific purpose ("Gebrauchsliteratur"), and scientific texts provide an overview of the German language development. A major advantage of the DTA is the preprocessing of the texts. They are tokenized, normalized, lemmatized, and POS-tagged albeit in a rather poor quality which complicates and even prevents automatic annotation.[7]

The DTA is the only database with such a high variety of genres that includes scientific, namely medical texts. Before the seventeenth century, German scientists used to publish their findings in Latin so that a German tradition of scientific writing in the native language of an individual developed only at the end of the ENHG period. Even then, the publishing process did not resemble the one we know today but consisted of letters to interested colleges. This puts this genre in the field of tension between different registers, namely written and oral discourse modes (Koch and Oesterreicher, 2007). Despite being a written form of communication, letters tend to be closer to the oral discourse mode than the written discourse mode. Typical examples are addressing the addressee or, according to the theory, placing more material in the PoF. At the same time, these authors might be influenced by the former Latin tradition with elaborate rules on how to write prose and might be influenced by that. Because (written) Latin does not have a sentence frame like German and has widespread dependencies that would strain the parsing capacities of a German native speaker, this might contradict the optimal distribution of information when a clause is written in a more Latin-like style at the beginning of seventeenth century. This strain between the letter style and the former Latin tradition might result in longer, intertwined clauses that decrease over the centuries. In the nineteenth century, however, texts might also resemble a more modern scientific style with shorter, less intertwined clauses. Therefore, it is also important to have a data basis that spreads over the centuries like the one provided by the DTA.

As we are not interested in grammatical but in lexical predictability,[8] lemmatization is a crucial factor for our analysis. Due to the non-standardized orthography in ENHG, normalization is an important step. The Language Model (see Section Language Model) would not capture the same word when it is spelled in different ways. Because words appear in different inflected forms, however, normalized data is not sufficient for the language models either, but we need lemmatized data to capture all instances of a given word in whatever form they appear and in whatever way they are written.

Our corpus from the DTA used in this study consists of the nine medical texts from 1650 to 1900 with 841,877 tokens[9]. The texts were chosen arbitrarily while translated texts were excluded. The 250-year time span was divided into 50-year-steps to account for possible changes in language use, orthography, and writing style preferences which are highly relevant for the calculation of the language model (section Language Model). The corpus under study consists of the following texts (**Table 1**).

Methods

Annotation

We want to emphasize that all annotations were made manually due to the poor POS-tagging of the DTA. We used WebAnno (Eckart de Castilho et al., 2016) for the annotation[10].

[7] The DTA-Project started more than ten years ago and is, therefore, not on a level we are used to in newer projects like the *Referenkorpus Frühneuhochdeutsch (ReF)*. We are aware of updates on the data but have not included possible improvements on the annotations because we manually annotated relevant information on downloaded versions of the texts. The time of the download was 2018, so this is the version of the corpus used here.

[8] As requested by one reviewer, we want to explain grammatical and lexical predictability briefly here. We understand grammatical predictability as the likelihood certain grammatical categories following each other. This could be measured using dependencies or POS-tags, e.g., to measure how likely it is for a relative pronoun to follow verbal material. This is not applicable to our current study for two reasons. First, our corpus is not dependency-parsed and has poor POS-tagging, preventing measuring grammatical predictability on our data. Second, grammatical predictability does not provide insights regarding how difficult the processing of clausal content is. This is measured using the likelihood of a certain lexical word in a context, e.g., how likely is "advice" following "medical."

[9] These texts are part of a larger corpus of 33 texts with 593,086 tokens which includes theological texts as well, created as part of the CRC. We have collected data from these texts but processed only the mentioned nine texts so far.

[10] We must thank Katrin Ortmann (RUB) for setting up and curating WebAnno and for trying to improve the POS-tagging at this point.

TABLE 1 | Corpus.

Period	References
1650–1700	Purmann, 1680; Abel, 1699
1700–1750	Unzer, 1746
1750–1800	Gall, 1791
1800–1850	Reil, 1803; Carus, 1820
1850–1900	Ludwig, 1852; Koch, 1878; Kraepelin, 1892

TABLE 2 | Language model.

Period	Training data (in token)	Test data (in token)	OOV-ratio	Number of RC (extraposed RC)
1650–1700	2,107,590	48,1693	8.93%	240 (116, 48%)
1700–1750	1,481,259	39,251	6%	680 (363, 53%)
1750–1800	2,572,263	26,325	14.72%	375 (130, 35%)
1800–1850	998,639	16,757	6.28%	1,023 (573, 56%)
1850–1900	1,270,561	29,060	12.13%	925 (467, 50%)

We manually annotated the following features in the corpus: the RC,[11] their position as described below, their antecedent, that is, the noun or pronoun the RC depends on, and their type (restrictive vs. non-restrictive). To annotate the RC type, we determined whether the RC is necessary to clearly identify its antecedent. The main criterion to determine the restrictiveness of the RC is whether the antecedent can be completely and uniquely identified without the RC. Certain hints at restrictiveness are, for example, given by certain determiners (e.g., *derjenige*, "the one"). In the case of non-restrictive relative clauses, we are confronted with the problem that we cannot be sure whether the insertion of "ja/eben" would have been marked for ENHG writers. Because of the language period, the annotation of restrictiveness was not possible in every case because we cannot reproduce the world knowledge of, for example, a seventeenth century writer. In these cases, the type was not annotated (NA). Furthermore, we annotated the Left and Right Sentence Brackets (LSB and RSB) following Wöllstein (2014). The categorization of the sentence brackets is necessary to determine whether an RC is extraposed or not. The length of the RC and the distance between antecedent and the first word of the RC were both calculated automatically and not manually annotated.

We only annotated RC and not whole sentences. We are aware that we should also look at the ID profile of the whole sentence, but again the DTA provides some disadvantages. Due to the rather irrelevant punctuation and the practice in ENHG to sometimes end a sentence with a semicolon, so not even a human reader can be sure whether that really marks the end of a sentence or just a clause, the automatic sentence recognition fails. As a result, some sentences are incomplete and WebAnno does not allow our annotation to continue over sentence boundaries, whereas others include several sentences and are marked as one. As we have not yet annotated the sentence boundaries manually, only the RC, themselves, are considered for the results. The number of RC we found in the corpus is given in **Table 2** in the following section.

The most relevant factor is, as mentioned before, the adjacency of RC and head nouns. When both are in the prefield or in the middle field framed by both sentence brackets they are clearly determined as embedded or *in situ* (8a). Also, when the RC (underlined) is behind the RSB and the head noun (bold) is either in the prefield or, more often, in the middle field, it is without a doubt an extraposed RC (8b). But there are also cases in which the determination of the RC position is not as easy. The RSB can remain empty but still build the end of the clause. Two special cases arise when the RC is at the end of the clause and adjacent to its head noun (8c), we called the RC ambiguous and excluded it from the analysis because we cannot rule out that there has not been a movement over the empty RSB. But when there is a material other than the RSB intervening between the head noun and RC we classified the RC as extraposed. While we can, strictly speaking, not be sure whether the RC is actually in the PoF (8d), the fact that the RC is no longer adjacent to the head is crucial and outweighs the uncertainty.

(8a) **Die alteration aber** / die aus dem kalten Waffer entftehet
The alteration but which out the cold waters results

/[gefchicht$_{LSB}$] auf folche Art:
happens in such way.

'But the alteration, which results from cold waters, happens in such a way.' (Abel, 1699, sentence 113).

b) Streng genommen [müsste$_{LSB}$] man dazu **alle**
strictly taken should one among that all

diejenigen Krankheiten [rechnen$_{RSB}$],
those diseases count

welche eine Folge von Verwundungen [. . .] sind
which a consequence of wounds are.

'Strictly speaking, one should count all those diseases which happen as a consequence of wounds among them.' (Koch, 1878, sentence 4).

c) Alfo [find$_{LSB}$] die Bruche eine gewaltfame [. . .]
So are the factures a violent

Zerfchmetterung **der harten Knochen** Ø$_{RSB?}$.
shattering of hard bones

fo aneinander hangen. Ø$_{RSB?}$. that to one another hang.

'So, the factures are a violent shattering of hard bones that hang close to each other.' (Purmann, 1680, sentence 169).

d) Peter traf einen Freund auf der Straße, den er lange nicht
Peter met a friend on the street who he long not

[11] This might include noun-related continuous RC as well, tough their position is not variable, as pointed out by a reviewer. If their exclusion significantly changes any of the results presented in section Results: Information Density and Length, will be topic of another study.

gesehen hatte.
seen had.

'Peter met a friend whom he had not seen for a while on the street.'

The software R (R Core Team, 2018) was used for further data processing. All sentences not including RC were excluded and punctuation marks were removed because rules for the placement of punctuation marks had not yet been established, and they were often placed according to personal preferences of the authors so that an additional meaning or advantage of their inclusion could not be found. Then, we calculated the skip-gram language model on every remaining word and checked for the influence of RC length, type, cumulative, and mean surprisal. Note that RC length was calculated automatically with R. We will provide more details regarding our motivation for the analysis in the following sections. Since the data is not very balanced, we perform the statistical analysis not only on the whole data which would not be feasible for the second hypothesis anyway but on every 50-year timespan separately (see Section RC per Period).

Language Model

In the next step, we calculated a Language Model (Hale, 2001) and a skip-gram Language Model with a 2-skip-bigram (Guthrie et al., 2006) for every 50 years on the lemma layer of the corpus using an SFB-intern tool. Skip-grams were chosen over bigrams because they do not only take immediately adjacent words for the model but allow tokens to be skipped to create trigrams, thus capturing the context better and achieving better coverage of the data. This is especially useful when the training data varies from the test data and increasing coverage of n-grams cannot be assumed (Guthrie et al., 2006, p. 1,223). The model was trained on those scientific texts in the DTA that were not included in the test data.[12]

Training data is used to gain estimated values over the following words given its context using a hidden Markov model. It states that the probability of a future unit can be predicted without looking too far into history (Mürmann, 2014). For languages, this means that not every linguistic utterance ever produced must be included in the calculation, but that a part of the linguistic utterances is sufficient to be able to make acceptable statements. The surprisal value of a word is obtained by calculating the probabilities of its occurrence and mapping them to the test data. This is done using the Maximum Likelihood Estimate: "the maximum likelihood estimate is so called because it is the choice of parameter values which gives the highest probability to the training corpus. [...] It does not waste any probability mass on events that are not in the training corpus, but rather it makes the probability of observed events as high as it can subject to the normal stochastic constraints." (Manning and Schütze, 1999, p. 198). Further smoothing methods are applied to enable the model to give an estimate to tokens unseen in the training data but are used in the test data.

The Language Models were calculated without punctuation marks since they are not meant for ENHG (see above). The following **Table 2** sums up the corpus including training data, out-of-vocabulary-token-ratio, and the number of RC.

Methodological Considerations About RC Extraposition and ID

The length of the RC is one of the most frequent factors used to explain extraposition. Various studies prove this for both German (e.g., Uszkoreit et al., 1998; Poschmann and Wagner, 2016) and English (e.g., Levy et al., 2012). At the same time, however, the factor of informativeness of the RC is also repeatedly used as an approach in theoretical and experimental studies. Intuitively, the two concepts do not contradict each other. The more words are available in a sentence, the more information it can contain. The more information there is, the more cognitive capacities are needed to process the sentence. However, if, at the same time, cognitive capacities are also used on other processing issues, such as the comprehension of a complex middle field of the matrix sentence, an RC occurring there could cause an overload of the available cognitive capacities. In this case, communication should fail. This approach is represented by the well-known theories on the extraposition of RC presented by Hawkins (1994) and Gibson (1998). Both, as mentioned above, limit themselves to measuring complexity by the number of words.

However, the information density approach of Shannon (1948) and more recent research by Levy and Jaeger (2007), Levy (2008), Jaeger (2010), and others show that an increase in length, i.e., the addition of words, does not necessarily equate to a significant increase in information if the information is understood as the predictability of a word in context. Both the immediate context of a word and the extended context can reduce the probability of occurrence of a word. This, in turn, would reduce the information content of the specific word and could eventually lead to a reduction in the overall information content of the sentence despite a higher number of words. A simple example (9) illustrates this:

(9) Die Stadt wurde von Caesar erobert.
The city was by Caesar conquered.

"The city was conquered by Caesar."

This sentence contains five words. Without a larger context, the information content of Caesar should be quite high. If you now add words at various points and thus increase the length of the sentence, you simultaneously reduce the informativeness of various words.

(10) Die Stadt Rom wurde von Gaius Julius Caesar erobert.
The city Rome was by Gaius Julius Caesar conquered.

"The city of Rome was conquered by Gaius Julius Caesar."

The sentence (10) is extended to nine words. At the same time, both the mention of Rome and the mention of first and gentil names of Caesar should ensure that the likelihood of "Caesar" increases enormously with the preceding "Gaius Julius" and that the negatively correlated surprisal value falls. Theoretically, but more difficult to prove, depending on the language model used, even the mention of Rome can cause the full name to

[12] The list of the texts and the downloads of these texts can be provided on request. The same holds for R-scripts and the corpus in its current form.

be assigned lower surprisal values since Caesar and Rome are closely connected. The added words can therefore ensure that the sentence is easier to process through the selective reduction of the information content, although it has become longer at the same time.

This effect was demonstrated by Levy and Jaeger (2007), among others, and, subsequently, many times by Jaeger (2010) when investigating optional elements in a sentence, namely, the optional use of *that* as an RC introducer. In less expectable contexts, the use of the relative pronoun can ensure that an overwhelmingly high processing load on the first word of the RC is reduced. It can therefore be stated at this point that informativeness and length do not necessarily have to be positively correlated with each other and that a separate consideration of the two is appropriate. These theoretical considerations lead to two possibilities for calculating the information density: We use cumulative and mean surprisal values.

The justification for the cumulative surprisal value lies in the parallel processing of information (e.g., McClelland and Elman, 1986). The entire information density theory of Shannon (1948) is based on the incremental approach. Words are processed one after the other and the likelihood of a word results from its context. Previous theories and experimental methods that measure processing difficulties mostly work with local phenomena. Bigram language models and, to a certain extent, skipgrams are strongly dependent on a narrowly defined context. Reading time studies measure delays on specific individual words and focus, simply put, on problems at individual points. These methods are not well-suited to determine the total processing effort of a sentence. Because other factors are also relevant for understanding such as parallel processing of grammatical structures or the inclusion of different sources (e.g., Cutler, 2008), it is important to find a model that approaches the total processing effort but is also usable for corpus data. The sum of all surprisal values in a clause, or even just a construction, can be understood as an approximation. The idea behind this is the following: the cognitive capacities are neither immediately free after processing a word nor are they immediately available again. Instead, they form a kind of pedestal that grows larger with each additional word depending on its surprisal. Only when the construction is completed does the full processing capacity become free again and the filling process of the pedestal can begin again at a low level.

However, the calculation of the sum leads to some problems. Even though it was argued above that more words do not automatically have to lead to more information on certain words and thus perhaps also in the total set, it can be assumed that the addition of surprisal values correlates with the length of the material studied. The more values are added, the larger the cumulative surprisal value can become. This would only not be the case if surprisal values are zero or negative, which would require perfect redundancy. However, this is not the case in languages (Shannon, 1948), which is why it is impossible to achieve a reduction in the cumulative surprisal value with an increase in length.

To reduce the influence of the length on the processing effort, the mean surprisal value must be calculated. The justification results from the calculation of the arithmetic mean value. A correlation between length and mean surprisal should no longer be found, length is practically factored out. Because the mean value is strongly influenced by outliers, an RC consisting of a few very surprising words could have a high average surprisal value which, according to our theory, should produce a higher processing effort and favor extraposition.

To illustrate, example (11a) shows an extraposed RC from 1,680 with a cumulative surprisal value of 24.13 but only six words, whereas (11b) shows an embedded RC from 1820 with 14 words and a cumulative surprisal value of 42.71, which is within the first quantile of cumulative surprisal values for embedded RC with more than 12 words.

(11a) so wird ein garftiges und schädliches Waffer herauflauffen
so will a nasty and harmful water out run

welches jederzeit rein abgewifcht
which always freshly wiped

werden muß
be must (Purmann, 1680, sentence 369)

"So a nasty and harmful water will run out which always has to be wiped clean."

b) [...] daß wahrend der Eroffnung des Muttermundes
that during the opening of the cervix

(deren allmahliges Vorfchreiten man am beften nach dem
of which gradual progress one the best after the

Durchmeffer der Oeffnung in
diameter of the opening in

Zollen beftimmt) gewohnlich die Rander des
inches measures usually the borders of the

Muttermundes kleine Einriffe erhalten
cervic small cracks get

(Gall, 1791, sentence 200)

"[...] that the borders of the cervix usually get small cracks while the cervix is opened the gradual process of which is measured best in inch according to the diameter of the opening."

Neither mean nor cumulative surprisal measurements have been previously used to explain RC extraposition. Both methods are somewhat interrelated and cannot be evaluated as better or worse suited to describe the processing effort for a construction purely based on preliminary theoretical considerations. Both involve the complete set of surprisal values, rather than focusing only on a local phenomenon and the increase or decrease of the likelihood of a word at that point. To find evidence for the previously postulated distinction between ID and length, a first section will evaluate some descriptive statistics before using linear regression (glm, R Core Team, 2018, Base-Package) to determine the best predictors.

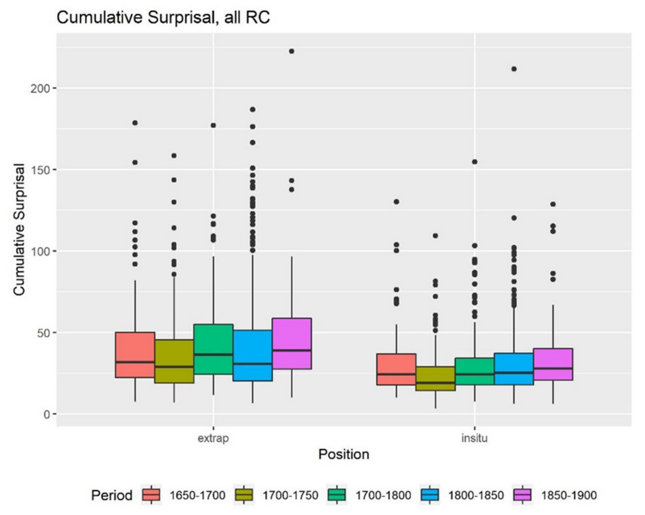


FIGURE 1 | Cumulative surprisal.

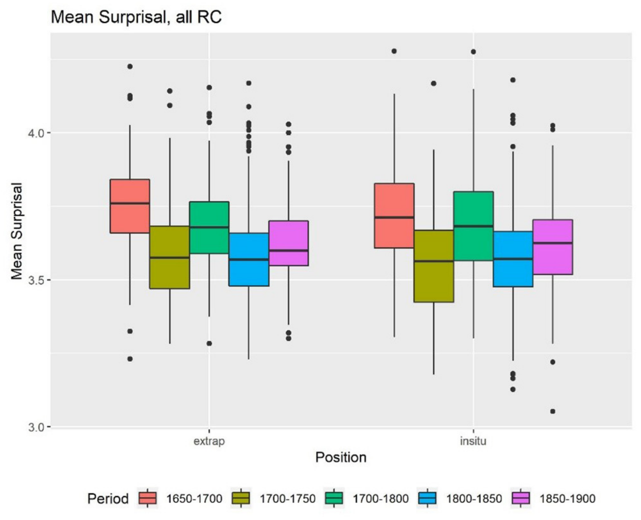


FIGURE 2 | Mean surprisal.

RESULTS: INFORMATION DENSITY AND LENGTH

Whole Data

The first factor, which is also most relevant for the hypothesis, is ID. First, we calculated the accumulated, mean Skip-gram surprisal values, and the length for all time periods. The descriptive statistics show that there is in fact a difference between the cumulative surprisal values of extraposed and embedded RC (**Figure 1**). In general, extraposed RC seem to have a higher cumulative surprisal value than embedded RC, which are labeled "*in situ*" in all graphs. The mean surprisal values for the RC in both positions do not appear to differ that much (**Figure 2**). In both cases, we find a lot of outliers but little differences within the centuries.

Problems arise when we check for the influence of length and the assumed correlation between length and cumulative surprisal values. The correlation value between length and cumulative surprisal values is 0.98, which suggests a very strong correlation. The longer an RC is, the higher are its surprisal values. But for the mean surprisal values, we do not find this correlation ($r =$ 0.00052). Thus, there is no correlation between the length of an RC and its mean surprisal value, and only an insignificant correlation between the two surprisal values ($r = 0.1283636$).

Checking the predictors using logistic regression (R Core Team, 2018)[13] and if writing styles (that is authors) do not influence extraposition, we only find an expected (cf. Section Methodological Considerations About RC Extraposition and ID) and slightly significant interaction between length and cumulative surprisal ($z = 2.571$, $p < 0.05$). All other predictors are not significant. In a second step, we removed the correlation between type and length, which does not change any of the parameters and does not lead to a better model. Next, the interaction between mean surprisal and length was removed, which presents the cumulative surprisal value as significant ($z = -2.417$, $p < 0.05$). The rational likelihood analysis conducted using ANOVA (R Core Team, 2018) shows that the model transformation is permissible. In the next step, we removed the interaction between cumulative surprisal and type, resulting in an interaction between cumulative and mean surprisal ($z = 1.982$, $p < 0.5$). The last interaction between type and mean surprisal was then cut along with the interaction between cumulative and mean surprisal value. The rational likelihood analysis granted this procedure as well. Our final model (**Table 3**) consists of the predictors cumulative surprisal ($z = -2.23$, $p < 0.05$), mean surprisal ($z = 1.511$, $p = 0.13$), length ($z = 0.56$, $p = 0.57$), type ($z = 1.74$, $p < 0.1$), and the interaction between cumulative surprisal and length ($z = 2.8$, $p < 0.001$). A further reduction of the model does not lead to a significantly better model. The interaction can be explained by the close connection between the calculation method and length. That makes it more difficult to determine whether length or surprisal is more influential. This result is interesting for several reasons. First, it shows a correlation between cumulative surprisal values and extraposition in a way we expected. But the influence of RC type, that is, restrictiveness contradicts previous statements in the literature. Restrictive RC are more likely to be embedded in our data whereas former research proposes the opposite.

TABLE 3 | Most influential effects in the final linear regression model (GLM) predicting position from surprisal values.

Predictor	*z*-value	*p*-value
Cumulative surprisal	−2.23	<0.05*
Type	1.74	<0.1
Cumulative surprisal: length	2.8	<0.001**

[13] Position ∼ (cumulative surprisal + mean surprisal + length+ restrictiveness)^2.

The removal of the interaction, though not covered by the rational likelihood analysis, lowers the p-values and marks length ($p = 0.38$) and mean surprisal ($p = 0.14$) as non-influential. If we drop length as well, cumulative surprisal seems to be the best predictor for extraposition ($z = -9.543$, $p < 0.001$), only followed by the RC type ($z = 1.74$, $p < 0.1$). We can therefore say that cumulative surprisal does seem to be highly correlated with extraposition, and we can therefore conclude that ID seems to be a better predictor than length. Still, one must be careful in making assumptions because this puts a time span of more than 200 years under consideration and the interaction in the model, which explains our data best, should not be forgotten. Therefore, the following sections will concentrate on the results for the 50-year timespan which were already used to calculate the Language Model. We can thus prevent the results from being skewed because of slightly imbalanced data.

RC per Period

Having established that ID seems to, indeed, have an influence on extraposition, the next step is to check whether this influence changes over the course of 250 years. Therefore, the corpus was split into five parts, each representing a 50-year timespan (**Table 4**).

In the first timespan (*1650–1700*), 240 RC were found, 116 (48%) of them, are extraposed. The cumulative surprisal values range from 6.697 to 242.55 with a mean of 42.07. Length is closely related to the cumulative surprisal value. The RC length differs between 3 and 58 words with a mean of 10.42. In the cases of very long RC with high cumulative surprisal values, the RC is not only complex in words but also in its grammatical complexity. The RC contains other subordinate clauses which are so closely linked to the content of the RC in question that it would be wrong to disregard the dependent subordinate clauses because this would not capture the whole message and its specific coding. This procedure was used for all other periods as well (12).

> 12) [...] wie ich offt der gleichen Patienten bekommen/[welche daß Schulterblat und den gantzen Arm voller Apoftemata gehabt/*[daß man es fchwerlich und mit groffer Muhe wieder zu rechte bringen konnen]*$_{\textit{dependend, subordinate clause}}$]$_{\text{RC|extraposed}}$
>
> “as I often had such patients [who had the scapula and the whole arm full of Staphylococcus-bacteria [so that it could hardy and with much effort be cured again]$_{\text{subordinate clause}}$]$_{\text{RC}}$”

As expected, the mean surprisal values do not have such a great variation. They only vary between 3.19 and 4.36 with a mean of 3.74. The distance between an extraposed RC and its head noun fluctuates between 1 and 10 with a mean of 2.38. It is interesting to notice that the material over which the RC is extraposed is mainly built by the RSB, one single constituent or one constituent, and the RSB. In the cases of a distance >4 words, we can still say that the RC is only moved over one constituent though this constituent contains a whole clause. Even when the head noun was in the prefield, only the sentence brackets and one other constituent interfered between it and the RC. In other cases, the large distance was caused by references when findings of other scientists were quoted. The distance was only calculated for extraposed RC. Thus, it is only included in the descriptive statistics because we are yet unable to reliably calculate the hypothetical distance over which embedded RC could be moved to land at the end of a clause due to the poor processing of DTA data and the uncertainty of clause boundaries as described in Section Annotation.

For the time span from 1700 to 1750, we find 680 RC in total, and 363 (53%) of them are extraposed ones. With 6.7, their smallest cumulative surprisal is slightly higher than the one from the 1650’s period while the largest cumulative surprisal is only 177.55 bits. Its mean is 34.99 bits. The closely related length varies between 2 and 50 with a mean of 9.7. The mean surprisal values differ from 3.19 to 4.41 with a mean of 3.67, and the distance varies between 1 and 17 with a mean of 2.08. Again, the large value of this variable is caused by interfering sentences such as parentheses. It becomes clear that the difference between the 1650’s and 1700’s RC is rather small. We find more RC, but their values mostly differ in the maximum cumulative surprisal value which might indicate a higher amount of information in RC in the late seventeenth century.

This changes again in the period of 1750 to 1800. We find slightly less RC with 375 and only 130 extraposed RC. That is the smallest percentage of RC in the whole corpus (35%). The smallest cumulative surprisal value is 7.17, the largest is 216.88, and the mean is 41.45. RC seems to be able to convey more information, compared to the previous period though not as much as in the first period. This is highly interesting because, at the same time, the range of length of RC decreases noticeably. Particularly, even the shortest RC contains six words while the longest on the other hand contains 13 words. The inner complexity of the RC decreases apparently in this period. The distance between the head noun and RC is smaller than in other periods as well. It ranges from 1 to 7 with a mean of 1.87. Once more, the mean surprisal values do not have a big variability. The smallest mean is 3.29, the biggest is 4.26, and the mean is 3.72.

The last two periods contain the highest number of RC. In the 1800 to 1850 period, 1,023 RCs were detected, among them 56% extraposed RC (573). The cumulative surprisal values range from 6.36 to 211.69 bits with a mean of 39.39. The length resembles the length of the early periods with a variety between 2 and 58, and an average of 10.42. The same holds for the distance between antecedent and RC. It varies again between 1 and 14. The longest distances are produced by interfering parentheses, clauses, and by references which were not excluded. The mean surprisal is rather constant again, ranging from 3.13 to 4.18.

The last period (1850 to 1900) contains 925 RC and 467 extraposed RC which corresponds to 50%. We find the second highest maximum cumulative surprisal values in this period (222.68) and the third highest minimal cumulative surprisal value (6.87). The average cumulative surprisal is 40.10. Another peak value is reached in the RC length, which ranges from 3 to 66 and achieves a mean of 11.58. The outlier RC of over 60 words is once more very complex and contains several dependent subordinate clauses. This period does not show any more extraordinary values in the distance which covers a span from 1 to 11 and is 1.74 words

TABLE 4 | Descriptive statistics.

Period	Number of RC (extraposed RC)	Min./Max. cumulative surprisal (mean)	Min./Max. mean surprisal (mean)	Min. /Max. length (mean)	Min./Max. distance (mean)
1650–1700	240 (116, 48%)	6.697/242.55 (42.066)	3.19/4.36 (3.74)	3/58 (10.42)	1/10 (2.38)
1700–1750	680 (363, 53%)	6.7/177.55 (34.99)	3.19/4.41 (3.67)	2/50 (9.7)	1/17 (2.08)
1750–1800	375(130, 35%)	7.17/216.88 (41.45)	3.291/4.260 (3.716)	3/37 (11.14)	1/7 (1.87)
1800–1850	1023 (573, 56%)	6.36/211.69 (39.39)	3.13/4.18 (3.57)	2/58 (10.42)	1/14 (1.78)
1850–1900	925 (467, 50%)	6.87/222.628 (40.10)	2.91/4.09 (3.62)	3/66 (11.58)	1/11 (1.74)

TABLE 5 | Most influential effects in the final GLM predicting position, 1650–1700.

Predictor	*z*-value	*p*-value
Cumulative surprisal	−2.669	<0.01**
Length	2.268	<0.05*

TABLE 6 | Most influential effects in the final GLM predicting Position, 1700–1750.

Predictor	*z*-value	*p*-value
Mean surprisal	−1.693	<0.1
Length	−3.961	<0.01**

long on average. The mean surprisal values vary between 2.91 and 4.09. The minimum mean surprisal value is the smallest in our corpus (2.91).

Having collected the data, the next step is to check which factor influences the RC position to which amount. The procedure for the regression analysis of the different timespans follows the procedure presented for the whole data. We included cumulative, mean surprisal, and the length of the material into a linear regression model (glm, R Core Team, 2018, Base Package)[14] and then conducted a backward model procedure using ANOVA (R Core Team, 2018). Restrictiveness was excluded since it was only marginally influential in the analysis of the whole data and could not be determined in many cases. Further explanations for the removal will be presented in section Discussion.

For the period **1650 to 1700**, the first model which includes all parameters and interactions does not show any significant predictors. This does not change until we remove all interactions and the mean surprisal values. Thus, the cumulative surprisal value is marginally significant ($z = -1.8$, $p < 0.1$) and claims that RC with higher cumulative surprisal values are more likely to be extraposed, whereas length is not only not significant but presents us with a value contradicting the idea that longer RC are placed in the post field (**Table 5**). Our data suggests the opposite. The first period, therefore, provides evidence for our first hypothesis.

The **period of 1700 to 1750** presents a slightly significant value for the mean surprisal values ($z = -1.71$, $p < 0.1$) in the model with all predictors. The backward model selection allows us to exclude the interactions between cumulative and mean surprisal, and the one between cumulative surprisal and length. The result improves the significance of the mean surprisal ($z = -2.076$, $p < 0.05$) and adds a slightly significant interaction between length and mean surprisal ($z = 1.718$, $p < 0.1$). Longer RC has higher surprisal values, but this interaction is only marginal. To remove this interaction from the model is possible, but the results will have insignificant values. Therefore, the interaction between mean surprisal and length is included in the model again, but the cumulative surprisal must be excluded.

The resulting model succeeds better in explaining the results. Having a model consisting of mean surprisal, length, and their interaction presents the following results: RC with a high mean surprisal value is more likely to be extraposed ($z = -2.147$, $p < 0.05$) and length gains in influence ($z = -1.686$, $p < 0.1$). The interaction shows a p-value over 0.1 now ($z = 1.541$, $p = 0.1234$). That is why the interaction is no longer included in the model. Our final model incorporates length and mean surprisal and is significantly better than a model without length ($p < 0.001$). Though mean surprisal values are still marginally influential ($z = -1.693$, $p < 0.1$), length is the best predictor for extraposition ($z = -3.961$, $p < 0.001$) in this case. This result stands in contrast to our finding for the first period and to our first hypothesis (**Table 6**). Further considerations on this period will be presented in Section Discussion.

The picture differs in the **period of 1750 to 1800**. As in the period of 1650 to 1700, the first model which incorporates all variables and interactions has no significant predictors. Models with interactions do not explain the phenomenon of extraposition sufficiently, and even the model with only length, cumulative and mean surprisal does not achieve this. We removed length as well in order to find a model which is able to explain the phenomenon. The result is highly significant for cumulative surprisal values. The higher the surprisal value the more likely the RC is to be extraposed ($z = -4.471$, $p < 0.001$). Mean surprisal values do not show this correlation ($z = 0.186$, $p = 0.052$). The backward model procedure shows that a model with mean surprisal does not explain the data significantly better ($p = 0.8079$). So, in this period, we find only a significant correlation between cumulative surprisal and extraposition and therefore evidence for the first hypothesis (**Table 7**).

[14] Position ∼ (cumulative surprisal + mean surprisal + length)^2.

TABLE 7 | Most influential effects in the final GLM predicting position, 1750–1800.

Predictor	*z*-value	*p*-value
Cumulative surprisal	−4.471	<0.01**
Mean surprisal	−0.186	<0.1

TABLE 8 | Most influential effects in the final GLM predicting position, 1800–1850.

Predictor	*z*-value	*p*-value
Cumulative surprisal	−5.474	<0.001***
Mean surprisal	0.853	=0.394

TABLE 9 | Most influential effects in the final GLM predicting position, 1850–1900.

Predictor	*z*-value	*p*-value
Cumulative surprisal	−1.8	<0.1
Mean surprisal	−1.736	<0.1
Cumulatvie surprisal: length	2.67	<0.05**

TABLE 10 | Most influential effects in the final GLM predicting position, after removing interactions, 1850–1900.

Predictor	*z*-value	*p*-value
Cumulative surprisal	−8.027	<0.001***
Mean surprisal	−1.835	<0.1

For the next **period of 1800 to 1850** similar findings can be presented. No variable in the model produces significant results when it is put in a model with all interactions or when the model includes all variables. As in the data from 1750 to 1800, we do not find significant results by incorporating cumulative and mean surprisal and length. In this model, length presents the highest p-value ($z = -0.096$, $p = 0.923$). Neither cumulative ($z = -0.397$, $p = 0.691$) nor mean surprisal ($z = 0.68$, $p = 0.492$) seem to be influential. We, therefore, exclude length and gain a model which presents a highly significant correlation ($z = -5.474$, $p < 0.001$) for the cumulative surprisal values and no correlation for mean surprisal ($z = 0.853$, $p = 0.394$). This slightly more complex model does not explain the data better than a model only including cumulative surprisal values. Again, we find evidence for our hypothesis: high cumulative surprisal values favor extraposition (**Table 8**).

The last **period (1850 to 1900)** is the first to present a significant interaction in the model with all variables and interactions. This interaction happens between cumulative surprisal values and length (2.057, $p < 0.05$). No other significant correlations or interactions are found. We, therefore, remove the interaction between mean surprisal and length. This reduces the interaction between cumulative surprisal and length to a slightly significant one (z =1.865, $p < 0.1$) and introduces a slightly significant cumulative surprisal value ($z = -1.768$, $p < 0.1$) as well. The following removal of the interaction between mean and cumulative surprisal values shows the influence of cumulative ($z = -1.8$, $p < 0.1$), mean surprisal ($z = 1.736$, $p < 0.1$), and a highly significant interaction between cumulative surprisal and length ($z = 2.67$, $p < 0.05$) (**Table 9**). A further reduction of the model does not lead to a model which explains the data any better. If we still take that step and exclude length, the only predictor in the model which does not show a significant correlation, the model results resemble those from other periods (**Table 10**) wherein cumulative surprisal is highly significant ($z = -8.027$, $p < 0.001$) and mean surprisal value marginally significant ($z = 1.835$, $p < 0.1$). But we must keep in mind that this model is not a significantly better model than the one including length and its interaction with cumulative surprisal values. In the last period, the influence of surprisal on extraposition seems to be only marginal but still stronger than the influence of length. But its strong interaction with cumulative surprisal might also influence these results.

We want to sum up our findings: For all periods except for the timespan 1700 to 1750, we find an influence of ID which exceeds the influence of length. For the timespan 1850 to 1900, our corpus does not allow a distinction between length and ID. Therefore, we must be careful with the data interpretation though removing length results in significant data for ID. All other periods provide evidence for our first hypothesis in which RC with higher cumulative surprisal values is more likely to be extraposed than RC with lower cumulative surprisal values. We can furthermore say that we also find evidence for the second hypothesis. ID does not lose its influence over time or at least until the late nineteenth century.

DISCUSSION

The research presented in this paper deals with the question of why RC is in the position they are found in, i.e., adjacent or extraposed. Using a corpus of RC from the late ENHG and early NHG, we investigated the frequently mentioned factors of length and restrictiveness of RC, on the one hand, and the ID of RC, on the other hand, to find out which factors are the most influential. ID was measured in this paper in terms of cumulative surprisal values based on a skip-gram Language Model.

The results of the investigation show that both types of RC occur in all investigated time periods. Also, the ratio of extraposed to embedded RC is balanced except for the period 1750–1800.

Looking at the factors for the positioning of RC, we find strong evidence for our hypothesis that high cumulative surprisal values are the strongest predictor for extraposition. This is in contrast to previous findings on RC extraposition being prevalent in literature.

Previous research on RC agrees that for both English and German, the length of the RC is the main criterion for whether it becomes extraposed or embedded (Shannon, 1992; Uszkoreit et al., 1998; Francis and Michaelis, 2012, 2014, 2017; Levy et al., 2012). The idea deals with the fact that longer relative clauses also influence the processability of the whole sentence. If they were placed in the middle field, their integration into the rest of the sentence would cause too much processing effort, which would jeopardize the processability of the sentence (Hawkins, 1994; Gibson, 1998). Length is thus synonymous with processing effort. While we cannot refute this idea, we can show that length does not directly equate to informativeness which is also highly connected to processing efforts [Levy (2008), among others]. We have shown in section Methodological Considerations About RC Extraposition and ID that, according to the concept of information theory, the information content of a word can be lowered by inserting further material into the sentence and thereby creating a drop in individual surprisal values on individual words. It is therefore possible to prevent very high surprisal values by increasing the sentence length and thus perhaps even reduce the overall processing effort. We showed, using German RC, that the information density of a sentence is a more meaningful approach to the extraposition of RC than sentence length.

In fact, a direct comparison shows that length predicts the position of the relative clause less well than information density. We found evidence for our hypothesis in general and showed that relative clauses with high cumulative surprisal values have a higher tendency to be extraposed than relative clauses with low cumulative surprisal values.

For the two time periods from 1700 to 1750 and from 1850 to 1900, however, further argumentation is needed to corroborate the hypothesis. The period from 1700 to 1750 is the only one that does not yield a significant result for the influence of ID on extraposition. Only length is a good predictor in this model. We attribute this result to the selection of the sub-corpus and the period itself. As our corpus only includes one text, Unzer (1746), the style of writing of the author determines the results. This author mainly uses RC with low informativeness but many words. Our hitherto unpublished analysis of other, albeit theological, texts from this period shows that length is not the main factor for extraposition. It can therefore be assumed that our result is at least partly due to the selection of the corpus for this period.

The time of text publication may be a reason. The sentence frame establishes itself in the eighteenth century and the justifications for post-field setting also begin to resemble those given for modern German (Konopka, 1996). Primarily, length and informational aspects such as the setting of two emphases are mentioned again in addition to dependency-related reasons such as the avoidance of too long distances. This is especially the case for middle fields that are too long when the distance between LSB and RSB becomes too great (Konopka, 1996, p. 131). This would argue for embedding short RC. Similar recommendations are also found among late seventeenth century grammarians, so one can conclude that this developmental process may have begun during this period. Therefore, the majority of the texts available to Unzer may have had rather short middle fields without long relative clauses with little information content, which may have influenced his own writing style. Nevertheless, even this does not fully clarify the facts found. Other research also shows an influence of length in earlier and later periods, which we cannot show. Of course, this in turn may also be influenced by the text type, which remains to be verified. It must be said that it is highly probable that the deviations in the period 1700–1750 are due to a weakness in the corpus selection and that further checks are therefore necessary.

The second time period for which an influence of the ID cannot be shown in the final analysis is the last in the corpus (1850–1900). Here, we found no correlation between length and extraposition. However, the interaction between length and cumulative surprisal cannot be excluded from the model without significantly degrading it. Therefore, it cannot be clearly concluded whether the cumulative surprisal value of a relative clause or its length exerts a stronger influence on extraposition. Yet, both surprisal calculation methods (mean and cumulative surprisal) exert a marginal influence on extraposition, while length with a p-value of 0.97 can be ruled out as an influencing factor in the combination. The influence thus seems to definitely be present, but it cannot be completely decoupled from the length. On the one hand, this could be an indication that length does have a decisive influence on the extraposition process and that the results of, e.g., Uszkoreit et al. (1998) would be just as confirmed in studies of modern texts as those of Levy et al. (2012) among others for English. We must, however, refer to the still insufficient research situation. Whether a change is actually initiated in the late nineteenth century would become clear if the same result could be reproduced for later texts.

Apart from these two periods, our results are very clear and provide strong evidence for our first hypothesis: Extraposition and embedding are influenced by the ID of the RC.

This observation is integrated into already existing theories of information density. High information content is co-indicated with processing difficulties [Levy (2008) among others]. This approach is also intuitively understandable. If a sentence contains a lot of information, it is more strenuous to understand it. Therefore, it is important to encode the complex content in a way that keeps the processing effort as small as possible otherwise the transmitted information might be lost. In the case of the RC studied here, this is done by moving them to another position in the sentence. According to the theories of Hawkins (1994) and Gibson (1998), this results in more free cognitive capacity since the matrix sentence to the RC has already been fully processed. It should be noted here that our Language Models only pick up the lexical information of the words in the RC since they have been trained on the lemmata. Grammatical information could not be included in the consideration of the RC extraposition due to the already mentioned bad POS tagging of the DTA texts. Grammatical information could bring an additional dimension, since not only the lexical information has to be processed, but also the parts of speech behind it could be included in the consideration. For example, it has already been shown that the insertion of a function word can weaken the information content of the following content word (Jaeger, 2010).

These observations from other studies (Jaeger, 2005, 2010; Frank and Jaeger, 2008) are closely related to the UID (Levy and Jaeger, 2007). However, the UID was mainly considered in case of local changes in the information profile. The most famous example is the reduction of the optional *that* [Levy and Jaeger (2007) among others], the presence of which leads to a too low information content on the onset of the RC. Such a differentiated approach to the UID is not possible with our data. Due to different spellings and the specific subject matter of the texts, a considerable number of words is still not contained in the lexicon of the Language Model (see section Corpus and Method), so that local approaches within the RC are not possible in a meaningful way.

Previous studies on the occupation of the postfield [Speyer (2011), Sapp (2014), Coniglio and Schlachter (2015) on even older language stages of German] report a decreasing influence of information structure on the occupation of the PoF. Interaction between information structure and ID can be assumed (Speyer and Lemke, 2017) but has also not yet been studied in detail for German. If we now look at the values available here from 1650 to 1850 and exclude the time period 1700–1750, this impression could also be somewhat confirmed with regard to ID. Although ID measurements are significant or even highly significant in each case, a minimal decreasing tendency can nevertheless be detected.

There are also factors that have proven to have little or no influence. These include, contrary to the opinion of the literature, the restrictiveness of RC. For Modern German, it is assumed that restrictive RC can be better extraposed than non-restrictive RC. The reason given for this is that RC is necessary to clearly identify its antecedent. The RC is, in other words, expected because the design of the head noun makes the presence of RC highly probable. The assumed surprisal for the construction should be small, even if it occurs later in the sentence. The indication for restrictiveness does not automatically allow predictions about the content of the RC[15]. The predictions about the position of the RC made in Hypothesis 1 still carry weight and the RC is extraposed in a more unpredictable content.

In our study, the data, as a whole, shows only a marginal influence of relative clause type on extraposition ($p < 0.1$). Since, in some cases, we could not determine the type with absolute certainty, as reported in section Information Density, there is a discrepancy between the level of knowledge of the annotators and the possible world knowledge of the text authors, which often led to the type not being determined. The possibilities for error in the determination are therefore present and not negligible. Moreover, the value tends to indicate that restrictive RC is embedded, which would contradict the existing literature on the correlation between extraposition and type. Correlations between cumulative or mean surprisal and type were also not found. So, even the marginal correlation that could be found cannot be attributed to processing effort. However, the expectation regarding the relative clause could be more due to grammatical factors, as already suggested above, and less to lexical content. In fact, the arguments regarding the extraposition of restrictive RC are never about whether the content is expectable. Only the existence of the RC is described as necessary. It could, therefore, also be worth combining part-of-speech with the lexical surprisal values for this partial aspect.

Before the final summary, we will take a brief look at the second hypothesis. We proposed that ID as a principle is valid over all time steps. In fact, there is no change in its influence on the RC position, except for the period of 1700–1750 we discussed previously. At least with the help of our calculation methods, it can be concluded that information density seems to have a constant influence on the design of sentences in early NHG. Also, the presence of other styles, such as the Latin syntax, which authors of scientific articles may have been familiar with does not influence the design of German sentences in a way that should violate the principles proposed by the ID. Efficient processability of sentences is a basic principle of sentence design at all time levels. We can therefore also consider our second hypothesis as confirmed.

CONCLUSION

We conclude that ID, measured as cumulative surprisal, is the best way to predict the position of a relative clause in the present corpus of medical texts from the seventeenth to nineteenth centuries. The length which was previously said to be the most influential predictor for extraposition can only present its influence in one period. This finding might be attributed to a poor choice of sub-corpus and should therefore be treated with caution. The same holds for restrictiveness. This factor does not yield significant results on the basis of this corpus. Furthermore, ID is a stable influencing factor in all time stages and can therefore be called a universal principle for the design of sentences even in earlier stages of German.

DATA AVAILABILITY STATEMENT

The annotated data set is available at https://github.com/SFB1102/C6Samples.

AUTHOR CONTRIBUTIONS

SV did research and writing. AS did an advisory contribution. Both authors contributed to the article and approved the submitted version.

FUNDING

This paper is based on material collected in the context of the CRC, Information Density and Linguistic Encoding, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–Project-ID 232722074–SFB 1102. Most of the material was presented at the Workshop *Rational Approaches in Language Science (RAILS) 2019*.

ACKNOWLEDGMENTS

Our thank goes to the organizers and the audience of this workshop for many helpful comments, furthermore to Katharina Wilhelm, Nils Dörr, and Katrin Ortmann for helping with the corpus creation and tools.

[15] The determiner *derjenige* (“that one”), for example, expresses the need for a restrictive RC but does not allow any conclusions about the content of that RC, because *derjenige* is semantically neutral.

REFERENCES

Abel, H. K. (1699). *Wohlerfahrner Leib-Medicus der Studenten*. Leipzig: Friedrich Groschuff.

Admoni, V. G. (1990). *Historische Syntax des Deutschen*. Tübingen: Niemeyer.

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* 47:10201. doi: 10.1177/00238309040470010201

Bader, M. (2014). *Defining Distance in Language Production*. Frankfurt: Ms. Goethe Universität Frankfurt.

Behagel, O. (1932). *Deutsche Syntax. Band 4*. Heidelberg: Carl Winters Universitätsbuchhandlung.

Birkner, K. (2008). *Relativ(satz)konstruktionen im gesprochenen Deutsch. Syntaktische, prosodische, semantische und pragmatische Aspekte*. Berlin: de Gruyter.

Carus, C. G. (1820). *Lehrbuch der Gynäkologie. Bd. 1*. Leipzig.

Chafe, W. (1976). "Givenness, contrastiveness, definiteness, subjects, topics and point of view," in *Subject and Topic*, ed C. N. Li (New York, NY: Academic Press).

Coniglio, M., and Schlachter, E. (2015). "Das Nachfeld im Deutschen zwischen Syntax, Informations- und Diskursstruktur," in *Das Nachfeld im Deutschen*, ed H. Vinckel-Roisin (Berlin: de Gruyter), 8. doi: 10.1515/9783110419948-008

Cutler, A. (2008). The abstract representations in speech processing. *Quart. J. Exp. Psychol.* 61, 1601–1619. doi: 10.1080/13803390802218542

Drach, E. (1937). *Grundgedanken der deutschen Satzlehre*. Frankfurt a.M.: Diesterweg.

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., et al. (2016). "A web-based tool for the integrated annotation of semantic and syntactic structures," In: *Proceedings of the LT4DH workshop at COLING 2016*. Osaka, Japan.

Eisenberg, P. (1999). *Grundriss der deutschen Grammatik*. Stuttgard/Weimar: Metzler. doi: 10.1007/978-3-476-03765-7

Francis, E. J., and Michaelis, L. A. (2012). "Effects of weight and definiteness on speakers' choice of clausal ordering in english," in *Lsa Annual Meeting Extended Abstracts, Vol. 3* (Washington, DC). doi: 10.3765/exabs.v0i0.577

Francis, E. J., and Michaelis, L. A. (2014). Why move? how weight and discourse factors combine to predict relative clause extraposition in English. *Compet. Motiv. Gram. Usage* 5, 70–87. doi: 10.1093/acprof:oso/9780198709848.003.0005

Francis, E. J., and Michaelis, L. A. (2017). When relative clause extraposition is the right choice, it's easier. *Lang. Cogn.* 9, 332–370. doi: 10.1017/langcog.2016.21

Frank, A., and Jaeger, F. T. (2008). "Speaking rationally: uniform information density as an optimal strategy for language production," in *Proceedings of the 30th Annual Meeting of the Cognitive Science Societey*.

Fritsch, W. J. (1990). *Gestalt und Bedeutung der deutschen Relativsätze*. München: Uni-Druck.

Gall, F. J. (1791). *Philosophisch-medizinische Untersuchungen über Natur und Kunst im kranken und gesunden Zustand des Menschen*. Wien.

Gallmann, P. (2005). *Der Satz. In: Duden. Die Grammatik 4*. Mannheim: Bibliographisches Institut and F.A. Brockhaus, 763–1056.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68:1. doi: 10.1016/S0010-0277(98)00034-1

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., et al. (2019). How efficiency shapes human language. *Trends Cogn. Sci.* 23:5. doi: 10.1016/j.tics.2019.09.005

Guthrie, D., Allison, B., Liu, W., Guthrie, L., and Wilks, Y. (2006). "A closer look at skip-gram modelling," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

Hale, J. (2001). "A probabilistic Early parser as a psycholinguistic model," in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computional Linguistics* (Pittsburgh, PA). doi: 10.3115/1073336.1073357

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

Helbig, G., and Buscha, J. (2001). *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. Berlin: Langenscheidt.

Hentschel, E., and Weydt, H. (2003). *Handbuch der deutschen Grammatik*. Berlin: de Gruyter.

Jaeger, T. F. (2005). "Optional that indicates production difficulty: evidence from disfluencies," in *Proceedings of DiSS'05. Disfluency in Spontanous Speech Workshop*. Aix-en-Provence.

Jaeger, T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cogn. Psychol.* 61:2. doi: 10.1016/j.cogpsych.2010.02.002

Jurafsky, D. (2003). "Probabilistic modeling in psycholinguistics: linguistic comprehension and production," in *Probabilistic Linguistics*, eds R. Bod, J. Hay, and S. Jannedy (Cambridge, MA: MIT Press).

Koch, P., and Oesterreicher, W. (2007). Schriftlichkeit und kommunikative Distanz. *ZGL* 35:24. doi: 10.1515/zgl.2007.024

Koch, R. (1878). *Untersuchung über die Aetiologie der Wundinfectionskrankheiten*. Leipzig. doi: 10.5962/bhl.title.101427

Konopka, M. (1996). *Strittige Erscheinungen der deutschen Syntax im 18. Jahrhundert*. Tübingen: Niemeyer. doi: 10.1515/9783110940039

Kraepelin, E. (1892). *Ueber die Beeinflussung einfacher psychischer Vorgänge durch einige Arzneimittel*. Jena.

Krifka, M. (2007). Basic notions of information structure. *Interdiscipl. Stud. Inform. Struct.* 6.

Lehmann, C. (1984). *Der Relativsatz: Typologie seiner Strukturen, Theorie seiner Funktionen, Kompendium seiner Grammatik*. Tübingen: Narr.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106:6. doi: 10.1016/j.cognition.2007.05.006

Levy, R., Fedorenko, E., Fedorenko, E., Breen, M., and Gibson, E. (2012). The processing of extraposed structures in English. *Cognition* 12, 12–36. doi: 10.1016/j.cognition.2011.07.012

Levy, R., and Jaeger, F. (2007). Speakers optimize information density through syntactic reduction. *Adv. Neural Inform. Process. Syst.* 19.

Ludwig, C. (1852). *Lehrbuch der Physiologie des Menschen. Bd. 1*. Heidelberg.

Manning, C., and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

McClelland, J. L., and Elman, J. L. (1986). The TRACE model of speech perception. *Cogn. Psychol.* 18, 1–86. doi: 10.1016/0010-0285(86)90015-0

Mürmann, M. (2014). *Wahrscheinlichkeitstheorie und stochastische Prozesse*. Berlin: Springer Spektrum. doi: 10.1007/978-3-642-38160-7

Nübling, D., Dammel, A., Duke, J., and Szczepaniak, R. (2013). *Historische Sprachwissenschaft des Deutschen. Eine Einführung in die Prinzipien des Sprachwandels*. Tübingen: Narr.

Ostermann, S. (2020). *Script knowledge for natural language understanding*. Saarbrücken. doi: 10.22028/D291-31301

Paul, H. (2007). *Mittelhochdeutsche Grammatik, 25 Edn*. Tübingen: Niemeyer. doi: 10.1515/9783110942354

Pfeifer, W. (1995). *Etymologisches Wörterbuch des Deutschen*. München: dtv.

Polenz, P. V. (2010). *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart*. Berlin: de Gruyter.

Poschmann, C., and Wagner, M. (2016). Relative clause extraposition and prosody in German. *Natural Lang. Linguist. Theory* 2016:8. doi: 10.1007/s11049-015-9314-8

Prince, E. F. (1981). "Toward a taxonomy of given-new information," in *Radical Pragmatics*, ed P. Cole (New York, NY: Academic Press, Inc.), 223–255.

Purmann, M. G. (1680). *Der rechte und wahrhafftige Feldscher*. Halberstadt.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/ (accessed October 12, 2021).

Reil, J. C. (1803). *Rhapsodieen über die Anwendung der psychischen Curmethode auf Geisteszerrüttungen*. Halle.

Sahel, S. (2015). "Zur Ausklammerung von Relativsätzen und Vergleichsphrasen im frühen Neuhochdeutschen (1650–1800)," in *Das Nachfeld im Deutschen*, ed Vinckel-Roisin (Berlin: de Gruyter), 9. doi: 10.1515/9783110419948-009

Sapp, C. (2014). Extraposition in middle and new high German. *J. Comparat. Germanic Linguist.* 17:6. doi: 10.1007/s10828-014-9066-6

Schildt, J. (1976). "Zur Ausbildung der Satzklammer," in *Zur Ausbildung der Norm in der deutschen Literatursprache auf der syntaktischen Ebene*, eds G. Kettmann and J. Schildt (Berlin: Akademieverlag), 235–284.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Technical J.* 27:tb01338. doi: 10.1002/j.1538-7305.1948.tb01338.x

Shannon, T. F. (1992). "Towards an adequate characterization of relative clause extraposition in modern German," in *On Germanic Linguistics: Issues and Methods*, ed I. Rauch (Berlin: de Gruyter).

Speyer, A. (2011). Die Freiheit der Mittelfeldabfolge im Deutschen - ein modernes Phänomen. *Beiträge Geschichte Deutschen Sprache Literatur* 133, 14–31.

Speyer, A. (2016). "Die Entwicklung der Nachfeldbesetzung in verschiedenen deutschen Dialekten," in *Zeitschrift für Dialektologie und Linguistik*. Beiheft, 165.

Speyer, A., and Lemke, R. (2017). *Information Density as a Factor for the Embedding of Relative Clauses*. Saarland University. Available online at: https://arxiv.org/abs/1705.06457 (accessed October 12, 2021).

Takada, H. (1998). *Grammatik und Sprachwirklichkeit von 1640-1700. Zur Rolle deutscher Grammatiker im schriftsprachlichen Ausgleichsprozeß*. Tübingen: Niemeyer. doi: 10.1515/9783110952223

Unzer, J. A. (1746). *Gedanken vom Einfluß der Seele in ihren Körper*. Halle (Saale).

Uszkoreit, H., Brants, T., Duchie, D., Krenn, B., Konieczny, L., Oepen, S., et al. (1998). Studien zur Performanzorientierten Linguistik. Aspekte der Relativsatzextraposition im Deutschen. *Kognitionswissenschaft* 7:65. doi: 10.1007/s001970050065

Vinckel-Roisin, H. (2015). "Facetten des Nachfelds im Deutschen," in *Das Nachfeld im Deutschen: Theorie und Empirie*, ed H. Vinckel-Roisin. doi: 10.1515/9783110419948

Voigtmann, S., and Speyer, A. (forthcoming). "Information density as a factor for systematic variation in Early New High German," in *Proceedings of Linguistic Evidence 2020* (Tübingen).

Wöllstein, A. (2014). *Topologisches Satzmodell, 2nd Edn*. Heidelberg: Winter.

Zifonun, G., Hoffmann, L., and Strecker, B. (1997). *Grammatik der deutschen Sprache. Bd 2*. Berlin: 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.

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 Psychology

ORIGINAL RESEARCH
published: 03 March 2022
doi: 10.3389/fpsyg.2021.672927

A Bayesian Approach to German Personal and Demonstrative Pronouns

Clare Patterson[1]*, *Petra B. Schumacher*[1], *Bruno Nicenboim*[2], *Johannes Hagen*[1] *and Andrew Kehler*[3]

[1] *Department of German Language and Literature I, Linguistics, University of Cologne, Cologne, Germany,* [2] *Department of Cognitive Science and Artificial Intelligence, Tilburg School of Humanities and Digital Sciences, Tilburg University, Tilburg, Netherlands,* [3] *Department of Linguistics, University of California San Diego, La Jolla, CA, United States*

When faced with an ambiguous pronoun, an addressee must interpret it by identifying a suitable referent. It has been proposed that the interpretation of pronouns can be captured using Bayes' Rule: P(referent|pronoun) ∝ P(pronoun|referent)P(referent). This approach has been successful in English and Mandarin Chinese. In this study, we further the cross-linguistic evidence for the Bayesian model by applying it to German personal and demonstrative pronouns, and provide novel quantitative support for the model by assessing model performance in a Bayesian statistical framework that allows implementation of a fully hierarchical structure, providing the most conservative estimates of uncertainty. Data from two story-continuation experiments showed that the Bayesian model overall made more accurate predictions for pronoun interpretation than production and next-mention biases separately. Furthermore, the model accounts for the demonstrative pronoun *dieser* as well as the personal pronoun, despite the demonstrative having different, and more rigid, resolution preferences.

Keywords: pronouns, demonstratives, Bayesian model, prominence, reference

Edited by:
Gerhard Jäger,
University of Tübingen, Germany

Reviewed by:
Jennifer E. Arnold,
University of North Carolina at Chapel Hill, United States
Stefan Kaufmann,
University of Connecticut, United States

Correspondence:
Clare Patterson
cpatters@uni-koeln.de

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

Received: *26 February 2021*
Accepted: *10 December 2021*
Published: *03 March 2022*

Citation:
Patterson C, Schumacher PB, Nicenboim B, Hagen J and Kehler A (2022) A Bayesian Approach to German Personal and Demonstrative Pronouns. Front. Psychol. 12:672927. doi: 10.3389/fpsyg.2021.672927

INTRODUCTION

The interpretation of anaphoric pronouns has provided a puzzle for many decades of linguistic research. Third-person anaphoric pronouns such as "she" in (1) are inherently ambiguous in that there are no rigid rules to determine the antecedent. The puzzle for the addressee, then, when faced with a pronoun, is to identify a suitable referent. Despite the ambiguity, this puzzle is solved with ease most of the time: in (1), for example, most people would assume that "she" refers to "the lawyer."

(1) The lawyer fascinated the judge. She was always so well prepared.

Despite this ease of interpretation, it has proven difficult to accurately describe how pronouns are resolved. It has, however, been possible to identify a range of individual factors which seem to influence resolution; for instance, there is evidence that referents mentioned from subject position are preferred to those mentioned from other positions (Crawley and Stevenson, 1990; Crawley et al., 1990; Gordon et al., 1993; Järvikivi et al., 2005); that referents mentioned first are preferred to those mentioned later (Clark and Sengul, 1979; Gernsbacher and Hargreaves, 1988; Järvikivi et al., 2005); that referents with an agentive thematic role are preferred to those with a patient thematic role (Stevenson et al., 1994; Schumacher et al., 2016); that referents which are topics are preferred to non-topics (e.g., Grosz et al., 1995). These factors also often overlap: in (1), "the lawyer" is both a subject and is mentioned first. The way in which individual factors work together, allowing the addressee to identify the correct referent, however, is still debated.

Describing pronoun resolution as a process that is influenced by a variety of factors allows us to describe certain general tendencies in the language, and can also give insights into the functions of pronouns. But it does not allow us to make precise quantitative predictions about how an addressee will interpret a pronoun in any given context. It is possible to come up with counter-examples for every factor listed above, and influencing factors can be overridden, or at least attenuated, by world knowledge or by coherence relationships between clauses or sentences.

A quite different approach to pronoun interpretation has been taken by Kehler et al. (2008) and Kehler and Rohde (2013). They put forward a simple probabilistic model, the Bayesian model for pronouns, which to a large extent sidesteps the (combination of) individual factors affecting pronoun resolution. Instead, the model makes predictions about how an addressee will interpret a pronoun in a particular linguistic context, by combining the *next-mention bias* with the *production bias*, as described below. Factors influencing the pronoun interpretation do so only indirectly, through their influence on either of the next-mention or production biases (or both).

According to the Bayesian model, addressees reverse-engineer speakers' intended referents following Bayesian principles:

$$P(referent|pronoun) = \frac{P(pronoun|referent)P(referent)}{\sum_{referent \in referents} P(pronoun|referent)P(referent)} \quad (2)$$

The posterior term P(referent|pronoun) represents the pronoun *interpretation bias*: upon hearing a pronoun (e.g., she), the probability that the addressee will resolve it to a particular referent. The likelihood term P(pronoun|referent) represents the pronoun *production bias*: the probability of the speaker choosing to use a pronoun to refer to an intended referent. Finally, the prior term P(referent) denotes the *next-mention bias*: the probability that a specific referent gets mentioned next by the speaker, regardless of the form of referring expression that they choose. According to this model, therefore, the interpretation and production models are not mirror images of each other, nor is there a simple combination of influencing factors. Instead, pronoun interpretation biases result from an addressee integrating their "top-down" predictions about the content of the ensuing message (particularly, who gets mentioned next) with the "bottom-up" linguistic evidence (particularly, the fact that the speaker opted to use a pronoun).

The performance of the Bayesian model – how well its predictions match actual interpretations – has been compared to the performance of two competing models derived and extended from the existing literature (Ariel, 1990; Gundel et al., 1993; Grosz et al., 1995; Arnold, 1998; inter alia; see Rohde and Kehler, 2014 for discussion). The first we refer to as the Expectancy model, according to which the addressee's interpretation bias toward a referent is (their estimate of) the probability that the referent is mentioned next in the context. The Expectancy model is inspired by Jennifer Arnold's claim that a referent's accessibility is influenced to a considerable extent by the hearer's estimate of the likelihood that it will be mentioned in the upcoming discourse (Arnold, 1998, 2001). Arnold further developed this insight into the Expectancy Hypothesis (Arnold et al., 2007; Arnold, 2010; Arnold and Tanenhaus, 2011). Arnold (2010) in particular suggests:

> *Under the communicative goal of referring, a plausible mechanism for expectancy is as a mechanism for discourse participants to coordinate accessibility. Expectancy describes how easily the comprehender will be able to retrieve the referent. Speakers could thus calculate expectancy as an estimate of accessibility to the listener.* (p. 193).

This characterization, which is couched in terms of reference production, does not go so far as to claim that pronoun comprehension can be completely equated to the next-mention bias, but it suggests that next-mention bias is a strong influencing factor on the accessibility or activation of a referent, and that this in turn should facilitate pronoun resolution. Our "Expectancy model" instead tests whether the next-mention bias alone guides the predicted interpretation bias, where the next-mention bias P(referent) is normalized by the probabilities of all possible referents that are consistent with the morphological features of the pronoun (e.g., gender, number). This model is mathematically expressed below using the assignment operator to emphasize the fact that this model does not follow normative probability theory.

$$P(referent|pronoun) \leftarrow \frac{P(referent)}{\sum_{referent \in referents} P(referent)} \quad (3)$$

The second competing model is what we call the Mirror model, according to which the interpretation bias toward a referent is proportional to the likelihood of the referent being pronominalized by the speaker, i.e., the *production bias*. Once again, the assignment operator in (4) reflects the fact that this model does not follow normative probability theory.

$$P(referent|pronoun) \leftarrow \frac{P(pronoun|referent)}{\sum_{referent \in referents} P(pronoun|referent)} \quad (4)$$

This model captures the idea that addressees will assign interpretations to pronouns by asking what entities the speaker is most likely to refer to using a pronoun instead of a competing referential form. The model is an operationalization of the assumption that pronoun production and pronoun comprehension coordinate on the same notion of entity prominence: that addressees reverse-engineer the speaker's referential intentions by estimating how likely the speaker is to use a pronoun for a particular referent given its perceived prominence in the discourse context. These estimates therefore rely on a strong correspondence between the form of a referential expression (pronoun, full noun phrase) and the accessibility of its referent. Though this assumption is not often explicitly stated in the psycholinguistics literature, it underlies the treatment of reference production scales being direct representations of mental states, from which assumptions can be made about the salience or accessibility of certain referents (e.g., Ariel, 1990; Gundel et al., 1993). This intuition is cached out by taking the addressee's estimate of the probability that a speaker will produce a pronoun for a particular referent, normalized by the sum of the probabilities for all compatible referents.

In the current study we assess the performance of the Bayesian model against the two competing models outlined above. Of particular importance is the novel quantitative method used for this assessment. Another novel aspect of this study is the extension of the Bayesian model to German demonstrative pronouns. These pronouns differ from personal pronouns in their resolution biases and therefore provide a good test of the generalizability of the Bayesian model. Furthermore, we go beyond previous assessments of the Bayesian model by testing not only implicit causality verbs (Experiment 2) but also dative-experiencer versus accusative verbs (Experiment 1), in order to explore the influence of grammatical versus thematic roles, which has implications for claims about the strong version of the Bayesian model. Below, we first introduce the strong version of the Bayesian model, and then go on to summarize previous quantitative assessments of the Bayesian model and highlight advantages of the current approach. We then present relevant background on German personal and demonstrative pronouns before stating the study aims.

Strong Bayesian Model

The primary claim of the Bayesian model is the central prediction underlying equation (2): that comprehenders reverse-engineer the speaker's referential intentions using Bayesian principles. That is, rather than interpreting pronouns by coordinating with the speaker via a single notion of entity prominence, comprehenders must engage with two types of prominence, one which underlies their estimates of the speaker's production biases (as captured by the likelihood) and one which underlies their estimates of the next-mention bias (as captured by the prior). It therefore predicts that if independent estimates of the prior, likelihood, and posterior probabilities are obtained, the equation in (2) would approximately hold. We refer to this claim as the *weak* form of the Bayesian model. The model has been successful, for instance, at explaining why in certain contexts, pronoun production biases strongly favor the subject but interpretation biases are more equivocal between potential referents (Source–Goal transfer-of-possession contexts) or even favor the grammatical object (object-biased implicit causality verbs; see Kehler and Rohde, 2013 for discussion).

Kehler et al. (2008) and Kehler and Rohde (2013) also suggested a STRONG version of the Bayesian model, in which the two terms in the numerator of (2) are conditioned by different types of contextual factors. On the one hand, early data had suggested that factors conditioning the next-mention bias P(referent) are primarily semantic and pragmatic in nature (e.g., verb type and coherence relations). On the other hand, the factors that condition the production bias P(pronoun|referent) appear to be grammatical and/or information structural (e.g., based on grammatical role obliqueness or topichood, both of which amount to a preference for sentential subjects). As alluded to above, the resulting prediction, therefore, is that a speaker's decision about whether or not to pronominalize a referent will be insensitive to a set of semantic and pragmatic contextual factors that the addressee will nonetheless bring to bear via the influence of the prior on interpretation.

Perhaps in the light of the strong, counterintuitive dissociation it posits, it has been the predictions of the strong form of the Bayesian hypothesis that have received the most attention in the literature. Whereas early studies have provided evidence to support it (Rohde, 2008; Fukumura and van Gompel, 2010; Rohde and Kehler, 2014; inter alia), some more recent studies, primarily by Arnold and colleagues, have found limited effects of semantic factors (thematic roles) on production (Rosa and Arnold, 2017; Zerkle and Arnold, 2019; Weatherford and Arnold, 2021; see also Arnold, 2001). These contradictory findings leave us with the looming questions of what the source of the disparities are, and of what type of model can explain the extant data as an ensemble, especially given that the identified effects of semantic factors on production are typically more limited or otherwise inconsistent than theories that rely on a singular notion of entity prominence would predict. It is not the goal of our work to settle this (big) question, but instead to add a new set of facts to the debate by examining the predictions of both the weak and strong models with respect to German personal and demonstrative pronouns.

Quantitative Assessment of the Bayesian Model

Rohde and Kehler (2014) present the first quantitative evaluation of the Bayesian model against the two competing models (Mirror and Expectancy). They conducted two story-continuation experiments. We describe the method and the materials in detail here, since they are relevant for several aspects of the current study. In a story-continuation experiment, participants are presented with incomplete text passages which they are asked to complete, like those shown in (5) and (6).

(5) a. John scolded Bill. __________
b. John infuriated Bill. __________
c. John chatted with Bill. __________

(6) a. John scolded Bill. He __________
b. John infuriated Bill. He __________
c. John chatted with Bill. He __________

Participants complete the passages, and judges then annotate their continuations. The examples in (5) are the FREE-PROMPT conditions, where just a blank line is presented and participants need to supply the entire sentence. The first referential expression in the participant's completion is annotated for reference (whether it refers to John or Bill or neither). The form of the referential expression is also annotated, that is, whether the expression itself is a pronoun, a full NP or some other expression. From the annotations of reference in the free-prompt condition, the next-mention bias can be calculated. From the annotation of form combined with reference information, the pronoun production bias for a particular referent (e.g., *John*) can be calculated. From the free-prompt data, then, predictions for all three models described in the previous section can be derived. The PRONOUN-PROMPT conditions are shown in (6). Here, instead of a blank line, a pronoun is presented in first position and the participant supplies the rest of the sentence. In these conditions the reference for the pronoun is annotated, yielding the actual interpretation bias for the pronoun. As such, the models' predicted interpretation bias as derived from the free-prompt data can be compared against actual interpretation bias measured from the pronoun-prompt data.

Using this method, Rohde and Kehler (2014; see also Rohde, 2008) tested whether the next-mention bias (i.e., the prior) and production bias (i.e., the likelihood) were sensitive to semantic biases arising from implicit causality (IC), that is, they tested the strong form of the Bayesian model. For example, a subject-biased IC verb such as *infuriate* as in (5b/6b) implies that the subject *John* is the cause of the infuriation event, while an object-biased IC verb such as *scold* as in (5a/6a) implies that the object *Bill* is the cause of the scolding event. In Rohde and Kehler's experiment, the IC verbs were compared to neutral (non-IC) verbs such as *chat with* as in (5c/6c). As predicted by the strong Bayesian hypothesis, the verb type affected both the next mention biases in the free condition (5) and the pronoun interpretation biases in the pronoun-prompt condition (6), with subject mentions in both prompt conditions being most frequent for subject-biased IC contexts, least frequent for object-biased IC contexts, and in between for non-IC controls. However, the difference in subject next-mention rate was not coupled with a difference in pronominalization rates for subject next-mentions in the free-prompt conditions. Instead, only the grammatical role of the referent's previous mention mattered: participants pronominalized references to the previous subject far more often than ones to the previous non-subject. To put a fine point on this, participants were no more likely to pronominalize a mention of the previous object in an object-biased IC context like (5a) than in a subject-biased IC one like (5b), and similarly no more likely to pronominalize a mention of the previous subject in a subject-biased context (5b) than in an object-biased one (5a).

For both experiments, predictions per participant and per item for the Bayesian, Mirror and Expectancy models were generated as described above. These predictions were correlated against per participant and per item actual observations from the pronoun-prompt condition and the correlations were evaluated using R^2. While the predictions of all the models were significantly correlated with the observed data, the Bayesian model consistently produced the strongest correlations.

Zhan et al. (2020) were able to improve on the assessment of model performance presented in Rohde and Kehler (2014) by combining R^2 with MSE and ACE metrics. MSE and ACE weigh different aspects of model performance; while ACE reflects discrepancies between predicted and observed behavior at extreme values, MSE reflects discrepancies throughout the range of values. A downside of their approach, however, is that the predictions are based on point estimates and do not take into account the uncertainty in the data. The measures of discrepancy ignore the inherent noisiness of the data that were used to make model predictions and might give overoptimistic estimates as a result[1]. In the analysis presented in this paper, we used Bayesian methods that propagate the uncertainty in the data to the predictions. Rather than point-values, we predict distributions of possible values. The width of the prediction distribution depends on the uncertainty (or variability) present in the data. This approach thus makes a new contribution to the assessment of pronoun interpretation models.

Cross-Linguistic Support for the Bayesian Model

The Bayesian model for pronouns has, for the most part, been developed and tested on English (Kehler et al., 2008; Kehler and Rohde, 2013; Rohde and Kehler, 2014), while cross-linguistic support is only now starting to emerge (Bader and Portele, 2019; Zhan et al., 2020). While there is nothing about the model's mechanics that make it specific to one language, it remains to be seen whether claims associated with the model are applicable in other languages. Zhan et al. (2020) tested subject-biased and object-biased IC verbs using the same story continuation task as Rohde and Kehler (2014). They replicated the effect of verb type on the next-mention bias and the effect of grammatical role (and not verb type) on the pronoun production biases, in line with the strong Bayesian model. Furthermore, their results also indicated that grammatical role rather than topichood affects the pronoun production biases, in contrast to Rohde and Kehler (2014).

It is also important to test the model in different pronoun systems. This was not a feature of the Zhan et al., study; while Mandarin Chinese has both null and overt pronouns, they appear to have largely overlapping resolution preferences. It is possible, for example, that the Bayesian model is better suited to making predictions for pronouns whose interpretation is quite flexible. It remains to be seen whether a pronoun with more rigid preferences can be accounted for equally well. We address this question by testing the Bayesian model on the German personal pronoun *er* and the demonstrative pronoun *dieser*. Below, we briefly outline the relevant properties of these pronouns and also consider the findings of Bader and Portele (2019), who incorporated aspects of the Bayesian model into their study on the German demonstrative *der*. We then set out the goals of this paper before reporting our experiments.

German Personal and Demonstrative Pronouns

German personal pronouns, for example *er* ("he"), are quite similar to English personal pronouns, but unlike English they can be used to refer to both animate and inanimate entities.

[1] This is similar to what happens when data are averaged for a *t*-test in comparison to using the "raw" data in a linear mixed model.

In addition, German has a rich set of demonstratives that can be used pronominally, for example *der*, *dieser*, *jener*, *derjenige*. When functioning as pronominals (as opposed to adnominals, e.g., *dieser Mann* "this man"), these demonstratives can refer to animate or inanimate entities just like personal pronouns[2].

When referring to animate entities, German personal and demonstrative pronouns tend to differ regarding both interpretative preferences and their influence on maintenance and shift of the sentence topic (see Schumacher et al., 2015, 2016; Portele and Bader, 2016; Fuchs and Schumacher, 2020). Most previous research on interpretive preferences has looked at *der* compared to *er*, while *dieser* has received far less attention. It has been claimed that the personal pronoun *er* has a bias toward subject referents (Bosch et al., 2003, 2007; Bouma and Hopp, 2006, 2007) while *der* has been described as object-biased (Kaiser, 2011) and as having an anti-topic bias (Bosch and Umbach, 2007; Wilson, 2009; Hinterwimmer, 2015; Bosch and Hinterwimmer, 2016). Nonetheless, the personal pronoun appears to be quite flexible; the demonstrative *der*, on the other hand, seems to be less flexible (Kaiser, 2011; Schumacher et al., 2015, 2016, 2017; Bader and Portele, 2019). Patil et al. (2020) examined the demonstrative *dieser* and found an anti-subject preference; they proposed that *dieser* is the formal counterpart of *der*. The contrast in flexibility of interpretation between personal and demonstrative pronouns allows us to explore whether the Bayesian model, in which the prior can move biases around, can also be applied to a more "rigid" pronoun.

Bader and Portele (2019), in a series of story-continuation experiments, found that subjecthood had the strongest impact on interpretation of *er*, while interpretation of *der* was influenced to some extent by subjecthood, topichood and linear order. They also used their data to assess the predictions of the Bayesian model. In a separate experiment participants were presented with the items from the first two experiments with just the free-prompt for story completion[3]. However, the experimental materials were more complex than in previous story-continuation experiments (e.g., Rohde and Kehler, 2014; Zhan et al., 2020), because items started with a context sentence in which a (feminine) referent was introduced before the critical sentence containing the two (masculine) entities which were potential referents for the pronouns tested. While the entity in the context sentence was not a potential referent for the pronoun in the pronoun-prompt conditions, it was nevertheless referred to in 49% of completions in the free-prompt condition (i.e., when the prompt contained no pronoun). This introduced an imbalance in the available observations. In fact, P(referent) was calculated using all observations (including references to the entity in the context sentence and to both entities) while the sum of production probabilities used in the Bayesian calculation was only from NP1 and NP2. We suspect this may have led to an imbalance in the calculation of predictions for the Bayesian model. They report a high R^2 value (0.95) for the correlation between predicted and observed values[4], but we think that this result should be interpreted with caution. Performance of competing models (Expectancy and Mirror) were not reported.

One further aspect of *der* (and *dieser*) demonstratives that should be highlighted is the potential role of agentivity. A series of studies by Schumacher et al. (2015, 2016, 2017) and Fuchs and Schumacher (2020) has shown that agentivity is an important factor for personal and demonstrative pronouns in German. This has been shown by contrasting verbs in which thematic roles and grammatical roles align with verbs in which they are not aligned. For example, in accusative verbs such as *ärgern* "annoy," agentivity and grammatical role are aligned because the subject of the verb has the proto-agent role and the object the proto-patient role[5]. In contrast, in dative-experiencer verbs such as *imponieren* "impress," agentivity and grammatical role are not aligned because the object has the proto-agent role and the subject has the proto-patient role (note also that in canonical order the object, not the subject, is in initial position). In other words, the grammatical role hierarchy (subject > object) and thematic role hierarchy (proto-agent > proto-patient) are aligned in the accusative verbs and not aligned in the dative-experiencer verbs. Pronoun interpretation in these experiments was affected to a greater degree by agentivity than by grammatical role, with personal pronouns tending to refer to the proto-agent and demonstratives to the proto-patient. Given this finding, we decided to exploit this verb-type contrast to explore the relative influence of agentivity and subjecthood on production biases in German. While the strong form of the Bayesian model specifies that subjecthood and/or topichood influences production likelihoods (Rohde and Kehler, 2014; Zhan et al., 2020), it is possible that in German agentivity also has an influence, in the light of Schumacher and colleagues' findings about the influence of agentivity on interpretation[6].

For the current study, we chose to focus on the demonstrative *dieser* as opposed to *der* for two reasons. First, *dieser* is better suited to a written experiment than *der*, which is perceived by some speakers to be slightly pejorative and is more appropriate in spoken, possibly less formal, contexts[7]. This

[2] In order to refer to propositional content (for example an aforementioned sentence) speakers of German use the neuter form of pronouns (*das*, *dies*), similar to English *this* and *that* (see Çokal et al., 2018).

[3] This methodology differs from previous story completion experiments testing the Bayesian model, because different sets of participants took part in the pronoun-prompt and free-prompt tasks.

[4] Predictions and observations were on a per-condition/pronoun basis rather than an item and/or participant basis, so a total of 16 observation pairs were used for the correlation.

[5] We follow Dowty's (1991) use of proto-roles.

[6] It should be noted, however, that Rohde and Kehler (2014) found no influence of thematic role on production likelihoods when testing active versus passive structures.

[7] Wiemer (1996, p. 85) indicates that pejorative use is a potential additional function of the *der*-type pronoun. Bethke (1990, p. 72) points out that the *der*-type pronoun is not only used in negatively connoted situations and claims that the pejorative use results from other linguistic and contextual factors. Corpus research reports very few cases of *der* with (mild) pejorative connotations: e.g., in the course books of Eurolingua 2 out of 936 instances of the *der*-type pronoun are pejorative (Ahrenholz, 2007, p. 338). The pejorative connotation might further be intertwined with contrast. Sometimes *dieser* is also associated with pejorative use.

is supported by Bader and Portele's (2019) experiments in which *dieser* was elicited far more frequently than *der* in the free-prompt conditions. Second, little is known about general interpretive preferences for *dieser* since most previous studies have looked at *der*; descriptions of *dieser* in German grammars are brief and empirically inadequate (but see Fuchs and Schumacher, 2020 for a recent comparison of *der* and *dieser*). It would therefore be useful to expand our understanding of how *dieser* differs from the personal pronoun in German.

CURRENT STUDY

The purpose of the current study is to assess the performance of the Bayesian model on German personal and demonstrative pronouns, in order to address the following main questions:

- Which model for pronouns (Bayesian, Expectancy or Mirror) best accounts for the interpretation of German personal and demonstrative pronouns?
- Is the resolution of demonstratives as rigid as some previous studies suggest, or is the interpretation influenced by the next-mention bias, as the Bayesian model would predict?
- Is there evidence for the strong form of the Bayesian model?

In the following, we present two text completion experiments that address these questions. We use the free-prompt data to generate predictions for the Bayesian, Mirror and Expectancy models and compare the predictions to the observations from the pronoun-prompt conditions. Model predictions are generated in a Bayesian statistical framework with a fully hierarchical structure and weakly informative priors. The hierarchical structure allows us to accommodate, for example, participant and item effects directly in our model predictions without having to average over them. In contrast to previous evaluations of model performance, the Bayesian statistical approach allows us to estimate the parameters of a distribution of predicted observations, allowing us to make more stable inferences about model performance.

EXPERIMENT 1

Experiment 1 was a text continuation task testing the next-mention, production and interpretation biases associated with the German personal pronoun *er* and the demonstrative pronoun *dieser*, in contexts with accusative verbs and dative-experiencer verbs. In addition to addressing the main questions set out above, our motivation for the verb-type contrast was to explore the relative contribution of agentivity and subjecthood to the production likelihoods. A strong influence of agentivity would be seen in higher pronoun production likelihoods for proto-agents than for proto-patients for personal pronouns, and the opposite pattern for demonstratives. Proto-agents are the first NP (henceforth NP1) in both verb types. A strong influence of subjecthood, in contrast, would result in higher personal production likelihoods for the grammatical subject, which is NP1 for accusative verbs and NP2 for the dative verbs. For the demonstrative, a grammatical role influence would result in higher production likelihoods for NP2 in accusative verbs and NP1 in dative verbs.

Participants

Fifty nine participants from the University of Cologne took part in Experiment 1. Nine participants were excluded because they did not complete the experiment (less than 75% of items completed); one participant was excluded for not following the task instructions and one participant was excluded for lack of German knowledge. Data from the remaining 48 participants (39 female, 7 male, 2 gender not indicated) were used in the analysis. All 48 participants indicated that they were German native speakers; 7 participants were bilingual. No participants reported language-related disorders.

Materials

Seventy two critical items were constructed, each in three prompt conditions: *er*, *dieser* or a free-prompt (blank line); see (7) and (8). A full list of items and fillers is available on OSF[8]. Critical items consisted of a context sentence followed by the prompt. The context sentences consisted of a main clause with two masculine animate arguments, starting with an adjunct (e.g., *vorletzte Nacht* "the night before last"). The main verb in the context sentences was either an accusative or a dative-experiencer verb (henceforth "dative"), always in the perfect tense (comprising a form of *sein* "to be" or *haben* "to have" plus a participle). Context sentences were always presented in canonical argument order (proto-agent before proto-patient, i.e., nominative–accusative for the accusative verbs and dative–nominative for the dative verbs). The 36 accusative items contained 36 different verbs, but the 36 dative items were limited to just four verbs which were re-used[9].

(7) Accusative items:

(a) *Er prompt:* Nach dem Fußballspiel hat der Franzose den Italiener gesehen. Er _________
(b) *Dieser prompt:* Nach dem Fußballspiel hat der Franzose den Italiener gesehen. Dieser _________
(c) *Free-prompt:* Nach dem Fußballspiel hat der Franzose den Italiener gesehen. _________

"After the football game the Frenchman *(nom.masc.)* saw the Italian *(acc.masc.)*. *He/DEM/...*"

(8) Dative items:

(a) *Er prompt:* Gestern ist dem Feuerwehrmann der Polizist aufgefallen. Er _________
(b) *Dieser prompt:* Gestern ist dem Feuerwehrmann der Polizist aufgefallen. Dieser _________
(c) *Free-prompt:* Gestern ist dem Feuerwehrmann der Polizist aufgefallen. _________

[8] http://osf.io/j5wtg

[9] This is because the number of dative verbs in German is restricted: in previous experiments (Fuchs, 2021) only four dative verbs (*gefallen* "to please," *missfallen* "to displease," *auffallen* "to notice," and *imponieren* "to impress") were interpreted correctly and hence used in the present experiment.

"Yesterday the firefighter *(dat.masc.)* noticed the police officer *(nom.masc.)*. *He/DEM/...*"

Role names (e.g., *Polizist* "police officer") with masculine gender were used for both entities introduced in the context sentence in all but two items, in which animals (also masculine) were used. Hierarchical relationships between the two roles (such as teacher–pupil) were avoided to prevent a prominence confound. The pronouns in the pronoun-prompt conditions always matched in gender with the entities in the context sentence so that both were potential referents for the pronoun. Note that feminine pronouns/referents were not tested, because the feminine personal pronoun *sie* in German is ambiguous in terms of case and number.

The 72 item-sets were mixed with 30 "true" fillers (25% gender-ambiguous, 50% gender-disambiguated and 25% items with one referent only) and 6 "catch" fillers (included to ensure that participants were paying attention to the task), and distributed over three lists in a Latin-square design. Ten of the fillers contained target sentences that began with a temporal adverbial (five items) or connector (five items), and ten contained target sentences with a connector followed by an auxiliary and a pronoun. Another ten filler items comprised personal or demonstrative pronoun-prompts in the style of the critical items. Two of the demonstrative filler items, which were presented among the first ten items, included an auxiliary or adverb after the pronoun-prompt (*Dieser ist* ____, "He.dem is"; *Trotzdem haben diese dann* ____, "Nevertheless they.dem have then") which forces a pronominal reading of the demonstrative. The aim was to prime the participants to produce a pronominal, as opposed to an adnominal, use of the demonstrative (Bader and Portele, 2019, reported very low uses of the demonstrative pronoun in the free-prompt condition, and Kaiser, 2011, reports 75.6% completions with an adnominal use of the demonstrative).

Procedure

The lists were presented to participants in a seminar setting as a paper questionnaire comprising 108 items. The first page contained study information and a consent form. Participants then answered a short series of biographical questions before starting the experimental task. Participants were instructed to complete every short story by supplying the second sentence, without making changes to the text presented. They were additionally instructed that the most obvious completion should be written and not the most creative or humorous one, and that completions should be kept short and precise.

Data Coding

The data was coded by two native speakers of German; one Linguistics Masters student and one technical assistant. Coder 1 identified missing and ungrammatical continuations which were excluded from the analysis. Both annotators made independent judgments about the intended referent of the first referential expression (in the pronoun-prompt conditions, the first referential expression was always the pronoun given in the prompt, i.e., *er* or *dieser*). The referent for the first referential expression was coded in five categories: NP1, NP2, both, neither, ambiguous. The two annotators agreed in 77% of observations, with a Cohen's (unweighted) Kappa of 0.669 ($z = 70.8$, $p \leq 0.001$). Observations where the annotators disagreed were resolved through discussion to produce a final data set for analysis. The first referential expression in the free-prompt data was also categorized. Data from 48 participants for 24 items (12 accusative, 12 dative) per prompt condition resulted in a total of 3456 observations; 1152 in each of the *er*-, *dieser*- and free-prompt conditions. The distribution of reference and the response categories for the first referential expression are given in **Supplementary Material**. For the following analyses, the dataset was reduced by dropping cases that were missing, ungrammatical, and references that were ambiguous, plural, complex (referring to a whole event or proposition), or where no referential expression occurred or cases where the first expression was an impersonal pronoun, leaving a total of 2390 observations for the analysis (679 free-prompt, 858 *er*-prompt and 853 *dieser*-prompt).

Data Analysis

For the data analysis and modeling, we use a Bayesian data analysis approach implemented in the probabilistic programming language *Stan* (Stan Development Team, 2020) in *R* (R Core Team, 2020)[10]. An important motivation for using the Bayesian approach is that it allows us to implement a fully hierarchical structure to any type of model (e.g., the so-called "maximal random effect structure"); a hierarchical structure provides the most conservative estimates of uncertainty (Schielzeth and Forstmeier, 2008). In all our models, we use regularizing priors, which we detail below. These priors are minimally informative and have the objective of yielding stable inferences (Gelman et al., 2008; Chung et al., 2015). Nicenboim and Vasishth (2016) and Vasishth et al. (2018) discuss the Bayesian approach in detail in the context of psycholinguistic and phonetic sciences research. We fit the models with four chains and 4000 iterations each, of which 1000 iterations were the burn-in or warm-up phase. In order to assess convergence, we verify that there are no divergent transitions, that all the $\hat{R}$ (the between- to within-chain variances) are close to one, that the number of effective sample size are at least 10% of the number of post-warmup samples, and we visually inspect the chains.

As we detail below, the models fit the produced referents (NP1 or NP2, discarding the ambiguous or other referents) with a Bernoulli likelihood, where its parameter θ is fitted in log-odds space, and/or the produced pronoun type (personal or demonstrative pronoun, or other expressions) with a categorical likelihood. The probability of a personal pronoun and "other" with respect to the reference category, demonstrative pronoun, is also fitted in log-odds space (that is, the categorical likelihood is composed of two equations that contrast the odds of

[10] We used: R (Version 4.0.3; R Core Team, 2020) and the R packages *bayesplot* (Version 1.7.2; Gabry et al., 2019), *cmdstanr* (Version 0.3.0; Gabry and Češnovar, 2020), *dplyr* (Version 1.0.2; Wickham et al., 2020), *ggplot2* (Version 3.3.3; Wickham, 2016), *kableExtra* (Version 1.3.1; Zhu, 2020*)*, *loo* (Version 2.3.1.9000; Vehtari et al., 2017; Yao et al., 2017), *matrixStats* (Version 0.57.0; Bengtsson, 2020), *posterior* (Version 0.1.3; Vehtari et al., 2020*)*, *purr* (Version 0.3.4; Henry and Wickham, 2020), and *tidyr* (Version 1.1.2; Wickham, 2020).

producing a personal pronoun or other expression instead of the reference category, demonstrative pronouns). For more details about categorical or multinomial logistic regression see Koster and McElreath (2017). For both the Bernoulli and the categorical regressions, we assume a hierarchical structure composed of an intercept denoted by α, a number of slopes denoted by β, and a number of by-participant and by-item adjustments to the intercept and slope, u and w, respectively. All these parameters have the following weakly regularizing priors:

- The intercepts of the Bernoulli (α) have priors in probability space: $logit^{-1}(\alpha) \sim Beta(1,1)$.
- The intercept of the equations in the categorical regression (α) have $Normal(0, 2)$ priors.
- All the slopes (β) have as a prior $Normal(0, 2)$.
- All the variance components of the by-group adjustments (or random effects) are $Normal_{+}(0, 2)$.
- The correlations between by-participants and by-items adjustments have each $lkj(2)$ as a prior.

For each model we report the mean estimates and 95% quantile-based Bayesian credible intervals of the main parameters. A 95% Bayesian credible interval has the following interpretation: it is an interval containing the true value with 95% probability given the data and the model (see, for example Jaynes and Kempthorne, 1976; Morey et al., 2016). We evaluate the fit of models graphically with holdout predictive check, and numerically using holdout validation (Vehtari and Ojanen, 2012). Crucially, we evaluate the performance of the different models with respect to their predictive accuracy on new data that is *never used to estimate the parameters*. An advantage of model comparison based on hierarchical Bayesian models is that the uncertainty of the models' parameters is propagated to the predictions that they make: This means that instead of point predictions, the models generate a distribution of predictions. For holdout validation, we compare the models based on their pointwise log predictive density[11].

Results

Raw proportions for the next-mention bias are shown in **Figure 1**. **Figure 2** shows the personal pronoun production likelihoods, and **Figure 3** the demonstrative pronoun production likelihoods[12]. When calculating likelihoods for the personal pronoun, both subject and non-subject personal pronouns were included. For the demonstrative pronoun, both subject and non-subject demonstrative pronoun *dieser*, and subject and non-subject demonstrative pronoun *der*, were included.

Modeling

Expectancy Model

The Expectancy model predicts that the probability of referring to NP1 in the pronoun-prompt data is determined by the prior probability of NP1 ($P(referent = NP1)$). This prior can be estimated from the free-prompt data. The Expectancy model was built in the following way and its parameters were estimated using only the free-prompt data:

$$\eta_i = \alpha_{NP1} + u_{NP1}\left[subj\_free[i]\right] + w_{NP1}\left[item\_free[i]\right]$$
$$+ vtype[i] \cdot \left(\beta_{vtype} + u_{vtype}\left[subj\_free[i]\right]\right)$$
$$P(NP1|\ldots) = P\big(referent = NP1|item\_free[i], subj\_free[i],$$
$$vtype_i\big) = logit^{-1}(\eta_i)$$
$$NP1_i \sim Bernoulli\,(P(NP1|\ldots)) \tag{9}$$

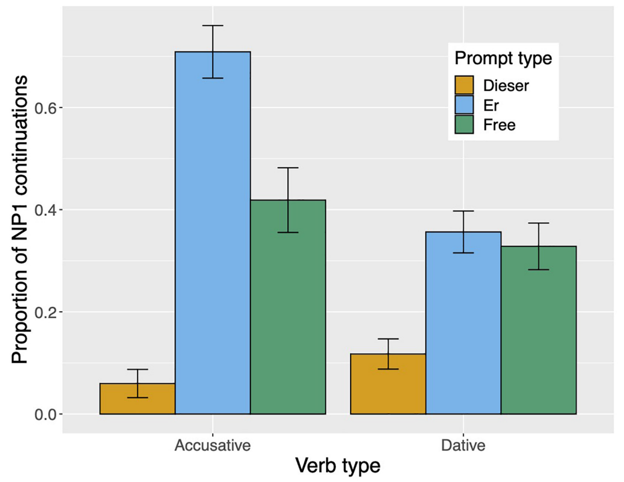


FIGURE 1 | Proportion of NP1 continuations per prompt condition (*er, dieser, and free*) and verb type in Experiment 1. The NP1 continuations represent the next-mention bias for NP1. Error bars represent 95% confidence intervals based on normalized data following Morey (2008) and Chang (2013).

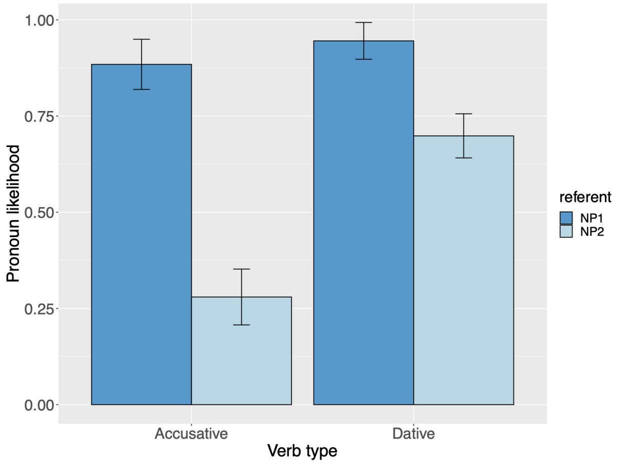


FIGURE 2 | Personal pronoun: probability of using a personal pronoun to refer to NP1 and NP2, per verb type, for Experiment 1. These bars represent the pronoun production likelihoods (and are therefore based on the free-prompt data). Error bars are by-participant and represent 95% confidence intervals based on normalized data following Morey (2008) and Chang (2013).

[11] The pointwise log predictive density is proportional to the MSE if the model is normal with constant variance, but it is also appropriate for models that are not normally distributed (Gelman et al., 2013, ch. 7).

[12] The tabulated data can be found in **Supplementary Material**.

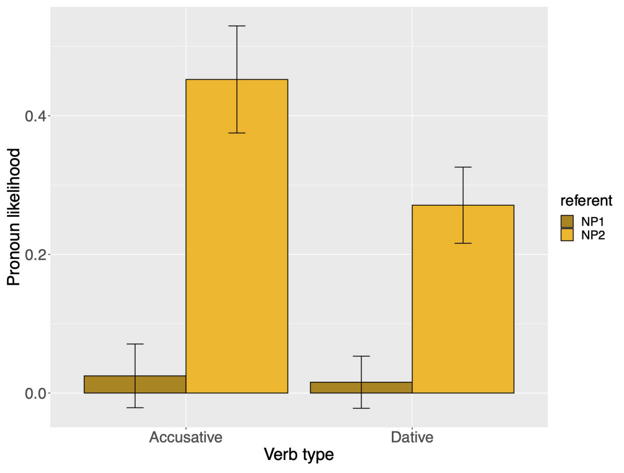


FIGURE 3 | Demonstrative pronoun: probability of using a demonstrative pronoun to refer to NP1 and NP2, per verb type, for Experiment 1. These bars represent the pronoun production likelihoods (and are therefore based on the free-prompt data). Error bars are by-participant and represent 95% confidence intervals based on normalized data following Morey (2008) and Chang (2013).

where *NP1* is 1 if the referent is NP1 and 0 if the referent is NP2, *i* indicates the observation of the free-prompt data, *vtype* is a vector that maps between observations and the corresponding verb type (accusative coded with 1 or dative coded with −1), *subj_free* and *item_free* are vectors that indicate the mapping between observations, and subjects and items, respectively, and *u* and *w* are the by-subject and by-items adjustments (or "random effects"). The three dots (...) symbolize all the information that the model is taking into account to estimate the probability of producing NP1 as a referent: the characteristics of the stimuli (i.e., intercept, beta, and by-item adjustments) and of the subject performing the free-prompt task (i.e., by-subject adjustments).

The parameters estimated with the free-prompt data were used to generate predictions for the pronoun-prompt data in the following way:

$$\begin{aligned}
&\eta_n = \alpha_{NP1} + u_{NP1}\left[subj\_pron[n]\right] + w_{NP1}\left[item\_pron[n]\right]\\
&+ vtype[n] \cdot \left(\beta_{vtype} + u_{vtype}\left[subj\_pron[n]\right]\right)\\
&P\left(NP1|\ldots\right) = P\left(referent = NP1|item\_pron[n], subj\_pron[n],\right.\\
&\left.vtype_n\right) = logit^{-1}\left(\eta_n\right)\\
&pred_{NP1_n} \sim Bernoulli\left(P(NP1|\ldots)\right)
\end{aligned} \tag{10}$$

where *n* indicates the observation of the pronoun-prompt data, *subj_pron* and *item_pron* are vectors that indicate the mapping between observations for subjects and items, respectively, and *u* and *w* are the by-subject and by-items adjustments. As before, the three dots (...) symbolize all the information that the model is taking into account to generate the predictions: the characteristics of the stimuli (i.e., intercept, beta, and by-item adjustments) and of the subject performing the pronoun-prompt task (i.e., by-subject adjustments).

Mirror Model

The Mirror model predicts that the probability of referring to *NP1* for pronoun-prompt data is determined by the *likelihood* of NP1 (*P(pronoun|referent = NP1)*) normalized to be a probability distribution by dividing the likelihood by the marginal probability distribution of the pronouns. This normalized *likelihood* can be estimated from the free-prompt data. The Mirror model was built in the following way:

$$\begin{aligned}
\log(\frac{\theta_{PP_i}}{\theta_{DP_i}}) &= \alpha_{PP} + u_{PP}[subj\_free[i]] + w_{PP}[item\_free[i]]\\
&+ vtype[i] \cdot (\beta_{PP,vtype} + u_{PP,vtype}[subj\_free[i]])\\
&+ ref\_free_i \cdot (\beta_{PP,ref} + u_{PP,ref}[subj\_free[i]]\\
&+ w_{PP,ref}[item\_free[i]]) + vtype[i] \cdot ref\_free_i \cdot (\beta_{PP,int}\\
&+ u_{PP,int}[subj\_free[i]] + w_{PP,int}[item\_free[i]])\\
log(\frac{\theta_{DP_i}}{\theta_{DP_i}}) &= 0\\
log(\frac{\theta_{other_i}}{\theta_{DP_i}}) &= \alpha_{other} + u_{other}[subj\_free[i]]\\
&+ w_{other}[item\_free[i]] + vtype[i] \cdot (\beta_{other,vtype}\\
&+ u_{other,vtype}[subj\_free[i]]) + ref\_free_i \cdot (\beta_{other,ref}\\
&+ u_{other,ref}[subj\_free[i]] + w_{other,ref}\ [item\_free[i]])\\
&+ vtype[i] \cdot ref\_free_i \cdot (\beta_{other,int}\\
&+ u_{other,int}[subj\_free[i]] + w_{other,int}[item\_free[i]])\\
pron_i &\sim Categorical(\theta_{PP_i}, \theta_{DP_i}, \theta_{other_i})
\end{aligned} \tag{11}$$

where *pron* is 1 if the free completion includes a personal pronoun, 2 if it includes a demonstrative pronoun, and 3 otherwise; *i* indicates the observation of the free-prompt data, *vtype* is vector that maps between observations and the corresponding verb type (accusative coded as 1 or dative coded as −1), *ref_free* indicates whether the referent of the completion is NP1 (coded with 1) or NP2 (coded with −1), and, just as for the Expectancy model, *subj_free* and *item_free* are vectors that indicate the mapping between the observations and subjects or items, respectively, and *u* and *w* are the by-subject and by-items adjustments. The parameters estimated with the free-prompt data were used to generate predictions for each observation *n* of the pronoun-prompt data as described below.

First, the likelihood of each referent is calculated. To simplify the equations, we define:

$$\begin{aligned}
P\left(PP|NP1, \ldots\right) &= P\left(pronoun = PP|referent = NP1,\right.\\
&\left.subj\_pron[n], item\_pron[n], vtype[n]\right)\\
P\left(PP|NP2, \ldots\right) &= P\left(pronoun = PP|referent = NP2,\right.\\
&\left.subj\_pron[n], item\_pron[n], vtype[n]\right)\\
P\left(DP|NP1, \ldots\right) &= P\left(pronoun = DP|referent = NP1,\right.\\
&\left.subj\_pron[n], item\_pron[n], vtype[n]\right)\\
P\left(DP|NP2, \ldots\right) &= P\left(pronoun = DP|referent = NP2,\right.
\end{aligned}$$

$$\begin{aligned}
&\quad subj\_pron[n], item\_pron[n], vtype[n]\big)\\
P\big(other|NP1, \ldots\big) &= P\big(pronoun = other|referent = NP1,\\
&\quad subj\_pron[n], item\_pron[n], vtype[n]\big)\\
P\big(other|NP2, \ldots\big) &= P\big(pronoun = other|referent = NP2,\\
&\quad subj\_pron[n]\ ,item\_pron[n], vtype[n]\big)\\
P\left(NP1|PP, \ldots\right) &= P\big(referent = NP1|pronoun = PP,\\
&\quad subj\_pron[n]\ ,item\_pron[n], vtype[n]\big)\\
P\left(NP1|DP, \ldots\right) &= P\big(referent = NP1|pronoun = DP,\\
&\quad subj\_pron[n]\ ,item\_pron[n], vtype[n]\big)
\end{aligned} \tag{12}$$

$$\begin{aligned}
&< P(PP|NP1, \ldots), P(DP|NP1, \ldots), P(other|NP1, \ldots) >\\
&= softmax(\\
&\quad \alpha_{PP} + u_{PP}[subj\_pron[n]] + w_{PP}[item\_pron[n]]\\
&\quad + vtype[n] \cdot (\beta_{PP,vtype} + u_{PP,vtype}[subj\_pron[n]])\\
&\quad + (\beta_{PP,ref} + u_{PP,ref}[subj\_pron[n]] + w_{PP,ref}[item\_pron[n]])\\
&\quad + vtype[n]\ \cdot\ (\beta_{PP,int} + u_{PP,int}[subj\_pron[n]]\\
&\quad + w_{PP,int}[item\_pron[n]]),\\
&\quad 0,\\
&\quad \alpha_{other} + u_{other}[subj\_pron[n]] + w_{other}[item\_pron[n]]\\
&\quad + vtype[n] \cdot\ \big(\beta_{other,vtype} + u_{other,vtype}[subj\_pron[n]]\big)\\
&\quad +\ \big(\beta_{other,ref} + u_{other,ref}[subj\_pron[n]]\\
&\quad + w_{other,ref}[item\_pron[n]])\\
&\quad + vtype[n]\ \cdot\ (\beta_{other,int} + u_{other,int}[subj\_pron[n]]\\
&\quad + w_{other,int}[item\_pron[n]]\\
&\quad)
\end{aligned} \tag{13}$$

$$\begin{aligned}
&< P(PP|NP2, \ldots), P(DP|NP2, \ldots), P(other|NP2, \ldots)\ >\\
&= softmax(\\
&\quad \alpha_{PP} + u_{PP}[subj\_pron[n]] + w_{PP}[item\_pron[n]]\\
&\quad + vtype[n] \cdot (\beta_{PP,vtype} + u_{PP,vtype}[subj\_pron[n]])\\
&\quad + (-1) \cdot (\beta_{PP,ref} + u_{PP,ref}[subj\_pron[n]]\\
&\quad + w_{PP,ref}[item\_pron[n]])\\
&\quad + vtype[n]\ \cdot (-1) \cdot (\beta_{PP,int} + u_{PP,int}[subj\_pron[n]]\\
&\quad + w_{PP,int}[item\_pron[n]]),\\
&\quad 0,\\
&\quad \alpha_{other} + u_{other}[subj\_pron[n]] + w_{other}[item\_pron[n]]\\
&\quad + vtype[n] \cdot\ \big(\beta_{other,vtype} + u_{other,vtype}[subj\_pron[n]]\big)\\
&\quad + (-1) \cdot \big(\beta_{other,ref} + u_{other,ref}[subj\_pron[n]]\\
&\quad + w_{other,ref}[item\_pron[n]]\big)\\
&\quad + vtype[n]\ \cdot (-1)\ \cdot (\beta_{other,int} + u_{other,int}[subj\_pron[n]]\\
&\quad + w_{other,int}[item\_pron[n]]\\
&\quad)
\end{aligned} \tag{14}$$

where:

$$softmax(y) = exp(y)/\sum^{k}(y_k) \tag{15}$$

Then, the probability of the referent NP1 is calculated conditioned on a personal pronoun and on a demonstrative pronoun:

$$P\left(NP1|PP, \ldots\right) = \frac{P\left(PP|NP1, \ldots\right)}{P\left(PP|NP1, \ldots\right) + P\left(PP|NP2, \ldots\right)} \tag{16}$$

$$P\left(NP1|DP, \ldots\right) = \frac{P\left(DP|NP1, \ldots\right)}{P\left(DP|NP1, \ldots\right) + P\left(DP|NP2, \ldots\right)} \tag{17}$$

These probabilities are used to predict each observation n conditional on the type of pronoun that was completed:

$$pred_{NP1_n} \sim Bernoulli(P(referent|pronoun_n, \ldots)) \tag{18}$$

As before, the . . . symbolize all the information that the model is taking into account generate the predictions: the characteristics of the stimuli (i.e., intercept, beta, and by-item adjustments) and of the subject performing the free-prompt task (i.e., by-subject adjustments). However, now the pronoun type of each observation affects the predictions of the model.

Bayesian Model

The Bayesian model predicts that the probability of referring to *NP1* for pronoun-prompt data is determined by its posterior distribution in the free-prompt data according to Bayes' rule: the *likelihood* of NP1 (*P(pronoun|referent = NP1)*) is multiplied by the *prior* probability of NP1 (*P(referent = NP1)*), normalized to be a probability distribution by dividing it by the marginal probability distribution of the pronouns. This *posterior* can be estimated by the free-prompt data.

The parameters of the Bayesian model are estimated using equations (9) from the Expectancy model and (11) from the Mirror model. This entails that the model contains the parameters $\beta_{vtypeNP1}$ and $\beta_{vtypePP}$. In addition, since the by-participants and by-items adjustments from both (9) and (11) are used, this model has six potentially correlated by-subject adjustments and three potentially correlated by-items adjustments. For this reason, the parameter estimates are not identical to the previous models. The parameters estimated with the free-prompt data were used to generate predictions for each observation *n* of the pronoun-prompt data as follows.

We calculate the prior *P(NP1)* based on equation (10) and the likelihoods depending on the pronoun type P(*pronoun*|*NP1*) based on equations (13) and (14). With these we calculate P(*NP1*|*pronoun*).

The posterior probability of the referent NP1 is calculated conditional on a personal pronoun and on a demonstrative pronoun:

$$P\left(NP1|PP, \ldots\right) = \frac{P\left(PP|NP1\right)P(NP1)}{P\left(PP|NP1\right)P(NP1) + P\left(PP|NP2\right)\left(1 - P\left(NP1\right)\right)} \quad (19)$$

$$P\left(NP1|DP, ..\right) = \frac{P\left(DP|NP1\right)P(NP1)}{P\left(DP|NP1\right)P(NP1) + P\left(DP|NP2\right)\left(1 - P\left(NP1\right)\right)} \quad (20)$$

These probabilities are used to predict each observation n conditional on the type of pronoun that was completed:

$$pred_{NP1_n} \sim Bernoulli(P(NP1|pronoun_n, \ldots)) \quad (21)$$

Parameter Estimates

Expectancy Model

Table 1 shows the mean estimate and credible interval for the parameters of the Expectancy model. Applying $logit^{-1}$ to the parameter values, we estimate the value of *P(NP1)* across verb types, as shown in **Table 2**.

Mirror Model

Table 3 shows the mean estimate and credible interval for the parameters of the Mirror model. Applying the *softmax* functions to the parameter values, we estimate the value of *P(NP1)* across verb type and pronoun type, as shown in **Table 4**.

Bayesian Model

Table 5 shows the mean estimate and credible interval for the parameters of the Bayesian model. Applying the *softmax* functions to the parameter values, we estimate the value of *P(NP1)* across verb type and pronoun type, as shown in **Table 6**.

Model Comparison

We compare the models numerically using the expected log-predictive density (elpd) score of the models, with a higher score indicating better predictive accuracy for the held out pronoun-prompt data, as shown in **Table 7**. There is a clear overall advantage in predictive accuracy for the Bayesian model.

When the difference between predictive density ("elpd_diff") is larger than four and the number of observations is larger than 100, then the normal approximation and the standard errors are quite reliable descriptions of the uncertainty in the difference. As a rule of thumb, differences larger than four are considered enough to differentiate the predictive performance of the models (Sivula et al., 2020). We also calculated the "weight" of the predictions of each model by averaging via stacking of predictive distributions. Stacking maximizes the potential elpd score by pulling the predictions of all the different models

TABLE 1 | Mean estimate and credible interval for the parameters of the Expectancy model, Experiment 1.

Parameter	Mean	q5	q95
α_{NP1}	−0.68	−0.99	−0.40
β_{vtype}	0.30	0.01	0.59

TABLE 2 | Value of *P*(*NP*1) across verb type for the Expectancy model, Experiment 1.

Variable	Mean	q5	q95
P(*NP*1|*verb = accusative*)	0.41	0.33	0.49
P(*NP*1|*verb = dative*)	0.28	0.19	0.37

TABLE 3 | Mean estimate and credible interval for the parameters of the Mirror model, Experiment 1.

Parameter	Mean	q5	q95
α_{PP}	2.57	2.01	3.21
α_{other}	−1.74	−3.00	−0.66
$\beta_{int_{PP}}$	0.24	−0.22	0.69
$\beta_{int_{other}}$	−0.35	−1.06	0.37
$\beta_{ref_{PP}}$	2.24	1.74	2.82
$\beta_{ref_{other}}$	1.01	0.11	1.87
$\beta_{vtype_{PP}}$	−0.62	−1.08	−0.16
$\beta_{vtype_{other}}$	1.05	0.35	1.83

TABLE 4 | Value of *P*(*NP*1) across verb type and pronoun type for the Mirror model, Experiment 1.

Variable	Mean	q5	q95
P(*NP*1|*pronoun = personal, verb = accusative*)	0.75	0.69	0.83
P(*NP*1|*pronoun = personal, verb = dative*)	0.57	0.54	0.60
P(*NP*1|*pronoun = demonstrative, verb = accusative*)	0.03	0.01	0.06
P(*NP*1|*pronoun = demonstrative, verb = dative*)	0.03	0.00	0.08

TABLE 5 | Mean estimate and credible interval for the parameters of the Bayesian model, Experiment 1.

Parameter	Mean	q5	q95
α_{NP1}	−0.66	−0.95	−0.38
α_{PP}	2.54	1.99	3.16
α_{other}	−1.70	−2.96	−0.58
$\beta_{int_{PP}}$	0.20	−0.28	0.65
$\beta_{int_{other}}$	−0.40	−1.12	0.31
$\beta_{ref_{PP}}$	2.24	1.75	2.79
$\beta_{ref_{other}}$	1.00	0.09	1.85
$\beta_{vtype_{PP}}$	−0.62	−1.09	−0.16
$\beta_{vtype_{other}}$	0.96	0.23	1.75
$\beta_{vtype_{NP1}}$	0.28	0.00	0.57

TABLE 6 | Value of *P*(*NP*1) across verb type and pronoun type for the Bayesian model, Experiment 1.

Variable	Mean	q5	q95
P(*NP*1|*pronoun = personal, verb = accusative*)	0.79	0.64	0.92
P(*NP*1|*pronoun = personal, verb = dative*)	0.22	0.10	0.36
P(*NP*1|*pronoun = demonstrative, verb = accusative*)	0.04	0.01	0.10
P(*NP*1|*pronoun = demonstrative, verb = dative*)	0.01	0.00	0.02

together. The values under the weight column represent the relative contribution of each model to the combined optimal model. In this case, the Bayesian model contributes almost 90% to the weighted predictions. In **Table 8**, we compare just the Mirror and Expectancy models. It is clear that the Mirror model has a predictive performance superior to the Expectancy model.

In **Table 9**, we show the difference in predictive density for the models split by verb type and pronoun type.

Figure 4 shows to what extent the predictions of the different models, depicted with violin plots, match the observed held out data from the participants. The predictions of the models are shown by means of their posterior predictive distribution: simulated datasets generated based on the posterior distributions of its parameters. The posterior predictive distribution shows what other possible datasets may look like. Because we show held-out data (in contrast with data used to "train" the model), we can compare the three models based on the extent to which the held out data looks more plausible under the predictive distributions. By-participant and by-item predictions of the models are depicted in **Supplementary Figures 1, 2** which can be found in the **Supplementary Material**.

From **Figure 4** and **Table 9**, we can see that the observed data are within the distribution of predictions of the Bayesian model in every condition, whereas the data cannot be accounted by the other models under all conditions. However, the Bayesian model is only clearly superior to the Mirror model for the personal pronoun in the dative contexts (while the Expectancy model performs much better here than it does in other conditions). The Mirror model comes close to the performance of the Bayesian model in the other three conditions, even though the Bayesian model is numerically superior.

Evaluating the Strong Form of the Bayesian Model

Here, we evaluate the claims of the strong form of the Bayesian model. First, to examine the influence of verb type on the prior and the production likelihoods, a model comparison was carried out comparing models with and without verb type in the prior and in the pronoun production likelihoods to assess the impact on predictive accuracy of the resulting models. **Tables 10**, **11** show the outcome of the model comparison.

The model comparison shows that the verb type has a large impact for the predictions of the model, and that the predictions of the Bayesian model deteriorate the most when the verb type information is removed from the prior. A model without verb type on the prior performs significantly worse than a full model (**Table 10**) and a model without verb type on the production likelihood (**Table 11**), demonstrating that the prior is influenced by verb type information, which is in line with the strong form of the Bayesian model. But removing verb type from the production likelihood also has a detrimental impact on predictive accuracy when compared to a full model. To explore this in more detail, we examine the influence of verb type on likelihoods for the personal and demonstrative pronouns separately.

We ran Bayesian multilevel models with the sum-coded factors Referent (proto-agent/NP1 versus proto-patient/NP2) and Verb Type (accusative versus dative) with random intercepts for participants and items, using the brms package (Bürkner, 2017) in RStudio (RStudio Team, 2019) on R version 3.6.1 (R Core Team, 2019)[13]. For the demonstrative pronouns, there was a clear effect of Referent (mean estimate −1.74, 95% CrI −2.24, −1.31), no effect of Verb Type (mean estimate 0.33, 95% CrI −0.11, 0.80) and no interaction between the two factors (mean estimate −0.12, 95% CrI −0.56, 0.34). This can be interpreted as follows: participants used a demonstrative pronoun to refer to the proto-patient (NP2) much more often than when referring to

TABLE 7 | Model comparison, Experiment 1.

	elpd_diff	se_diff	elpd	se_elpd	weight
Bayesian	0	0	−728	27	0.89
Mirror	−132	14	−860	27	0.00
Expectancy	−238	24	−966	16	0.11

The table is ordered by the expected log-predictive density (elpd) score of the models, with a higher score indicating better predictive accuracy. The highest scoring model is used as a baseline for the difference in elpd and the difference standard error (SE). "Weight" represents the weights of the individual models that maximize the total elpd score of all the models.

TABLE 8 | Comparison of Mirror and Expectancy models.

	elpd_diff	se_diff	elpd	se_elpd	weight
Mirror	0	0	−860	27	0.67
Expectancy	−106	31	−966	16	0.33

The table is ordered by the expected log-predictive density (elpd) score of the models, with a higher score indicating better predictive accuracy. The highest scoring model is used as a baseline for the difference in elpd and the difference standard error (SE). "Weight" represents the weights of the individual models that maximize the total elpd score of all the models.

TABLE 9 | Difference in expected log-predictive density (elpd_diff) of the models assessed for four subsets of the data from Experiment 1, depending whether the verb is accusative or dative, and whether the pronoun shown is personal (PP) or demonstrative (DP).

Model	elpd_diff	se_diff	weight
PP – accusative			
Bayesian	0.0	0.0	0.11
Mirror	−11.1	6.5	0.63
Expectancy	−56.8	11.9	0.26
DP – accusative			
Bayesian	0.0	0.0	1.00
Mirror	−2.5	1.7	0.00
Expectancy	−136.7	11.2	0.00
PP – dative			
Bayesian	0.0	0.0	0.48
Mirror	−113.9	10.0	0.00
Expectancy	0.2	4.2	0.52
DP – dative			
Bayesian	0.0	0.0	0.76
Mirror	−4.1	5.7	0.00
Expectancy	−44.3	16.8	0.24

[13] Full model specification and outputs can be found in **Supplementary Material**.

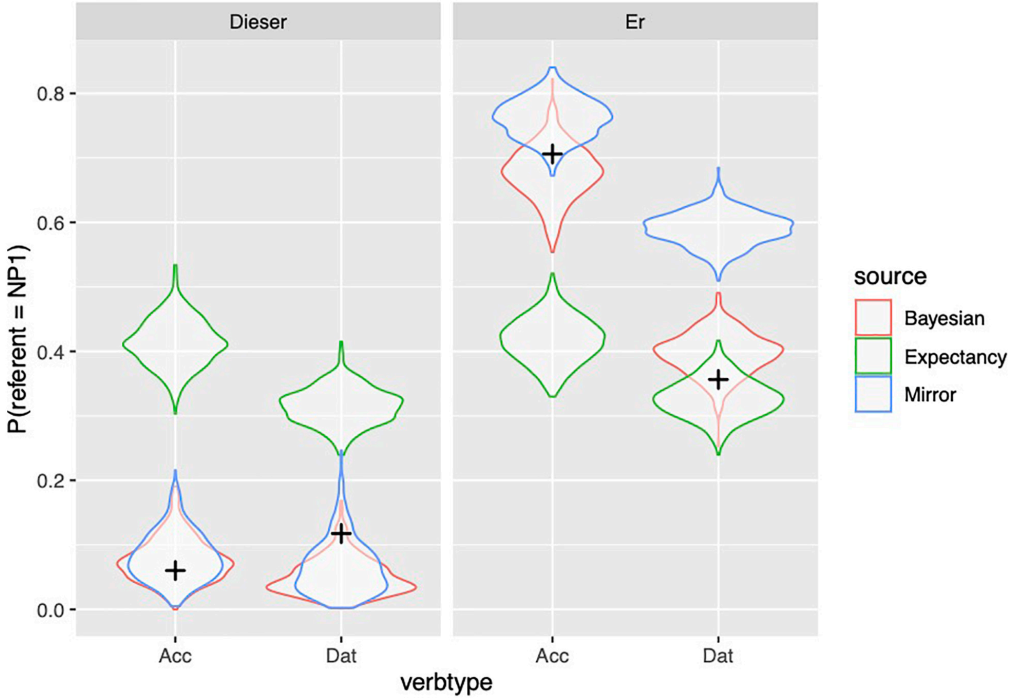


FIGURE 4 | Observed proportion of responses (from held out data, Experiment 1) are depicted with black crosses; distribution of simulated proportions based on the model predictions are depicted with violin plots.

the proto-agent (NP1), across both accusative and dative verbs. For the personal pronouns, the model showed a clear effect of Referent in the opposite direction (mean estimate 1.46, 95% CrI 1.18, 1.78). The model also showed an effect for Verb Type, but the estimate here was closer to zero (mean estimate −0.75, 95% CrI −1.04, −0.45). There was no interaction between Referent and Verb Type (mean estimate 0.24, 95% CrI −0.03, 0.52). This shows that participants used a personal pronoun to refer to the proto-agent (NP1) more often than when referring to the proto-patient (NP2). The overall rate of pronominalization for the personal pronoun was higher for the dative verbs compared to the accusative verbs, but the relative (NP1–NP2) production bias was not influenced by verb type.

Discussion

In Experiment 1, the Bayesian model clearly outperforms both the Mirror model and Expectancy model overall. Additionally, the model is able to account better for both the personal and demonstrative pronouns than the competing models when performance is assessed separately for each pronoun in all but two comparisons, although the degree of difference between models does vary (see **Table 9**)[14]. The Mirror model comes close to the performance of the Bayesian model in two out of four conditions, and for demonstratives in the accusative contexts Bayesian and Mirror model performance is indistinguishable. The Expectancy model is outperformed by both the Mirror and the Bayesian models except for personal pronouns in dative contexts, where Bayesian and Expectancy are indistinguishable and both far outperform the Mirror model. The variation over the different conditions demonstrates, however, that the Bayesian model is more powerful for taking into account elements of both other models, i.e., movement in the prior (Expectancy) and production likelihoods (Mirror), while neither element alone can capture behavior across the conditions.

We also tested the predictions of the strong Bayesian model. In our analysis, verb type had a larger influence on the prior than on the likelihoods, which is in line with the strong Bayesian model. But removing verb type from the likelihood also had negative

TABLE 10 | Model comparisons after removing verb type from the likelihood and the prior.

	elpd_diff	se_diff	elpd	se_elpd	weight
Full Bayesian	0	0.0	−728	27	0.97
No verb type in likelihood	−24	7.0	−753	28	0.03
No verb type in prior	−55	7.5	−783	27	0.00

The table is ordered by the expected log-predictive density (elpd) score of the models, with a higher score indicating better predictive accuracy. The highest scoring model is used as a baseline for the difference in elpd (elpd_diff) and the difference standard error (se_diff). 'Weight' represents the weights of the individual models that maximize the total elpd score of all the models.

[14] Indeed, it is not surprising that the Mirror model is a lot closer to the Bayesian model and the actual data for the demonstrative pronouns than the Expectancy model. The Expectancy model resolves the pronoun to the referent that is most expected; one of the functions of demonstrative pronouns is to highlight a less expected referent. Hence, no-one would claim that the Expectancy model as it is implemented here can be applied to demonstratives. Nevertheless, looking only at personal pronouns, the Bayesian model still outperforms the Expectancy model.

TABLE 11 | Model with no verb type in the likelihood compared to a model with no verb type on the prior.

	elpd_diff	se_diff	elpd	se_elpd	weight
No verb type in likelihood	0	0.0	−753	28	0.87
No verb type in prior	−30	9.2	−783	27	0.13

The table is ordered by the expected log-predictive density (elpd) score of the models, with a higher score indicating better predictive accuracy. The highest scoring model is used as a baseline for the difference in elpd (elpd_diff) and the difference standard error (se_diff). "Weight" represents the weights of the individual models that maximize the total elpd score of all the models.

impact on predictive accuracy. However, given that the verb type contrast in this experiment encompasses a change in position of the subject (NP1 in accusative verbs and NP2 in dative verbs), the constructions are perhaps not directly comparable.

The second test of the strong Bayesian model was examining the pattern of results in the pronoun production likelihoods separately for personal and demonstrative pronouns. Here we saw no interaction of verb type with referent; this is in line with the strong Bayesian model which states that likelihoods should not be influenced by verb type, although it should be noted that the verb type under examination here is of a different nature than the verb contrasts normally examined. Additionally, we were interested in the relative influence of subjecthood and agentivity, because the two factors make contrasting predictions for the effect of Referent (NP1 versus NP2) across the two verb types. In previous studies, subjecthood (and/or topichood) influenced production likelihoods. Our results were as follows: demonstrative pronouns were much more likely to be produced when referring to NP2 versus NP1 across both verb types. The NP2 was the proto-patient in both accusative and dative verbs, suggesting a strong influence of non-agentivity rather than non-subjecthood. Personal pronoun likelihoods, on the other hand, showed a less clear pattern. For the accusative verbs there was a clear NP1 (subject/proto-agent) advantage. There was a weaker advantage for NP1 (object/proto-agent) in the dative verbs, but the difference in NP1 advantage was not confirmed statistically. Overall participants were more likely to produce a pronoun following dative verbs compared to accusative verbs. The proto-agent advantage speaks for an influence of agentivity rather than subjecthood, but the pattern in the dative verbs is nevertheless puzzling and prevents us from drawing strong conclusions here.

It is certainly the case that the verb type contrast examined here (accusative versus dative verbs) is of a different nature than the contrasts tested previously. While an IC contrast, exemplified in (5) and (6), represents a difference in expected continuations, the accusative–dative contrast represents a difference in the assignment of argument roles. It is therefore perhaps not surprising that the patterns in Experiment 1 are different from previous studies in which an IC contrast was used. For this reason, we carried out Experiment 2, using an IC contrast to make our results more comparable to previous studies. This experiment also gives us a chance to replicate our findings with respect to overall model performance and represents a more straightforward test of the predictions of the strong Bayesian model.

EXPERIMENT 2

Experiment 2 was a text continuation task with the German personal pronoun *er* and the demonstrative pronoun *dieser* using an IC-based verb-type contrast, more closely reflecting materials in previous studies (Kehler et al., 2008; Rohde and Kehler, 2014). Specifically, we used stimulus–experiencer (SE) and experiencer–stimulus (ES) verbs (see Bott and Solstad, 2014 for an overview of the semantic properties of these verbs). In addition, the contrast allows us to look again at the contribution of agentivity and subjecthood. Recall that we pursue the proto-role approach (Dowty, 1991), where thematic roles are characterized by features associated with proto-agents and proto-patients. Experiencers are typically considered agent-like because they entail sentience.

In ES constructions, subjects and experiencers (as the highest thematic role) are aligned, potentially yielding a higher production likelihood for NP1. In SE constructions, NP1 outranks NP2 with respect to subjecthood but NP2 outranks NP1 with respect to agentivity[15]. A subset of dative items from Experiment 1 were also included in an attempt to verify the pattern in the production likelihoods from Experiment 1[16].

Participants

Forty participants (18–67 years) were recruited on the online platform Prolific.ac to take part in Experiment 2. Data from all 40 participants (15 female, 25 male) were used in the analysis. All participants indicated that they were German native speakers; 8 participants were bilingual. No participants reported language-related disorders. All participants gave their consent and received a small fee for participation.

Materials

Thirty six critical items were constructed, 18 SE items and 18 ES items. 28 verbs were taken from Bott and Solstad (2014) who systematically tested the semantics of implicit causality verbs in a set of German verbs; additional verbs were pretested according to the "*that*-clause replacement test" and the "*absichtlich*-test" (adverbial "*deliberately*" being added to transitive verb frames) following Bott and Solstad (2014). In order to avoid effects of polarity, each of the two groups of 18 critical items included nine verbs related to negative perception (e.g., *schockieren* "shock," SE; *verachten* "despise," ES) and nine that were positive (e.g., *faszinieren* "fascinate," SE; *respektieren* "respect," ES). The critical items consisted of a context sentence which contained a nominative argument, the main verb in present tense and an accusative argument, and a prompt sentence which was either a personal pronoun-prompt (*er*), a demonstrative pronoun-prompt (*dieser*) or a free-prompt (blank line). In both SE and ES items, contexts were presented in canonical order (subject verb object). Example items are given in (22) and (23).

[15] But see Dowty (1991, p. 579) for competition between agentive features in SE contexts where the stimulus entails the proto-agent property causation and the experiencer entails sentience.

[16] Analysis and outcome for the dative items can be found on OSF (osf.io/j5wtg). The overall pattern for the dative items was similar to Experiment 1.

(22) SE items:

(a) *Er prompt:* Der Jurist faszinierte den Richter. Er __________
(b) *Dieser prompt:* Der Jurist faszinierte den Richter. Dieser __________
(c) *Free-prompt:* Der Jurist faszinierte den Richter. __________

"The lawyer *(nom.masc.)* fascinated the judge *(acc.masc.)*. *He/DEM/...*"

(23) ES items:

(a) *Er prompt:* Der Christ respektierte den Moslem. Er __________
(b) *Dieser prompt:* Der Christ respektierte den Moslem. Dieser __________
(c) *Free-prompt:* Der Christ respektierte den Moslem. __________

"The Christian *(nom.masc.)* respected the Muslim *(acc.masc.)*. *He/DEM/...*"

For both NPs, role names were chosen following the same criteria as in Experiment 1. 22 filler items were also created: six dative-experiencer contexts from Experiment 1, four nominative-accusative IC verb contexts followed by a connector, three contexts with a single NP followed by a connector, seven catch fillers (included to ensure that participants were paying attention to the task), and two *dieser* items to prime a pronominal reading (see Experiment 1 for a description). The filler set included a mix of feminine and masculine pronouns and referents to counterbalance the large number of masculine referents in the critical items. The items were distributed over three lists in a Latin-square design.

Procedure

Based on a short description of the task, participants could choose to take part in the study via the Prolific.ac application. Participants gave their consent and answered a short series of biographical questions before starting the experimental task. Task instructions were the same as for Experiment 1.

Data Coding

Data was coded in the same way as for Experiment 1. The two annotators agreed in 86% of observations, with a Cohen's (unweighted) Kappa of 0.78 ($z = 45.5$, $p < 0.0001$). Data from 40 participants for 12 items (6 SE, 6 ES) per prompt condition resulted in a total of 1440 observations; 480 each in the *er*-prompt, *dieser*-prompt and free-prompt conditions. The distribution of reference and the response categories for the first referential expression are given in **Supplementary Material**. For the following analyses, the dataset was reduced in the same way as in Experiment 1, leaving a total of 1221 observations for the analysis (352 free-prompt, 430 *dieser*-prompt and 439 *er*-prompt).

Data Analysis

A data analysis plan and accompanying predictions were registered in advance of carrying out this experiment on aspredicted.org. The registration can be found in **Supplementary Material**. While the data collection followed the registered plan, the data analysis was in the end superseded by the Bayesian statistical analysis presented here. This type of analysis was a late addition to the project that we did not foresee at the time of data collection. Data analysis and models are the same as in Experiment 1, with the exception that the verb types are ES (coded as 1) and SE (coded as −1).

Results

Raw proportions for the next-mention bias are shown in **Figure 5**. **Figure 6** shows the personal pronoun production likelihoods, and **Figure 7** the demonstrative pronoun production likelihoods[17]. When calculating likelihoods for the personal pronoun, both subject and non-subject personal pronouns were included. For the demonstrative pronoun, both subject and non-subject demonstrative pronoun *dieser*, and subject and non-subject demonstrative pronoun *der*, were included.

A follow-up rating experiment was also conducted: see discussion below. Method and results for this rating experiment can be found in **Supplementary Material**.

Parameter Estimates

Expectancy Model

Table 12 shows the mean estimate and credible interval for the parameters of the Expectancy model. Applying $logit^{-1}$ to the

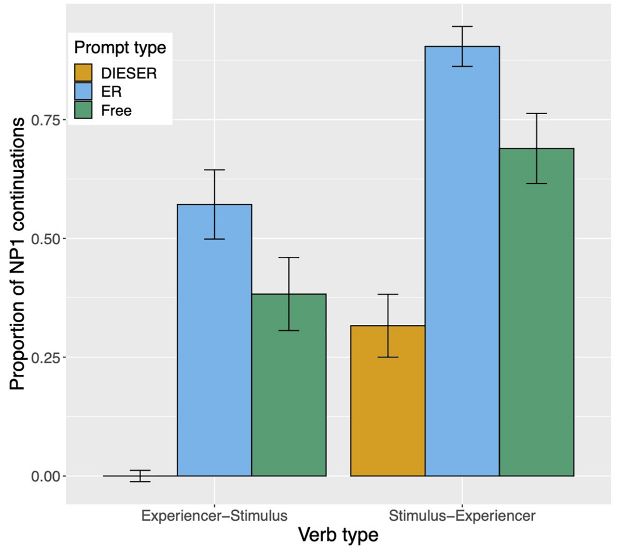


FIGURE 5 | Proportion of NP1 continuations per prompt condition (*er, dieser, and free*) and verb type in Experiment 2. The NP1 continuations represent the next-mention bias for NP1. Error bars represent 95% confidence intervals based on normalized data following Morey (2008) and Chang (2013).

[17] The tabulated data can be found in **Supplementary Material**.

parameter values, we estimate the value of *P(NP1)* across verb types, as shown in **Table 13**.

Mirror Model

Table 14 shows the mean estimate and credible interval for the parameters of the Mirror model. Applying *softmax* to the parameter values, we estimate the value of *P(NP1)* across verb type and pronoun type, as shown in **Table 15**.

Bayesian Model

Table 16 shows the mean estimate and credible interval for the parameters of the Bayesian model. Applying *softmax* to the parameter values, we estimate the value of *P(NP1)* across verb type and pronoun type, as shown in **Table 17**.

Model Comparison

As before, we compare the models numerically using the elpd score of the models, as shown in **Table 18**. In **Table 19** we show the elpd score of the models split by verb type and pronoun type. There is again a clear overall advantage in predictive accuracy for the Bayesian model (**Table 18**). The Bayesian model contributes 90% to the weighted predictions in the overall comparison.

Figure 8 shows to what extent the predictions of the different models, depicted with violin plots, match the observed held out data from the participants, as per Experiment 1. By-participant and by-item predictions of the models are depicted in **Supplementary Figures 3, 4** which can be found in the **Supplementary Material**.

From **Figure 8** and **Table 19**, it is clear that the observed data are well within the distribution of predictions of the Bayesian model, whereas the data cannot be accounted by the other models under all conditions. Unlike in Experiment 1, here the Bayesian model outperforms the Mirror model in all conditions except for demonstrative pronouns in the ES contexts, where performance of the two models is indistinguishable. The Bayesian model outperforms the Expectancy model in all conditions.

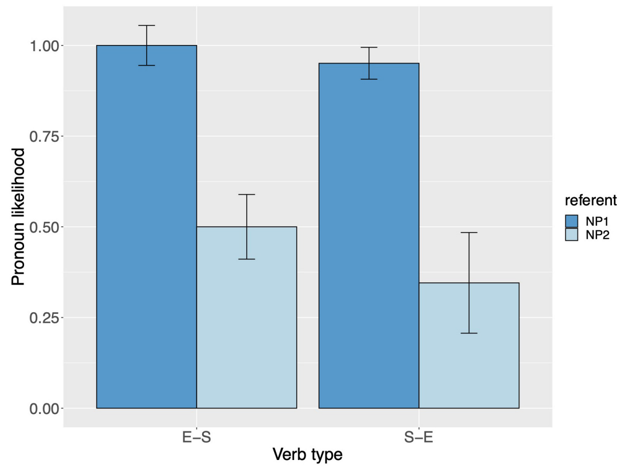


FIGURE 6 | Personal pronoun: probability of using a personal pronoun to refer to NP1 and NP2, per verb type, for Experiment 2. These bars represent the pronoun production likelihoods (and are therefore based on the free-prompt data). Error bars are by-participant and represent 95% confidence intervals based on normalized data (Morey, 2008; Chang, 2013).

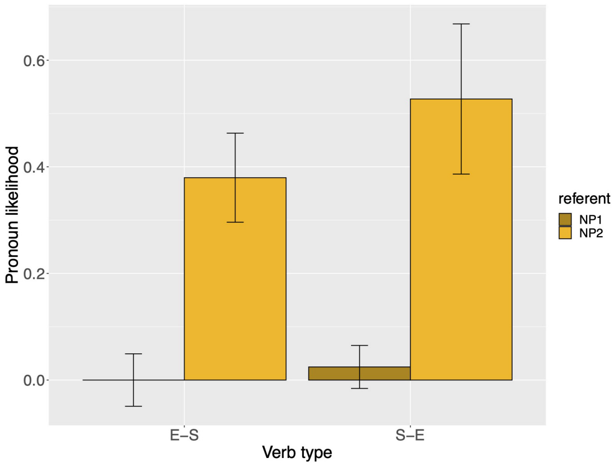


FIGURE 7 | Demonstrative pronoun: probability of using a demonstrative pronoun to refer to NP1 and NP2, per verb type, for Experiment 2. These bars represent the pronoun production likelihoods (and are therefore based on the free-prompt data). Error bars are by-participant and represent 95% confidence intervals based on normalized data (Morey, 2008; Chang, 2013).

TABLE 12 | Mean estimate and credible interval for the parameters of the Expectancy model, Experiment 2.

Parameter	Mean	q5	q95
α_{NP1}	0.19	−0.29	0.66
β_{vtype}	−0.94	−1.51	−0.42

TABLE 13 | Value of *P*(*NP*1) across verb type for the Expectancy model, Experiment 2.

Variable	Mean	q5	q95
P(*NP*1|*verb* = *ES*)	0.33	0.18	0.48
P(*NP*1|*verb* = *SE*)	0.75	0.61	0.87

Evaluating the Strong Form of the Bayesian Model

Here, we evaluate the claims of the strong form of the Bayesian model by again examining (i) the influence of verb type on the prior (i.e., the next-mention bias) and (ii) the influence of verb type and the relative contribution of agentivity and subjecthood on the pronoun production likelihoods (i.e., on P(pronoun|referent)). First, to examine the influence of verb type, a model comparison was carried out comparing models with and without verb type in the prior and in the pronoun production likelihoods to assess the impact on predictive accuracy of the resulting models. **Tables 20**, **21** show the outcome of the model comparison.

The model comparison shows that the verb type has a large impact for the predictions of the model, and that the predictions of the Bayesian model deteriorate the most when the verb type information is removed from the prior. A model without verb type on the prior performs significantly worse than a

TABLE 14 | Mean estimate and credible interval for the parameters of the Mirror model, Experiment 2.

Parameter	Mean	q5	q95
α_{PP}	3.67	2.51	5.0
α_{other}	−3.46	−5.39	−1.6
$\beta_{int_{PP}}$	0.55	−0.41	1.6
$\beta_{int_{other}}$	−0.47	−2.11	1.1
$\beta_{ref_{PP}}$	3.68	2.61	4.9
$\beta_{ref_{other}}$	−0.04	−1.69	1.5
$\beta_{vtype_{PP}}$	1.00	0.04	2.1
$\beta_{vtype_{other}}$	0.08	−1.44	1.6

TABLE 15 | Value of $P(NP1)$ across verb type and pronoun type for the Mirror model, Experiment 2.

Variable	Mean	q5	q95
$P(NP1\|pronoun = personal, verb = ES)$	0.64	0.56	0.74
$P(NP1\|pronoun = personal, verb = SE)$	0.73	0.60	0.87
$P(NP1\|pronoun = demonstrative, verb = ES)$	0.00	0.00	0.01
$P(NP1\|pronoun = demonstrative, verb = SE)$	0.01	0.00	0.03

TABLE 16 | Mean estimate and credible interval for the parameters of the Bayesian model, Experiment 2.

Parameter	Mean	q5	q95
α_{NP1}	0.17	−0.30	0.66
α_{PP}	3.60	2.51	4.82
α_{other}	−3.41	−5.39	−1.58
$\beta_{int_{PP}}$	0.56	−0.39	1.62
$\beta_{int_{other}}$	−0.52	−2.07	0.96
$\beta_{ref_{PP}}$	3.61	2.58	4.78
$\beta_{ref_{other}}$	−0.07	−1.74	1.49
$\beta_{vtype_{PP}}$	0.96	0.03	2.00
$\beta_{vtype_{other}}$	0.06	−1.45	1.59
$\beta_{vtype_{NP1}}$	−0.94	−1.50	−0.42

TABLE 17 | Value of $P(NP1)$ across verb type and pronoun type for the Bayesian model, Experiment 2.

Variable	Mean	q5	q95
$P(NP1\|pronoun = personal, verb = ES)$	0.67	0.31	0.92
$P(NP1\|pronoun = personal, verb = SE)$	0.71	0.35	0.95
$P(NP1\|pronoun = demonstrative, verb = ES)$	0.00	0.00	0.01
$P(NP1\|pronoun = demonstrative, verb = SE)$	0.02	0.00	0.06

TABLE 18 | Model comparison, Experiment 2.

	elpd_diff	se_diff	elpd	se_elpd	weight
Bayesian	0	0	-368	19	0.9
Mirror	-98	13	-467	24	0.0
Expectancy	-209	23	-578	16	0.1

The table is ordered by the expected log-predictive density (elpd) score of the models, with a higher score indicating better predictive accuracy. The highest scoring model is used as a baseline for the difference in elpd (elpd_diff) and the difference standard error (se_diff). "Weight" represents the weights of the individual models that maximize the total elpd score of all the models.

full model (**Table 20**) and worse than a model without verb type on the production likelihood (**Table 21**), demonstrating that the prior is influenced by verb type information, in line with the strong form of the Bayesian model. Removing verb type from the production likelihood also has a detrimental impact on predictive accuracy when compared to a full model, demonstrating that overall production likelihoods are also to some extent influenced by verb type; this is explored further by examining the factors affecting the production likelihoods for personal and demonstrative pronouns separately. This was done using Bayesian multilevel models with the same set up as in Experiment 1.

For the demonstrative pronouns, there was a clear effect of Referent (mean estimate −2.37, 95% CrI −3.13, −1.71), no effect of Verb Type (mean estimate −0.49, 95% CrI −1.22, 0.17) and no interaction between the two factors (mean estimate −0.18, 95% CrI −0.89, 0.45). This shows that participants used a demonstrative pronoun to refer to the object (NP2) much more often than when referring to the subject (NP1), across both SE and ES verbs. For the personal pronouns, the model showed a clear effect of Referent (mean estimate 2.35, 95% CrI 1.77, 3.05). The model also showed an effect for Verb Type, but the lower bound of the Credible Interval is almost at zero (mean estimate 0.62, 95% CrI 0.06, 1.27). There was no interaction between Referent and Verb Type (mean estimate 0.35, 95% CrI −0.23, 1.00). This shows that participants used a personal pronoun to refer to the subject (NP1) more often than when referring to the object (NP2), regardless of verb type. The overall rate of pronominalization for the personal pronoun may be slightly higher for the ES verbs compared to the SE verbs, but this effect should be interpreted with caution. We discuss the implications for the relative contributions of subjecthood and agentivity below.

Discussion

In Experiment 2 the Bayesian model again clearly outperforms both the Mirror model and Expectancy model overall. The Bayesian model is able to account better for both the personal and demonstrative pronouns than the competing models when performance is assessed separately for each pronoun (see **Table 19**), although the degree of difference between models does vary as before. The Bayesian model outperforms the Mirror model in three out of four conditions. The caveat about the Expectancy model predictions for the demonstrative still applies, but again the Bayesian model outperforms the Expectancy model for the personal pronouns.

One surprising pattern in Experiment 2 is the high number of NP1 continuations with the *dieser* prompt for the SE verbs (see **Figure 5**). The Bayesian model does a good job of predicting this pattern, although the predicted values are spread out, indicating less certainty about the prediction (see **Figure 8**). Nevertheless, the high number of NP1 interpretations here is not expected, given the more rigid tendencies of demonstratives. We suspected that this could be due to the experimental design: SE contexts strongly bias toward continuations about the stimulus subject

TABLE 19 | Difference in expected log-predictive density (elpd_diff) of the models assessed for four subsets of the data from Experiment 2, depending whether the verb is stimulus–experiencer (SE) or experiencer–stimulus (ES), and whether the pronoun shown is personal (PP) or demonstrative (DP).

Model	elpd_diff	se_diff	weight
PP – SE			
Bayesian	0.00	0.00	1.00
Mirror	−25.92	3.86	0.00
Expectancy	−25.91	5.15	0.00
DP – SE			
Bayesian	0.00	0.00	0.69
Mirror	−48.74	8.12	0.00
Expectancy	−53.10	19.28	0.31
PP – ES			
Bayesian	0.00	0.00	0.37
Mirror	−23.84	8.40	0.28
Expectancy	−14.24	7.34	0.35
DP – ES			
Bayesian	0.00	0.00	0.19
Mirror	0.06	0.45	0.81
Expectancy	−115.91	6.24	0.00

(NP1). At the same time, demonstratives would normally avoid reference to a subject. As such, being presented with a *dieser* prompt in SE contexts presents something of a challenge to participants who may be conflicted about continuing with a less preferred referent (experiencer in this case) but working with the *dieser* bias, or working against the *dieser* bias but satisfying the bias to talk about the stimulus. In order to check whether our suspicion was correct, we carried out a follow-up rating experiment which is described in **Supplementary Material**. We predicted that the completions in which *dieser* refers to NP1 in the SE condition should be less felicitous than SE completions in which *dieser* refers to NP2, since only the latter works with the grammatical bias associated with *dieser*, and less felicitous than SE completions in which *er* refers to NP1, because *er* does not have a bias against NP1 reference. Our predictions were borne out; a cumulative link model showed that both *dieser*–NP2 completions and the *er*–NP1 completions were significantly more likely to elicit better ratings than *dieser*–NP1 completions (z = 11.52 for *dieser*–NP2 and 12.28 for *er*–NP1)[18]. Given this result, it is striking that the Bayesian model is able to capture the actual data from the SE *dieser*–NP1 completions, and at the same time reflect the uncertainty about the predictions in this condition, which is also reflected in the rating data from the follow-up experiment.

Finally, we again tested the predictions of the strong form of the Bayesian model. As in Experiment 1, the model comparisons for Experiment 2 showed that removing verb type from the prior had a more detrimental effect on the predictive accuracy of the model than removing it from the likelihood, underlining the influence of verb type on prior as found by Rohde and Kehler (2014). While predictive accuracy was also affected by removing verb type from the likelihoods, there was no Verb Type by Referent interaction when the likelihoods were examined separately for each pronoun; this finding provides further support for the strong Bayesian model.

Turning to the relative influence of subjecthood and agentivity on the likelihoods, this was again tested via an effect of Referent. Here, we saw strong effects for personal and demonstrative pronouns, in opposite directions[19]. The pattern shows a strong influence of subjecthood for personal pronouns and an objecthood bias in the likelihoods for demonstrative pronouns, regardless of the thematic role of the subjects and objects. We return to these findings in the general discussion.

GENERAL DISCUSSION

In this study we set out to test the following questions:

- Which model for pronouns (Bayesian, Expectancy or Mirror) best accounts for the interpretation of German personal and demonstrative pronouns?
- Is the resolution of demonstratives as rigid as some previous studies suggest, or is the interpretation influenced by the next-mention bias, as the Bayesian model would predict?
- Is there evidence for the strong form of the Bayesian model?

We evaluated overall model performance by using data from the free-prompt conditions in two text-continuation experiments to generate predictions for the Bayesian, Expectancy and Mirror models. These predictions were compared to actual interpretations from the pronoun-prompt conditions, allowing us to assess the predictive accuracy of the models. Overall results from Experiments 1 and 2 show convincingly that the Bayesian model outperforms both the Expectancy and the Mirror models. When the performance was evaluated per verb type and pronoun type separately, the Mirror model performed almost as well as the Bayesian model in three conditions of Experiment 1 but not in Experiment 2, where the Bayesian model outperformed the Mirror model in three out of four conditions. The Bayesian model was even able to predict behavior that was somewhat unexpected, as in the higher-than-expected number of interpretations of *dieser* as NP1 in the SE condition. The fact that the Bayesian model outperforms the Mirror and Expectancy models is further confirmation of the findings from Rohde and Kehler (2014) and Zhan et al. (2020), and in fact the model performance of the Bayesian model as evaluated here (in particular for Experiment 2) is actually better than in those studies, where the Bayesian and Mirror models showed a similar performance in certain conditions. This validates the approach in Kehler et al. (2008) and

[18] Furthermore, we tested whether participants had the same interpretations of the SE *dieser*–NP1 completions as our annotators; participants agreed with our annotations (by choosing NP1) on average 71% of the time. The probability of NP1 choice, calculated per item, did not have a significant influence on the ratings of the SE *dieser*–NP1 completions ($z = -0.26$).

[19] As a reminder, in ES constructions subjects and experiencers (as the highest thematic role) are aligned, while in SE constructions NP1 outranks NP2 with respect to subjecthood but NP2 outranks NP1 with respect to agentivity. Thus if agentivity and subjecthood both have a strong influence on likelihoods, the NP1–NP2 difference should be stronger in the ES constructions than in the SE constructions.

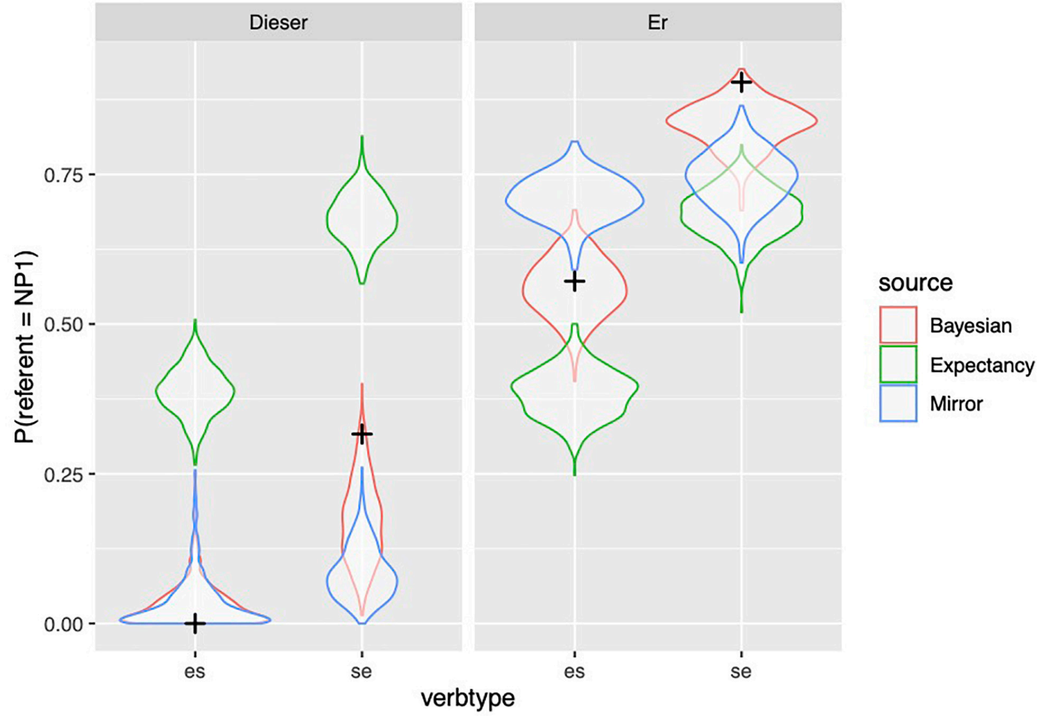


FIGURE 8 | Observed proportion of responses (from held out data, Experiment 2) are depicted with black crosses; distribution of simulated proportions based on the model predictions are depicted with violin plots.

Kehler and Rohde (2013) of applying simple Bayesian principles to the complex problem of pronoun resolution.

The fact that the Mirror model was more competitive with the Bayesian model in Experiment 1 than Experiment 2 can be attributed to the verb type contrasts under investigation. The IC contrast in Experiment 2 represents a difference in expected continuations, i.e., a contrast in the prior. Given that the Mirror model does not include the prior, it is not surprising that it does not capture all the data here. On the other hand, the accusative–dative contrast in Experiment 1 represents a difference in the assignment of argument roles, which does not entail such extreme movement of the prior, allowing the Mirror model to perform better. Nonetheless, the overall performance of the Mirror model in Experiment 1 was not as good as the performance of the Bayesian model.

The three models were implemented for the first time in a Bayesian statistical framework, which goes beyond the modeling in previous studies (Rohde and Kehler, 2014; Bader and Portele, 2019; Zhan et al., 2020) and has a number of advantages. The fully hierarchical structure allowed us to accommodate participant and item effects directly into our predictions without averaging. In contrast to previous evaluations of model performance, the Bayesian statistical approach allows us to estimate the parameters of a distribution of predicted observations, and as such we can make more stable inferences about model performance. In the Bayesian statistical approach there is no requirement for additive smoothing (as in Zhan et al., 2020) because the uniform Beta prior over the intercept ensures that probability estimates cannot be zero or one; this is especially important for the demonstrative pronouns where the less flexible interpretation leads to zeros in some cells of the design. Our modeling approach also allowed us to evaluate claims about the strong form of the Bayesian model in a new way, by removing verb type from model components and evaluating the impact on the predictive accuracy of the models. This revealed that removing verb type from the prior had a detrimental impact on the predictive accuracy of the model, which is in line with the strong Bayesian model.

TABLE 20 | Model comparisons after removing verb type from the likelihood and the prior for Experiment 2.

	elpd_diff	se_diff	elpd	se_elpd	weight
Full Bayesian	0	0.0	−368	19	1
No verb type in likelihood	−35	6.5	−404	23	0
No verb type in prior	−61	9.0	−429	22	0

The table is ordered by the expected log-predictive density (elpd) score of the models, with a higher score indicating better predictive accuracy. The highest scoring model is used as a baseline for the difference in elpd (elpd_diff) and the difference standard error (se_diff). "Weight" represents the weights of the individual models that maximize the total elpd score of all the models.

In this study we examined German personal and demonstrative pronouns, a new contribution to the evidence about the Bayesian model for pronouns. It also provides a new perspective on the interpretation of demonstrative pronouns. German demonstratives have been long neglected in literature on pronoun resolution, and have only recently

TABLE 21 | Model with no verb type in the likelihood compared to a model with no verb type on the prior, Experiment 2.

	elpd_diff	se_diff	elpd	se_elpd	weight
No verb type in likelihood	0	0	−404	23	0.82
No verb type in prior	−25	9	−429	22	0.18

The table is ordered by the expected log-predictive density (elpd) score of the models, with a higher score indicating better predictive accuracy. The highest scoring model is used as a baseline for the difference in elpd (elpd_diff) and the difference standard error (se_diff). "Weight" represents the weights of the individual models that maximize the total elpd score of all the models.

gained more attention. Here too the main debate has focussed on identifying particular factors that influence resolution and that can distinguish preferences for the resolution of demonstratives from those of personal pronouns. The current study may shift the debate toward understanding the resolution of demonstratives in the context of a speaker and an addressee, where predictions about the message content are combined with the estimation of the speaker's choice of referential form.

It was noted that the Expectancy model (i.e., our operationalization of Expectancy as P(referent)) was not predicted to capture demonstrative pronouns because of their tendency to refer to less "expected" referents. Indeed, the modeling results for Expectancy model in the demonstratives confirm that this approach is not successful. The Bayesian model, too, makes use of expectations about the upcoming referent (i.e., the next-mention bias), and it was possible therefore that the Bayesian model would not be so successful for demonstratives. But our results show that this is not the case. The combination of next-mention bias with production likelihoods is a powerful modification that makes the Bayesian model flexible enough to accommodate pronoun types with quite different resolution tendencies. As such, our study has shown that the Bayesian model is not limited to just one type of pronoun.

That being said, broad cross-linguistic evidence for the Bayesian model is lacking, having previously been evaluated fully only on flexible personal pronouns in English and pronouns in Mandarin Chinese (where null and overt pronouns appear to overlap in resolution preferences). While the current study allows us to incorporate demonstrative pronouns into the Bayesian model without revising its basic assumptions, it remains to be seen whether this is the case for a wider variety of pronouns or indeed other types of anaphora. A broader exploration of pronoun systems and languages would therefore be welcomed, as well as studies presenting more than two potential referents for a pronoun.

In addition to the modeling outcomes, the current study reveals some general patterns in the resolution of *dieser*, which has not been extensively empirically tested. In Experiment 1 the *dieser* prompt showed a strong resolution to the NP2/proto-patient, even when the proto-patient was a subject as in the dative contexts. In Experiment 2, conversely, *dieser* was resolved exclusively to the NP2/object in the ES conditions and showed a tendency to the NP2/object in the SE conditions. Taking both experiments together, *dieser* appears not to follow an anti-subject bias nor an anti-agent bias, contra several claims in the literature (Fuchs and Schumacher, 2020; Patil et al., 2020). The overall pattern is a strong preference to refer to the NP2, regardless of grammatical or thematic role. Indeed, the follow-up experiment to Experiment 2 showed that resolving *dieser* to NP1 was less felicitous than resolving to NP2, and less felicitous than resolving *er* to NP1, underlining a preference for the second-mentioned referent (at least in the limited set of contexts presented in our study). But the contrast in interpretations for *dieser* between ES and SE contexts does point to the interpretation being affected by the next-mention bias. However, the outcome of the follow-up experiment also underlines the challenge of testing pronouns with less flexible interpretation preferences: this can create conflict in some conditions when context biases and pronoun biases clash, leading to productions and/or interpretations that would not normally be considered by participants.

Finally, in this study we attempted to assess the relative contribution of agentivity and subjecthood to pronoun production likelihoods. The strong form of the Bayesian model claims that likelihoods should be affected by subjecthood (and/or topichood); studies of German pronouns have shown that agentivity is important for interpretation, but until now it has not been demonstrated whether agentivity influences expectations about an upcoming referent or acts on the likelihood of producing a pronoun. The pattern for production likelihoods in Experiment 1 revealed an agentivity influence on likelihoods (proto-agents for personal pronouns and proto-patients for demonstratives), but the pattern for personal pronouns was unclear. Particularly striking were the production biases seen in dative contexts, where personal pronouns were the preferred referential forms for both potential referents – an effect not seen in previous studies. Whereas previous work has argued that subjecthood leads to a strong pronominalization bias, this study is the first to show that this bias applies to subjects that are not NP1 in argument structure. In Experiment 2 the production likelihoods were only affected by grammatical role (personal pronouns produced for subjects, demonstratives for objects) and there was no evidence of agentivity having an influence. It should be noted that the contrast in agentivity features between experiencers and stimuli (i.e., between NP1 and NP2 in SE and ES contexts) is not very large, certainly not as clear as the contrast between proto-agents and proto-patients in accusative and dative verbs. Some research has suggested that in SE contexts the stimulus is more "agent-like" than the experiencer (Dowty, 1991). This could have led to the agentivity influence not being detectable in Experiment 2. However, having the two factors, grammatical role and agentivity, being manipulated via verb type makes it harder to interpret the claims about a lack of verb type influence on likelihoods, as would be predicted under the strong form of the Bayesian model. Overall, the results from both experiments make it difficult to draw firm conclusions about the relative contributions of these factors. This aspect should be tested further in future experiments with an altered design.

Our study makes the following contributions: by assessing performance in a Bayesian statistical framework, we have strengthened the quantitative evidence for the Bayesian model for pronouns. By testing German personal and demonstrative pronouns, we have extended cross-linguistic support for the Bayesian model and also applied it to a type of pronoun with a more rigid interpretation bias, showing the model's flexibility, while at the same time providing new insights into the comprehension of the German demonstrative *dieser*. The study also provides evidence in favor of the strong form of the Bayesian model, with verb type affecting the prior but not the production likelihoods of personal and demonstrative pronouns separately. However, given that overall, model performance was negatively affected by the removal of verb type information from the production likelihoods, there is room for speculation that the dissociation of factors in the strong form of the Bayesian model could be moderated. Finally, the study was set up to provide clearer evidence about the role of agentivity versus subjecthood on the pronoun production likelihoods, but we are unable to draw strong conclusions here, and leave this question for future research.

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: Center for Open Science (OSF): osf.io/j5wtg.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethikkommission der Deutschen Gesellschaft für Sprachwissenschaft. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CP contributed to data curation, formal analysis, investigation, methodology, project administration, supervision, validation, writing – original draft, and writing – review and editing. PS contributed to conceptualization, formal analysis, funding acquisition, project administration, supervision, and writing – review and editing. BN contributed to formal analysis, methodology, visualization, and writing – review and editing. JH contributed to data curation, investigation, methodology, resources, validation, and writing – original draft. AK contributed to conceptualization, methodology, and writing – review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

The research for this study has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 281511265 – SFB 1252 “Prominence in Language” in the project C07 “Forward and backward functions of discourse anaphora” at the University of Cologne, Department of German Language und Literature I, Linguistics.

ACKNOWLEDGMENTS

The authors wish to thank the following people for their assistance: Claudia Kilter for annotation; Brita Rietdorf, Markus Philipp, and Ilka Huesmann for assistance with data collection; Magdalena Repp for setting up the follow-up experiment; and Julia Veth for assistance with manuscript preparation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.672927/full#supplementary-material

REFERENCES

Ahrenholz, B. (2007). *Verweise mit Demonstrativa im Gesprochenen Deutsch: Grammatik, Zweitspracherwerb und Deutsch als Fremdsprache*. Berlin: De Gruyter.

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. London: Routledge.

Arnold, J. E. (1998). *Reference Form and Discourse Patterns*. Ph.D. thesis. Stanford, CA: Stanford University.

Arnold, J. E. (2001). The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Process.* 31, 137–162.

Arnold, J. E. (2010). How speakers refer: the role of accessibility. *Lang. Linguist. Compass* 4, 187–203.

Arnold, J. E., and Tanenhaus, M. K. (2011). “Disfluency effects in comprehension: how new information can become accessible,” in *The Processing and Acquisition of Reference*, eds E. Gibson and N. Perlmutter (Cambridge: MIT Press).

Arnold, J. E., Brown-Schmidt, S., and Trueswell, J. (2007). Children's use of gender and order-of-mention during pronoun comprehension. *Lang. Cogn. Process.* 22, 527–565. doi: 10.1080/01690960600845950

Bader, M., and Portele, Y. (2019). The interpretation of German personal pronouns and d-pronouns. *Z. Sprachwiss.* 38, 155–190. doi: 10.1515/zfs-2019-2002

Bengtsson, H. (2020). *matrixStats: Functions That Apply to Rows and Columns of Matrices (and to Vectors)*. Available online at: https://CRAN.R-project.org/package=matrixStats (accessed February 26, 2021).

Bethke, I. (1990). *Der, die, das als Pronomen*. Munich: Iudicium Verlag.

Bosch, P., and Hinterwimmer, S. (2016). “Anaphoric reference by demonstrative pronouns in German,” in *Empirical Perspectives on Anaphora Resolution*, eds A. Holler, C. Goeb, and K. Suckow (Berlin: De Gruyter), 193–212. doi: 10.1515/9783110464108-010

Bosch, P., and Umbach, C. (2007). “Reference determination for demonstrative pronouns,” in *Proceedings of the Conference on Intersentential Pronominal Reference in Child and Adult Language*, Vol. 48, eds D. Bittner and N. Gagarina (Berlin: Zentrum für Allgemeine Sprachwissenschaft).

Bosch, P., Katz, G., and Umbach, C. (2007). “The non-subject bias of German demonstrative pronouns,” in *Anaphors in Text: Cognitive, Formal and Applied Approaches to Anaphoric Reference*, eds M. Schwarz-Friesel, M. Consten, and M. Knees (Amsterdam: John Benjamins), 145–164. doi: 10.1075/slcs.86.13bos

Bosch, P., Rozario, T., and Zhao, Y. (2003). “Demonstrative pronouns and personal pronouns: German der versus er,” in *Proceedings of the EACL 2003 Workshop on the Computational Treatment of Anaphora*, Budapest.

Bott, O., and Solstad, T. (2014). “From verbs to discourse: a novel account of implicit causality,” in *Psycholinguistic Approaches to Meaning and Understanding Across Languages*, eds B. Hemforth, B. Mertins, and C. Fabricius-Hansen (Cham: Springer International Publishing), 213–251. doi: 10.1007/978-3-319-05675-3_9

Bouma, G., and Hopp, H. (2006). “Effects of word order and grammatical function on pronoun resolution in German,” in *Proceedings of the Ambiguity in Anaphora*

Workshop Proceedings, eds R. Artstein and M. Poesio (Essex: University of Essex), 5–13.

Bouma, G., and Hopp, H. (2007). "Coreference preferences for personal pronouns in German," in *Intersentential Pronominal Reference in Child and Adult Language*, Vol. 48, eds D. Bittner and N. Gagarina (Berlin: Zentrum für Allgemeine Sprachwissenschaft), 53–74. doi: 10.21248/zaspil.48.2007.354

Bürkner, P.-C. (2017). brms: an R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* 80, 1–28. doi: 10.18637/jss.v080.i01

Chang, W. (2013). *Plotting Means and Error Bars (ggplot2). Cookbook for R*. Available online at: http://www.cookbook-r.com/Graphs/Plotting_means_and_error_bars_(ggplot2)/ (accessed February 26, 2021).

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., and Dorie, V. (2015). Weakly informative prior for point estimation of covariance matrices in hierarchical models. *J. Educ. Behav. Stat.* 40, 136–157. doi: 10.3102/1076998615570945

Clark, H. H., and Sengul, C. J. (1979). In search of referents for nouns and pronouns. *Mem. Cogn.* 7, 35–41. doi: 10.3758/BF03196932

Çokal, D., Sturt, P., and Ferreira, F. (2018). Processing of it and this in written narrative discourse. *Discourse Process.* 55, 272–289. doi: 10.1080/0163853X.2016.1236231

Crawley, R. A., and Stevenson, R. J. (1990). Reference in single sentences and in texts. *J. Psycholinguist. Res.* 19, 191–210. doi: 10.1007/BF01077416

Crawley, R. A., Stevenson, R. J., and Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. *J. Psycholinguist. Res.* 19, 245–264. doi: 10.1007/BF01077259

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language* 67, 547–619.

Fuchs, M. (2021). *Demonstrative Pronouns and Attention-Orienting*. Ph.D. thesis. Köln: University of Cologne.

Fuchs, M., and Schumacher, P. B. (2020). "Referential shift potential of demonstrative pronouns—Evidence from text continuation," in *Demonstratives in Discourse*, eds A. Næss, A. Margetts, and Y. Treis (Berlin: Language Science Press), 185–213.

Fukumura, K., and van Gompel, R. P. (2010). Choosing anaphoric expressions: do people take into account likelihood of reference? *J. Mem. Lang.* 62, 52–66. doi: 10.1016/j.jml.2009.09.001

Gabry, J., and Češnovar, R. (2020). *cmdstanr: R Interface to 'CmdStan'*.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019). Visualization in Bayesian workflow. *J. R. Stat. Soc. A* 182, 389–402. doi: 10.1111/rssa.12378

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B., et al. (2013). *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall. doi: 10.1201/b16018

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* 2, 1360–1383. doi: 10.1214/08-AOAS191

Gernsbacher, M. A., and Hargreaves, D. J. (1988). Accessing sentence participants: the advantage of first mention. *J. Mem. Lang.* 27, 699–717. doi: 10.1016/0749-596X(88)90016-2

Gordon, P. C., Grosz, B. J., and Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cogn. Sci.* 17, 311–347. doi: 10.1207/s15516709cog1703_1

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: a framework for modeling the local coherence of discourse. *Comput. Linguist.* 21, 203–225.

Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language* 69, 274–307. doi: 10.3389/fpsyg.2021.623648

Henry, L., and Wickham, H. (2020). *purrr: Functional Programming Tools*. Available online at: https://CRAN.R-project.org/package=purrr (accessed February 26, 2021).

Hinterwimmer, S. (2015). "A unified account of the properties of German demonstrative pronouns," in *Proceedings of the Workshop on Pronominal Semantics at NELS 40*, eds P. Grosz, P. Patel-Grosz, and I. Yanovich (Amherst, MA: GLSA Publications), 61–107.

Järvikivi, J., van Gompel, R. P. G., Hyönä, J., and Bertram, R. (2005). Ambiguous pronoun resolution: contrasting the first-mention and subject-preference accounts. *Psychol. Sci.* 16, 260–264. doi: 10.1111/j.0956-7976.2005.01525.x

Jaynes, E. T., and Kempthorne, O. (1976). "Confidence intervals vs Bayesian intervals," in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, eds W. L. Harper and C. A. Hooker (Dordrecht: Springer Netherlands), 175–257. doi: 10.1007/978-94-010-1436-6_6

Kaiser, E. (2011). "On the relation between coherence relations and anaphoric demonstratives in German," in *Proceedings of Sinn und Bedeutung*, Vol. 15, eds I. Reich, E. Horch, and D. Pauly (Saarbrücken: Saarland University Press).

Kehler, A., and Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theor. Linguist.* 39, 1–37. doi: 10.1515/tl-2013-0001

Kehler, A., Kertz, L., Rohde, H., and Elman, J. L. (2008). Coherence and coreference revisited. *J. Semantics* 25, 1–44. doi: 10.1093/jos/ffm018

Koster, J., and McElreath, R. (2017). Multinomial analysis of behavior: statistical methods. *Behav. Ecol. Sociobiol.* 71:138. doi: 10.1007/s00265-017-2363-8

Morey, R. D. (2008). Confidence intervals from normalized data: a correction to Cousineau (2005). *Quant. Methods Psychol.* 4, 61–64. doi: 10.20982/tqmp.04.2.p061

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., and Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychon. Bull. Rev.* 23, 103–123. doi: 10.3758/s13423-015-0947-8

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

Patil, U., Bosch, P., and Hinterwimmer, S. (2020). Constraints on German diese demonstratives: language formality and subject-avoidance. *Glossa A J. Gen. Linguist.* 5:14.

Portele, Y., and Bader, M. (2016). Accessibility and referential choice: personal pronouns and D-pronouns in written German. *Discours Rev. Linguist. Psycholinguist. Inform. A J. Linguist. Psycholinguist. Comput. Linguist.* 18, 1–39.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rohde, H. (2008). *Coherence-Driven Effects in Sentence and Discourse Processing*. Ph.D. thesis. San Diego, CA: University of California.

Rohde, H., and Kehler, A. (2014). Grammatical and information-structural influences on pronoun production. *Lang. Cogn. Neurosci.* 29, 912–927. doi: 10.1080/01690965.2013.854918

Rosa, E. C., and Arnold, J. E. (2017). Predictability affects production: thematic roles can affect reference form selection. *J. Mem. Lang.* 94, 43–60. doi: 10.1016/j.jml.2016.07.007

RStudio Team (2019). *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc.

Schielzeth, H., and Forstmeier, W. (2008). Conclusions beyond support: overconfident estimates in mixed models. *Behav. Ecol.* 20, 416–420. doi: 10.1093/beheco/arn145

Schumacher, P. B., Backhaus, J., and Dangl, M. (2015). Backward- and forward-looking potential of anaphors. *Front. Psychol.* 6:1746. doi: 10.3389/fpsyg.2015.01746

Schumacher, P. B., Dangl, M., and Uzun, E. (2016). "Thematic role as prominence cue during pronoun resolution in German," in *Empirical Perspectives on Anaphora Resolution*, eds A. Holler and K. Suckow (Berlin: De Gruyter), 213–240. doi: 10.1515/9783110464108-011

Schumacher, P. B., Roberts, L., and Järvikivi, J. (2017). Agentivity drives real-time pronoun resolution: evidence from German er and der. *Lingua* 185, 25–41. doi: 10.1016/j.lingua.2016.07.004

Sivula, T., Magnusson, M., and Vehtari, A. (2020). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *arXiv* [Preprint]. arXiv:2008.10296.

Stan Development Team (2020). *Stan Modeling Language Users Guide and Reference Manual, Version 2.25*. Available online at: https://mc-stan.org (accessed February 26, 2021).

Stevenson, R. J., Crawley, R. A., and Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Lang. Cogn. Process.* 9, 519–548. doi: 10.1080/01690969408402130

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., and Kong, E. (2018). Bayesian data analysis in the phonetic sciences: a tutorial introduction. *J. Phonet.* 71, 147–161. doi: 10.1016/j.wocn.2018.07.008

Vehtari, A., and Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Stat. Surveys* 6, 142–228. doi: 10.1214/12-SS102

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* 27, 1413–1432. doi: 10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2020). Rank-normalization, folding, and localization: an improved Rhat for assessing convergence of MCMC. *Bayesian Anal.* 16, 667–718.

Weatherford, K. C., and Arnold, J. E. (2021). Semantic predictability of implicit causality can affect referential form choice. *Cognition* 214:104759. doi: 10.1016/j.cognition.2021.104759

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Wickham, H. (2020). *tidyr: Tidy Messy Data*. Available online at: https://CRAN.R-project.org/package=tidyr (accessed February 26, 2021).

Wickham, H., François, R., Henry, L., and Müller, K. (2020). *dplyr: A Grammar of Data Manipulation*. Available online at: https://CRAN.R-project.org/package=dplyr (accessed February 26, 2021).

Wiemer, B. (1996). Die Personalpronomina er. Vs. Der. Und ihre textsemantischen Funktionen. *Deutsche Sprache* 24, 71–91.

Wilson, F. (2009). *Processing at the Syntax–Discourse Interface in Second Language Acquisition.* Ph.D. thesis. Edinburgh: University of Edinburgh.

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2017). Using stacking to average Bayesian predictive distributions. *Bayesian Anal.* 13, 917–1007. doi: 10.1214/17-BA1091

Zerkle, S. A., and Arnold, J. E. (2019). Does planning explain why predictability affects reference production? *Dialogue Discourse* 10, 43–55.

Zhan, M., Levy, R., and Kehler, A. (2020). Pronoun interpretation in Mandarin Chinese follows principles of Bayesian inference. *PLoS One* 15:e0237012. doi: 10.1371/journal.pone.0237012

Zhu, H. (2020). *kableExtra: Construct ComplexTable With 'kable' and Pipe Syntax*. Available online at: https://CRAN.R-project.org/package=kableExtra (accessed February 26, 2021).

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership