Cognition 230 (2023) 105265

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

A chimpanzee by any other name: The contributions of utterance context and information density on word choice

Cassandra L. Jacobs [a,b,*], Maryellen C. MacDonald [b]

[a] Department of Linguistics, University at Buffalo, Buffalo, NY, United States of America

[b] Department of Psychology, University of Wisconsin-Madison, Madison, WI, United States of America

ARTICLE INFO

Keywords:
Language production
Surprisal
Neural network
Reduction
Lexical selection

ABSTRACT

An important feature of language production is the flexibility of lexical selection; producers could refer to an animal as *chimpanzee, chimp, ape, she*, and so on. Thus, a key question for psycholinguistic research is how and why producers make the lexical selections that they do. Information theoretic approaches have argued that producers regulate the uncertainty of the utterance for comprehenders, for example using longer words like *chimpanzee* if their messages are likely to be misunderstood, and shorter ones like *chimp* when the message is easy to understand. In this work, we test for the relative contributions of the information theoretic approach and an approach more aligned with psycholinguistic models of language production. We examine the effect on lexical selection of whole utterance-level factors that we take as a proxy for register or style in message-driven production accounts. Using a modern machine learning-oriented approach, we show that for both naturalistic stimuli and real-world corpora, producers prefer words to be longer in systematically different contexts, independent of the specific message they are trying to convey. We do not find evidence for regulation of uncertainty, as in information theoretic approaches. We offer suggestions for modification of the standard psycholinguistic production approach that emphasizes the need for the field to specify how message formulation influences lexical choice in multiword utterances.

1. Introduction

When we want to speak, sign, or write, complex machinery is engaged to promote successful language production. From the smallest variation in how we pronounce words, to the way we tell stories, language producers are influenced by what they have already said, what others have said, and what they want to say next. One aspect of language use—lexical choice—is influenced both by short-term demands on the production system, such as the phonological forms of words that have been recently produced (Dell & O'Seaghdha, 1992; Sevald & Dell, 1994), as well as by longer-ranging, discourse-level factors such as whether a referent has already been mentioned (Bard et al., 2000; Clark & Marshall, 1981). A lifetime of exposure to linguistic structural regularities helps producers of a language be successful, and abundant research has investigated how these patterns are learned, refined, and evolve over time (e.g., Dell & Chang, 2014; MacDonald, 2013). One central question is how both lifelong patterns of language use and the current context shape the form of utterances, especially how producers arrive at a given lexical choice and what forms that lexical choice may take. In this work, we specifically probe the degree to which properties of an utterance at different levels influence lexical choice. We first review some factors shaping lexical choice and other variation in utterance form.

1.1. Forces underlying variation in utterance form

Understanding how lexical choice unfolds is critical for being able to explain why producers' utterances have the character they do. Researchers have paid perhaps the greatest attention to how producers select the forms of referring expressions – such as “the cat” or “Fluffy” – depending on different contextual factors, such as the producers' own knowledge and the discourse status of the referent and/or other potential referents. Referring expressions can vary widely in both their length and the degree to which a label uniquely identifies a referent (Arnold & Griffin, 2007; Slevc, Wardlow Lane, & Ferreira, 2007). For example, one area of investigation has focused on when producers choose to produce a pronoun, such as the selection of the words *they*, *she*, or *these*, as opposed to a full noun phrase, such as *the doctor*, *the magician,* or *the green candies on the table* (Grosz, Joshi, & Weinstein, 1983). In general, producers tend

[*] Corresponding author at: Department of Linguistics, 609 Baldy Hall, University at Buffalo, North Campus, Buffalo, NY 14260-6420, United States of America.
E-mail address: cxjacobs@buffalo.edu (C.L. Jacobs).

https://doi.org/10.1016/j.cognition.2022.105265
Received 21 October 2021; Received in revised form 16 August 2022; Accepted 24 August 2022
Available online 12 September 2022

to produce longer referring expressions when a referent is relatively new to the discourse (Gundel, Bassene, Gordon, Humnick, & Khalfaoui, 2010; Gundel, Hedberg, & Zacharski, 1993), when the referent is unpredictable (Rosa & Arnold, 2017; Weatherford & Arnold, 2021), or when they need to contrast between a current referent and a prior referent (Hint, Nahkola, & Pajusalu, 2020; Watson, 2010). Importantly, pronouns provide one potential dimension of linguistic *reduction*, or the shortening of the forms of referring expressions, which may take place at several levels of linguistic representation, which we turn to below.

In addition to varying word choice or syntactic structure, producers can also adjust the physical form of referring expressions. For example, acoustic reduction is the shortening of aspects of the acoustic signal, such as a vowel's intensity or a word's duration, which often occurs when entities have been mentioned previously (Bard et al., 2000; Kahn & Arnold, 2012; Lam & Watson, 2010), or when a word or sound is statistically probable (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Cohen Priva, 2017; Seyfarth, 2014). A related phenomenon known as syntactic reduction occurs within larger structures such as full utterances, as in constructions such as the English "optional *that*", which permits the omission of a relative pronoun (Bresnan, 1972; Ferreira, 2003; Ferreira & Dell, 2000), especially when upcoming information is predictable from the preceding context (Jaeger, 2010).

The degree of reduction in the form of a linguistic choice can also be modulated by stylistic or interpersonal factors, such as their register, accent, or dialect. For example, shortened, phonologically reduced forms (i.e., "wanna" and "gonna") may be more commonly used around friends than in a university classroom, where discourse tends to be more formal and pronunciations may be closer to those given in dictionaries (Bybee & Thompson, 1997; Pavlick & Tetreault, 2016). The style that a producer uses at any given moment depends on what their audience knows and understands (Bell, 1984) as well as socio-indexical factors (i.e., aspects of the producer such as their age, gender, etc. or aspects of the addressee; Tagg & Seargeant, 2014; Kemper, 1994), which can influence the entire vocabulary and grammatical structures they have access to.

Together, these findings suggest that producers have abundant flexibility in both the choice of referring expressions and their phonological and acoustic realization, though both aspects may be influenced by the context in which the utterance is produced. We note that despite abundant work identifying the many forces shaping utterance form, an account of how producers weigh these various factors is still needed.

1.2. Statistical context and utterance form

Our entry into producers' choice of referring expressions in the present study follows a clever approach developed by Mahowald, Fedorenko, Piantadosi, and Gibson (2013), who examined reduction using closely related terms that can refer to the same entity, specifically morpho-phonologically related pairs of lexical alternatives such as the shorter *chimp* and the longer *chimpanzee*. These pairs of words are interesting from a psycholinguistic standpoint because they allow for a direct comparison between words that are highly orthographically and semantically similar, potentially permitting insight into why producers prefer a particular term in different contexts. While pairs like *chimp* and *chimpanzee* could in principle alternate for any of the reasons already discussed, Mahowald et al. proposed a specific account of use of short vs. long forms that appears to diverge from at least some accounts of language production. We describe the Mahowald et al. approach in more detail below, but it is first useful to view the alternative proposals in broad form.

In brief, Mahowald et al. (2013) propose an information theoretic account of language production (Aylett & Turk, 2004; Jaeger, 2010; Piantadosi, Tily, & Gibson, 2011), in which the statistical properties of an utterance, often properties of the words previously produced, modulate the choice of a subsequent word, including choices between short/long forms like *chimp*/*chimpanzee*. Under smooth signal (Aylett & Turk, 2004) and uniform information density accounts (Jaeger, 2010; Levy & Jaeger, 2007; Manin, 2006), the form properties of words or other linguistic structures should vary with their probability in context. Producers are assumed to select words in a highly incremental fashion (Jaeger, 2010), which allows for recently produced words to affect the form of upcoming words, such as whether *chimpanzee* or *chimp* is produced. More concretely, these theoretical accounts state that producers attempt to *smooth out* the unexpectedness of a "message" by changing the forms of what is to be produced to compensate for the expectedness of that message. On this view, words that are less expected in the context of recently produced words should be encoded as longer forms in order to ensure that the producer's message will be properly understood (Aylett & Turk, 2004; Jaeger, 2010; Piantadosi et al., 2011), whereas "it is most efficient to assign shorter codes [words] to those elements that convey less information" (Mahowald et al., 2013, p. 314). Mahowald et al.'s approach follows directly from Shannon's (1948) noisy channel model of communication, which formulates constraints on the optimal forms to ensure signal transmission. This idea is illustrated in Fig. 1A, where the probability of the upcoming message (e.g., the referent) following previously produced words in an utterance ("*Primatologists hypothesize that*") is thought to directly affect producers' choice of the next word, perhaps favoring *chimpanzees* over *chimps*.

Information theoretic approaches to language production inhabit Marr's (1982) computational level of analysis. The statement that the form of a signal is optimized to some extent in context is fundamentally a description of statistical phenomena evident in corpus analyses and laboratory studies. However, the claim that producers "actively control information rate" (Mahowald et al., 2013, p. 317), requires an algorithmic explanation (again in Marr's terms) characterizing the producer's cognitive processes to achieve this control over information rate. Such an explanation would need to specify what linguistic representations are used in the language production systems, how they are being processed, and in what order.

An alternative view that is inspired by psycholinguistic production models (Chang, Dell, & Bock, 2006; Dell, 1986; Levelt, Roelofs, & Meyer, 1999; Oppenheim, Dell, & Schwartz, 2010) is shown in Fig. 1B. Typically structured as neural network connectionist models, these language production models encode some intended meaning representations (e.g., a label for an image), which are used as input to subsequent levels of processing (Levelt et al., 1999), and which ultimately result in signed or spoken utterances. At the level of production of individual words, which are the dominant unit of study in psycholinguistic tasks (Schriefers, Meyer, & Levelt, 1990), the producer is assumed to be engaged in conceptualization of a *message* that they aim to convey. However, at the level of an entire utterance, conceptualization must directly drive the parallel selection of many words that are suitable to convey a more complex message (Dell & O'Seaghdha, 1992; Kuiper, Van Egmond, Kempen, & Sprenger, 2007; Levelt, 1989; McCauley et al., 2021).

When messages are allowed to vary with the context, a producer's specific lexical choice will depend not only a semantic match to the message (e.g., *chimpanzee* is a better match for some intended message than *gorilla*) as in the standard model, but also other more wide-reaching properties of the discourse, such as the degree of shared knowledge between the interlocutors (Yoon & Brown-Schmidt, 2014) or discourse register (Bentum, Ten Bosch, Van den Bosch, & Ernestus, 2019; Biber, 2012; Brooke & Hirst, 2014; Levelt, 1989; Pavlick & Tetreault, 2016). For a producer aiming to convey a message in a more formal register, aspects of the message promote certain lexical forms over others at every point—*primatologists* rather than *scientists, hypothesize* over *believe,* and *chimpanzees* instead of *chimps*. This parallel selection of discourse-appropriate words will naturally lead to correlations between their properties (their formality, frequency, etc.). If standard models of language production can be modified to accommodate broader message types than single word production, then these models may be appropriate for an algorithmic-level explanation (Marr, 1982) of the language production process, which is critical for building a mechanistic account

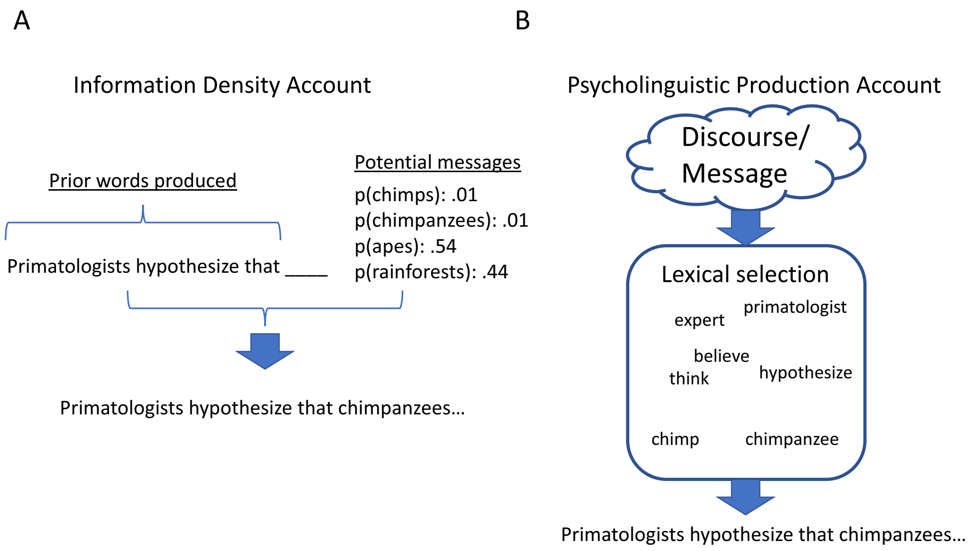

Fig. 1. A. In a Uniform Information Density account, the choice of a word is affected by its predictability in context. When a message, or concept to be communicated, is unpredictable in the context of recently uttered words, long words should be preferred. In this example, if the terms *chimp* and *chimpanzee* are unexpected, then *Primatologists* and *hypothesize* promotes the longer, higher-information word *chimpanzee* over the shorter alternative *chimp*.
B. In psycholinguistic accounts of lexical selection (e.g., Levelt et al., 1999), qualities of the message to be conveyed drive the formulation of an utterance plan, which leads to the parallel activation and selection of the words used to convey the message. In this example, a highly academic message promotes use of words like *primatologist, hypothesize, and chimpanzee.*

of *how* linguistic regularities that are believed to drive lexical choice are involved in the time course of message planning and production.

Fig. 1B shares several similarities with the discourse-sensitive, multiword utterance proposal laid out in Chapter 4 of Levelt (1989). In particular, Levelt characterizes the language production process as the transformation of a set of intentions, a subset of which are communicative, and a further subset of which are illocutionary. These intentions correspond to the "message" which contains a macro plan composed of a series of speech acts, within which individual words are planned. In this framework, the high-level intention of an utterance is thought to directly influence lexical selection for each of the physical outputs. In contrast to the account laid out in Levelt et al. (1999), however, this account (and our informal characterization of the production process) is a verbal model and lacks a clear computational basis. In the General Discussion, in light of the results we present in Experiments 1 and 2, we propose a path toward a more mechanistic account of language production that permits the joint selection of lexical items that continues in the same tradition as early connectionist models of single word production and which may bring the field closer to an algorithmic theory of production, manifesting the account of Levelt (1989) or others who have incorporated high level information such as discourse or illocution into their verbal models.

Thus, the key difference between the approaches is that in Fig. 1A, predictability and signal-smoothing have a direct effect on the form of subsequent words, whereas in the account in 1B, lexical choices stem from properties of the entire utterance that the speaker is planning. Note that we have not yet specified for either approach exactly what features of the prior words or utterance plan exert an influence on lexical choices – we simply use this figure to illustrate the different character of the two hypotheses. In this paper, we focus on the contrast between these approaches, specifically the degree to which lexical choices depend on the local probability of a message being understood (as in Fig. 1A) vs. factors that could be traced to the producer's message (Fig. 1B). With this question in mind, we turn to the specifics of the Mahowald et al. study next.

1.3. Short vs. long word choices

Mahowald et al. (2013) conducted a corpus analysis and experiment to test whether uniform information density principles could inform accounts of lexical choice among short/long wordforms like *chimp/chimpanzee*. Their corpus study tested whether an information theoretic measure known as information content was systematically different between the short and long words in a pair. To compute the information content of a word (e.g., *chimp*), Mahowald et al. followed the approach of Piantadosi et al. (2011) and extracted all of the three-word sequences (*trigrams*) containing critical target words from the Google n-grams corpus (Brants & Franz, 2006) and computed the conditional probability of that word given the two immediately preceding words (i.e., $p(w_3 \mid w_1 w_2)$), from which they derived *trigram surprisal* (the negative log transform of this conditional probability). Trigram surprisals for a given word (e.g., *chimp*) in context were then averaged to determine the information content of a word. Consistent with prior work on length-information content relationships (Manin, 2006; Piantadosi et al., 2011; Wimmer, Köhler, Grotjahn, & Altmann, 1994), Mahowald et al. found that longer words (*chimpanzee*) typically had higher information content relative to the shorter word (*chimp*). While this correlation between the statistical properties of the recently produced words and the length of the target word is consistent with Mahowald et al.'s claim that producers should prefer shorter forms in predictable contexts, the correlation may also be consistent with the message-based account illustrated in Fig. 1B. On this view, the nature of the producer's multiword-level message will influence many word choices in the utterance, creating correlations among word properties (Evert, 2005) that need not stem from producers' attempts to modulate the information density of an utterance.

These message considerations are relevant to the second phase of Mahowald et al.'s (2013) investigation, a survey of participants' preferences in a binary forced choice sentence completion task containing sentences that either had high ("supportive") or low ("neutral") contextual predictability of the terms. Their sentence stimuli contained a preamble to be completed by one of two targets as the final word. Supportive contexts such as, "Susan was very bad at algebra, so she hated…" had specific intended completions, such as *math* or *mathematics.* In neutral sentences such as, "Susan introduced herself to me as someone who loved…", the identity of the final word was far less constrained. Supportive and neutral sentences were defined predominantly by the cloze probabilities of the critical final words (collected via a separate group of participants): supportive sentences had a summed short+long cloze probability of 52%, while neutral sentences had a summed short+long cloze probability of 2%. Mahowald et al. did not relate participants' short vs. long completion choices to trigram surprisal as in the corpus analysis discussed above. Mahowald et al. found a relationship between this contextual support measure and preferences for long vs. short forms in the forced choice task: participants preferred the short form in the high cloze predictive sentences (short form preferred 67% of the time) compared to low cloze neutral contexts (short preferred 56% of the time). Mahowald et al. concluded that people prefer short forms when the message behind the next word is highly predictable, e.g., when the sum of the predictability of *chimp* and *chimpanzee* is high, such that the perceiver could easily predict upcoming reference to this type of primate. This position is in line with information theoretic accounts of reduction in the face of predictable messages (Aylett & Turk, 2004; Jaeger, 2010). This result has been replicated in similar stimuli involving different types of words and in other languages (Zarcone & Demberg, 2021).

An important concern in this work is the claim in information theoretic studies that alternative wordforms are fully equivalent, including the short/long alternatives like *chimp* and *chimpanzee*, but also overt vs. omitted optional words such as *that* in English, pronouns versus full noun phrases, etc. (Hale, 2003; Jaeger, 2010). For smoothing processes to work, it is critical that selecting between long vs. short forms like *chimpanzee* and *chimp* modulate only the information content (surprisal) of the utterance, without substantially distorting the producer's message or obscuring the producer's communicative goals (Mahowald et al., 2013). This view conflicts with the perspective in usage-based approaches in linguistics and psycholinguistics, which holds that different forms entail different meanings. For example, most short forms are derived through regular morpho-phonological processes from the long forms (e.g., *chimp* is considered a clipping of *chimpanzee*; Harley, 2017). Producers appear to use these devices to invent shortened terms for interpersonal communicative effect (Bauer, 2012; Fandrych, 2004; Marchand, 1969; Mattiello, 2013). As a language community adopts shortened forms of a word, the new forms acquire different senses, resulting from their contextual differentiation (Berg, 2011). Some contextual factors affecting a producer's use of short and long wordforms may include the intended register, as noted above, or the effect an utterance is meant to have on the emotional state of the addressee (Bauer, 2012; Fandrych, 2004; Marchand, 1969; Mattiello, 2013).

Computational approaches within natural language processing also typically assume that distinct word forms should have distinct (though possibly highly similar) representations for the two words (Firth, 1957; Landauer & Dumais, 1997; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). In modern distributional semantics models, when words like *chimp* and *chimpanzee* typically occur in distinct linguistic contexts, their representations are typically further away from each other in their semantics (Lund & Burgess, 1996; Landauer & Dumais, 1997; Mikolov et al., 2013; Peters et al., 2018; Ethayarajh, Duvenaud, & Hirst, 2019). These computational results support the proposal that short vs. long forms are used to express different ideas. Thus, aspects of both the natural language processing and linguistic literatures on these morpho-phonologically related alternatives suggests that short/long forms are not fully interchangeable. On this view, lexical choices between short and long forms stem from aspects of the producer's message; therefore, the dominant contributing factor to lexical choice may not stem from any active decision on the part of the producer to modulate information density.

The present work addresses these questions and provides four broad scientific contributions. First, we aimed to conduct a more thorough investigation of word length preferences using a greater variety of linguistic contexts for each pair of lexical alternatives, with both more participants and more sentences for judgment than the Mahowald et al. study. Second, we created multiple supportive sentence contexts for *both* the short and long forms of each word pair to better test how contextual factors mediate participants' ratings of the goodness of short vs. long forms. Third, this work leverages a modern neural language model and other computational approaches to test Mahowald et al.'s predictability-length claims and also to test claims for the message-driven approach to word choice. These first three methods are brought to bear in Experiment 1, where we investigate the extent to which participants' preferences for short vs. long forms are predicted from estimates of lexical predictability (neural surprisal; Frank & Bod, 2011), as would be consistent with Mahowald et al.'s approach. We also use a neural language model to assess semantic properties of the global sentence contexts as a proxy for the discourse context and register of our sentences, and we ask whether this utterance-based measure predicts word length preferences, as would be predicted by a more global account of lexical selection. Finally, we extend theses analyses of word choice to existing corpora. In Experiment 2, we apply our models of participants' judgments to real-world language from corpora and show out-of-sample validity for our method, further testing message-based approaches.

2. Experiment 1

If shortened forms of words (e.g., *chimp*) convey different messages than the source words that they are created from (i.e., *chimpanzee*), or if short and long forms are used for other discourse or socio-indexical reasons, then it is important to reexamine some assumptions about alternations between morpho-phonologically related pairs in the Mahowald et al. studies. In a behavioral experiment, we assessed participants' preferences for short or long forms of alternating pairs across different sentence contexts, which were provided in the form of a rating task. In addition to statistical measures such as surprisal taken from a modern neural language model (Ng et al., 2019), we also test whether properties of the surrounding linguistic context predict participants' ratings using a machine learning approach. Specifically, we built classifiers to predict whether participants' preferences for short vs. long forms could be accounted for by general properties of the entire utterance, which we take as a proxy for the producer's message. We then compare this approach to one in which participants' preferences depend on statistical measures of predictability given the immediately preceding context.

2.1. Method

2.1.1. Participants

Ninety-one self-reported native speakers of English from the University of Wisconsin-Madison participant pool were recruited to take part in this study in exchange for extra credit in their introductory psychology class. As part of a demographic survey included in the study, one participant reported being a non-native producer of English and was excluded from analyses, leaving 90 participants for analysis.

2.1.2. Stimulus design and construction

To understand the role of context on lexical choice, we designed 10 sentences for each of 38 of the 39 original alternating pairs from Mahowald et al. (2013), after determining that one pair (*porn/pornography*) was sensitive material and therefore inappropriate for inclusion in the present study. We included all of the pairs that Mahowald et al. had identified in either their corpus study (Experiment 1) or their behavioral study (Experiment 2) for completeness and increased power for detecting the effect of context on lexical preferences. These alternating pairs included forward clippings, such as *chimp* and *chimpanzee*, backward clippings, such as *phone* and *telephone*, initialisms such as *identification* and *ID*, more complex shortenings with accompanying phonological change such as *bicycle* and *bike*, as well as some phrases shortened to acronyms (e.g., *air conditioning* to *A/C*).

We aimed to include a wide variety of sentences for rating to address two potentially critical issues with the Mahowald et al. stimuli. First, it is unclear how much the Mahowald et al. sentences may have been biased toward the specific short form of a target word rather than predicting the general message (the summed predictability of short and long forms). Second, the number of stimulus and sentences derived from these that were used in their study is unspecified, with a potential lower bound of 39 pairs obtained only by reconstructing from their figures. With only one sentence context for each pair, there is a concern that results are affected by idiosyncrasies of the exact contexts. To address these concerns, we created ten distinct sentence contexts for each alternating pair, five of which were designed to support the *longer* form, and five supporting the *shorter* form. We allowed target words to vary in their positions within the sentences, rather than restricting them to sentence-final position, in contrast to Mahowald et al. (2013), to allow us to test potential effects of material following the target short/long word.

In the first round of stimulus creation, three research assistants wrote sentences for a unique third of the target sentences, with a specific target word (e.g., *chimpanzee*) in mind. Then, two other research assistants rated the sentences for naturalness and the degree to which the sentence felt strongly biasing toward the intended completion. Following this, all research assistants met together to revise the sentences and agree on a final score. Sentences that were not clearly biasing toward the appropriate form were rewritten. This process was repeated separately for those stimuli with higher initial naturalness and bias ratings, resulting in 380 total sentences.

After completion of the second phase of stimulus revisions, the authors and research assistants further refined the sentences. We marked for additional revision those candidate sentences that were highly similar to other stimuli in order to prevent sentences from having substantial lexical or semantic overlap (e.g., two sentences discussing getting a milkshake or hamburger at Culver's, a fast-food restaurant chain located in parts of the US). As described in Appendix 1, we used a neural language model (DistilBERT; Sanh et al., 2019) to identify potential stimulus sentences that were highly similar to each other. Based on the distribution of sentence-sentence similarity scores, we set a threshold cosine similarity level (0.75 in a range from −1 to 1) and the first author revised highly similar sentences to depict more distinct events. Finally, to ensure that participants would read the sentences and to control for target word position across short- and long-biasing sentences, we revised sentences if the target words were very near the left boundary in a sentence by adding material to the beginnings of some sentences (e.g., "*To my surprise*, the billboard …"). These changes helped to equate the mean relative position of the critical words, which helps to normalize comparison of target word placement for sentences of differing lengths to a relative sentence position value between 0 and 1 ($\widehat{\mu}_{\text{short}} = 0.613$, $\widehat{\mu}_{\text{long}} = 0.624$). Sentences biasing long wordforms were slightly longer ($\widehat{\mu}_{\text{long}}$ 14.9 words vs. $\widehat{\mu}_{\text{short}} = 12.8$ words) and long wordforms typically occurred one word later in the sentence ($\widehat{\mu}_{\text{short}} = 8.46$ vs. $\widehat{\mu}_{\text{long}} = 9.50$). There was no difference in the relative placement (i.e., word position divided by the length of the entire sentence) of the blanks within the sentences for the rating task (one-sided $t(378) = 0.06$, p = n.s.).

2.2. Procedure

The experiment was presented as a survey on Qualtrics and took about 15 min to complete. First, each participant provided informed consent and then completed a short demographic questionnaire. As part of the study of interest, each participant's survey contained a different random sample of the sentences, one from each of the pairs of lexical alternatives, resulting in 38 total sentences rated by each individual participant. As in Mahowald et al. (2013), our participants were asked to rate the relative goodness of two lexical alternatives as part of a fill-in-the-blank task. However, unlike Mahowald et al. (2013), the present study allowed for a continuous, whole integer-valued scale ranging from −10 to +10, which was presented to participants as a continuous slider (without visible numerical values). At either end of the slider, one option (e.g., *chimp*) was presented as an endpoint on the left and the other (e.g., *chimpanzee*) on the right. The initial slider value was set at 0 for each trial. The endpoint position for the short and long words for a pair was randomized, such that short words and long words did not systematically appear on one side of the scale. In all cases, the short alternative was always coded to −10, and the long alternative always corresponded to +10. Because sentences were sampled randomly within each word pair, participants saw one sentence for each word pair but were not guaranteed to see an even split of short- and long-biasing sentences.

2.3. Results

We aimed to test two proposals in this study. First, following Mahowald et al. (2013), we sought to test whether producers prefer to use short words when surprisal is low, in keeping with information density accounts of language production (Aylett & Turk, 2004; Jaeger, 2010). Second, we aimed to develop utterance-level predictors related to a producer's message, specifically a formal or didactic register that might promote long words vs. a more casual register, which might promote short words. For example, the frequencies of words in the context (Biber, 1992; Nini, 2019) and the lengths of words in a document are all predictive of the degree of formal register use (Brooke & Hirst, 2014; Pavlick & Tetreault, 2016). We initially approximate this register factor by taking the median log lexical frequency of all other words in the utterance (Baayen, 2002). However, we make additional gains when we use a computational method built on a neural network language model to generate sentence-specific predictions about producers' preferences to test whether properties of messages modulate word form preferences.

2.4. Predicting short versus long ratings

We first analyzed participants' ratings for the short vs. long words in the 380 stimulus sentences; recall that lower values reflect greater preference for the short form in our integer-valued −10 to +10 coding. We found that sentence stimuli that we designed to be short-biased were rated lower (by-participants mean = −3.63; by-sentence mean = −3.59) than items we designed to be long-biased (by-participants mean = −0.29; by-sentence mean = −0.42). The difference in mean ratings across short- and long-biased sentences was significant both by participants (one-sided $t(178) = 9.38$; Fig. 2) and by sentences (one-sided $t(378) = 5.60$). These results show that the passages were effective in biasing ratings toward both the short and long alternatives. These data also show that participants are biased toward preferring the short form for most sentences we provided, even for sentences that we designed to encourage "long" preferences. This pattern was evident in the Mahowald et al. data as well.

We additionally observed that participants' ratings were skewed toward the endpoints of the rating scale (see Fig. 3 below), showing that many sentences elicited strong preferences. As a result, for all of our critical analyses of participants' ratings presented below, we dichotomized ratings into long-preferred (ratings from 0 to 10, inclusive) and short-preferred ratings (−10 to −1, inclusive), independent of whether a stimulus item was originally designated as long- or short-biased. With these binarized rating data, we fit logit mixed effects models implemented in `lmerTest` (version 3.0–1; R version 3.4.3; Kuznetsova, Brockhoff, & Christensen, 2017) with random intercepts and slopes by participants, random intercepts by Sentence, and nested random intercepts and slopes by Word nested within Long Form. Unless reported otherwise, full models converged with the standard `glmer` parameters. We use the `bobyqa` algorithm (Powell, 2009) when faced with convergence issues; if issues persisted, we simplified the random effects structure to remove high correlations between the random slopes.

2.5. Statistical and linguistic properties of the sentence context

2.5.1. Local neural surprisal

Our first goal is to test for potential contributions of information theoretical measures in predicting participants' ratings. Mahowald et al. (2013) showed that participants more strongly preferred the shorter word in more strongly constraining sentences than in more neutral ones, which they interpreted to support information theoretic accounts of lexical choice and specifically information density proposals that more predictive contexts should lead to the selection of shorter wordforms. Building on their approach, we can quantify constraint using surprisal estimated by a neural network (Frank, 2009; Goodkind & Bicknell, 2018; Monsalve, Frank, & Vigliocco, 2012), which is an information

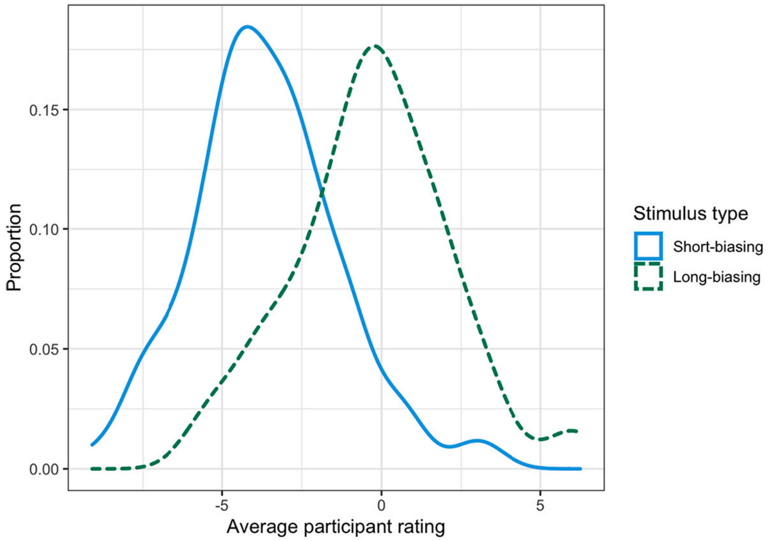

Fig. 2. By-participant average ratings across stimulus types. Participants typically rated long words as more appropriate in long-biasing sentences than in short-biasing sentences.

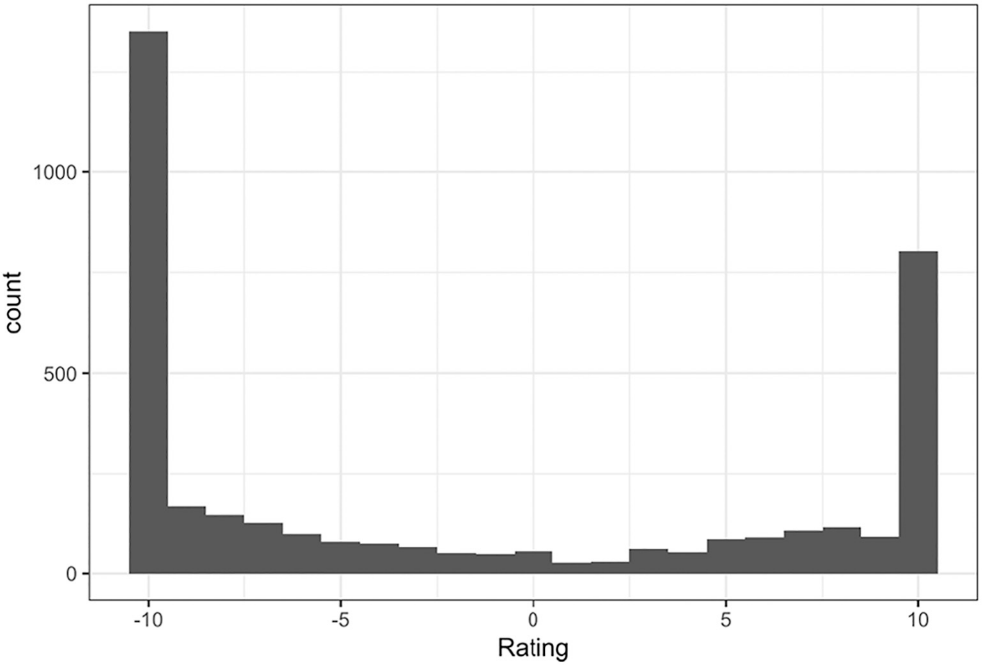

Fig. 3. Skew of participant ratings. The negative (left) side of the scale reflects preference for short forms of words and the positive (right) side reflects a long form preference.

theoretic estimate of the probability of a word given a context. Neural surprisal is now the most common automatic means of quantifying predictability and produces next-word probabilities (e.g., Frank, 2009; van Schijndel & Linzen, 2018), and it has been shown to align closely with cloze probabilities in human behavioral data (Eisape, Zaslavsky, & Levy, 2020; Jacobs & McCarthy, 2020), in addition to providing the closest fit to self-paced reading time data and eyetracking data among available surprisal measures (Goodkind & Bicknell, 2018). While it is theoretically feasible to obtain trigram surprisals as Mahowald et al. (2013) used in their corpus analysis, we note that Mahowald et al. (2013) did not use surprisal to predict participant short/long judgments; instead they used binarized cloze from other participants to predict short/long judgments. Additionally, for our naturalistic stimuli, we find trigram coverage to be quite poor, with only 43% percent of critical word contexts being represented in the Google n-gram corpus.

As a means of advancing beyond trigram surprisal estimates, we extracted surprisal values associated with both the short (*chimp*) and long (*chimpanzee*) terms of a pair using a neural network language model that can process the whole sentence up to the target word (i.e., for a target word at position *i*, $w_1 \ldots w_{i-1}$) unlike trigram language models (Frank, 2009; Goodkind & Bicknell, 2018). Following Jacobs and McCarthy (2020), we use a left-to-right language model that was initially trained to predict the next word using a cloze task-like task (Ng et al., 2019) to more closely approximate how human participants may approach these sentences. We discuss implementational details in Appendix 2.

In keeping with numerous studies using surprisal as a proxy for information density (e.g., Jaeger, 2010), we tested whether predictability leads to a preference for the shorter term. For each sentence, we created a summed predictability index (SPI) by summing the probabilities that the model assigned to both short and long completions (i.e., $p(\textit{chimp} \mid w_1 \ldots w_{i\text{-}1})$ and $p(\textit{chimpanzee} \mid w_1 \ldots w_{i\text{-}1})$) and computed surprisal as the negative log probability of that aggregate sum. That is, if the conditional probability of *chimp* was 0.0001 and the conditional probability of chimpanzee was 0.000001, then the summed predictability index would be -log(0.0001 + 0.000001). We included the SPI as fixed and random effects and found that local surprisal did not significantly predict participants' response preferences. In fact, including the SPI resulted in worse fit to participants' responses compared to a null model ($\chi^2(7) = 6.74$, p = n.s.), suggesting that predictability does not meaningfully account for producers' word preference between short and long forms in these materials. We summarize these results below in Table 1.

2.5.2. Word frequencies in the context

We next sought to expand our analyses into broader, message-level statistical predictors that may capture discourse register (Biber, 1992; Nini, 2019) and other factors potentially influencing word length. Of course, we do not have direct access to a producer's message or discourse register, and so our strategy here is to use analyses of the contexts surrounding the short/long target word to approximate these variables.

We first focus on an approximation of formal vs. informal discourse register using word frequency, an established feature used in discourse analysis (Biber, 1992; Nini, 2019). We conducted two analyses to assess the relationship between word frequency in the rest of the sentence and judgments of appropriateness in context. First, we computed the median log SUBTLEX frequency (Brysbaert & New, 2009) of all words *preceding* the critical word in our stimulus items in order to capture the idea that higher-frequency words will generally appear in more common trigrams (Evert, 2005) and also appear in more informal or casual discourse contexts, which might tend to promote the use of short forms of words. We computed the *median* of the log transform of the frequencies of words in the rest of the sentence in order to account for skew and reduce the importance of extreme frequency values. The frequency of the words upstream in the sentence was a significant predictor of participants' ratings on the judgment task, such that the higher the median log word frequency of the upstream context, the more likely participants were to rate the shorter word as sounding better ($\beta = -0.17$, Z = -2.07, $p < .05$).

We next examined whether this relationship between ratings and word frequency of the prior context also extended to the rest of the sentence after the target word. An effect of frequency of the downstream context is predicted if ratings vary with overall discourse register, which should carry through the whole sentence and not be limited to the words preceding the target word. A model using the median log word frequency of all words in the sentence excluding the target word was also strongly predictive of participants' ratings ($\beta = -0.39$, Z = -4.81, $p < .001$). We plot this result in Fig. 4 below. Importantly, the median word frequency of the full context provided a significantly better fit to the rating data than the median word frequencies of only the words in the preceding context ($\chi^2(0) = 21.73$, p < .001).

We also sought to verify that our stimuli did not contain strong contextual cues as to the missing target word's frequency. We correlated on a per-item basis, for all sentences containing single word alternations (e.g., chimp/chimpanzee but not multiword items such as US/United States), the correlation between the target word frequencies and the frequencies of the preceding words or full sentence. We found that the preceding context word frequencies were not correlated with target word frequency ($r = -0.05$, $t(287) = -0.89$, $p = .37$). The full context frequency likewise only showed a marginally significant correlation with the target word frequency ($r = 0.13$, $t(287) = 1.88$, $p = .06$). Taken together, these results suggest that participants' preferences for long vs. short form are predicted by a very broad effect of context that approximates the formal/informal nature of discourse registers, and preferences are not particularly influenced by the statistical properties of the words themselves.

Table 1
Model predicting participants' word form preferences using local neural surprisal.

Form type (long vs. short) ~
Nested random intercepts for pairs of alternates (Long Form / Word) +
Random intercepts and slopes for Sentence, Participant, Long Form/Word +
Summed predictability index

Name	β	*SE*	*Z*	p
Intercept	−0.69	0.27	−2.58	< 0.01
Summed predictability index (SPI)	0.09	0.11	0.83	n.s.

2.6. Representing the whole sentence context

We next sought to test how other linguistic properties of the whole utterance may be valid predictors of participants' responses. Whereas word frequency likely reflects aspects of formal vs. informal discourse register, there are other ways in which a producer's goals may affect choices of short vs. long words. For example, these utterance properties might be reflected in the syntactic structures used in the utterance (Ferreira, 1996; Peters et al., 2018; Haskell & MacDonald, 2003), the grammatical role of the target word and of other words in the sentence (McDonald, Bock, & Kelly, 1993), accessibility and plausibility (Bock & Warren, 1985), and so on. Additional insight on the relationship between messages and lexical choices may therefore be gained by further study of the relationship between utterance properties and lexical choices.

Our next analysis is most concerned with characterizing how utterance-level properties beyond word frequency constrain lexical choice. Specifically, to test the hypothesis that the linguistic properties of the entire sentence, not merely local surprisal, predict participants' ratings of short vs. long forms, we employed a neural network model trained on large quantities of English text, which can capture syntactic, semantic, and lexical information simultaneously (Rogers, Kovaleva, & Rumshisky, 2020). Contemporary neural networks can take strings of text and output a dense, low-dimensional vector known as an embedding, corresponding to the hidden states of the model. We do not claim that these models are cognitively plausible models of language processing; rather, we simply use the model to characterize our sentences on a high level, consistent with a view in which more abstract, message-level factors influence lexical choice throughout an utterance. If discourse register or other utterance-level factors influence producers' preferences for words of different lengths, then there should be common properties across utterances (reflecting register or other global factors) that tend to support long forms vs. those that tend to support and short wordforms, independent of the word or words in question. In other words, high-level utterance properties that predict preferences for *chimp* vs. *chimpanzee,* for example, should be relevant to long vs. short biases for other word pairs. Because of the complexity of the representations we extracted from these models, we developed a machine learning pipeline that enables us to understand this generalizability in a familiar mixed effects modeling context. In Fig. 5 below, we summarize the pipeline that we use to obtain predicted probabilities of participants' responses preferring either the short or long form of a word. Next, we provide a general description of our analysis; details are presented in Appendix 3.

First, we took the 380 sentences from our stimulus set and edited each sentence to hide the critical word of interest (Fig. 5, Panel 1). We segmented each sentence into its component words (Panel 2) for the neural language model to ingest. We then used a state-of-the-art neural network model (Panel 3; RoBERTa; Liu et al., 2019) to obtain vector representations (also known as *embeddings*) of all non-target words in a stimulus sentence and averaged these embeddings together to create a single embedding for that sentence. Then, for each of the 38 critical target pairs in our stimuli (e.g., *chimp/chimpanzee*), we built a unique regularized logistic regression model trained to predict participants' ratings from sentence embeddings using ratings by all participants for

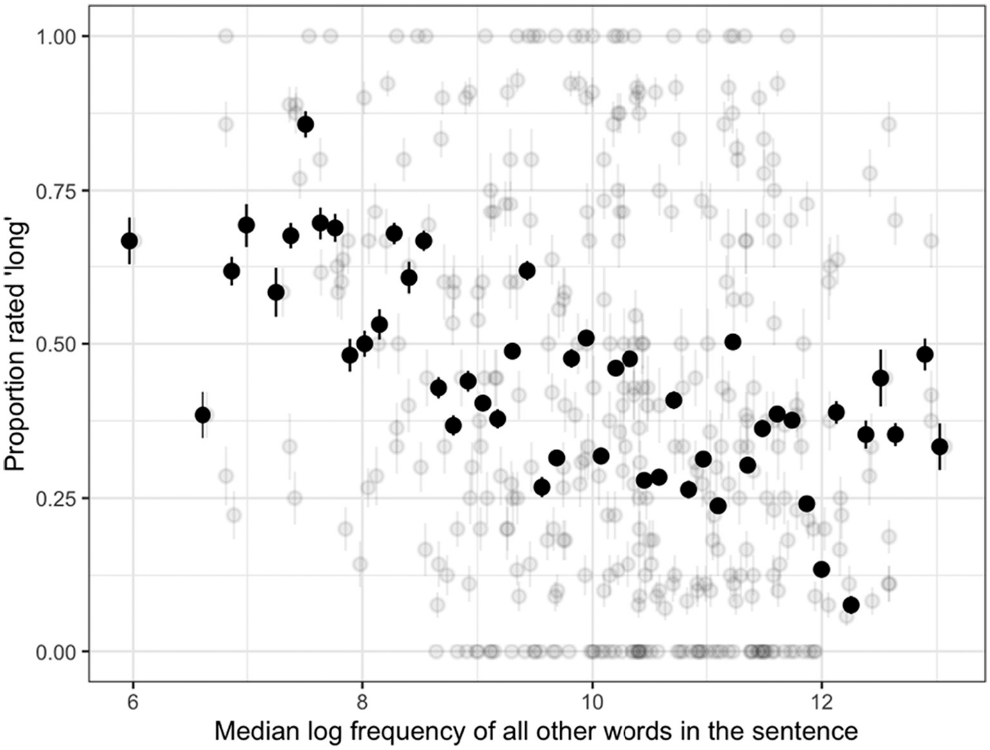

Fig. 4. Relationship between median word frequency for the entire sentence context and participant ratings for short/long word alternatives. Light points reflect one mean (proportion) for each sentence in stimulus set. Error bars reflect one bootstrapped standard error of a by-item mean. Dark points reflect means binned for visualization only using the ggplot2::stat_summary_bin function.

the other 370 sentences (= 37 pairs × 2 forms × 5 sentences per form), withholding the 10 unique sentences for a given pair (i.e., all chimp/chimpanzee stimuli). This leave-one-out procedure, in which we train only on all the other pairs of words, allows us to see whether the properties of sentences that contain long words generalize to unseen pairs (Panel 4). These trained models can then be used to output a predicted probability that participants will prefer the short or long form of a word for the other 10 sentences that were held out (Panel 5). Once all the models were trained, we then used the predicted probabilities for all 380 sentences as a predictor in a logit mixed effects model accounting for participant responses.

We tested whether the classifiers trained in the previous section over the full 768-dimensional embedding are able to account for held-out judgments in a mixed effects modeling framework. We included the classifier's predicted probability as a fixed effect and random slope for Participants and Word nested within Long Form. The model also included random intercepts by Participant, Sentence, and Word nested within Long Form. We found that the classifier's predicted probability was a significant positive predictor of participants' binarized ratings. That is, the more the classifier predicted that participants would prefer a long word in context given their responses to other sentences, the more often participants actually preferred the long word in that specific sentence. This result is consistent with the claim that discourse register, operationalized as global sentence context, contributes to speakers' preference for short vs. long word forms. We summarize these results below in Table 2. We plot the relationship between the (binned) classifier predicted probability and participants' ratings below in Fig. 6.

In a final analysis of the rating data, we tested for the potential joint contribution of our embedding-derived length preference predictor and surprisal for explaining variance in participants' decisions about word length preferences. Note that the information theoretic and message-driven accounts shown in Fig. 1A and B need not be mutually exclusive; there could be message and discourse factors influencing lexical choices (Fig. 1B), with concurrent smoothing of information density across the utterance (Fig. 1A). Additionally, it is possible that the models trained on sentence embeddings partially encode surprisal and context word frequency information. Indeed, the model predicted probability may in fact encode some of this information already. For example, the median context word frequency and the probability generated by the classifier trained on RoBERTa embeddings are strongly correlated (r = 0.39, $t(379) = -8.24$, $p < .001$). On the other hand, the correlation of model predicted probability with the summed predictability index (SPI; combined short and long surprisal) was very modest ($r = 0.08$, $t(379) = 1.72$, $p = .08$). An analysis assessing any joint influence of surprisal and the new embedding-derived predictor tests for this possibility of shared information across measures.

To assess any joint contribution of these factors, we built the same mixed effects model as presented in Table 2 but added the SPI as a fixed effect and random slope by participants. The results of this more complex model revealed the same pattern as before: surprisal was not a significant predictor of participants' responses. Moreover, the addition of surprisal worsened model fit to the data relative to the model containing only the classifier-based predictor (log likelihoods −1938 and −1941, respectively), suggesting again that surprisal is not a major component in producers' word choice, or at the very least is not a major factor in participants' judgments of our sentences. We include this joint model in Appendix 5.

3. Discussion

These results suggest that the participants' judgment of short and long referring expressions in the lab appear to vary with a wide variety of statistical and structural factors. In this experiment, we demonstrated that lexical surprisal obtained from a neural language model, reflecting the information theoretic approach advocated by Mahowald et al. (2013), was not particularly predictive of lexical preference in context. By contrast, the median log frequency of the surrounding words in a sentence, consistent with known effects of discourse register, was

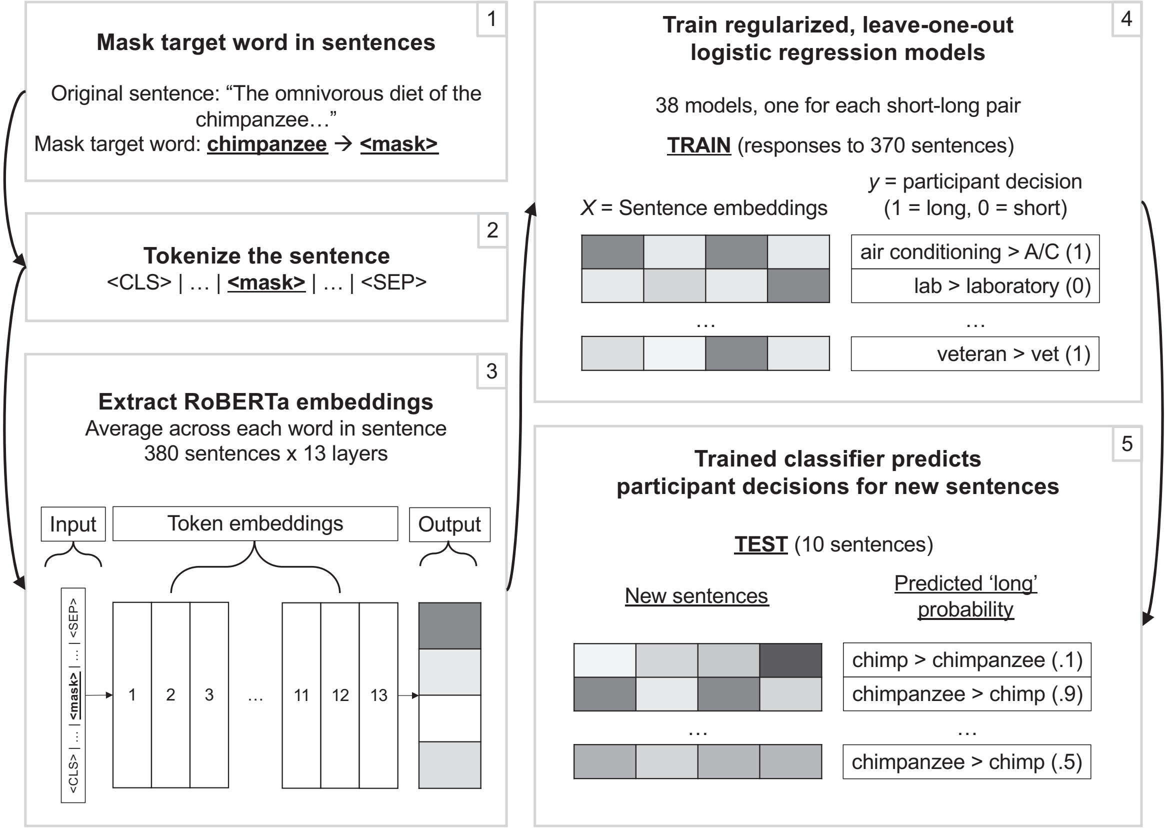
1
Mask target word in sentences
Original sentence: "The omnivorous diet of the chimpanzee…"
Mask target word: chimpanzee → <mask>
2
Tokenize the sentence
<CLS> | … | <mask> | … | <SEP>
3
Extract RoBERTa embeddings
Average across each word in sentence
380 sentences x 13 layers
Input
Token embeddings
Output
<CLS> | … | <mask> | … | <SEP>
1
2
3
…
11
12
13
4
Train regularized, leave-one-out logistic regression models
38 models, one for each short-long pair
TRAIN (responses to 370 sentences)
X = Sentence embeddings
y = participant decision (1 = long, 0 = short)
air conditioning > A/C (1)
lab > laboratory (0)
…
veteran > vet (1)
5
Trained classifier predicts participant decisions for new sentences
TEST (10 sentences)
New sentences
Predicted 'long' probability
chimp > chimpanzee (.1)
chimpanzee > chimp (.9)
…
chimpanzee > chimp (.5)

Fig. 5. Pipeline for extracting probabilistic word length preferences from sentences. <CLS> stands for the start-of-sentence token, and < SEP> the end-of-sentence token, which are obligatory additional symbols to mark sentence boundaries.

Table 2
Model predicting participants' word form preferences using out-of-sample word embedding-based model predicted score.

Form type (long vs. short) ~
Nested random intercepts for pairs of alternates (Long Form / Word) +
Random intercepts and slopes for Sentence, Participant +
Classifier predicted probability + Median log context word frequency

Name	β	SE	Z	p
Intercept	−0.62	0.27	−2.34	< 0.05
Classifier predicted probability	0.55	0.11	4.84	< 0.001

predictive of whether producers preferred short or long words. We then showed that latent factors that characterize our sentence stimuli could be used to predict producers' preference for shorter or longer words, showing that the choice of word forms is influenced by broader factors than the upstream context. The results of Experiment 1 demonstrate that distributional statistics of whole sentences, rather than local or upstream context, mediate producers' preferences for short or long words, consistent with influences of messages on lexical choices. Moreover, the successes of these models in accounting for lexical preferences are *not* lexically specific, because all models were tested on out-of-sample lexical pairs. That is, because we hid the target words from our models and used a leave-one-out testing procedure, our models generated probabilities for brand new sentences containing different terms, providing a robust test of whether the linguistic properties that affect wordform preferences generalize to novel utterances and independent of what the target words are. It is possible that the neural network also encodes the frequencies of words in the context; however, this is not necessarily a problem for our approach. As we have noted, lexical frequency is a strong predictor of different registers (Biber, 1992; Nini, 2019) and may both constrain and be constrained by syntactic structure, phonological properties of the sentence, and so on. While the classifier-generated predictions do correlate with more concrete variables such as the lexical frequencies of context words, the additional explanatory power of the classifier predicted probability is capable of capturing aspects of the utterance that are less immediately obvious. Future work will need to assess whether these predictors capture other psycholinguistically-relevant variables to lexical choice, such as syntactic structure, other types of long-distance dependencies between words, or predictability effects not immediately captured by surprisal estimates.

The most conservative conclusion from these results is that the selection of alternating wordforms varies together with variations in the content of the surrounding context. Counter to information density accounts of language production and in contrast to Mahowald et al. (2013), we do not find a contribution of surprisal to participants' preferences. Instead, we find that producers seem to prefer long words to occur in contexts containing more low-frequency words. Additionally, an analysis of judgments using measures derived from arguably high-level properties in our sentence stimuli showed that there are general factors that influence length preferences, independent of lexical identity. These results are broadly in line with a vast literature that shows that discourse structure, morphological structure, phonological structure, and the social communicative goals of the producer all influence lexical choice independent of lexical statistics, and inconsistent with findings suggesting that lexical choice is strongly influenced by predictability. In Experiment 2, we aim to confirm and generalize these findings by extending our analysis to naturally occurring sentences.

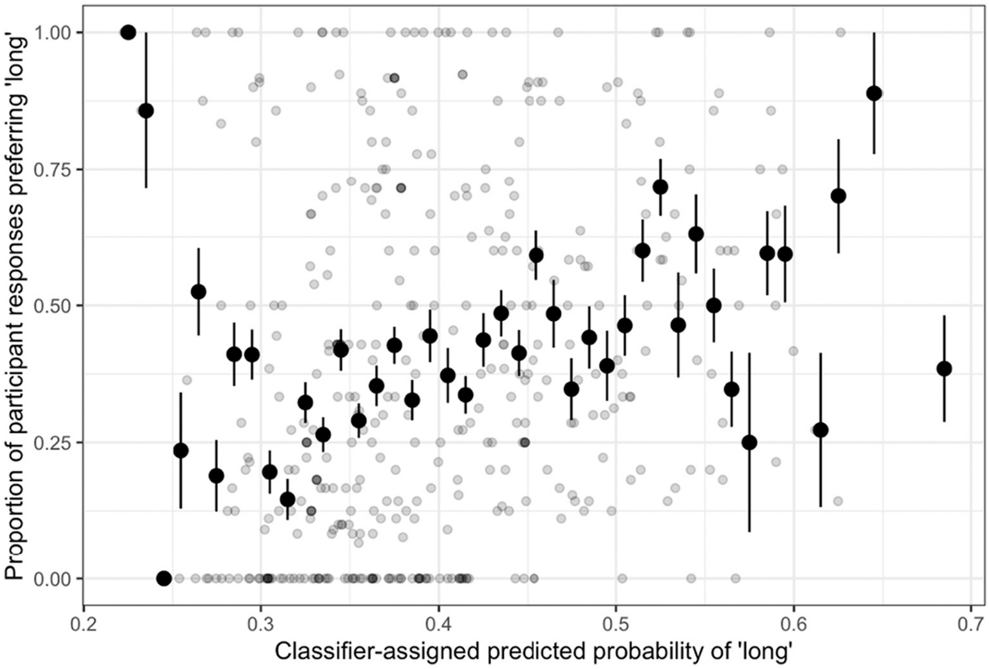

Fig. 6. Relationship between model predicted out-of-sample ratings and participants' wordform preferences. Binning for visualization only using the ggplot2::stat_summary_bin function.

4. Experiment 2: Generalizing participant judgments to natural corpora

One possible concern about Experiment 1 is that in designing materials to promote the short or long form of a word, we may have crafted materials that are unnatural in some way that influenced participants' ratings. Participants in behavioral experiments can often identify the manipulations embedded in a psycholinguistic task (Klein et al., 2012; Nichols & Maner, 2008), which could have amplified specific regularities that we used in the stimulus design phase, rather than organic linguistic regularities. Of course, the development of tailored sentence materials has a long tradition in psycholinguistics and can provide useful data, but in all cases, it is beneficial to examine more naturalistic material. The goal of Experiment 2 is to understand the degree to which participants' judgments in Experiment 1 reflect naturally occurring linguistic properties of our sentences. We addressed this question in a two-part analysis, first extracting probabilities from the previously trained classifiers from Experiment 1 for real-world sentences taken from the Corpus of Contemporary American English (COCA, Davies, 2009), then comparing the probabilities of the model decisions to the empirical lexical outcome (short vs. long) in the corpus materials. This allows us to see how much generic, broad, utterance-level properties of our stimuli in Experiment 1 are present in real-world sentences. In addition, an analysis of the correspondence between these probabilistic predictors and real-world sentence outcomes would suggest that participants' judgments reflect general statistical knowledge that some structural regularities license long words in long contexts and short words in short contexts. We thus sought to test whether the models trained on participants' responses could generate predictors that could account for lexical choice in natural sentences.

4.1. Method

4.1.1. Data

We tested whether participants' linguistic preferences in our rating task reflected their knowledge gleaned from language experience by relating their ratings to real-world materials. We selected a random sample of up to 200 sentences for each term in our set of our 38 short-long word pairs (76 words in total) from the 1990–2015 portion of COCA (Davies, 2009). We selected COCA because it is a balanced corpus and includes news, magazine, academic, fiction, and spoken domains, therefore containing variability in discourse register. Most importantly, this corpus has a high degree of coverage for our terms, with no missing vocabulary terms.

4.1.2. Analysis

To test for generalization from the stimuli of Experiment 1 to naturally occurring, non-experimental sentences, we again applied the neural network model RoBERTa (Liu et al., 2019) in the same manner as in Experiment 1 to obtain embeddings of up to 200 randomly sampled COCA sentences for each of two words in a pair (i.e., *chimp* or *chimpanzee*). It was critical to up-sample sentences because not all target words had 200 full observations – some had as few as 13, and others as many as 10,000 tokens. All sentences were used re-sampled if <200 were available in the corpus, resulting in up to 400 total test sentences for each short/long pair. As in Experiment 1, we hid the target words of interest from the model (e.g., “But not the same <mask> seen frolicking with Liz in that other picture”) to isolate the effect of context. As before, we applied the classifiers trained on ratings from the 37 item pairs (370 sentences total) from Experiment 1 to obtain a probability that participants would have preferred the long or short form of the word in the sentence, if they had seen this in the experiment instead. As before, we did not use a classifier trained on pairs that were specific to our target pairs. Instead, we used the models that were trained on only ratings by our participants in Experiment 1 from 37 of the pairs. For example, to predict usage of *chimp* vs. *chimpanzee* in sentences in the COCA corpus we used the classifiers trained on participants' ratings of the 370 non-chimp/chimpanzee sentences from Experiment 1. Thus, the task is a close analogue to our behavioral prediction task, but with naturally occurring language and a greater number of sentences.

As with the human judgment prediction task, we obtained a predicted probability of participants' preferences for the word at the mask position being either the “short” or “long” form in a pair, which we used as a (fixed effect) predictor in a logit mixed effects model. However, instead of predicting additional ratings, as in Experiment 1, the mixed effects model in Experiment 2 aimed to predict whether the actual form of the word in these naturally occurring sentences was the short or long form of an alternating pair. The mixed effects model contained random intercepts and maximal random slopes by item (Pair). We found that the higher the probability the classifier assigned to the long form for a given sentence, the more likely the target word, which was hidden from the neural network model, was actually the long word in the corpus sentences. We summarize these results in Table 3 below and visualize these results in Fig. 7.

4.2. Discussion

Experiment 2 provides strong, complementary evidence to Experiment 1. We showed here that a model of producers' preferences for the word lengths of alternating, morpho-phonologically related pairs of words (e.g., *chimp*, *chimpanzee*) occurring in constructed stimuli generalizes to real-world linguistic examples of these words. These results show that many of the factors that influence participants' judgments also influence the selection of short versus long wordforms in corpora. The results of Experiment 2 are encouraging given that the materials from Experiment 1 were designed to be read as isolated sentences, whereas the sentences extracted from COCA were parts of a broader text or spoken discourse. Given the results of Experiment 2, we conclude that participants' behavior in Experiment 1 is driven at least in part by sentence-wide distributional cues, such as register as approximated by surrounding word frequency, and the latent structural factors that characterize sentences containing longer words over sentences that typically contain shorter ones (Wimmer et al., 1994). As in Experiment 1, the results of Experiment 2 are broadly consistent with accounts of referring expression production that are sensitive to message-level phenomena, including register, communicative intent, or the effect that a producer hopes to have on the listener.

5. General discussion

Producers are faced with many complex and interlocking choices when they begin to plan an utterance, and accounting for those choices will lead to insights about language use and production processes. Information density approaches to language production (Aylett & Turk, 2004; Jaeger, 2010; Levy & Jaeger, 2007; Zarcone & Demberg, 2021, among others) present a thought-provoking account of how lexical choice is sensitive to the prior context, stating that the predictability of a message will constrain the forms that producers select (Jaeger, 2010), largely standing in contrast to standard production theories of lexical selection (Levelt et al., 1999). Below we draw on the results of the present study to argue that both information theoretic accounts and the standard psycholinguistic production model of lexical production are insufficient to explain lexical choice, with an eye toward the high-level accounts we presented in Fig. 1 and returning to the conceptual foundations laid out in Levelt (1989). With the limitations of these contemporary accounts in mind, we present some thoughts toward development of an alternate perspective to both information density and classic production accounts of lexical choice. Our account aims to integrate many levels of linguistic abstraction into the production process, permitting both apparent signal smoothing phenomena and context effects, though precise mathematical details must still be formally specified in future work.

Table 3
Model predicting word form from participant ratings.

Form type in COCA sentence (long vs. short) ~
Random intercepts for pair of alternates (Long Form) +
Classifier predicted probability

Name	*β*	*SE*	*Z*	p
Intercept	−0.07	0.05	−1.54	n.s.
Classifier predicted probability	0.24	0.08	2.90	< 0.01

The information theoretic approach proposes that the forms of our utterances – including lexical selection – can be optimized, or smoothed, for comprehension by avoiding overly high or low degrees of message predictability (Aylett & Turk, 2004; Jaeger, 2010; Levy & Jaeger, 2007). The results of the current study present a challenge to this approach, because they show that the selection of an individual word is related to properties of a much broader surrounding context that are not particular to the words themselves. Instead, properties of the *surrounding words*, such as their frequency or grammatical properties, strongly predict lexical preferences in our analyses. In Experiment 1, we first demonstrated that the surprisal of a word given the preceding words did not significantly predict ratings of short/long forms like *chimp/chimpanzee*, either on their own or in conjunction with the frequencies of words in the context. This suggests that the effect of predictability on lexical choice is less straightforward than has been assumed (Mahowald et al., 2013; Zarcone & Demberg, 2021). As for other properties of utterances that might influence linguistic preferences, the frequency of upstream words was predictive of ratings, but the word frequencies of the full sentence context (left and right context) were a better predictor of lexical preferences than the frequencies of words in the left context only. This is an important result, as it shows that there are relationships among word choices throughout the sentence.

Using a more abstract approach to the question of how context influences lexical choice, we also showed that a machine learning model trained to predict ratings from a neural network model-based representation of context found commonalities across sentences: Some contextual factors affect the selection of longer words like *chimpanzee/television/bicycle*, while others promote the selection of shorter words like *chimp/TV/bike*. Stated another way, Experiment 1 showed that it is possible to predict whether a long word will be used from general properties of the context alone, without considering the meanings of the target words at all. This result is at least partially consistent with rational approaches to language production (Goodman & Frank, 2016), in that the context is predictive of the word forms that producers select. However, our results also show the precise words involved are not necessarily very important, which poses a problem for information density proposals that rely on computing the probabilities of specific wordforms in context. However, from our perspective, it is not critical that we used an embedding to demonstrate the importance of broader linguistic factors on wordform preferences. For example, other ways of encoding the syntactic, semantic, and pragmatic context may also have been appropriate; however, we opted to use a neural language model because interactions between linguistic variables like these are a critical aspect of language processing (e.g., Hsiao, Gao, & MacDonald, 2014). A better understanding of these context effects is a major goal for future research and should be useful for several theoretical accounts.

Why have others found effects of predictability on form selection (Mahowald et al., 2013; Zarcone & Demberg, 2021), but the present study failed to? Evidence that predictability or information density affects lexical selection comes largely from corpora or cloze data. While many large-scale analyses of corpora have used nuanced statistical models with increasingly sophisticated techniques that are capable of predicting language processing (e.g., Goodkind & Bicknell, 2018; Hollenstein et al., 2021), it is possible that the surprisal values from the model that we used (Ng et al., 2019) are not the most accurate for predicting alternations between individual words; neural language models, though they typically align closely with human predictions, have been demonstrated to have some deficiencies in predicting (psycho-)linguistic data (Dudy & Bedrick, 2020; Eisape et al., 2020; Jacobs & McCarthy, 2020). Another possible explanation for this null effect of surprisal is that our sentence materials did not manipulate predictability, while others did (Mahowald et al., 2013; Zarcone & Demberg, 2021). Therefore, any estimates we have may not show the full breadth of predictive power that stimuli that vary in their cloze probabilities of the targets would provide. Nevertheless, the results of Experiment 1 give us good reason to believe that the appropriateness of a longer word to complete a sentence has a reliable

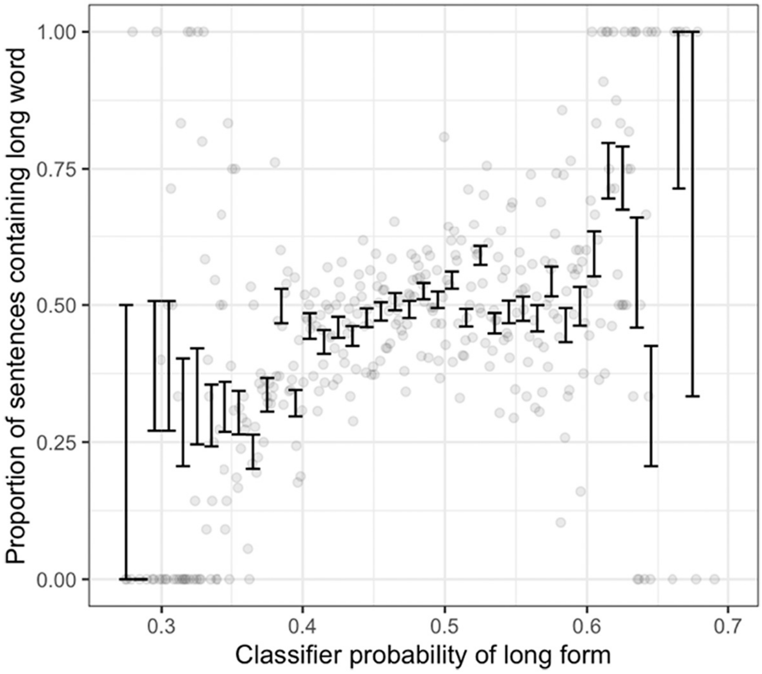

Fig. 7. Relationship between classifier trained on RoBERTa embeddings to predict participant ratings and corpus probabilities. X axis is the binned predicted probability (for visualization purposes only using the ggplot2::stat_summary_bin function) that participants would have preferred a "long" completion in that utterance had this been a sentence they saw in Experiment 1.

impact on lexical choice, promoting the use of longer words in general. Such an effect is directly predicted by accounts of language production that allow for broad constraints driven by whole-utterance level messages that allow for words to influence each other during message formulation (Dell & O'Seaghdha, 1992; Hsiao et al., 2014).

6. Consequences for production accounts

It is our perspective that a computational account is critical to meaningfully advance our understanding of language production (Guest & Martin, 2021; van Rooij & Baggio, 2020; van Rooij & Blokpoel, 2020). We therefore consider the available computational and algorithmic theories at our disposal. First, it is possible to conceptualize the results of the present study as fitting within a modification of the standard model of (lexical) language production (Levelt et al., 1999) that extends the definition of "message" at the message formulation stage to include broader aspects of the utterance plan. Indeed, in his original discussion of the language production process, Levelt (1989) highlights the importance of discourse contextual factors on lexical selection, stating, "Establishing agreement on the discourse type may require explicit negotiation at the outset, but usually the type of discourse is *invoked* by the way the talk is conducted (Schegloff, 1987). For instance, it is in the way one person talks like a doctor (i.e., speaking of a 'hematoma' instead of a 'bruise') that the interlocutor recognizes that the discourse is of the doctor-client type." (p. 112). However, the verbal model of Levelt (1989) was never converted into a computational framework. Instead, the dominant computational model of language production (WEAVER++; Levelt et al., 1999; Roelofs, 1997) is largely concerned with single word production that is most appropriate for tasks where the "message" (i.e., the semantics of a concept to be named) is known, either in the form of a picture that speakers must name (Huettig, Rommers, & Meyer, 2011; Schriefers et al., 1990) or a word form that is cued by another word (Meyer, 1990, 1991; Roelofs, 1996; Roelofs & Meyer, 1998). In both cases, a single word "message" is already selected, and the model's job is simply to produce the word that corresponds the best to that message. WEAVER++ (Roelofs, 1997) puts lexical selection after message planning, which itself leads to the selection of a lemma, or a basic word form.

It is unclear whether the choice between *chimp* and *chimpanzee* in WEAVER++ is ultimately determined at the highest level of planning (conceptualization), with cascading effects to lower levels. The proposals laid out in Levelt et al. (1999) and Levelt (1989) therefore must specify where this process should take place, and for that, greater experimentation and higher-level computational models of the language production process are needed. In either case, however, it is clear that a single-word view on language production is insufficient to the broader range of phenomena in language production (Dell, Chang, & Griffin, 1999) and that the field should begin to integrate discourse level factors into production models, a level of linguistic abstraction which has received little attention in the 30 years since Levelt (1989)'s original publication (Meyer, Roelofs, & Brehm, 2019).

It is our view that modern neural network approaches can continue in the spirit of classic language production models to provide a unifying account with information density findings. A rich literature in natural language generation, for example, has wrestled with issues of representation and how to enforce linguistic structure on productions without pathological behavior (Holtzman, Buys, Du, Forbes, & Choi, 2019; Kulikov, Welleck, & Cho, 2021). Many approaches leverage so-called abstract meaning representations (Banarescu et al., 2013), which contain a set of propositions that characterize an event or state of the world. These can be fed into neural network models to produce simple and complex utterances (e.g., Dušek, Howcroft, & Rieser, 2019; Mager et al., 2020) to varying degrees of success (Demberg, Hoffmann, Howcroft, Klakow, & Torralba, 2016; Manning, Wein, & Schneider, 2020; van Miltenburg et al., 2021). Importantly, many modern systems are able to seamlessly integrate stylistic factors into their models or induce them with priming-like mechanisms (Ficler & Goldberg, 2017; Kabbara & Cheung, 2016; Oraby et al., 2018), which theoretically enables the contribution of broad utterance context to lexical choice and provides one step toward realizing Levelt (1989)'s vision for a high-level cascading model of language production at the utterance level, in which a number of factors modulate lexical selection.

We visualize this possibility below in Fig. 8.

Figs. 1 and 8 both highlight the joint importance of many confluent

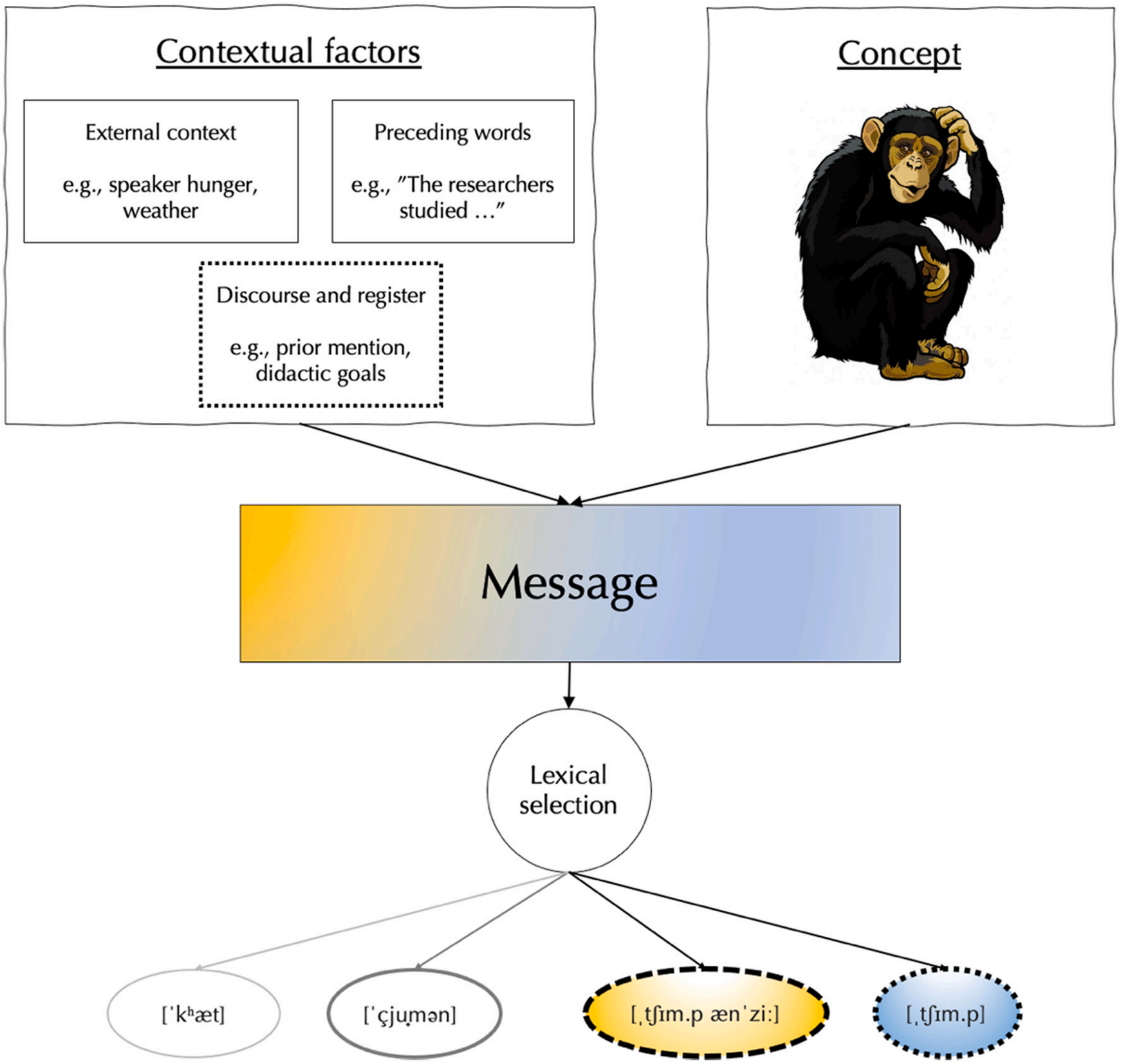

Fig. 8. Schematic of message formulation process. In it, contextual factors, ranging from upstream words, to situational factors, to the discourse status of referents, all impact the realization of an utterance associated with the name or description of a concept we wish to convey. Different forces may push producers to one outcome (e.g., "chimp") over another (e.g., "chimpanzee"). In this architecture, message formulation constrains lexical choice in a manner consistent with both standard connectionist and modern information theoretic approaches.

linguistic (and extralinguistic) factors that influence lexical choice. Indeed, stochastic factors such as noise in the state of the language production system surely influence word choice. Additionally, the existence of short and long forms that alternate does not always imply that distinct shades of meaning are the only factors influencing lexical selection. Rather, based on distributional arguments alone, we show that short and long forms tend to occur in systematically different contexts, which has been taken as evidence in usage-based linguistics that the two forms typically express different ideas. We also note that from the perspective of the classifier approach taken in Experiments 1 and 2, there is in fact gradience between short and long forms. That is, there is a graded relationship between sentence properties and the odds of using a particular word form in a given context.

Of course, building systems that on the one hand embrace the dynamic properties of neural networks while simultaneously aiming for interpretability is a major ongoing scientific endeavor. Understanding the neural network "black box" is a challenge for psycholinguistic theory development, because existing models are not particularly transparent and require clever linguistic probes. We hope ongoing scientific work in this area will bring us closer to understanding human language production as well. For example, a model like the Sentence Gestalt-style model (John & McClelland, 1990) can be controlled more tightly by the experimenter, unlike large language models trained on internet text, learning entirely from natural language propositional structures. However, a valuable future direction could be to use real corpora to account for psycholinguistic data using this modeling framework (e.g., Rabovsky & McClelland, 2020).

Another potential concern of embeddings-based approaches is that they may come to rely on regularities in the data that are not of linguistic interest but which are predictive of some feature (e.g., ungendered-to-gendered pronoun mistranslation). Additionally, a neural language model's inductive biases for learning from text data may be different from human inductive biases for learning from language data. Even if a statistical language model and a human being might rely on different cues, the successes of our modeling approaches suggest some relevant signal can predict lexical choice. Further work should couch this modeling problem in well-understood, cognitively motivated representations.

7. Consequences for information theoretic accounts

With this proposal in mind for extending language production to the full utterance level to incorporate register and stylistic factors, we can return to the assumptions of information theoretic approaches.

Mahowald et al. (2013) have used a particular form of analysis in investigating the nature of language distributions, in which aspects of linguistic context are used to predict a particular linguistic behavior, such as word length (Piantadosi et al., 2011), presence/absence of optional words (Jaeger, 2010) or morphemes (Zarcone & Demberg, 2021). We have followed a similar methodological path, again using aspects of linguistic context to predict a linguistic behavior, specifically word length. As corpus analyses become more common, including ones with an information theoretic perspective, it is useful to consider what types of explanations these analyses are able to provide, and how the results of statistical analyses can be explained by broader consideration of the cognitive processes that are the source of the language data we use.

The statistical analyses predicting linguistic behavior in one part of a corpus (e.g., use of *chimp* vs. *chimpanzee*) from another part of the corpus, such as prior context or whole sentence context, are essentially correlations between different parts of a corpus, or between a context and a behavioral result, like judgments about short vs. long words. At Marr's (1982) functional level of analysis, these relationships are extremely interesting, as they reveal that utterances have robust dependencies that will need to be accounted for in theories of language knowledge and processing. Indeed, psycholinguists who are interested in understanding producers' choices in real-world scenarios should continue to pay attention to corpus analyses that can reveal regularities that drive semantic, syntactic, and morphological variation across discourse registers. Moreover, information theoretic accounts may prove extremely useful in characterizing statistical differences in discourse registers (e.g., Bentum et al., 2019).

The interpretation of these analyses at a more algorithmic level (in Marr's sense, i.e., a theory of language production or comprehension processes) is less straightforward, however, because correlations between linguistic context and some behavior do not themselves license claims about the causes of these relationships. Some proponents of uniform density accounts of utterance form have argued that the density of information in a linguistic string is fairly uniform because producers consciously or unconsciously try to make it uniform to aid communication. This causal claim is clearly articulated in Mahowald et al. (2013): "…the correlation between word length and informativeness is likely influenced by language production phenomena, where users actively prefer to convey meanings with short forms when the meanings are contextually predictable." In contrast, we have argued that the correlations between the words in different parts of the utterance emerge from several forces that can be viewed as part of the producer's message, including the discourse register, which shapes lexical selection throughout the utterance. While we do not necessarily identify a role for surprisal in modulating lexical preferences in the present study, the previous findings of information density's influence on wordforms (e.g., phonetic duration) suggest that an additional constraint as illustrated in Fig. 8 could include noisy channel principles, or incorporate other types of penalties, which could in principle be consistent with multiple constraint satisfaction accounts of language production. For example, Piantadosi, Tily, and Gibson's (2012) information theoretic analysis of the existence of lexical ambiguity in languages suggests that re-use of words is efficient for the speaker, and the resulting ambiguity is tolerable for the comprehender, yielding overall efficiency gains from ambiguity. A related possibility is that lexical selection, thought to be a very early process in production planning, is less affected by pressures to smooth the signal or avoid misunderstanding in a noisy channel than are later-occurring processes of phonological and articulator planning.

More generally, the current results encourage us to consider another aspect of uniform information density approaches to communication efficiency, namely, efficiency for whom? At Marr's computational level, the scientific question concerns an efficient communication system; but in more algorithmic terms, researchers often suggest that producers are aiming to help comprehenders, as in the Mahowald et al. quotes above and in the introduction. By contrast, MacDonald (2013) has argued because production is more demanding than comprehension (e.g., Boiteau, Malone, Peters, & Almor, 2014), efficiencies that benefit the producer will tend to better contribute to general communication efficiency more than further tailoring the input for the comprehender's benefit (see also Piantadosi et al., 2012). A variant of these claims is Good Enough Production, where the producer implicitly weighs difficulty and message factors in choices of utterance form (V. Ferreira & Griffin, 2003; Goldberg & Ferreira, 2022; Koranda, Zettersten, & MacDonald, 2022). For example, the syntactic reduction phenomenon discussed in the introduction, in which producers insert an optional *that* before relative clauses (*the dog [that] Shauna* adopted) and complement clauses (*I know [that] the dog…*) have been described within both information theoretic accounts and production-based accounts. Jaeger (2010), argued that the distribution of *that* use and omission in complement clauses obtains because "speakers prefer utterances that distribute information uniformly across the signal" (p. 25), so that speakers produce *that* more often when the upcoming material has high information content. A more production-centric approach holds that producers use *that* to regulate planning time in incremental production, in which producers are engaged in overt production while simultaneously planning upcoming material; the *that* acts something like a pause to permit more planning time for upcoming material, which is more needed when the upcoming material is difficult (Ferreira & Dell, 2000; Race & MacDonald, 2003). It should be clear here that both accounts offer an approach to managing the timing of language production, and that one possibility is that the information theoretic relations may obtain because of mechanistic processes benefitting the producer. For example, Jaeger (2010) argues that the information theoretic approach is the best fit in his complement clause analyses, but he also notes that the information theoretic approach is consistent with probabilistic production accounts, including ones we cite above, e.g., Chang et al. (2006).

We see several steps as necessary to move forward toward further distinguishing and investigating these processing accounts of information distribution in utterances. For the word length effects we have investigated here, an important next step is further understanding of discourse register and other factors that modulate selection of short vs. long forms. Corpus studies clearly show that different registers vary in lexical choices, collocations, and lexical predictability (Bentum et al., 2019; Biber, 2012). And in experiment-based approaches, both Mahowald et al. (2013) and the present study, showed that manipulation of sentence context affects judgments. However, the sources of that influence are not fully clear and should receive further investigation. Our results suggest that the factors that modulate length are quite general and go beyond the particular words in a sentence, but more specificity and insight are likely possible.

Second, we expect that other factors may constrain producers' choice of short vs. long wordforms, via forces that dictate whether a long word does or doesn't have a commonly accepted shorter alternative. Frequency of use has been widely recognized as a factor in shortening (Zipf, 1949; see also Piantadosi et al., 2011, 2012), but other factors may also be important, including articulatory difficulty. For example, English has the short form *chimp* for *chimpanzee* and the short form *orang* for *orangutan,* but there is no short form for *gorilla*, which like *chimpanzee,* is a three-syllable word referring to a great ape species. This small sample points to a possible role for articulatory difficulty as a force in speakers' invention and choice of short word alternatives: the words *chimpanzee* and *orangutan* both are unusual for English words in both their stress patterns and phonotactics, while *gorilla* appears easier to articulate by virtue of being more aligned with other English words and is part of a phonological neighborhood—*manila, vanilla*, *flotilla,* etc. If the development and lexical selection of shortened forms are driven in part by articulatory difficulty, this would constitute one of relatively few examples to date where phonological-level factors affect lexical selection level processes in language production (Ferreira & Griffin, 2003; Harmon & Kapatsinski, 2017).

Third, is also clear that language production choices are not limited solely to two-word or two-structure possibilities. Indeed, the choice between short vs. long forms could be expanded to include a third choice, such as pronouns, or even a fourth choice, null pronouns in languages that allow them (e.g., Hint et al., 2020). Similarly, it will be important to move beyond lexical choice in single sentences to examine larger texts or discourses, where place in the discourse and the forms of previous mentions may influence choices of referring expression (Grosz et al., 1983).

For all these forms of variation in language production, an important question concerns the viability of message-based and information smoothing processes to characterize producers' behavior in online language production. For the former, more attention is needed to the mapping between message and utterance form. There is abundant work in this area at the lexical level from single word production studies, but far less is known about how a message is formulated for more complex utterances in more varied situations beyond typical studies in the lab. If that work proves fruitful, then in our view, the message-driven approach is attractive because correlations between different sentence components are emergent from the processes that have long been thought to drive lexical selection. Because algorithmic-level accounts of language production will always require some mechanisms of this sort, the intentional, deliberate version of information smoothing accounts may be superfluous in this domain, and at least in the case of our surprisal analyses, they do not provide a good account of length judgments. Moreover, claims that producers are actively smoothing the signal for comprehenders is not consistent with language production evidence that elaborate attention to comprehenders' needs is computationally costly, not necessary for good comprehension, and rarely attempted (Brown-Schmidt & Heller, 2018).

Finally, the relationship between register and language production requires more research. While there is abundant work in sociolinguistics and corpus linguistics identifying registers and characterizing language patterns at different registers (Biber, 1992, 2012; Nini, 2019), this work is largely divorced from research addressing how register affects lexical selection, word order, phonological realization or other online production processes. In addition to broadening the field of language production, studies of how register influences lexical and other choices will also be important to information theoretic and other accounts of language use.

Funding

This research was supported by Grant 1849236 from the National Science Foundation.

CRediT authorship contribution statement

Cassandra L. Jacobs: Conceptualization, Methodology, Software, Visualization, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Maryellen C. MacDonald:** Supervision, Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Funding acquisition.

Data availability

Link to data/code is in Supplementary Materials section of manuscript

Acknowledgements

We thank our dedicated research assistants Molly Hoferle, Liz Hellrung, and Tia Sadlon for their hard work constructing stimuli and piloting data, as well as Loïc Grobol and Aniello de Santo for their feedback on the manuscript.

Appendix 1. DistilBERT

In order to revise stimuli that were highly similar to others, we used DistilBERT[1] (Sanh, Debut, Chaumond, & Wolf, 2019) to quickly extract vector representations of all of the tokens in the candidate sentences. We then averaged across all tokens to obtain a single vector for each sentence, from which we computed a cosine similarity matrix from each sentence vector to all other stimuli. We considered highly similar sentences to have cosine distance ≥ 0.75, as we noted that these sentences often used the same syntactic constructions or depicted similar events – potential across terms (e.g., *chimp* and *rhino*). This criterion led to the revision of 69 sentences (18%).

References

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *Proceedings of EMCˆ2 @ NeurIPS 2019.*

Appendix 2. Surprisal

We used the monolingual English model from Ng et al. (2019), a neural transformer-based language model, to obtain surprisal values for our stimuli. To obtain a surprisal value for a given wordform, we feed the neural language model the entire upstream linguistic context prior to the critical target word (e.g., *chimp* or *chimpanzee*). We then assess the activation (log probabilities; Frank, 2009) of subsequent outcomes, which provide an estimate of neural surprisal. These models operate over "word pieces" rather than words (Sennrich, Haddow, & Birch, 2016; though see Bostrom and Durrett, 2020 for a review), which requires us to choose some way of combining the probabilities associated with each piece. We chose to sum the surprisal values (therefore a multiplicative probability), but others have taken the activations associated with the final word piece, the average, or simply the first piece. We did not find that different aggregations of the surprisals of word pieces affected the general pattern of our results in any way.

References

Bostrom, K., & Durrett, G. (2020, November). Byte pair encoding is suboptimal for language model pretraining. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (pp. 4617-4624).

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., & Edunov, S. (2019, August). Facebook FAIR's WMT19 News Translation Task Submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)* (pp. 314-319).

[1] Transformers implementation; 4.1.1; Python 3.7.0

Sennrich, R., Haddow, B., & Birch, A. (2016, August). Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers* (pp. 371-376).

Appendix 3. Logistic regression models

Here, we used the RoBERTa model (Liu et al., 2019; `transformers` implementation; 4.1.1; Python 3.7.0; Wolf et al., 2019), which has obtained high levels of performance on benchmark tasks that quantify linguistic knowledge. RoBERTa is a *masked* language model, which allows us to hide, or *mask*, the target words (e.g., *chimp* or *chimpanzee*) in our sentences from the model before we assess the similarity between sentences. That is, it is possible for us to completely omit the target words (e.g. *chimp, chimpanzee*) from all sentences, leaving only a mystery word that cannot contaminate the sentence-level representation (embedding) of the surrounding context. Like many modern neural network models, RoBERTa produces embeddings at several different layers (up to 13) which increase in linguistic abstraction, from predominantly form-based representations in the lower layers to higher-level propositional structure in the higher layers (Jawahar, Sagot, & Seddah, 2019). To obtain a vector representation of the whole sentence at a specific layer, we average all of the word vectors for that sentence while excluding the masked word token (for a similar approach, see Hawkins, Frank, & Goodman, 2020).

We used the layerwise sentence embeddings directly in a regularized (lasso) logistic regression model implemented in `scikit-learn` (version; 0.23.2; Pedregosa et al., 2011) to predict whether participants rated long words as more acceptable than short words in a given context (a binary variable). To ensure that the model was capable of generalization to novel sentence contexts and that it could not memorize properties about any specific topic (e.g., words exclusively seen in *chimp* or *chimpanzee* sentences), we masked all target words for all sentences and trained one model for each pair of words with the training input being participants' judgments for all 37 other word pairs. To generate predictors for downstream inferential statistics, which we describe in the next section, we obtained the model's predicted probability for the held-out ("out-of-sample") sentences. As a result of our analyses of the different layers, all results we report are conducted with the 7th layer of the RoBERTa outputs.

As in our mixed effects models, we binarized participant decisions into short or long preferences. Therefore, our dichotomization of participants' numerical ratings into a binary judgment for mathematical convenience obscures the true relationship between ratings and model accuracy and we report accuracy merely for completeness. Models were generally capable of predicting participants' judgments, with the embeddings coming from the best performing layer accurately predicting 62% of the ratings. Higher layers (10−13) typically better encoded the factors affecting participants' judgments (62% correct), with the poorest discrimination (58% correct) at the lowest layers (1–3). This result is consistent with other work that has that the highest-order, message-sensitive layers are critical for many natural language understanding tasks (Jawahar et al., 2019). The majority class baseline (the base rates of participants producing a short form) occurs at approximately 59.5% of responses. However, the fact that classifier accuracy improves only modestly from the baseline is not necessarily a problem for our approach.

We hoped to demonstrate nuance in the probabilities assigned by the model, similar to surprisal analyses and other analyses showing linearity in log probability space (e.g., Smith & Levy, 2013), so we opted to use the predicted probability from the classifiers directly in a model of participants' behavior rather than the predicted label for a specific response. We also explored an unsupervised version of this approach, which resulted in low-dimensional vector representations of each sentence embedding; we report these analyses in Appendix 4.

References

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the dynamics of learning in repeated reference games. *Cognitive Science*, *44*, e12845.

Jawahar, G., Sagot, B., & Seddah, D. (2019, July). What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3651-3657).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825-2830.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Appendix 4. UMAP embeddings

Despite the successes of the classifier predicted probabilities in accounting for judgments, another valid approach is to conduct unsupervised dimensionality reduction on the sentence stimuli directly, which allows us to evaluate the sentence representations on their own, without contamination from classifier biases. That is, classifier predicted probabilities are the product of "filtered" sentence representations, as they are influenced by the coefficients learned by the classifiers, which were trained directly on rating data. Consequently, we were interested in knowing how much latent vector properties would be directly capable of accounting for lexical preferences. Therefore, we performed dimensionality reduction to compress each 768-dimensional embedding into two dimensions by using an algorithm known as UMAP (McInnes, Healy, Saul, & Großberger, 2018), which learns the correlational structure of the sentence vectors. Once trained, the two-dimensional representation can be included in a mixed effects model of behavioral responses instead of probabilities. Below, we demonstrate that the learned UMAP dimensions closely correspond to the classifier probabilities from the model that we trained on participant responses in the previous section, despite being trained on different objectives. This visualization indicates potential viability of using UMAP dimensions to characterize aspects of our stimuli that may affect participant behavior.

Further inspection of the relationship between the different UMAP predictors and participants' ratings show that some layers within RoBERTa, such as layers 1, 3, and 8, show a continuous relationship between UMAP dimensions and participants' rating preferences. Others, such as 4, 5, 6, and 12, appear more categorical. While the precise information encoded at each layer is unclear, the layers in these models do typically represent different sources of linguistic information (Jawahar et al., 2019; Rogers et al., 2021), and so finding continuous and categorical relationships between latent dimensions and participant ratings is somewhat surprising. Future work directly manipulating sentence structure should include breakpoint analyses (e.g., Brehm and Goldrick, 2017) to understand whether neural language models response categorically or continuously to different linguistic structural factors.

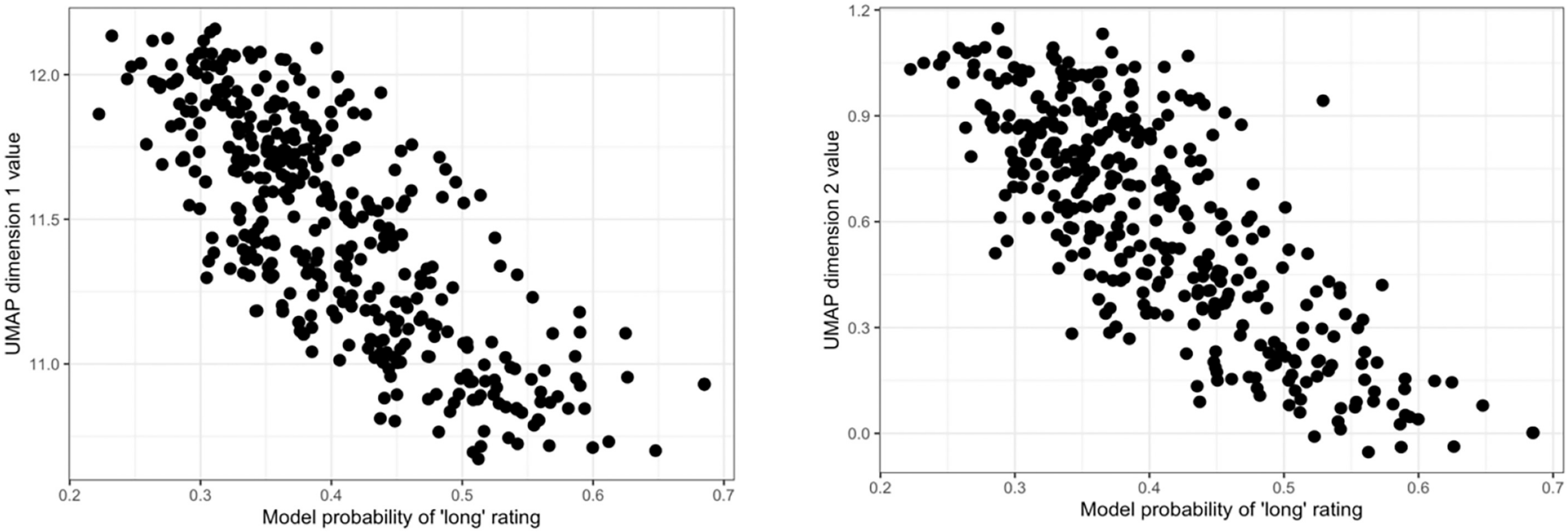

Fig. A4.1. Relationship between first and second UMAP dimensions and ratings classifier predicted probabilities.

Using the two latent UMAP dimensions, we constructed a mixed effects model with random intercepts and slopes by participants and random intercepts by Sentence and show that these latent dimensions strongly predict participants' lexical preferences, though the specific direction and significance of these latent dimensions varies by layer, though we do not present those analyses here. Instead, we focus on the top layer (Layer 13) for our analyses, which we present below in Table A4.1. There is a strong relationship between the first UMAP dimension (UMAP_x) and participants' response preferences; UMAP_x and the second dimension, UMAP_y, are highly collinear and the estimate for the slope on UMAP_y is not significant.

Table A4.1
Model predicting participants' word form preferences from UMAP scores.

Form type (long vs. short) ~
Random intercepts for pair of alternates (Long Form) +
Classifier predicted probability

Name	*β*	*SE*	*Z*	p
Intercept	−0.72	0.28	−2.52	< 0.05
UMAP_x	−0.50	0.16	−3.24	< 0.01
UMAP_y	−0.16	0.12	−1.39	n.s.

References

Brehm, L., & Goldrick, M. (2017). Distinguishing discrete and gradient category structure in language: Insights from verb-particle constructions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1537-1556.

Jawahar, G., Sagot, B., & Seddah, D. (2019, July). What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3651-3657).

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software*, *3*, 29.

Appendix 5. Combined surprisal and classifier-predicted probability model

Table A5.1
Model predicting participants' word form preferences from UMAP scores.

Form type (long vs. short) ~
Nested random intercepts for pairs of alternates (Long Form / Word) +
Random intercepts for Sentence +
Random intercepts and slopes for Participant (no correlation estimates) +
Classifier predicted probability +
Summed predictability index (neural surprisal)

Name	*β*	*SE*	*Z*	p
Intercept	0.21	0.33	0.63	n.s.
Summed predictability index	0.01	0.01	1.37	n.s.
Classifier-predicted probability	1.52	0.24	6.25	< 0.001

Appendix 6. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105265.

References

Arnold, J. E., & Griffin, Z. M. (2007). The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language, 56*, 521–536.

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

Baayen, R. H. (2002). *Word frequency distributions*. Germany: Springer Netherlands.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., … Schneider, N. (2013, August). Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse* (pp. 178–186).

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language, 42*(1), 1–22.

Bauer, L. (2012). Blends: Core and periphery. *Cross-Disciplinary Perspectives on Lexical Blending*, 11–22.

Bell, A. (1984). Language style as audience design. *Language in Society, 13*, 145–204.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language, 60*, 92–111.

Bentum, M., Ten Bosch, L., Van den Bosch, A., & Ernestus, M. (2019). Do speech registers differ in the predictability of words? *International Journal of Corpus Linguistics, 24*, 98–130.

Berg, T. (2011). *Structure in language: A dynamic perspective*. n.p: Taylor & Francis.

Biber, D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes, 15*, 133–163.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory, 8*, 9–37.

Bock, J. K., & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition, 21*, 47–67.

Boiteau, T. W., Malone, P. S., Peters, S. A., & Almor, A. (2014). Interference between conversation and a concurrent visuomotor task. *Journal of Experimental Psychology. General, 143*(1), 295–311. https://doi.org/10.1037/a0031858

Brants, T., & Franz, A. (2006). *Web 1T 5-gram Ver. 1. LDC2006T13*. Philadelphia: Linguistic Data Consortium.

Bresnan, J. W. (1972). *Theory of complementation in English syntax*. Doctoral dissertation. Massachusetts Institute of Technology.

Brooke, J., & Hirst, G. (2014). Supervised ranking of co-occurrence profiles for acquisition of continuous lexical attributes. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers* (pp. 2172–2183).

Brown-Schmidt, S., & Heller, D. (2018). Perspective-taking during conversation. In *Oxford Handbook of Psycholinguistics* (pp. 551–574).

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977–990.

Bybee, J., & Thompson, S. (1997, September). Three frequency effects in syntax. In *Annual meeting of the Berkeley Linguistics Society* (Vol. 23, No. 1, pp. 378–388).

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113*, 234–272.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshe, B. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding*. Cambridge: Cambridge University Press.

Cohen Priva, U. (2017). Informativity and the actuation of lenition. *Language, 93*, 569–597.

Davies, M. (2009). The 385+ million word Corpus of contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*, 159–190.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*, 283–321.

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 369*, 20120394.

Dell, G. S., & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition, 42*, 287–314.

Dell, G. S., Chang, F., & Griffin, Z. M. (1999). Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science, 23*, 517–542.

Demberg, V., Hoffmann, J., Howcroft, D. M., Klakow, D., & Torralba, A. (2016). Search challenges in natural language generation with complex optimization objectives. *KI-Künstliche Intelligenz, 30*, 63–69.

Dudy, S., & Bedrick, S. (2020, November). Are some words worth more than others?. In *Proceedings of the first workshop on evaluation and comparison of NLP systems* (pp. 131–142).

Dušek, O., Howcroft, D. M., & Rieser, V. (2019). Semantic noise matters for neural natural language generation. In *Proceedings of the 12th international conference on natural language generation* (pp. 421–426).

Eisape, T., Zaslavsky, N., & Levy, R. (2020, November). Cloze distillation improves psychometric predictive power. In *Proceedings of the 24th conference on computational natural language learning* (pp. 609–619).

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019, July). Towards understanding linear word analogies. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3253–3262).

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations*. Doctoral dissertation. Universität Stuttgart. https://doi.org/10.18419/opus-2556.

Fandrych, I. M. (2004). *Non-morphematic word-formation processes: A multi-level approach to acronyms, blends, clippings and onomatopoeia*. Doctoral dissertation. University of the Free State.

Ferreira, V. S. (1996). Is it better to give than to donate? Syntactic flexibility in language production. *Journal of Memory and Language, 35*, 724–755.

Ferreira, V. S. (2003). The persistence of optional complementizer production: Why saying "that" is not saying "that" at all. *Journal of Memory and Language, 48*, 379–398.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology, 40*, 296–340.

Ferreira, V. S., & Griffin, Z. M. (2003). Phonological influences on lexical (mis) selection. *Psychological Science, 14*(1), 86–90.

Ficler, J., & Goldberg, Y. (2017, September). Controlling linguistic style aspects in neural language generation. In *Proceedings of the workshop on stylistic variation* (pp. 94–104).

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In *Studies in linguistic analysis*.

Frank, S. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 31, No. 31).

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science, 22*, 829–834.

Goldberg, A. E., & Ferreira, F. (2022). Good-enough language production. *Trends in Cognitive Sciences, 26*(4), 300–311. https://doi.org/10.1016/j.tics.2022.01.005

Goodkind, A., & Bicknell, K. (2018, January). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)* (pp. 10–18).

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences, 20*, 818–829.

Grosz, B., Joshi, A., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st annual meeting of the association for computational linguistics*. Association for Computational Linguistics.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science, 16*, 789–802.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 274–307.

Gundel, J. K., Bassene, M., Gordon, B., Humnick, L., & Khalfaoui, A. (2010). Testing predictions of the givenness hierarchy framework: A crosslinguistic investigation. *Journal of Pragmatics, 42*, 1770–1785.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research, 32*, 101–123.

Harley, H. (2017). *English words: A linguistic introduction*. United Kingdom: Wiley.

Harmon, Z., & Kapatsinski, V. (2017). Putting old tools to novel uses: The role of form accessibility in semantic extension. *Cognitive Psychology, 98*, 22–44. https://doi.org/10.1016/j.cogpsych.2017.08.002

Haskell, T. R., & MacDonald, M. C. (2003). Conflicting cues and competition in subject–verb agreement. *Journal of Memory and Language, 48*, 760–778.

Hint, H., Nahkola, T., & Pajusalu, R. (2020). Pronouns as referential devices in Estonian, Finnish, and Russian. *Journal of Pragmatics, 155*, 43–63.

Hollenstein, N., Chersoni, E., Jacobs, C. L., Oseki, Y., Prévot, L., & Santus, E. (2021, June). CMCL 2021 shared task on eye-tracking prediction. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 72–78).

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019, September). The curious case of neural text degeneration. In *Proceedings of the international conference on learning representations*.

Hsiao, Y., Gao, Y., & MacDonald, M. C. (2014). Agent-patient similarity affects sentence structure in language production: Evidence from subject omissions in mandarin. *Frontiers in Psychology, 5*, 1015.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*, 151–171.

Jacobs, C. L., & McCarthy, A. D. (2020, July). The human unlikeness of neural language models in next-word prediction. In *Proceedings of the the fourth widening natural language processing workshop*.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology, 61*, 23–62.

John, M. F. S., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46*, 217–257.

Kabbara, J., & Cheung, J. C. K. (2016, November). Stylistic transfer in natural language generation systems using recurrent neural networks. In *Proceedings of the workshop on uphill battles in language processing: Scaling early achievements to robust methods* (pp. 43–47).

Kahn, J. M., & Arnold, J. E. (2012). A processing-centered look at the contribution of givenness to durational reduction. *Journal of Memory and Language, 67*, 311–325.

Kemper, S. (1994). Elderspeak: Speech accommodations to older adults. *Aging and Cognition, 1*, 17–28.

Klein, O., Doyen, S., Leys, C., de Saldanha, M., da Gama, P. A., Miller, S., … Cleeremans, A. (2012). Low hopes, high expectations: Expectancy effects and the replicability of behavioral experiments. *Perspectives on Psychological Science, 7*, 572–584.

Koranda, M. J., Zettersten, M., & MacDonald, M. C. (2022). Good enough production: Speakers choose easy words over more precise ones. *Psychological Science*. https://doi.org/10.1177/09567976221089603

Kuiper, K., Van Egmond, M. E., Kempen, G., & Sprenger, S. (2007). Slipping on superlemmas: Multi-word lexical items in speech production. *The Mental Lexicon, 2*, 313–357.

Kulikov, I., Welleck, S., & Cho, K. (2021). *Mode recovery in neural autoregressive sequence modeling. arXiv preprint arXiv:2106.05459*.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1–26.

Lam, T. Q., & Watson, D. G. (2010). Repetition is easy: Why repeated referents have reduced prominence. *Memory & Cognition, 38*, 1137–1146.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*, 1–38.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Proceedings of the 19th international conference on neural information processing systems* (pp. 849–856).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692*.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments. *& Computers, 28*, 203–208.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology, 4*, 226.

Mager, M., Astudillo, R. F., Naseem, T., Sultan, M. A., Lee, Y. S., Florian, R., & Roukos, S. (2020, July). GPT-too: A language-model-first approach for AMR-to-text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1846–1852).

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition, 126*, 313–318.

Manin, D. (2006). Experiments on predictability of word in context and information rate in natural language. *Journal of Information Processes, 6*, 229–236.

Manning, E., Wein, S., & Schneider, N. (2020, December). A human evaluation of AMR-to-English generation systems. In *Proceedings of the 28th international conference on computational linguistics* (pp. 4773–4786).

Marchand, H. (1969). *The categories and types of present-day English word-formation. A synchronic-diachronic approach*. Munich, Germany: Beck.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: Freeman.

Mattiello, E. (2013). *Extra-grammatical morphology in English*. De Gruyter Mouton.

McCauley, S. M., Bannard, C., Theakston, A., Davis, M., Cameron-Faulkner, T., & Ambridge, B. (2021). Multiword units lead to errors of commission in children's spontaneous production: "What corpus data can tell us?*". *Developmental Science*. https://doi.org/10.1111/desc.13125

McDonald, J. L., Bock, K., & Kelly, M. H. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology, 25*, 188–230.

Meyer, A. S. (1990). The time course of phonological encoding in language production: The encoding of successive syllables of a word. *Journal of Memory and Language, 29*, 524–545.

Meyer, A. S. (1991). The time course of phonological encoding in language production: Phonological encoding inside a syllable. *Journal of Memory and Language, 30*, 69–89.

Meyer, A. S., Roelofs, A., & Brehm, L. (2019). Thirty years of Speaking: An introduction to the special issue. *Language, Cognition and Neuroscience, 34*, 1073–1084.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012, April). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th conference of the European Chapter of the Association for Computational Linguistics* (pp. 398–408).

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., & Ed.unov, S. (2019, August). Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the fourth conference on machine translation (Volume 2: Shared task papers, Day 1)* (pp. 314–319).

Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology, 135*, 151–166.

Nini, A. (2019). The multi-dimensional analysis tagger. In T. Berber Sardinha, & M. Veirano Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 67–94). London; New York: Bloomsbury Academic.

Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition, 114*, 227–252.

Oraby, S., Reed, L., Tandon, S., Sharath, T. S., Lukin, S., & Walker, M. (2018, July). Controlling personality-based stylistic variation with neural natural language generators. In *Proceedings of the 19th annual SIGdial meeting on discourse and dialogue* (pp. 180–190).

Pavlick, E., & Tetreault, J. (2016). An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics, 4*, 61–74.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT* (pp. 2227–2237).

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences, 108*, 3526–3529.

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

Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. In *Cambridge NA report NA2009/06* (pp. 26–46). Cambridge: University of Cambridge.

Rabovsky, M., & McClelland, J. L. (2020). Quasi-compositional mapping from form to meaning: A neural network-based approach to capturing neural responses during human language comprehension. *Philosophical Transactions of the Royal Society B, 375*, 20190313.

Race, D. S., & MacDonald, M. C. (2003). The use of "that" in the production and comprehension of object relative clauses. In *, 25. Proceedings of the annual meeting of the Cognitive Science Society*. Cognitive Science Society.

Roelofs, A. (1996). Serial order in planning the production of successive morphemes of a word. *Journal of Memory and Language, 35*, 854–876.

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition, 64*, 249–284.

Roelofs, A., & Meyer, A. S. (1998). Metrical structure in planning the production of spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 922–939.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics, 8*, 842–866.

Rosa, E. C., & Arnold, J. E. (2017). Predictability affects production: Thematic roles can affect reference form selection. *Journal of Memory and Language, 94*, 43–60.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *Proceedings of EMC^2 @ NeurIPS 2019*.

Schegloff, E. A. (1987). Analyzing single episodes of interaction: An exercise in conversation analysis. *Social Psychology Quarterly*, 101–114.

Schriefers, H., Meyer, A. S., & Levelt, W. J. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language, 29*, 86–102.

Sevald, C. A., & Dell, G. S. (1994). The sequential cuing effect in speech production. *Cognition, 53*, 91–127.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition, 133*, 140–155.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379–423.

Slevc, L. R., Wardlow Lane, L., & Ferreira, V. S. (2007). Pronoun production: Word or world knowledge. In *, 53. MIT working papers in linguistics* (pp. 191–203).

Tagg, C., & Seargeant, P. (2014). Audience design and language choice in the construction and maintenance of translocal communities on social network sites. In *The language of social media* (pp. 161–185). London: Palgrave Macmillan.

van Miltenburg, E., Clinciu, M., Dušek, O., Gkatzia, D., Inglis, S., Leppänen, L., Mahamood, S., Manning, E., Schoch, S., Thomson, C., & Wen, L. (2021, August). Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th international conference on natural language generation* (pp. 140–153).

van Rooij, I., & Baggio, G. (2020). Theory development requires an epistemological sea change. *Psychological Inquiry, 31*, 321–325.

van Rooij, I., & Blokpoel, M. (2020). Formalizing verbal theories: A tutorial by dialogue. *Social Psychology, 51*, 285–298.

van Schijndel, M., & Linzen, T. (2018). A neural model of adaptation in reading. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4704–4710).

Watson, D. G. (2010). The many roads to prominence: Understanding emphasis in conversation. In *, Vol. 52. Psychology of learning and motivation* (pp. 163–183). Academic Press.

Weatherford, K. C., & Arnold, J. E. (2021). Semantic predictability of implicit causality can affect referential form choice. *Cognition, 214*, Article 104759.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics, 1*, 98–106.

Yoon, S. O., & Brown-Schmidt, S. (2014). Adjusting conceptual pacts in three-party conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 919–937.

Zarcone, A., & Demberg, V. (2021). A bathtub by any other name: The reduction of German compounds in predictive contexts. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43, No. 43).

Zipf, G. (1949). *Human behavior and the principle of least effort*. New York: Addison-Wesley.