Strong Prediction: Language model surprisal explains multiple N400 effects

James A. Michaelov[1], Megan D. Bardolph[1], Cyma K. Van Petten[2], Benjamin K. Bergen[1], and Seana Coulson[1]

[1]Department of Cognitive Science, University of California San Diego, La Jolla, California, USA

[2]Department of Psychology, Binghamton University, Binghamton, New York, USA

Author Note

James A. Michaelov https://orcid.org/0000-0003-2913-1103

Cyma K. Van Petten https://orcid.org/0009-0007-6931-5149

Benjamin K. Bergen https://orcid.org/0000-0002-9395-9151

Seana Coulson https://orcid.org/0000-0003-1246-9394

Correspondence concerning this article should be addressed to James A. Michaelov, Department of Cognitive Science, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093. E-mail: j1michae@ucsd.edu

Abstract

Theoretical accounts of the N400 are divided as to whether the amplitude of the N400 response to a stimulus reflects the extent to which the stimulus was predicted, the extent to which the stimulus is semantically similar to its preceding context, or both. We use state-of-the-art machine learning tools to investigate which of these three accounts is best supported by the evidence. GPT-3, a neural language model (LM) trained to compute the conditional probability of any word based on the words that precede it, was used to operationalize contextual predictability. In particular, we used an information theoretical construct known as surprisal (the negative logarithm of the conditional probability). Contextual semantic similarity was operationalized by using two high-quality co-occurrence-derived vector-based meaning representations for words: GloVe and fastText. The cosine between the vector representation of the sentence frame and final word was used to derive Contextual Cosine Similarity (CCS) estimates. A series of regression models were constructed, where these variables, along with cloze probability and plausibility ratings, were used to predict single trial N400 amplitudes recorded from healthy adults as they read sentences whose final word varied in its predictability, plausibility, and semantic relationship to the likeliest sentence completion. Statistical model comparison indicated GPT-3 surprisal provided the best account of N400 amplitude and suggested that apparently disparate N400 effects of expectancy, plausibility and contextual semantic similarity can be reduced to variations in the predictability of words. The results are argued to support predictive coding in the human language network.

Keywords: distributional semantics, ERPs, N400, neural language models, predictive coding

Strong Prediction: Language model surprisal explains multiple N400 effects

Introduction

Fedorenko and Thompson-Schill (2014) note that the brain systems that support language processing are better described at the level of interactive networks than individual brain regions, arguing that investigations into the functional significance of neural activity are best directed at large-scale distributed neural networks, that is, a set of interconnected brain regions acting in concert. This may explain why language researchers have found event-related brain potentials (ERPs) to be such a useful method for probing the neurobiology of language, despite known limitations in the spatial resolution of the technique (see Federmeier et al., 2016 for a review). EEG reflects post-synaptic potentials generated mainly in cortical pyramidal cells (Luck, 2014). Moreover, brain activity cannot be detected at the scalp unless large numbers (on the order of 10 million) of neurons are simultaneously active (Woodman, 2010). The identification of any scalp recorded potentials whose amplitude is systematically modulated by language processing demands is thus likely to reflect activity in the very sort of interactive neural networks Fedorenko and Thompson-Schill (2014) propose.

One ERP component of particular interest to language researchers is the N400, a monophasic negativity peaking approximately 400ms after the onset of a visually presented word. The N400 was first reported in a study that compared ERPs elicited by the last word of sentences that made sense (*He takes his coffee with cream and* ***sugar***) versus those that did not *(He takes his coffee with cream and* ***dog***; Kutas and Hillyard, 1980). However, it soon became clear that the N400 is not only observed at the end of sentences; it is elicited by all words, written, spoken, or signed, and that its amplitude is modulated by factors such as contextual congruity, frequency of usage, and category membership, all thought to affect the difficulty of retrieving information in semantic memory (for review see Kutas & Federmeier, 2011).

Here we consider the adequacy of two proposals regarding the functional significance of the N400 that differ in their implications for the underlying neurocognitive mechanisms. The first is that N400 amplitude is sensitive to the conditional probability of words in their linguistic contexts as driven by a predictive coding mechanism. This account is referred to below as *predictive preactivation*. The second is that N400 amplitude is driven by a context-sensitive retrieval mechanism and as such indexes the semantic similarity of incoming words to the semantic features of prior words in the context. This is referred to below as *contextual semantic similarity*. We briefly review empirical support for each of these proposals as well as that for a combined account.

One reason for the continued dispute on this issue is that advocates of each account have mostly focused on a subset of N400 effects, discounting the relevance of less amenable phenomena and arguing that they are potentially explicable given a suitable operationalization of either expectancy or semantic similarity. Whereas advocates of predictive processing focus on expectancy effects (DeLong, Troyer, et al., 2014; Kuperberg & Jaeger, 2016; Bornkessel-Schlesewsky & Schlesewsky, 2019; Kuperberg et al., 2020), advocates of contextual similarity and combined accounts focus on the way that N400 amplitude is modulated by the presence of semantically related words in the immediate context (Ettinger et al., 2016; Federmeier, 2021; Lau et al., 2013; Uchida et al., 2021). By contrast, the present study examines manipulations of the expectancy, plausibility, and the relatedness of sentence final words to the words that precede them.

Noting how researchers in the neurobiology of language have struggled to operationalize the theoretical constructs proposed to drive the N400, we turn instead to tools from computational linguistics. The 21st century has seen immense progress in the utility of *language models*, statistical tools to characterize the probability of words in texts (Berger & Packard, 2022; Jurafsky & Martin, 2021). Trained on large corpora to compute the probability distribution over a vocabulary of words, language models are used in applications such as information retrieval, speech recognition, machine translation, and chatbots. Although language models are not proposed as cognitive models per se, we suggest that the data-driven estimates they provide serve as excellent metrics for the theoretical constructs proposed to drive the N400. We utilize three state-of-the-art language models to provide metrics for the predictability and the contextual semantic similarity of our sentence-final words and compare their adequacy in accounting for N400 effects of expectancy, plausibility, and relatedness in human participants.

Predictive Preactivation Account

One prominent account of the N400 is that it reflects the activation of semantic features associated with the eliciting word (Kutas & Federmeier, 2011). According to this account, contextual congruity effects occur because elements of the prior context have already activated some of these associated features. If relevant features associated with a word have been activated by the preceding context—whether these be semantic features (Federmeier, 2021; Kuperberg et al., 2020) or a combination of semantic, grammatical, and phonological features (as supported by the work of DeLong et al., 2005; Fleur et al., 2020; Nicenboim et al., 2020; Otten et al., 2007; Urbach et al., 2020; Van Berkum et al., 2005)—they need not be newly activated when the word is encountered, and thus the amplitude of the N400 is less than when words are encountered alone or in less supportive contexts.

The most obvious source of support for predictive preactivation lies in the close relationship between N400 amplitude and the expectancy metric known as *cloze probability* (the proportion of people to fill in the relevant gap in a sentence with a given word; Taylor, 1953, 1957). A higher-cloze continuation of a sentence elicits a smaller (i.e., more positive) N400 response, while a lower-cloze continuation elicits a larger (more negative) N400 (Kutas & Federmeier, 2011; Kutas & Hillyard, 1984). In fact, in previous work the two variables have been reported to have a Pearson correlation coefficient r of -0.9 or more (Kutas & Federmeier, 2011; Kutas & Van Petten, 1994). As the cloze task requires participants to predict an upcoming word, cloze probability has often been argued to reflect how predictable a word is in context (Brothers & Kuperberg, 2021; Fischler &

Bloom, 1979; Kuperberg et al., 2020; Kutas et al., 2011; Kutas & Hillyard, 1984; Luke & Christianson, 2016; Tannenbaum et al., 1965; Van Petten & Luka, 2012). Moreover, the negative correlation between N400 amplitude and cloze probability tells us that N400 amplitude is not simply a categorical indicator of surprise, but reflects the predictability of the eliciting word in a more fine-grained way.

Beyond the graded predictability effect, the predicted preactivation account is supported by the way that N400 amplitude is modulated by sentence context. Research has shown that words elicit a large N400 when presented alone, a large N400 when presented in sentence frames that render them unexpected, and a progressively smaller N400 in more supportive sentence contexts, suggesting that what reduces the amplitude of the response is the activation of neural representations associated with the stimulus before the stimulus is encountered (Dambacher et al., 2006; Payne et al., 2015; Van Petten, 1993; Van Petten & Kutas, 1990, 1991; for discussion, see Federmeier, 2021; Van Petten & Luka, 2012). Second, unlikely sentence continuations elicit similar sized N400 in constraining contexts in which there is a highly salient alternative (e.g., **month** in *The bill was due at the end of the* ***hour***) and in open-ended contexts in which there is not (e.g., *He kicked himself when he realized that he forgot the* ***key***; see DeLong & Kutas, 2020; Federmeier, 2021; Kuperberg et al., 2020; Van Petten & Luka, 2012).

This sensitivity to the contextual fit of the actual word encountered rather than the predictability of potential alternatives has been interpreted as suggesting that rather than the registration of surprise, the N400 reflects the activation of semantic (and possibly other) features associated with the word presented. In this account, cloze probability effects occur because the greater the extent of preactivation for a word's features, the smaller the N400 elicited by the word (DeLong & Kutas, 2020; DeLong, Quante, et al., 2014; Federmeier, 2021; Kuperberg et al., 2020; Kutas et al., 2011; Kutas & Federmeier, 2011; Van Petten & Luka, 2012).

In addition to cloze, the amplitude of the N400 is also correlated with other metrics of predictability. Research has found that predictions of language models, computational systems designed to predict the probability of a word in context based on the surface-level statistics of language, are correlated with the N400 response to these words (Aurnhammer & Frank, 2019; Frank et al., 2015; Merkx & Frank, 2021). Specifically, such studies find that the surprisal, the negative logarithm of the conditional probability of a word, is a significant predictor of N400 amplitude (Aurnhammer & Frank, 2019; Ettinger, 2020; Frank et al., 2015; Merkx & Frank, 2021; Michaelov & Bergen, 2020; Michaelov et al., 2022; Parviz et al., 2011; Szewczyk & Federmeier, 2022).

Research also shows that language model surprisal can be used to model N400 effects—in many cases, where we find a significant difference in N400 amplitude between stimuli from two experimental conditions, we also find a significant difference in surprisal in the same direction (Michaelov & Bergen, 2020). Further, this computational approach fits into a larger body of work showing that N400 amplitude is sensitive to the statistics of language—for example, more frequent words elicit smaller N400 responses (Dambacher et al., 2006; Fischer-Baum et al., 2014; Kutas & Federmeier, 2011; Rugg, 1990; Van Petten, 1993; Van Petten & Kutas, 1990). These results together suggest that the N400 component reflects a neural process that veridically tracks the conditional probability of upcoming words. Note that the definition of conditional probability here is not restricted to that calculated by a traditional n-gram model, only based on actual co-occurrences of lexical items; language models are designed to generalize based on their training data when making predictions, and humans are also thought to do so (DeLong & Kutas, 2020; DeLong, Troyer, et al., 2014; Kuperberg et al., 2020).

Contextual Semantic Similarity

An alternative explanation of the neural activity underlying the N400 is contextual semantic similarity. Under this account, as we comprehend a sentence, the semantic features of each word are activated and briefly maintained, thereby reducing the neural activity required in response to words with overlapping features (Federmeier, 2021). While this feature-based account is compatible and indeed central to some prediction-based accounts of the N400 (e.g. Kuperberg et al., 2020), the key difference is that the activations are limited to semantic features of previously encountered words. That is, there is no additional spreading activation to related words or semantic features, and, crucially, no prediction. Some investigators have suggested that contextual semantic similarity accounts for all variation in N400 amplitude (Ettinger et al., 2016; Uchida et al., 2021), while others suggest semantic similarity acts in concert with a prediction mechanism (see, e.g. Federmeier, 2021; Frank & Willems, 2017; Lau et al., 2013).

Several previous ERP studies have examined the impact of semantically related words within sentences or sentence-like word strings, with results that suggest the N400 component is sensitive to semantic similarity among the individual words that comprise sentences along with factors that are difficult to accommodate within a pure similarity account. For instance, an early experiment found that the relationship between the two terms of a statement about category membership influenced the N400, whereas the truth or falsity of the statement had no impact, so that *a robin is a **bird*** and *a robin is not a **bird*** were equivalent and both led to smaller N400s than *a robin is/is not a **vehicle*** (Fischler et al., 1984). Similarly, Kounios and Holcomb (1992) found no impact of quantifiers *all*, *some*, and *no* on statements about category membership. However, a more recent study on this topic reports N400 effects both for relationships between words (viz., *farmers* primes *crops* more than *farmers* primes *worms*) as well as a small N400 effect of quantifiers, that is, the final word of the more plausible sentence *farmers often grow **crops*** elicited a smaller N400 than *farmers rarely grow **crops*** (Urbach & Kutas, 2010).

Outside the realm of negation and quantification, initial studies showed that the presence of a strongly related word within either a meaningful sentence (e.g., *When the **moon** is full, it is hard to see many **stars** or the Milky Way*) or a grammatically legal but meaningless word string (e.g., *When the **moon** is rusted, it is available to buy many **stars** or the Santa Ana*) leads to a smaller N400 to **stars** than if the prior context does not include a related word (Van Petten, 1993; Van Petten et al., 1997). However, other studies indicate that N400 is not driven solely by an automatic semantic comparison process during sentence comprehension. Coulson and colleagues found much smaller N400s to the second words of related (*tin/**aluminum***) than unrelated (*tin/**disposal***) word pairs when the pairs were presented by themselves (Coulson et al., 2005). The word pairs were then embedded in sentences that were compatible or incompatible with the word-pair relationship, like the quartet below.

(1) (a) Coke cans used to be made out of tin but now they use **aluminum**.

(b) Paul heard a loud grinding noise when someone put a tin can right down the garbage **aluminum**.

(c) Paul heard a loud grinding noise when someone put a tin can right down the garbage **disposal**.

(d) Coke cans used to be made out of tin but now they use **disposal**.

In the incongruous sentences, the presence of a semantically related word continued to reduce the amplitude of the N400 elicited by the final words—condition (b) smaller than (d)—but this difference was dramatically smaller and shorter in duration than when the word pairs were presented in isolation. In contrast, the impact of overall sentence congruity—conditions (a) and (c) versus (b) and (d)—dwarfed the impact of a single related word earlier in the sentence.

Camblin et al. (2007) similarly pitted overall plausibility against lexical relationships by embedding strongly related word pairs (*arms* / *legs*) in discourse contexts that were more or less compatible with the word-pair relation (skin irritation from a sunburn would be likely to affect both arms and legs, but irritation from a wool sweater would not). Much like Coulson et al. (2005), they found smaller N400s for the second words of semantically similar pairs than their unrelated controls, but that this effect was substantially smaller when opposed by the global discourse context.

As for the prediction account, the contextual semantic similarity account is supported by work with computational models. N400 amplitude, for example, has been found to correlate with the degree of semantic similarity between prime and target word (Chwilla & Kolk, 2005; Van Petten, 2014), as operationalized by Latent Semantic Analysis (LSA), a measure of semantic distance derived from word co-occurrence frequencies in written corpora (Dumais et al., 1988; Dumais, 2004; Landauer et al., 1998). This is also true for words in sentence contexts—N400 amplitude is correlated with the LSA distance between a target word and the words that precede it (Chwilla et al., 2007; Parviz et al., 2011), and with other statistically derived metrics of word similarity (Broderick et al., 2018; Ettinger et al., 2016; Frank & Willems, 2017; Parviz et al., 2011; Uchida et al., 2021; Van Petten, 2014).

Multiple Systems Accounts

A number of investigators have suggested the brain activity underlying the N400 reflects both predictive preactivation and contextual semantic similarity. Some of these suggest that the contextual semantic similarity system operates by default, and the predictive system is engaged under conditions of increased attention (Federmeier, 2021), or when predictions are more likely to be successful, as when a high proportion of word pairs are semantically related (Holcomb, 1988; Lau et al., 2013). Some studies have shown that conditions that foster prediction result in N400 effects with an earlier onset latency than conditions that do not, such as those with little time between words (Anderson & Holcomb, 1995; Luka & Van Petten, 2014), or a small proportion of related word pairs (Lau et al., 2013).

According to other accounts, both systems are constantly active but implemented in different brain circuits. In one fMRI experiment, Frank and Willems (2017) found that contextual semantic similarity was correlated with activations in the anterior middle temporal sulcus, the precuneus, and bilateral angular gyri, whereas predictability was correlated with activations in the left inferior temporal sulcus, left posterior fusiform gyri, bilateral superior temporal gyri, and bilateral amygdalae. In view of the limited temporal resolution of fMRI, however, it is also possible that these findings reflect a disparate impact of contextual similarity and predictability at distinct stages of language processing.

Finally, one well-replicated result appears challenging to accommodate in single-system accounts, whether predictive or similarity-based. Kutas and Hillyard (1984) first reported that generally poor (unexpected) sentence completions elicited smaller N400s if they were semantically related to the most expected completion than if not, so that *He liked lemon and sugar in his* ***coffee*** led to a less negative ERP than an equally unexpected word (***dog***) that is semantically dissimilar to the expected completion (***tea***). The finding that words related to the best completion elicit significantly less negative N400 responses than their unrelated counterparts has been replicated many times, and occurs regardless of whether the related words comprise congruous or anomalous continuations of a sentence (Amsel et al., 2015; DeLong et al., 2019; Federmeier & Kutas, 1999; Ito et al., 2016; Kutas, 1993; Kutas & Hillyard, 1984; Kutas et al., 1984; Thornhill & Van Petten, 2012). One might imagine that this effect (relationship-to-best-completion, or RBC) arises from predicting a sentence completion, followed by an assessment of the similarity between that prediction and the actually delivered word, but no study has suggested that the RBC effect is temporally delayed relative to simple sentence congruity effects. Because an RBC condition is included in the present study, we return to theoretical accounts and attempts to computationally model it in the Discussion.

The Present Study

In the present study we explore whether the brain activity underlying the scalp-recorded N400 component is driven by predictability, contextual semantic similarity, or a combination of the two. To do so, we recorded EEG as participants read sentences whose final words were designed to elicit three kinds of N400 effects: predictability, plausibility, and relatedness to the best completion (RBC). Based on the stimuli used by Thornhill and Van Petten (2012), our materials were sentence frames with four different kinds of sentence-final words. As in the original study, the predictability manipulation was guided by results from a cloze task. The Best Completion condition was thus the word with the highest cloze probability. The Related completions were low-cloze completions semantically related to the best completions, as determined by Thornhill and Van Petten (2012). Likewise the Unrelated completions were low cloze completions unrelated to the best completions. Finally, to investigate the plausibility effect, we included Implausible completions, completions with a cloze probability of zero that were also implausible.

(2) (a) Best Completion: On his vacation, he got some much needed **rest**.

(b) Related: On his vacation, he got some much needed **relaxation**.

(c) Unrelated: On his vacation, he got some much needed **sun**.

(d) Implausible: On his vacation, he got some much needed **airlines**.

We then use state-of-the-art language models to calculate the predictability and contextual similarity of our stimuli and investigate how well these metrics predict the single-trial N400 amplitudes elicited by the stimuli. To operationalize predictability, we used the transformer neural network language model, GPT-3. Research has shown that in general, larger language models trained on more data provide the best fits to human data, and that transformer neural networks are the architecture best suited to predicting N400 data (Merkx & Frank, 2021). However, rather than using the conditional probabilities assigned by GPT-3 to our stimuli, we instead utilize *surprisal* scores, the negative logarithm of the probability assigned by the language model to a given word in context. Previous work has shown that when directly compared, language model surprisal is a better predictor of N400 amplitude than raw probability (Szewczyk & Federmeier, 2022; Yan & Jaeger, 2020).

Contextual semantic similarity is generally calculated as the cosine distance between a vector representation of the stimulus word (often referred to as an embedding) and the mean vector across each word in the context, where the vector representations are based on the statistics of language. To operationalize contextual semantic similarity, we took advantage of two different tools for obtaining vectors for word meanings, GloVe (Pennington et al., 2014) and fastText (Bojanowski et al., 2016; Mikolov et al., 2018). GloVe (Pennington et al., 2014) is an unsupervised learning algorithm trained on global, aggregated word-word co-occurrence statistics that yields vector representations for words. The fastText library (Bojanowski et al., 2016) is an updated version of word2vec (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013), which has been used in previous work investigating the effect of contextual semantic similarity (Ettinger et al., 2016; see also Frank and Willems, 2017; Nieuwland et al., 2020 for related approaches). Both models are driven by language statistics, but GloVe embeddings are derived from co-occurrence statistics from a whole corpus (Pennington et al., 2014), while fastText embeddings are retrieved from a neural network (known as a continuous bag-of-words model) trained to predict a word based on the other words occurring in a given sentence (Bojanowski et al., 2016; Mikolov et al., 2018).

We expect that our experimental manipulation of predictability, plausibility, and relatedness to the best completion will replicate each of these well-documented effects on the N400, as would be evidenced by an effect of experimental condition. In particular, we expect the Best completions to elicit the least negative (most positive) N400, the Implausible completions to elicit the most negative N400, and the Related and Unrelated completions to fall in between the two. Despite the fact that the Related and Unrelated completions are matched for cloze probability and plausibility, the Related completions are expected to elicit smaller N400 than Unrelated completions.

Next we use our metrics of predictability and contextual semantic similarity to model single-trial N400 data using linear mixed effects regressions. If the brain activity underlying the N400 reflects predictive preactivation, we expect regressions incorporating surprisal to provide the best account of the data. Alternatively, if the brain activity underlying the N400 reflects contextual semantic similarity, we expect regressions incorporating one of our cosine similarity measures to provide the best account of the data. Finally, if the N400 reflects the operation of both a predictive preactivation mechanism and one for contextual similarity, the best account of the data will lie in regressions that incorporate measures both for surprisal and cosine similarity.

Materials and Methods

Participants

50 UCSD volunteers participated for course credit or payment. Participants were right-handed, fluent English speakers with normal or corrected-to-normal vision with no history of neurological or psychiatric disorders. Participants ranged in age from 18 to 31 years old.

Materials

Our stimuli were based on the original stimuli of the experiment carried out by Thornhill and Van Petten (2012). These stimuli were of the form given in Table 1. For each sentence frame, stimuli fall under four conditions—Best Completions, the completions with the highest cloze probability; Related Completions, low-cloze completions that are semantically related to the best completions (as determined by Thornhill and Van Petten, 2012); and Unrelated Completions, low-cloze completions that are unrelated to the best completions. Thornhill and Van Petten (2012) found that these stimuli elicit both a predictability and RBC effect in human comprehenders. In order to also investigate the plausibility effect, we added a fourth experimental condition of Implausible Completions.

Sentences were normed via online surveys using the same participant pool we used to recruit participants for the EEG study. First, cloze probability measures were collected from UCSD students such that each sentence frame was completed by at least 35 participants. In this survey, participants were provided with sentence frames and instructed to produce the last word of the sentence. Average cloze probability and standard deviation for each condition are shown in Table 1.

All sentences were also rated for plausibility by a separate group of at least 30 students. In this survey, participants read one sentence at a time and were asked to rate each on a scale from 1 (very plausible) to 5 (very implausible). Multiple stimulus lists were employed so that each participant viewed only one of the four versions of each sentence frame. Average plausibility ratings for each experimental condition are shown in Table 1. All sentences in the Implausible condition had ratings above 3.5, with an average rating of 4.3. By contrast, the other conditions all had ratings below 2, suggesting participants found them plausible.

These stimuli were initially constructed as part of a larger study. In order to use the computational tools required to test our hypotheses, we opted to analyze a subset of the data such that critical words of all sentence stimuli appeared as whole tokens in GPT-3, GloVe, and fastText—that is, critical words were present as whole words in the vocabularies of these models. We then further selected stimuli such that, as in Thornhill and Van Petten (2012), there was no overall difference in cloze probability between the related and unrelated completions. We also additionally ensured that there was no overall difference in plausibility. Thus, the two conditions differed only in how related they were to the Best Completion for that sentence. This resulted in a final stimulus set of 125 sentence frames in 4 conditions, for a total of 500 items. The stimuli were presented along with 165 other sentences that were part of the larger study and thus similar in character to the experimental sentences. As for the experimental sentences, these additional stimuli were equally likely to end with the Best, Related, Unrelated, or Implausible completion for the sentence frame as each participant saw approximately 41 non-experimental stimuli in each condition—in addition to the approximately 31 experimental sentences in each condition.

Table 1

Descriptive Statistics for Sentences: Mean and standard deviation of cloze probabilities and plausibility ratings (1 = very plausible; 5 = very implausible) for each experimental condition.

		Cloze		Plausibility	
Condition	Example Stimulus	Mean	SD	Mean	SD
Best	*It's hard to admit when one is* ***wrong****.*	49.8%	27.3%	1.4	0.3
Related	*It's hard to admit when one is* ***incorrect****.*	2.3%	3.3%	1.5	0.4
Unrelated	*It's hard to admit when one is* ***lonely****.*	2.3%	3.9%	1.5	0.3
Implausible	*It's hard to admit when one is* ***screened****.*	0%	0%	4.3	0.4

Procedure

Testing consisted of a single experiment session, with words presented centrally using RSVP presentation. For each sentence, participants first saw a break screen, then pressed a key to display the sentence. A fixation character remained on the screen while words were presented for 300ms, followed by a 200ms blank screen. The final word was displayed for 1200ms. After some sentences, participants saw a question about the content of the previous sentence (e.g., "Was the previous sentence about banking?") and responded Yes or No with a button press.

EEG Recording and Analysis

The electroencephalogram was recorded from 29 electrodes in an Electro-cap organized in the International 10–20 configuration. Additional electrodes were placed below the eye and near the external canthi to detect eye movements and blinks. Scalp electrodes were referenced on-line to an electrode on the left mastoid, and later re-referenced to an average of the left and right mastoid electrodes. The EEG was amplified using an SA Instrumentation bioelectric amplifier, digitized online at 250 Hz.

EEG was time locked to the onset of each sentence final word. Mean voltage during the 100ms interval preceding each word's appearance was used to baseline epochs spanning 100ms before until 900ms after word onset. Trials containing artifacts due to blinks, eye movements, or amplifier saturation were removed prior to analysis. As discussed in Materials, the data used in the present study were collected as part of a larger experiment involving additional stimuli constructed to cover the same four conditions. We analyze all the data for stimuli that fulfilled the requirements stated in Materials, namely, stimuli where all critical words existed as whole words in all language models' and word embeddings models' vocabularies and Related and Unrelated words were matched for Cloze and Plausibility.

N400 amplitude was operationalized as the mean voltage 300-500ms post-onset recorded from nine centro-parietal electrodes: C3, Cz, C4, CP3, CPz, CP4, P3, Pz, and P4. All graphs and statistical analyses were run in *R* (R Core Team, 2022) using *Rstudio* (RStudio Team, 2020) and the *tidyverse* (Wickham et al., 2019), *lme4* (Bates et al., 2015), *lmerTest* (Kuznetsova et al., 2017), *corrr* (Kuhn et al., 2022), *colorspace* (Zeileis et al., 2020; Zeileis et al., 2009), *gridExtra* (Auguie, 2017), and *cowplot* (Wilke, 2020) packages. All figures use colorblind-friendly palettes (Chang, 2022; Jackson, 2016; Zeileis et al., 2020). All reported *p*-values are corrected for multiple comparisons based on false discovery rate (Benjamini & Yekutieli, 2001).

Computational Metrics

In this paper, we derive three computational metrics based on the statistics of language—GPT-3 surprisal, GloVe cosine similarity, and fastText cosine similarity. While the pretrained models we used differ in a number of ways, we did attempt to match some of their properties as much as possible. For example, GPT-3, GloVe, and fastText are all trained on Common Crawl data (https://commoncrawl.org/), albeit using subsets of different sizes. GPT-3 is trained on 300 billion tokens, GloVe on 840 billion, and fastText on 600 billion tokens. In spite of these differences, at a minimum the corpus is the same and the training set is the same order of magnitude for all three models. Further, to ensure that all the models are equally able to capture the relationships between the stimuli and their contexts, stimuli were chosen such that critical words existed as whole words in all models' vocabularies. For this reason, we use the version of fastText that does not include sub-word information in its representations, as the other models do not have access to sub-word information. More details on how each metric was calculated are provided below.

GPT-3 Surprisal

The OpenAI API (OpenAI, 2021) was used to access the predictions of the largest original GPT-3 model (*Davinci*), which has 175 billion parameters (Brown et al., 2020). Each sentence stimulus was input into the API and GPT-3 was used to calculate the probability of the final word given its preceding context. This figure was then used to calculate the log-probability of each critical word. Since these log-probabilities used the natural exponent as a base, they were converted to the logarithm of base two and multiplied by negative one. The resultant surprisal values are thus measured in bits (see, e.g., Futrell et al., 2019, for discussion).

GloVe Cosine Similarity

To obtain the measure of contextual similarity we refer to as GloVe Contextual Cosine Similarity, we used the GloVe (Pennington et al., 2014) vectors made available through the GloVe website (https://nlp.stanford.edu/projects/glove/)—specifically, the version with a 2.2 million word vocabulary and 300-dimensional vectors trained on 840 billion tokens from the Common Crawl corpus. We took the mean vector of all the words preceding the stimulus word and then used SciPy (Virtanen et al., 2020) to calculate the cosine similarity between this vector and the vector corresponding to the stimulus word. Because cosine similarity is based on the angle between two vectors and is not affected by the overall magnitude, this approach is equivalent to taking the sum of the context vectors as in Frank and Willems (2017).

We also calculate the similarity between the best completion (i.e., highest-cloze sentence completions) and each critical word in each sentence frame, which we refer to as GloVe Best Completion Cosine Similarity or GloVe BCCS.

fastText Cosine Similarity

To calculate fastText Contextual Cosine Similarity, we utilized the fastText (Bojanowski et al., 2016) vectors made available through the fastText website (https://fasttext.cc/)—specifically, the version with a 2 million word vocabulary, 300-dimensional vectors, and no sub-word information trained on 600 billion tokens from the Common Crawl corpus. As with the GloVe vectors, we calculated the cosine similarity between the vector corresponding to the stimulus word and the mean vector of the preceding context. In addition to calculating fastText Contextual Cosine Similarity, we also calculate fastText Best Completion Cosine Similarity or fastText BCCS.

Results

Figure 1 shows grand average ERP waveforms for words in each of the four conditions (Best Completion, Related, Unrelated, and Implausible) along with topographic maps. By convention, negative voltage is plotted upwards making it apparent that, as predicted, the Implausible condition elicited the largest (most negative) N400, and the Best Completions elicited the smallest (most positive) N400. The Unrelated condition fell in between these two extremes, and, as predicted, elicited more negative ERPs than did the Related condition (which was virtually overlapping the Best Completion condition, despite the large difference in their average cloze probability). The topographic maps were formed by first calculating point-by-point difference waves obtained by subtracting the amplitude of ERPs recorded at each electrode in the Best Completion condition from their counterparts in the Related, Unrelated, and Implausible conditions, respectively. The mean amplitude 300-500ms was then measured on each difference wave and plotted on the scalp to visualize the relative pattern of positive and negative voltage. The posterior negativity apparent in all three plots is characteristic of N400 ERP effects reported in sentence reading paradigms like the one used here.

Figure 2 presents normalized (z-scored; and in the case of surprisal and plausibility, multiplied by -1) values in each experimental condition for the outcome variable (N400) and for each of our predictors. Note that the human derived metrics of cloze probability and plausibility reflect our experimental design. The Best Completions were intended to be predictable, while the Related, Unrelated, and Implausible conditions were designed to be unexpected, with Related and Unrelated conditions equated for cloze probability. Similarly, Best Completions, Related, and Unrelated conditions were all intended to be plausible, whereas the Implausible condition was intended to be implausible. Figure 1 indicates that all of the computational metrics were associated with differences between Best Completions and Implausible endings. Related and Unrelated conditions were quite similar on some metrics—such as GloVe Contextual Cosine Similarity (CCS) and fastText CCS—and differed on others, such as GPT-3 surprisal and both measures of Best Completion Cosine Similarity (BCCS).

Figure 3 presents a heatmap of correlations between the various predictors used in the regression analyses below. Recall that Contextual Cosine Similartity (CCS) is the cosine of the angle between the vector for each word and the mean of the vectors for each of the words in the preceding sentence context and serves as an operationalization of contextual semantic similarity. Best Completion Cosine Similarity (BCCS) is the cosine of the angle between the vector for each word and the vector for the word that is the best completion for the sentence frame and is relevant for some multiple systems accounts. Although the two kinds of embeddings (GloVe and fastText) yielded virtually identical estimations of similarity between pairs of words—as reflected in the 0.98 correlation between GloVe BCCS and fastText BCCS—they differed somewhat in their estimates of contextual semantic similarity (CCS) as GloVe CCS and fastText CCS had a correlation coefficient of 0.66. Relative to GloVe CCS, fastText CCS was more associated with cloze probability (0.39 versus 0.32), GPT-3 surprisal (-0.61 versus -0.46), and plausibility (-0.56 versus -0.37). Relative to GloVe CCS, the fastText CCS measure also showed more sensitivity to the semantic relationship between each unexpected ending and the best completion, as evidenced by a greater correlation with fastText BCCS (0.52 versus 0.33) and even with GloVe BCCS (0.54 versus the 0.4 correlation between GloVe CCS and Glove BCCS).

GPT-3 exhibited similar correlations with cloze probability (-0.33) as did the CSS measures described above. Moreover, GPT-3 surprisals were highly correlated with human measures of plausibility (0.85), a level far greater than any of the other measures. As noted above, GPT-3 surprisal exhibited moderate negative correlations with both measures of CCS (-0.61 for fastText and -0.46 for GloVe). GPT-3 exhibited even higher correlations with the measures of BCCS (-0.71 for fastText and -0.73 for GloVe), presumably due to the way BCCS implicitly incorporates the predictions of the best completion.

Figure 1

ERP scalp maps and waveforms. Panel A shows the topography of the mean amplitude 300-500ms of the difference wave for the RBC and Best Completion conditions (top), Unrelated and Best Completion (middle), and Implausible and Best Completion (bottom) using a spherical spline interpolation. Panel B shows the ERP waveforms for each condition (Best Completion, Related, Unrelated, Implausible) as measured at the Centroparietal Electrode Cluster used in the regression models.

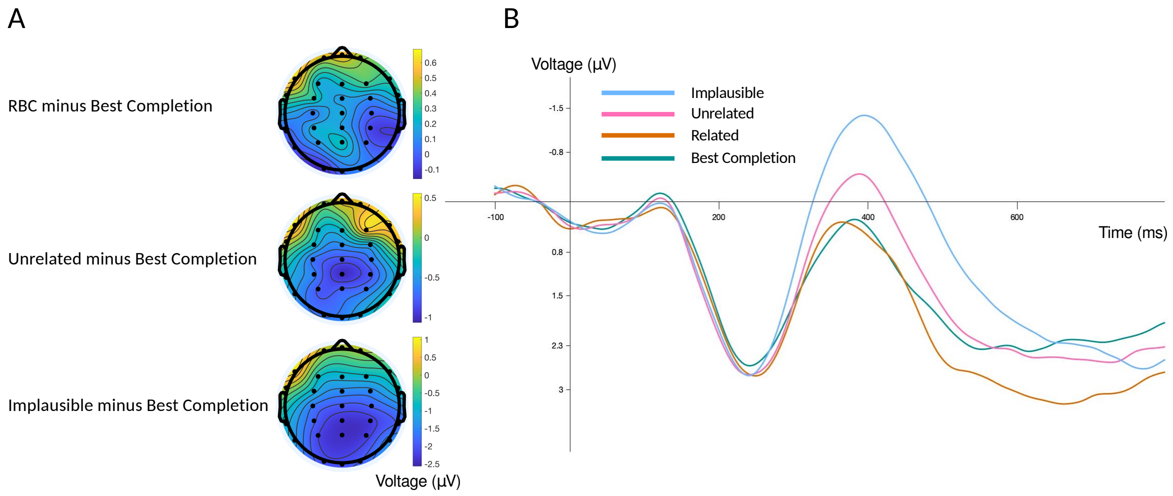


Figure 2

Average values of all predictors under each experimental condition. For easier comparison across predictors, we plot negative surprisal and plausibility, and the values of all predictors were z-scored. For easier comparison to the N400 waveform, the y-axis is reversed, with negative values plotted upwards. Error bars show the standard error.

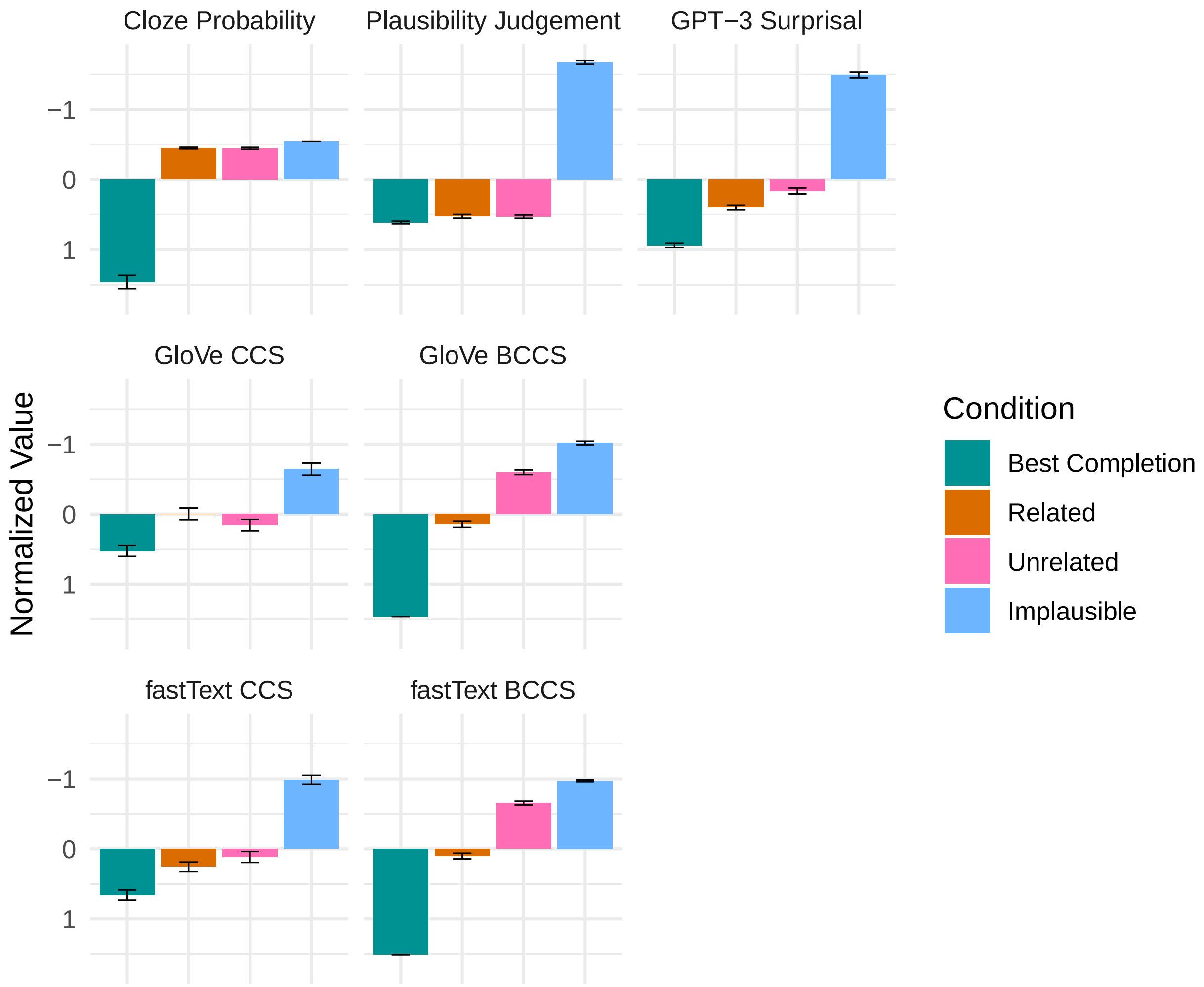
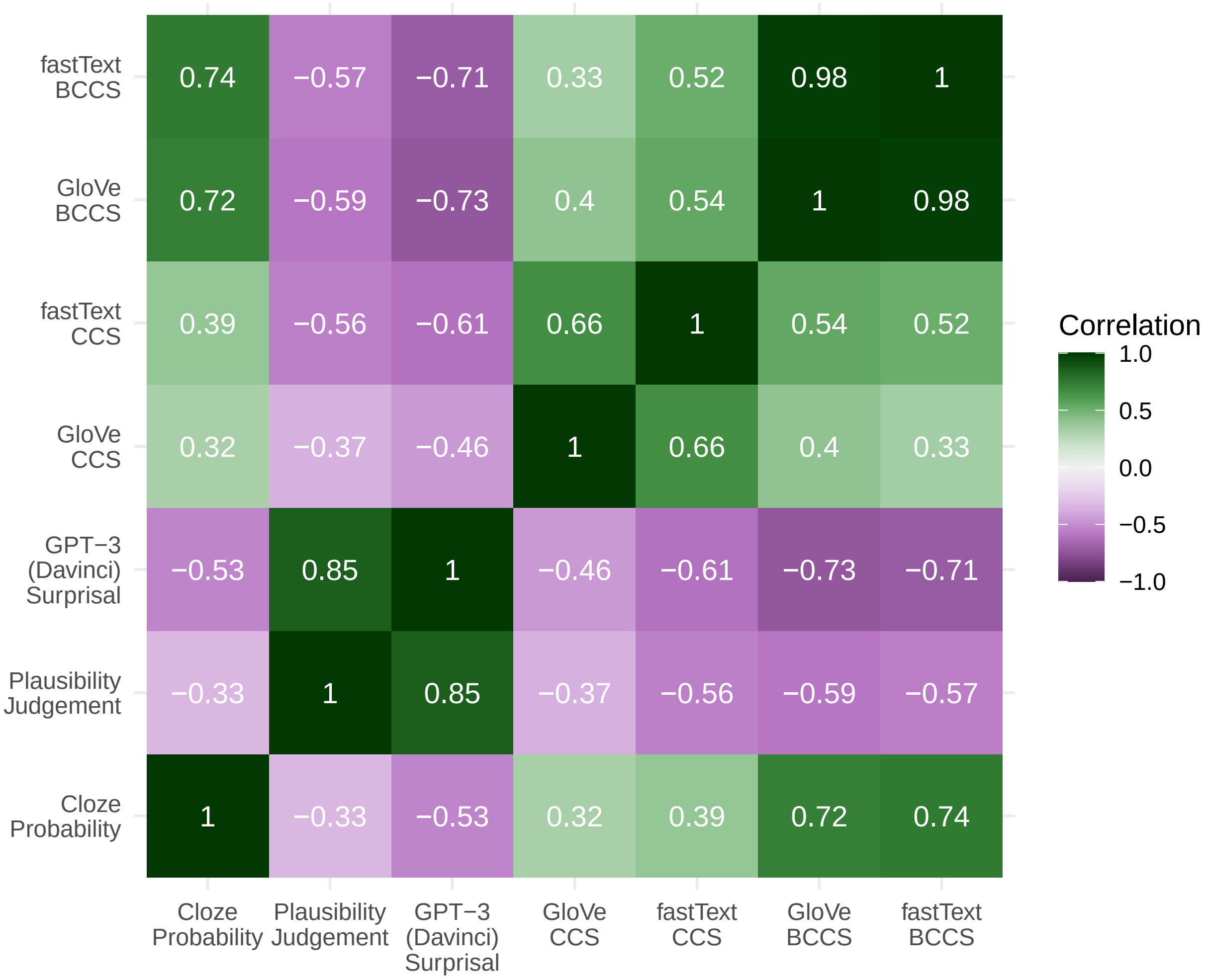


Figure 3

Heatmap of correlations between predictors



	Cloze Probability	Plausibility Judgement	GPT−3 (Davinci) Surprisal	GloVe CCS	fastText CCS	GloVe BCCS	fastText BCCS
fastText BCCS	0.74	−0.57	−0.71	0.33	0.52	0.98	1
GloVe BCCS	0.72	−0.59	−0.73	0.4	0.54	1	0.98
fastText CCS	0.39	−0.56	−0.61	0.66	1	0.54	0.52
GloVe CCS	0.32	−0.37	−0.46	1	0.66	0.4	0.33
GPT−3 (Davinci) Surprisal	−0.53	0.85	1	−0.46	−0.61	−0.73	−0.71
Plausibility Judgement	−0.33	1	0.85	−0.37	−0.56	−0.59	−0.57
Cloze Probability	1	−0.33	−0.53	0.32	0.39	0.72	0.74

Single Factor Accounts

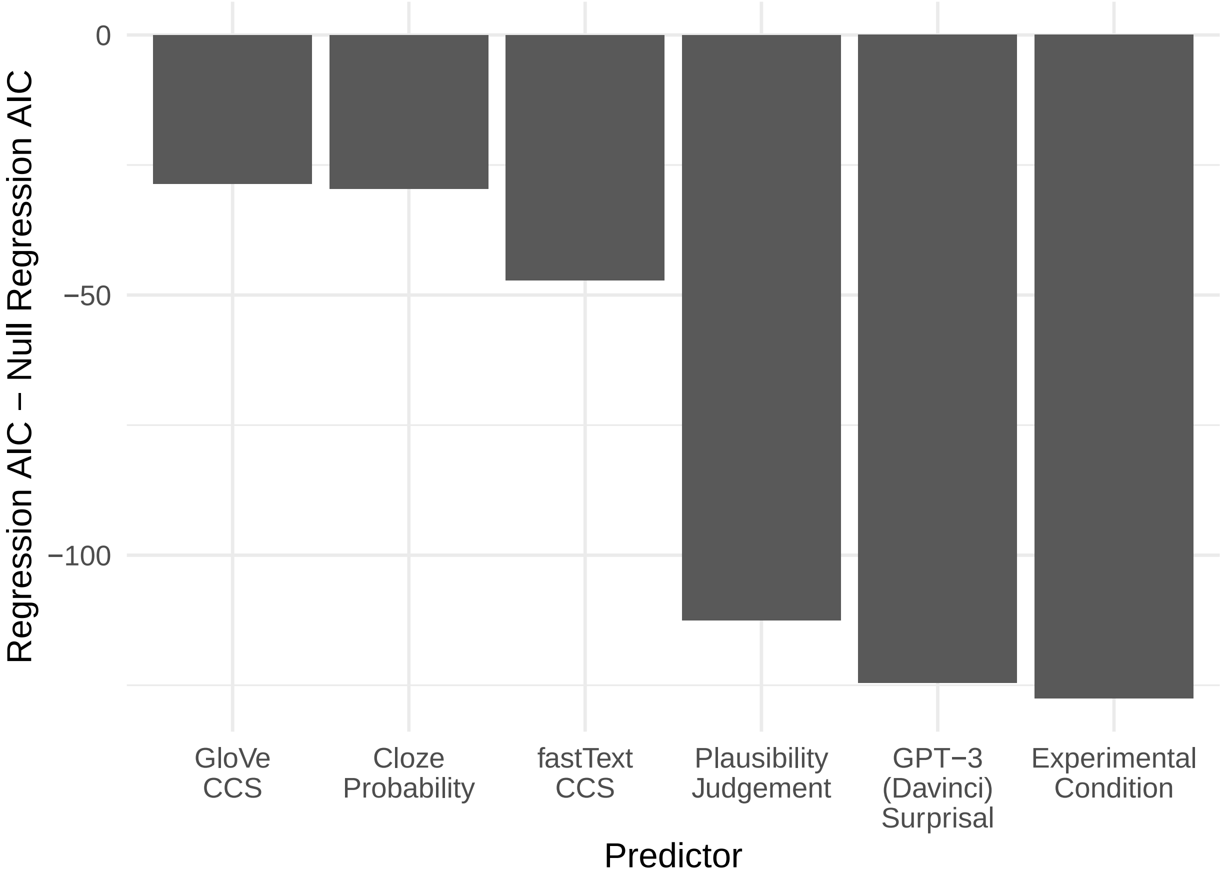
To begin our investigation, we evaluate how well each metric predicts N400 amplitude, allowing us to both validate our statistically-derived metrics (surprisal and cosine similarity) against the more traditional human-derived metrics (cloze probability and plausibility judgements), and to directly compare the former in their ability to predict N400 amplitude.

In order to compare these predictors, we constructed linear mixed-effects regression models with with each variable of interest as a fixed effect and used Akaike's Information Criterion (AIC; Akaike, 1973) to compare the regressions' fits of the neural data. Each regression had a fixed effect of either cloze probability, plausibility judgement, GloVe Contextual Cosine Similarity, fastText Contextual Cosine Similarity, GPT-3 surprisal, and experimental condition. Note that we use cloze probability rather than cloze surprisal (i.e., log-transformed cloze probability) because previous work has not shown any clear evidence that the latter is a better predictor of N400 amplitude (see Michaelov et al., 2022; Szewczyk and Federmeier, 2022). In addition, one experimental condition (Implausible) was entirely made up of stimuli where critical words had a cloze probability of zero, which cannot be log-transformed; and 'smoothing' such zero values to allow log-transformation by assigning them a very low probability also introduces problems for analysis (Nieuwland et al., 2018).

Because the inclusion of random slopes often leads to problems with convergence and singular fits, we chose to utilize a parsimonious random effects structure (Bates et al., 2018) in our regressions. Consequently, model comparison always involves regression models with the same random effects structures, which allows for comparison across models with different predictors. All regressions had random intercepts of sentence frame, subject, and electrode, as well as fixed effects of word frequency (calculated using the *wordfreq* Python package; Speer et al., 2018) and orthographic neighborhood size as operationalized by Coltheart's N (Coltheart et al., 1977; calculated using MCWord; Medler and Binder, 2005). We also included a random intercept for each critical word because critical words often occurred in more than one condition.

Figure 4

The AICs of the regressions resulting from the single factor analyses. CCS refers to Contextual Cosine Similarity.



The AIC of each regression, normalized by the AIC of the null regression (which includes the same random effects structure as the other regressions, and only word frequency and orthographic neighborhood size as fixed effects) is presented in Figure 4.

Of the continuous predictors, Figure 4 indicates that the best-fitting regression is that including GPT-3 surprisal as a main effect, suggesting GPT-3 surprisal is the best predictor of N400 amplitude. GPT-3 surprisal is followed by human plausibility judgements, which are followed by fastText CCS, which in turn is followed by cloze probability and GloVe CCS. It is generally accepted that a difference in AIC of 4 indicates a substantial difference (Burnham & Anderson, 2004), and the difference between cloze probability and GloVe CCS is only 0.9; thus it is not clear from our analysis which is the better predictor.

Figure 4 also indicates that the regression including experimental condition (a categorical variable with four levels: Best Completion, Related, Unrelated, and Implausible) has a lower AIC than the GPT-3 surprisal regression. However, experimental condition should not be considered to reflect a single variable in the way that the other individual predictors do because it includes information about predictability, plausibility, and relatedness to the best completion. Additionally, the experimental condition regression has an AIC of only 3 less than the GPT-3 surprisal regression; thus it is not clear that experimental condition is in fact a better predictor than GPT-3 surprisal.

We also ran likelihood ratio tests on each of the predictors listed in Figure 4, comparing each regression to a null regression, i.e., one without the predictor of interest but all other fixed and random effects. All variables were significant predictors of N400 amplitude (GloVe CCS: $\chi^2(1) = 30.6, p < 0.001$; Cloze: $\chi^2(1) = 31.6, p < 0.001$; fastText CCS: $\chi^2(1) = 49.1, p < 0.001$; Plausibility: $\chi^2(1) = 114.5, p < 0.001$; GPT-3 Surprisal: $\chi^2(1) = 126.6, p < 0.001$; Condition: $\chi^2(3) = 133.6, p < 0.001$).

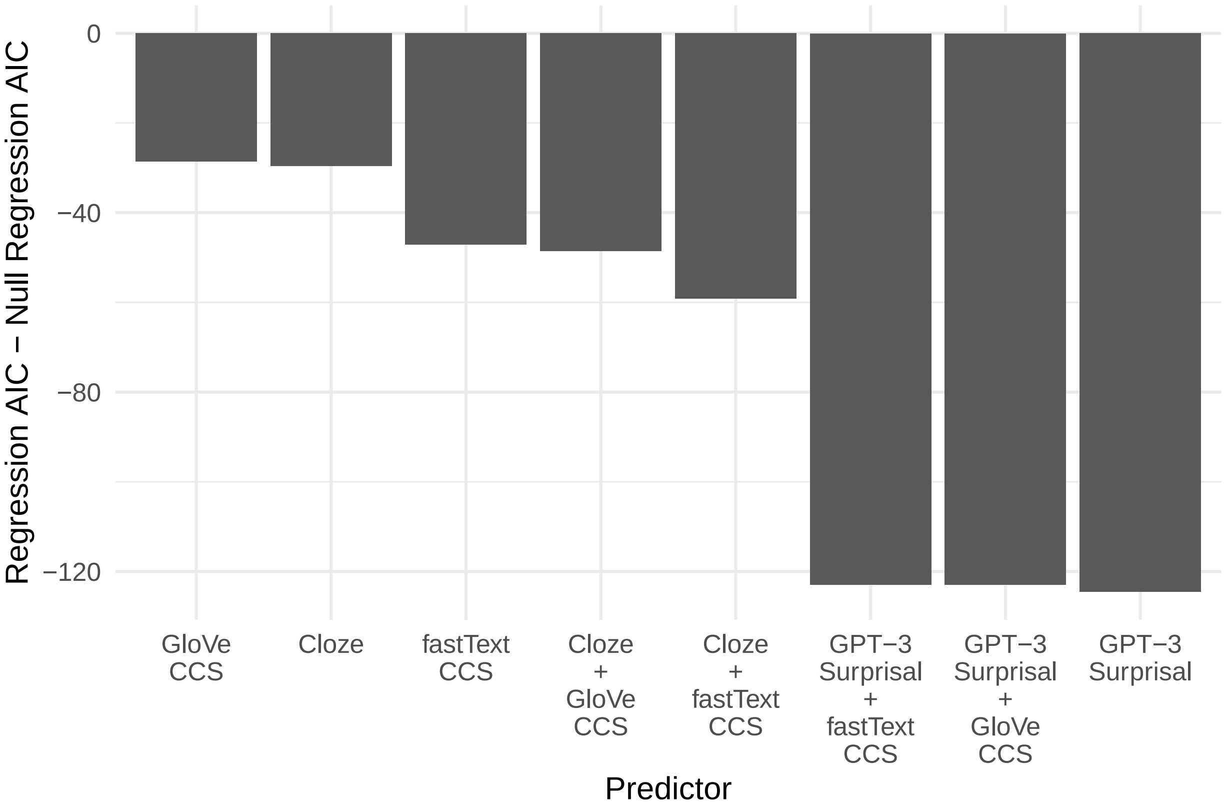
Combined Accounts

The GPT-3 surprisal metric was chosen to model a prediction-based account of the N400, and GloVe and fastText contextual cosine similarity (CCS) were chosen to model the contextual semantic similarity accounts. As noted above, some authors have suggested the N400 indexes neurocognitive systems sensitive both to the predictability of a word and to its similarity to the semantic context. To investigate the viability of such combined accounts, we compare the AICs of regressions including a single variable corresponding to either prediction or contextual semantic similarity, with the AICs of regressions also including one of the other. Thus, we look at all combinations of prediction (viz., Cloze Probability and GPT-3 Surprisal) with CCS metrics. The results are presented in Figure 5.

A comparison of the AICs suggests that cloze probability and the two CCS metrics explain variance in N400 amplitude not explained by the other. This is borne out by the likelihood ratio tests: after correcting for multiple comparisons the cloze probability regression is improved by adding either GloVe ($\chi^2(1) = 21.0, p < 0.001$) or fastText CCS ($\chi^2(1) = 31.6, p < 0.001$) as a predictor; and conversely, the GloVe ($\chi^2(1) = 22.0, p < 0.001$) and fastText ($\chi^2(1) = 14.0, p < 0.001$) regressions are each improved by adding cloze probability as a predictor. This suggests cloze probability and the CCS metrics explain non-overlapping portions of the variance in N400 amplitude. However, the same is not true of GPT-3 surprisal—while adding GPT-3 surprisal improves both the GloVe ($\chi^2(1) = 96.3, p < 0.001$) and fastText ($\chi^2(1) = 77.8, p < 0.001$) CCS regressions, the GPT-3 surprisal regression is not improved by adding either GloVe ($\chi^2(1) = 0.4, p = 1.000$) or fastText CCS ($\chi^2(1) = 0.4, p = 1.000$). Thus GPT-3 explains variance left unexplained by the CCS measures, while the information provided by CCS was largely redundant with that provided by GPT-3.

Figure 5

The AICs of the regressions resulting from the two-variable analyses corresponding to combined accounts. CCS refers to Contextual Cosine Similarity.



The plausibility effect

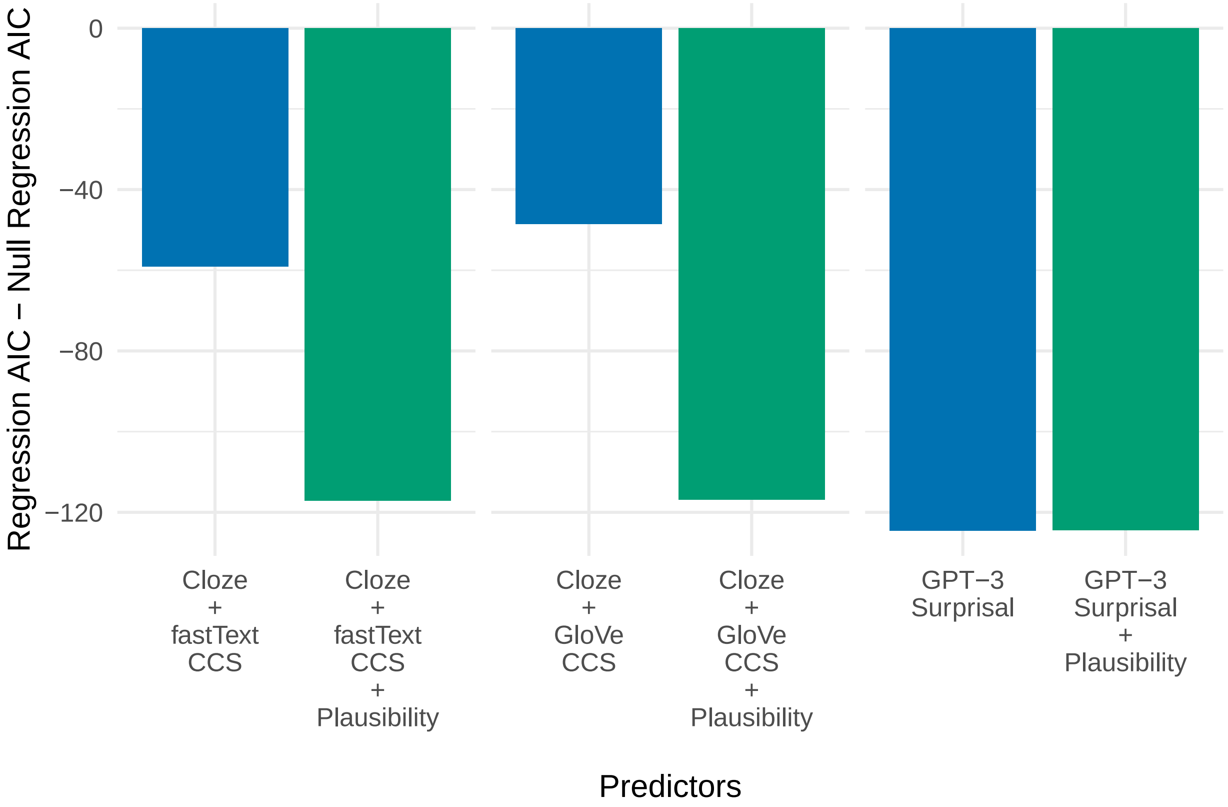
To test how well our metrics explain the variance in N400 amplitude traditionally explained by plausibility judgements, here we investigate whether the addition of plausibility as a predictor improves the GPT-3 surprisal regression, the cloze + GloVe CCS regression, and the cloze + fastText CCS regression. These regressions were selected because they were the models including each of our original three statistically-derived metrics (that is, for predictability and for contextual semantic similarity) that performed the best in accounting for observed variance in N400 amplitude. Of these, we can consider the GPT-3 surprisal regressions as relevant to the predictive preactivation account of the

N400 and the cloze + CCS regressions as relevant to multiple systems accounts.

Shown in Figure 6, the results indicate that even when combined with cloze probability (and thus, when part of a combined model that takes into account predictability as well as contextual similarity), the AICs of the regressions including GloVe ($\chi^2(1) = 70.3, p < 0.001$) and fastText ($\chi^2(1) = 60.0, p < 0.001$) CCS are improved by the addition of plausibility as a predictor. By contrast, the GPT-3 surprisal regression is not improved by adding plausibility as a predictor ($\chi^2(1) = 1.9, p = 0.715$). Whereas neither CCS metric can model the N400 plausibility effect—even when combined with cloze—variance attributable to plausibility was captured by GPT-3 surprisal. Thus, predictability alone (operationalized by GPT-3 surprisal) can explain the apparent effect of plausibility on N400 amplitude.

Figure 6

The AICs of the regressions resulting from the analyses investigating whether the single-factor and combined models account for the effect of plausiblity. CCS refers to Contextual Cosine Similarity.



The relatedness to the best completion effect

Finally, we explore the extent to which relatedness to the best completion is captured by our three metrics. As with plausibility, we look at whether adding a metric of relatedness to the best completion improves regression fit, where we operationalize relatedness to the best completion as the cosine distance between the word embeddings of the best completion for each sentence frame and the critical word in each of the other conditions, a metric we name best completion cosine similarity (BCCS). We used both GloVe and fastText to derive measures of BCCS.

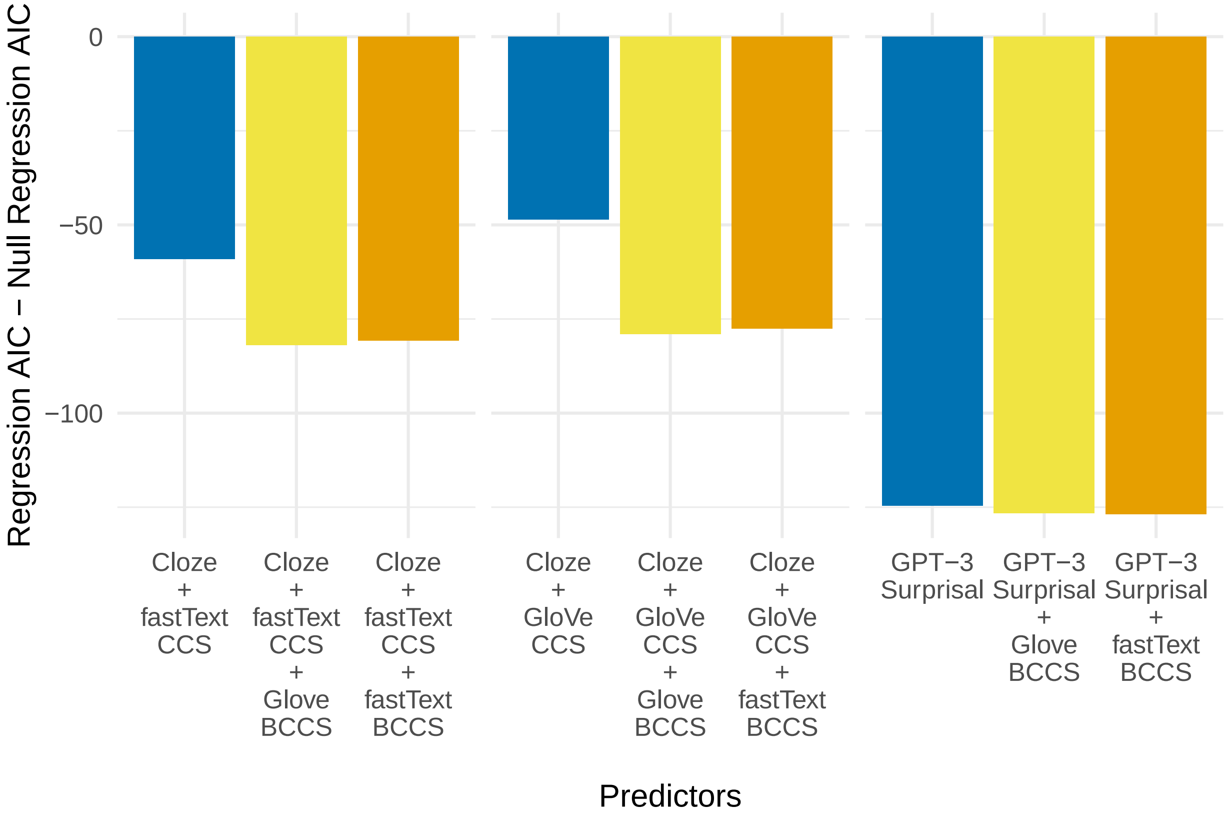
As with plausibility, we investigate whether our previous best regressions for each of our three statistical metrics—that is, GPT-3 surprisal, cloze + GloVe CCS, and cloze + fastText CCS—are improved by the addition of BCCS to the model. The results are shown in Figure 7. The addition of GloVe BCCS to either cloze + CCS regressions led to improvements in model performance (Cloze + GloVe CCS: $\chi^2(1) = 32.4, p < 0.001$; Cloze + fastText CCS: $\chi^2(1) = 24.8, p < 0.001$); likewise the addition of fastText BCCS to either cloze + CCS regression led to significant improvements (Cloze + GloVe CCS: $\chi^2(1) = 31.0, p < 0.001$; Cloze + fastText CCS: $\chi^2(1) = 23.6, p < 0.001$). These results show that even when combined with cloze, contextual similarity cannot explain the relatedness to best completion effect.

On the other hand, adding GloVe BCCS to the GPT-3 surprisal regression only reduces the AIC by 2, and adding fastText BCCS only reduces the AIC by 2.3; far from a clear improvement. When we run likelihood ratio tests, neither is found significantly improve regression fit after controlling for multiple comparisons (GloVe BCCS: $\chi^2(1) = 4.0, p = 0.192$; fastText BCCS: $\chi^2(1) = 4.3, p = 0.175$). However, unlike all our other tests, this result is dependent on controlling for multiple comparisons—before this step, both BCCS metrics do appear to have a significant effect (GloVe BCCS: $p = 0.044$; fastText BCCS: $p = 0.039$). Thus, both when comparing AICs and testing using likelihood ratio tests, while BCCS metrics may appear to improve model fit, they do not do so reliably.

One possible concern is that the extent to which the BCCS metrics predict N400 amplitude above and beyond surprisal may be undermined by the fact that for one condition (Best Completion), all BCCS values are, by definition, 1, as the critical word *is* the best completion. For this reason we also ran the same analysis excluding all data for Best Completions. The results were qualitatively the same: after correction for multiple comparisons, neither GloVe BCCS ($\chi^2(1) = 5.0, p = 0.118$; uncorrected $p = 0.025$) nor fastText BCCS ($\chi^2(1) = 5.7, p = 0.087$; uncorrected $p = 0.017$) significantly improved the regression already including GPT-3 surprisal.

Figure 7

The AICs of the regressions resulting from the analyses investigating whether the single-factor and combined models account for the effect of the relatedness to the best completion. CCS refers to Contextual Cosine Similarity and BCCS refers to Best Completion Cosine Similarity.



Discussion

The aim of this paper was to use current state-of-the-art language models to compare the predictions of two accounts of the neural activation underlying the N400 response—predictive preactivation versus contextual semantic similarity. To do this, we investigated how well GPT-3 surprisal—our best approximation of the kinds of predictions neurocognitive systems may make based on the statistics of language—predicts N400 amplitude. We compared this with the performance of GloVe and fastText contextual cosine similarity, our two best approximations of contextual semantic similarity based on the statistics of language. Finally, we compared this with the performance of combined models including both kinds of metrics. Based on this approach, we found that the predictive preactivation account explains more variance in N400 amplitude than the two models of contextual semantic similarity.

Below we consider the adequacy of predictive preactivation, contextual semantic similarity, and combined systems to account for the three kinds of N400 effects examined in the present study: expectancy effects, plausibility effects, and relatedness to best completion (RBC). In each case, predictive preactivation provides a better account of N400 amplitude variation than does either a pure contextual similarity account or a multiple systems account. We end with a consideration of how the features of the deep learning language systems we used here relate to those of the language network in the brain.

Expectancy Effects

While the close association between measures of contextual predictability and N400 amplitude is most naturally accounted for by the predictive preactivation account, advocates of contextual semantic similarity have argued that expectancy effects on the N400 arise because highly expected words share more semantic features with their context than do less expected words. This is demonstrated in computational modeling work by Ettinger et al. (2016), who uses the similarity between word2vec (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013) representations of stimulus words and their contexts to account for the N400 amplitude differences between the best completions and their lower cloze counterparts in a widely cited study by Federmeier and Kutas (1999). Similarly, using wikipedia2vec embeddings (Yamada et al., 2020), Uchida et al. (2021) show that high cloze sentence continuations from a number of ERP language studies are more similar to their contexts than their less predictable counterparts.

In the present study, we likewise find that contextual similarity as measured both by

GloVe CCS and fastText CCS is greater for best completions than for the other less expected endings. However, in a direct comparison of how well various measures of predictability versus contextual similarity account for variance in N400 amplitude, predictability as indexed by GPT-3 surprisal was the clear winner, providing a better account of the data than either GloVe CCS or fastText CCS. Moreover, the finding that regressions using both CCS measures improved when combined with cloze probability suggests these measures of contextual semantic similarity were unable to fully capture expectancy effects on the N400.

Of course, this same finding—that regressions with CCS measures are improved by the cloze probability factor—replicates work that supports the multiple systems account of the N400 (Federmeier, 2021; Lau et al., 2013). However, GPT-3 surprisal out-performed even these regressions (see Figure 5), suggesting that the predictive preactivation account of N400 is superior to both a pure contextual semantic similarity account and to a combined systems account.

Plausibility Effects

GPT-3 surprisal also accounts for more variance in N400 amplitude than our human-derived measure of cloze probability (in line with Michaelov et al., 2022), presumably due to its ability to capture subtle differences between highly unexpected items. Indeed, as Nieuwland et al. (2020) note, plausibility effects on the N400 might result because less plausible stimuli are also less predictable. Because cloze probability measures are limited in the extent to which they can adequately capture the predictability of highly improbable words, plausibility ratings may serve as a proxy for their predictability, allowing us to differentiate *very* low-probability completions from *extremely* low-probability ones. Of course, plausibility effects can also be accounted for in principle via contextual semantic similarity, since we would expect less plausible stimuli to be less related to their context.

Results of the present study, however, argue against the latter possibility as we find that even when combined with cloze probability, regressions including measures of contextual semantic similarity could not fully account for the plausibility effect. This finding serves as a conceptual replication of Nieuwland et al. (2020) who found that plausibility explains amplitude variance in the N400 not explained by either cloze probability or a contextual similarity metric derived from word2vec. However, unlike Nieuwland et al. (2020), we find that one metric of predictability—namely, GPT-3 surprisal—can successfully model the plausibility effect. In fact, it explains all the variance that plausibility judgements do. Thus, in contrast to the findings of Nieuwland et al. (2020), the results of the present study suggest that a single neurocognitive process—predictive preactivation—may be able to account for both predictability and plausibility effects on the N400. Whether this also applies to analyses across individual time-steps within the N400 time window (of the kind carried out by Nieuwland et al., 2020) is a question for further research.

Relatedness to Best Completion

As described in the Introduction, the RBC effect is not trivially explained by either predictability or contextual similarity; however, in principle it can be accommodated by either account, and there is some evidence for each. Under a predictability perspective, if semantic prediction is taking place, then we should expect words with a similar meaning to the best completion to be preactivated along with the best completion (DeLong et al., 2019). Consistent with this account, the predictions of computational language models have been used to successfully model the RBC effect (Michaelov & Bergen, 2020). Specifically, Michaelov and Bergen (2020) report that two language models (Gulordava et al., 2018; Jozefowicz et al., 2016) find related words to be more predictable than unrelated overall when modeling the stimuli from an experiment by Ito et al. (2016), and that one of these language models also shows the same pattern for stimuli from Kutas (1993).

According to the contextual semantic similarity account, the RBC effect results because words related to the best completion share semantic features with it. Thus, related words elicit reduced N400 for much the same reason the best completions do — their features have been preactivated because they are semantically related to the sentence context. This has also been successfully modeled computationally: Ettinger et al. (2016) finds that the similarity between the *word2vec* (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013) representation of a stimulus word and its preceding context demonstrates the RBC effect found by Federmeier and Kutas (1999)—words related to best completions were more semantically similar to the preceding context than were unrelated words.

The present study provides a conceptual replication of results reported both by Michaelov and Bergen (2020) and by Ettinger et al. (2016). Using GPT-3 surprisal we find that our Related completions were more predictable than the Unrelated ones (in line with Michaelov & Bergen, 2020); using fastText CCS we find that Related completions were more similar to the preceding context than were the Unrelated ones (in line with Ettinger et al., 2016). However, results in Figure 2—like those in both Michaelov and Bergen (2020) and Ettinger et al. (2016)—only demonstrate that overall, there is a significant difference in the predictability and in the contextual semantic similarity of Related and Unrelated completions as estimated by these computational language models; there is no direct comparison with human data.

The strength of the present study lies in our efforts to do just this. Direct comparison with the human N400 data suggests that the predictability metric from GPT-3 explains more variance in N400 amplitude than does either metric of semantic similarity to the context. Moreover, in our attempts to probe how well each metric captures the RBC effect, we utilized two computational measures of the semantic similarity between each best completion and the other three completions for the sentence frame: GloVe BCCS and fastText BCCS. As both the graphs in Figure 2 and the high correlation coefficient in Figure 3 suggest, the two BCCS measures were virtually identical with each other and both captured the human intuition that Related words were closer in meaning to the Best Completions than Unrelated words.

Regression models of N400 data indicate that the addition of either GloVe or fastText BCCS metrics to models already including cloze probability and GloVe or fastText CCS improves model fit (see Figure 7). This suggests that neither of our contextual semantic similarity metrics could fully account for the RBC effect—even when combined with cloze probability. On the other hand, the GPT-3 surprisal regression of N400 data was not substantially improved by the addition of either BCCS metric (see Figure 7), suggesting the variance associated with our measure of similarity to the best completion was largely redundant with that captured by GPT-3 surprisal. Moreover, the regression model including only GPT-3 surprisal out-performed all of the regression models with additive combinations of CCS, cloze probability, and BCCS. GPT-3 surprisal provides a better account of the RBC effect than does either a pure contextual semantic similarity account or a combination of prediction and contextual similarity.

While the superiority of GPT-3 over the contextual similarity measures is unambiguous, there is a bit of uncertainty regarding whether GPT-3 is improved by the addition of the BCCS metrics. In our statistical model comparisons, we do not consider regressions with a difference in AIC of less than 4 to differ meaningfully in their fit (following Burnham and Anderson, 2004). However, it is the case that numerically, the regressions including both GPT-3 surprisal and either GloVe or fastText BCCS have a lower AIC than that *only* including surprisal. Unfortunately, the outcome of the relevant likelihood ratio tests was also somewhat equivocal on this matter. After correcting for multiple comparisons, neither GloVe nor fastText BCCS explain a significant amount of the variance in N400 amplitude above and beyond what is explained by GPT-3 surprisal. Before correction, however, those comparisons were both significant at the 0.05 level. It is thus important to consider what might explain this (marginally) better fit to the data.

One straightforward explanation can be arrived at by further inspection of Figure 2. As can be seen, GPT-3 surprisal provides a good account of the difference in the expectancy between Best Completions and the Unrelated condition, and a good account of the difference between the Unrelated and the Implausible condition—impressions borne out by the analyses comparing surprisal to human-derived metrics of cloze probability and plausibility. The disconnect between GPT-3 surprisal and N400 data lies mainly in failing to fully capture the similarity in N400 amplitude between the Best Completions and the Related condition, as the latter elicit more positive N400 in humans than the GPT-3 regression model fits suggest. Thus, the addition of another variable that captures the difference between Related and Unrelated completions—variance not present in cloze probability or plausibility, and unreliable in the CCS metrics—may explain the improved fit with the addition of BCCS metrics. This may also explain the slightly lower AIC of the regression including the categorical variable of experimental condition in Figure 4.

Crucially, however, even if GPT-3 does not fully account for the RBC effect, the RBC effect observed here supports predictive preactivation as at least a partial account of the brain activity underlying the N400. If words semantically related to the best completion are facilitated in virtue of being related to the best completion, this presupposes the preactivation of information related to the best completion (DeLong and Kutas, 2020 see also Kuperberg et al., 2020). For example, it may be the case that the reason for the greater facilitation for related than unrelated words is that predictive processing involves the preactivation of conceptual semantic features rather than lexical items (Thornhill & Van Petten, 2012). Alternatively, it may be that there is a separate associative mechanism that activates words related to the best completion. In the first case, the preactivation of the related word occurs as part of a single predictive process; in the second, as a consequence. Both possibilities require the preactivation of the best completion—either the lexical item itself or its semantic features. Regardless, the present study clearly shows that, as operationalized here, predictive preactivation provides a better account of the RBC N400 effect than does contextual semantic similarity (see Figure 7).

Overall, in addition to being the best metric of predictability tested (in line with the results of Michaelov et al., 2022), GPT-3 surprisal also appears to successfully account for additional reported N400 effects, namely, that more plausible completions elicit smaller N400 responses than less plausible completions, and that words that are semantically related to the best (highest-cloze) completion elicit smaller N400 responses than unrelated words. In sum, with a good enough operationalization of contextual predictability, we can reduce all effects observed during the temporal interval associated with the N400 to this single factor. The most parsimonious interpretation is that apparent effects of expectancy, plausibility, and RBC all index sensitivity to contextual predictability—and predictability derived from the statistics of language at that—suggesting N400 effects are due to a predictive preactivation process.

Implications for Neural Mechanisms

Although we do not here treat any of the computational models used in this study as cognitive models, it is important to consider what the differences in the way that they work imply about that language network in the human brain. GPT-3 is a neural language model trained to optimize its estimates of the probability of upcoming words and how these values change with different amounts of linguistic context. Moreover, GPT-3 surprisal was the single best numerical predictor of N400 amplitude. On the other hand, GloVe and fastText, which model the relations between words, performed worse overall at predicting N400 amplitude. In this way, our results are highly compatible with predictive coding theories that suggest neural systems are constantly generating and updating an internal model of the environment (Allen & Tsakiris, 2018; Bendixen et al., 2012; Clark, 2013; Friston, 2010; Huang & Rao, 2011; McRae et al., 2019; Rao & Ballard, 1999; Shipp et al., 2013).

Applied to language, such approaches typically take the form of neural systems that generate predictions regarding upcoming words, using the word encountered at the next time step to generate a learning signal known as a prediction error (e.g., Elman, 1990). Indeed, something that we believe has been under-appreciated in this regard is that the loss function used to train language models such as GPT-3, cross-entropy, is equivalent to surprisal (see Jurafsky and Martin, 2021, pp. 149-150). The close relationship we observed here between GPT-3 surprisal and N400 amplitude is perfectly in line with the suggestion that the N400 reflects a prediction-error based update of an internal language model (Bornkessel-Schlesewsky and Schlesewsky, 2019; Fitz and Chang, 2019; Hodapp and Rabovsky, 2021; Kuperberg, 2021; Kuperberg et al., 2020; Lewis and Bastiaansen, 2015; Rabovsky, 2020).

As Kuperberg et al. (2020) note, this account does not fit neatly into either retrieval (e.g. Brouwer et al., 2017; Brouwer & Hoeks, 2013; Kutas & Federmeier, 2000; Kutas et al., 2006; Lau et al., 2008; Van Berkum, 2009, 2010) or integration (e.g. Hagoort et al., 2009; van den Brink & Hagoort, 2004) accounts of the N400. Under our predictive coding account of the N400, the N400 is a measure of the neural activation elicited by a stimulus that was not already activated by prediction based on the preceding context. In this way, it indexes retrieval difficulty—the effort required to fully activate the neural representations needed to process the stimulus, which is reduced if some of these representations are already activated. By contrast, N400 amplitude could also be considered to index integration in that words that are easier to integrate with the preceding context are likely to be more strongly predicted (see, e.g., Kuperberg et al., 2020; Kuperberg & Jaeger, 2016). However, this only encompasses a limited subset of what could be considered integration difficulty—words that are highly anomalous, violate thematic roles, or lead to a substantial shift in the meaning of the preceding context instead appear to elicit later positivities (Coulson & Lovett, 2004; DeLong & Kutas, 2020; Kuperberg et al., 2020).

Our results are compatible in principle with a two-system account involving both contextual semantic similarity and predictive preactivation (as in Federmeier, 2021; Frank and Willems, 2017; Lau et al., 2013). However, given that the former does not explain any additional variance in the neural data, a predictive-preactivation-only account is more parsimonious. Further, in view of the correlation between GPT-3 surprisal and the CCS metrics (GloVe: $r = -0.46$; fastText: $r = -0.61$), it is possible that N400 effects previously explained as resulting from contextual semantic similarity may be an artifact of its correlation with the contextual predictability of words. Indeed, direct evidence of a neurocognitive process implementing contextual semantic similarity-based activation would require demonstrating an effect of contextual semantic similarity that cannot be linked to its contextual predictability.

One possible candidate for an effect that would help to test this is the finding that in some contexts, highly anomalous words that violate thematic roles (A. Kim & Osterhout, 2005; Kuperberg et al., 2003; Nieuwland & Van Berkum, 2005) or temporal event structure (Delogu et al., 2019) do not elicit a larger N400 response than non-violating stimuli. For example, Kuperberg et al. (2003) find no significant difference in N400 amplitude between *For breakfast the eggs would only* ***eat*** and *For breakfast the boys would only* ***eat***, and Delogu et al. (2019) do not find a significant difference between *John entered the restaurant. Before long, he opened the* ***menu*** and *John left the restaurant. Before long, he opened the* ***menu***. In both cases, the critical word's relation to the preceding context appears to nullify the increase in N400 amplitude one might expect from the degree of semantic anomaly. To the best of our knowledge, only one study (Michaelov & Bergen, 2020) has attempted to model this effect using the stimuli from A. Kim and Osterhout (2005), finding that the surprisal elicited by stimuli such as *The hearty meal was devouring* is significantly higher than that elicited by either *The hearty meal was* ***devoured*** or *The hungry boy was* ***devouring***, which differs from N400 amplitude where the three were not significantly different. This would indeed suggest that predictability, and thus prediction, cannot fully account for the N400 effect. However, it is important to note that this study used recurrent neural networks, whose predictions have been found to correlate far less with N400 amplitude than contemporary transformer language models (Merkx & Frank, 2021; Michaelov et al., 2022). Thus, whether this effect can be accounted for by contextual predictability alone is still an open question, and we believe a fruitful avenue for future research.

The results of using a language model to model the study carried out by A. Kim and Osterhout (2005) may also be valuable in better understanding the content of the preactivation underlying the N400 response. For example, a number of accounts argue that the preactivation underlying the N400 response is at the level of the semantic features of words (Federmeier, 2021; Kuperberg et al., 2020). While there is evidence that N400 amplitude is sensitive to phonological and grammatical features (DeLong et al., 2005; Fleur et al., 2020; Nicenboim et al., 2020; Otten et al., 2007; Urbach et al., 2020; Van Berkum et al., 2005), it may be that the shared semantic features between, for example, *devouring* and *devoured*, are sufficient to preactivate both words equally. Thus a semantically-augmented language model may be able to better model the effect.

Alternatively, or in addition, it may be that the preactivation underlying the N400 operates at the morphemic level either in general (as proposed by Smith and Levy, 2013), or in cases where the redundant derived forms of words are not stored (for discussion, see Hanna and Pulvermüller, 2014). It may be that it is *devour* that is activated, and any additional activation conferred by *-ing* or *-er* suffixes is is so subtle as to be undetectable in the scalp-recorded N400. This suggestion is in line with the finding that N400 amplitude is most sensitive to the predictability of content words (Frank et al., 2015). This could be investigated by testing language models with different tokenization schemes, for example, those where tokenization schemes are implemented that make tokens correspond more closely to morphemes (for discusion and attempts, see Bostrom and Durrett, 2020; Hofmann et al., 2021; Klein and Tsarfaty, 2020; Mohebbi et al., 2021; Yehezkel and Pinter, 2023).

Finally, it may be the case that surprisal measures derived from language models relate to aspects of the brain response to words in sentences beyond the N400. For example, predictions of the recurrent neural networks tested by Michaelov and Bergen (2020) were better correlated with post-N400 positivities than the N400. The adequacy of different neural language models in fitting various aspects of the ERP waveform (such as those discussed in DeLong and Kutas, 2020; Kuperberg et al., 2020) is thus a promising area of further research, and may help to shed light on language processing in the human brain.

A further intriguing question is the role played by the statistics of language.GPT-3 is trained using only linguistic data, meaning its predictions are solely based on the statistical patterns available in their language input. By contrast, under the majority of contemporary accounts of the N400, world experience plays a key role in shaping the semantic representations that are activated during language comprehension (e.g., Amsel et al., 2015; Chwilla and Kolk, 2005; Federmeier, 2021; Hagoort et al., 2004; Kutas and Federmeier, 2011; Metusalem et al., 2012; Paczynski and Kuperberg, 2012). For this reason, it may be surprising that a model deriving its semantics solely from language is able to predict words in a way that so closely appears to match the activation of words in humans. One possible conclusion to draw from this is that humans, too, base their linguistic predictions on the statistics of language.

While there is evidence that both humans (Bedny et al., 2019; J. S. Kim et al., 2021; Marmor, 1978; Saysani et al., 2018) and language models (Abdou et al., 2021; Li et al., 2021; Piantadosi & Hill, 2022) can learn a wide range of semantic information based on language input alone, language models have also been found to have limitations. Specifically, language models trained only on language data struggle to learn perceptual properties of entities (Forbes et al., 2019) and are limited in the kinds of novel affordances they can infer for objects (Jones et al., 2022). By contrast, N400 amplitude is sensitive to people's understanding of the sensorimotor properties of the referents of words (Amsel et al., 2015; Amsel et al., 2013, 2014; Wu & Coulson, 2011). Perhaps most importantly, language alone drives the probability estimates of GPT-3, whereas the N400 is sensitive to the contextual congruity of faces, gestures, images, environmental sounds, and action sequences (see Kutas and Federmeier (2011) for review). Further work is needed to determine how other, non-linguistic sources of information influence the N400 response.

Data and Code Availability Statements

The data, code, and analysis scripts used for the present study are available at https://osf.io/pysbc.

Acknowledgements

For their helpful comments and valuable discussion, we would like to thank the anonymous reviewers, the attendees of the 26th Architectures and Mechanisms for Language Processing Conference and the 43rd Annual Meeting of the Cognitive Science Society, as well as the Language and Cognition Laboratory, Brain and Cognition Laboratory, and Kutas Cognitive Electrophysiology Laboratory members and lab meeting attendees. We would also like to thank the San Diego Social Sciences Computing Facility Team for their technical assistance. The RTX A6000 used for this research was donated by the NVIDIA Corporation.

References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., & Søgaard, A. (2021). Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 109–132. https://doi.org/10.18653/v1/2021.conll-1.9

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petrov & F. Csáki (Eds.), *Second international symposium on information theory* (pp. 267–281). Akadémiai Kiadó. https://doi.org/10.1007/978-1-4612-1694-0_15

Allen, M., & Tsakiris, M. (2018). *The body as first prior: Interoceptive predictive processing and the primacy of self-models*. Oxford University Press. Retrieved June 4, 2021, from https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198811930.001.0001/oso-9780198811930-chapter-2

Amsel, B. D., DeLong, K. A., & Kutas, M. (2015). Close, but no garlic: Perceptuomotor and event knowledge activation during language comprehension. *Journal of Memory and Language*, *82*, 118–132. https://doi.org/10.1016/j.jml.2015.03.009

Amsel, B. D., Urbach, T. P., & Kutas, M. (2013). Alive and grasping: Stable and rapid semantic access to an object category but not object graspability. *NeuroImage*, *77*, 1–13. https://doi.org/10.1016/j.neuroimage.2013.03.058

Amsel, B. D., Urbach, T. P., & Kutas, M. (2014). Empirically grounding grounded cognition: The case of color. *NeuroImage*, *99*, 149–157. https://doi.org/10.1016/j.neuroimage.2014.05.025

Anderson, J. E., & Holcomb, P. J. (1995). Auditory and visual semantic priming using different stimulus onset asynchronies: An event-related brain potential study. *Psychophysiology*, *32*(2), 177–190. https://doi.org/10.1111/j.1469-8986.1995.tb03310.x

Auguie, B. (2017). *gridExtra: Miscellaneous functions for "Grid" graphics*. Manual. https://CRAN.R-project.org/package=gridExtra

Aurnhammer, C., & Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, *134*, 107198. https://doi.org/10.1016/j.neuropsychologia.2019.107198

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). Parsimonious Mixed Models. *arXiv:1506.04967 [stat]*. Retrieved March 30, 2022, from http://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bedny, M., Koster-Hale, J., Elli, G., Yazzolino, L., & Saxe, R. (2019). There's more to "sparkle" than meets the eye: Knowledge of vision and light verbs among congenitally blind and sighted individuals. *Cognition*, *189*, 105–115. https://doi.org/10.1016/j.cognition.2019.03.017

Bendixen, A., SanMiguel, I., & Schröger, E. (2012). Early electrophysiological indicators for predictive processing in audition: A review. *International Journal of Psychophysiology*, *83*(2), 120–131. https://doi.org/10.1016/j.ijpsycho.2011.08.003

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, *29*(4), 1165–1188. Retrieved May 4, 2021, from https://www.jstor.org/stable/2674075

Berger, J., & Packard, G. (2022). Using natural language processing to understand people and culture. *American Psychologist*, *77*, 525–537. https://doi.org/10.1037/amp0000882

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arXiv:1607.04606 [cs]*. https://doi.org/10.48550/arXiv.1607.04606

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2019). Toward a Neurobiologically Plausible Model of Language-Related, Negative Event-Related Potentials. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00298

Bostrom, K., & Durrett, G. (2020). Byte Pair Encoding is Suboptimal for Language Model Pretraining. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4617–4624. https://doi.org/10.18653/v1/2020.findings-emnlp.414

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Current Biology*, *28*(5), 803–809.e3. https://doi.org/10.1016/j.cub.2018.01.080

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, *116*, 104174. https://doi.org/10.1016/j.jml.2020.104174

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive Science*, *41*(S6), 1318–1352. https://doi.org/10.1111/cogs.12461

Brouwer, H., & Hoeks, J. C. J. (2013). A time and place for language comprehension: Mapping the N400 and the P600 to a minimal cortical network. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00758

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. Retrieved August 16, 2021, from https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Camblin, C. C., Gordon, P. C., & Swaab, T. Y. (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language*, *56*(1), 103–128. https://doi.org/10.1016/j.jml.2006.07.005

Chang, W. (2022). Colors (ggplot2). Retrieved October 3, 2022, from http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/

Chwilla, D. J., & Kolk, H. H. J. (2005). Accessing world knowledge: Evidence from N400 and reaction time priming. *Cognitive Brain Research*, *25*(3), 589–606. https://doi.org/10.1016/j.cogbrainres.2005.08.011

Chwilla, D. J., Kolk, H. H. J., & Vissers, C. T. W. M. (2007). Immediate integration of novel meanings: N400 support for an embodied view of language comprehension. *Brain Research*, *1183*, 109–123. https://doi.org/10.1016/j.brainres.2007.09.014

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the Internal Lexicon. In *Attention and Performance VI*. Routledge.

Coulson, S., Federmeier, K. D., Van Petten, C., & Kutas, M. (2005). Right Hemisphere Sensitivity to Word- and Sentence-Level Context: Evidence From Event-Related Brain Potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 129–147. https://doi.org/10.1037/0278-7393.31.1.129

Coulson, S., & Lovett, C. (2004). Handedness, hemispheric asymmetries, and joke comprehension. *Cognitive Brain Research*, *19*(3), 275–288. https://doi.org/10.1016/j.cogbrainres.2003.11.015

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, *1084*(1), 89–103. https://doi.org/10.1016/j.brainres.2006.02.010

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

DeLong, K. A., Chan, W.-h., & Kutas, M. (2019). Similar time courses for word form and meaning preactivation during sentence comprehension. *Psychophysiology*, *56*(4), e13312. https://doi.org/10.1111/psyp.13312

DeLong, K. A., & Kutas, M. (2020). Comprehending surprising sentences: Sensitivity of post-N400 positivities to contextual congruity and semantic relatedness. *Language, Cognition and Neuroscience*, *35*(0), 1044–1063. https://doi.org/10.1080/23273798.2019.1708960

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*, 150–162. https://doi.org/10.1016/j.neuropsychologia.2014.06.016

DeLong, K. A., Troyer, M., & Kutas, M. (2014). Pre-Processing in Sentence Comprehension: Sensitivity to Likely Upcoming Meaning and Structure. *Language and Linguistics Compass*, *8*(12), 631–645. https://doi.org/10.1111/lnc3.12093

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

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '88*, 281–285. https://doi.org/10.1145/57167.57214

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, *38*(1), 188–230. https://doi.org/10.1002/aris.1440380105

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Ettinger, A. (2020). What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48. https://doi.org/10.1162/tacl_a_00298

Ettinger, A., Feldman, N., Resnik, P., & Phillips, C. (2016). Modeling N400 amplitude using vector space models of word representation. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. https://cogsci.mindmodeling.org/2016/papers/0256/

Federmeier, K. D. (2021). Connecting and considering: Electrophysiology provides insights into comprehension. *Psychophysiology*, *n/a*(n/a), e13940. https://doi.org/10.1111/psyp.13940

Federmeier, K. D., & Kutas, M. (1999). A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, *41*(4), 469–495. https://doi.org/10.1006/jmla.1999.2660

Federmeier, K. D., Kutas, M., & Dickson, D. S. (2016). Chapter 45 - A Common Neural Progression to Meaning in About a Third of a Second. In G. Hickok & S. L. Small (Eds.), *Neurobiology of Language* (pp. 557–567). Academic Press. https://doi.org/10.1016/B978-0-12-407794-2.00045-6

Fedorenko, E., & Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in Cognitive Sciences*, *18*(3), 120–126. https://doi.org/10.1016/j.tics.2013.12.006

Fischer-Baum, S., Dickson, D. S., & Federmeier, K. D. (2014). Frequency and regularity effects in reading are task dependent: Evidence from ERPs. *Language, Cognition and Neuroscience*, *29*(10), 1342–1355. https://doi.org/10.1080/23273798.2014.927067

Fischler, I., & Bloom, P. A. (1979). Automatic and attentional processes in the effects of sentence contexts on word recognition. *Journal of Verbal Learning and Verbal Behavior*, *18*(1), 1–20. https://doi.org/10.1016/S0022-5371(79)90534-6

Fischler, I., Bloom, P. A., Childers, D. G., Arroyo, A. A., & Perry, N. W. (1984). Brain potentials during sentence verification: Late negativity and long-term memory strength. *Neuropsychologia*, *22*(5), 559–568. https://doi.org/10.1016/0028-3932(84)90020-4

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, *111*, 15–52. https://doi.org/10.1016/j.cogpsych.2019.03.002

Fleur, D. S., Flecken, M., Rommers, J., & Nieuwland, M. S. (2020). Definitely saw it coming? The dual nature of the pre-nominal prediction effect. *Cognition*, *204*, 104335. https://doi.org/10.1016/j.cognition.2020.104335

Forbes, M., Holtzman, A., & Choi, Y. (2019). Do Neural Language Representations Learn Physical Commonsense? *The 41st Annual Meeting of the Cognitive Science Society.*

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

Frank, S. L., & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, *32*(9), 1192–1203. https://doi.org/10.1080/23273798.2017.1323109

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. *Proceedings of the 2019 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 32–42. https://doi.org/10.18653/v1/N19-1004

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless Green Recurrent Networks Dream Hierarchically. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1195–1205. https://doi.org/10.18653/v1/N18-1108

Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (Fourth, pp. 819–836). MIT Press. Retrieved November 3, 2021, from https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_64579

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of Word Meaning and World Knowledge in Language Comprehension. *Science*, *304*(5669), 438–441. https://doi.org/10.1126/science.1095455

Hanna, J., & Pulvermüller, F. (2014). Neurophysiological evidence for whole form retrieval of complex derived words: A mismatch negativity study. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00886

Hodapp, A., & Rabovsky, M. (2021). The N400 ERP component reflects a learning signal during language comprehension. *bioRxiv*, 2021.03.25.436922. https://doi.org/10.1101/2021.03.25.436922

Hofmann, V., Pierrehumbert, J. B., & Schütze, H. (2021). Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words. *arXiv:2101.00403 [cs]*. Retrieved June 3, 2021, from http://arxiv.org/abs/2101.00403

Holcomb, P. J. (1988). Automatic and attentional processing: An event-related brain potential analysis of semantic priming. *Brain and Language*, *35*(1), 66–85. https://doi.org/10.1016/0093-934X(88)90101-0

Huang, Y., & Rao, R. P. N. (2011). Predictive coding. *WIREs Cognitive Science*, *2*(5), 580–593. https://doi.org/10.1002/wcs.142

Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, *86*, 157–171. https://doi.org/10.1016/j.jml.2015.10.007

Jackson, R. (2016). 15-level colorblind-friendly palette – Jackson Lab. Retrieved October 3, 2022, from https://jacksonlab.agronomy.wisc.edu/2016/05/23/15-level-colorblind-friendly-palette/

Jones, C. R., Chang, T. A., Coulson, S., Michaelov, J. A., Trott, S., & Bergen, B. (2022). Distrubutional Semantics Still Can't Account for Affordances. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). Retrieved October 4, 2022, from https://escholarship.org/uc/item/44z7r3j3

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the Limits of Language Modeling. *arXiv:1602.02410 [cs]*. Retrieved April 8, 2020, from http://arxiv.org/abs/1602.02410

Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing* (Third). [Online Draft]. Retrieved September 5, 2020, from https://web.stanford.edu/~jurafsky/slp3/

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, *52*(2), 205–225. https://doi.org/10.1016/j.jml.2004.10.002

Kim, J. S., Aheimer, B., Montané Manrara, V., & Bedny, M. (2021). Shared understanding of color among sighted and blind adults. *Proceedings of the National Academy of Sciences*, *118*(33), e2020192118. https://doi.org/10.1073/pnas.2020192118

Klein, S., & Tsarfaty, R. (2020). Getting the ##life out of living: How Adequate Are Word-Pieces for Modelling Complex Morphology? *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 204–209. https://doi.org/10.18653/v1/2020.sigmorphon-1.24

Kounios, J., & Holcomb, P. J. (1992). Structure and process in semantic memory: Evidence from event-related brain potentials and reaction times. *Journal of Experimental Psychology: General*, *121*, 459–479. https://doi.org/10.1037/0096-3445.121.4.459

Kuhn, M., Jackson, S., & Cimentada, J. (2022). *Corrr: Correlations in R*. Manual. https://CRAN.R-project.org/package=corrr

Kuperberg, G. R. (2021). Tea With Milk? A Hierarchical Generative Framework of Sequential Event Comprehension. *Topics in Cognitive Science*, *13*(1), 256–298. https://doi.org/10.1111/tops.12518

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A Tale of Two Positivities and the N400: Distinct Neural Signatures Are Evoked by Confirmed and Violated Predictions at Different Levels of Representation. *Journal of Cognitive Neuroscience*, *32*(1), 12–35. https://doi.org/10.1162/âĂŃjocn_a_01465

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, *17*(1), 117–129. https://doi.org/10.1016/S0926-6410(03)00086-7

Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language and Cognitive Processes*, *8*(4), 533–572. https://doi.org/10.1080/01690969308407587

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), *Predictions in the brain: Using our past to generate a future* (pp. 190–207). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195395518.003.0065

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470. https://doi.org/10.1016/S1364-6613(00)01560-6

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205. https://doi.org/10.1126/science.7350657

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. https://doi.org/10.1038/307161a0

Kutas, M., Lindamood, T. E., & Hillyard, S. A. (1984). Word expectancy and event-related brain potentials during sentence processing. In S. Kornblum & J. Requin (Eds.), *Preparatory states and processes* (pp. 217–237). Lawrence Erlbaum.

Kutas, M., & Van Petten, C. (1994). Psycholinguistics electrified: Event-related brain potential investigations. In M. A. Gernsbacher (Ed.), *Handbook of Psycholinguistics* (First, pp. 83–143). Academic Press.

Kutas, M., Van Petten, C. K., & Kluender, R. (2006). Psycholinguistics Electrified II (1994–2005). In M. Gernsbacher & M. Traxler (Eds.), *Handbook of Psycholinguistics* (Second, pp. 659–724). Elsevier. https://doi.org/10.1016/B978-012369374-7/50018-3

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259–284. https://doi.org/10.1080/01638539809545028

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 Effects of Prediction from Association in Single-word Contexts. *Journal of Cognitive Neuroscience*, *25*(3), 484–502. https://doi.org/10.1162/jocn_a_00328

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. https://doi.org/10.1038/nrn2532

Lewis, A. G., & Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex*, *68*, 155–168. https://doi.org/10.1016/j.cortex.2015.02.014

Li, B. Z., Nye, M., & Andreas, J. (2021). Implicit Representations of Meaning in Neural Language Models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1813–1827. https://doi.org/10.18653/v1/2021.acl-long.143

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique* (Second). A Bradford Book.

Luka, B. J., & Van Petten, C. (2014). Prospective and retrospective semantic processing: Prediction, time, and relationship strength in event-related potentials. *Brain and Language*, *135*, 115–129. https://doi.org/10.1016/j.bandl.2014.06.001

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60. https://doi.org/10.1016/j.cogpsych.2016.06.002

Marmor, G. S. (1978). Age at onset of blindness and the development of the semantics of color names. *Journal of Experimental Child Psychology*, *25*(2), 267–278. https://doi.org/10.1016/0022-0965(78)90082-6

McRae, K., Brown, K. S., & Elman, J. L. (2019). Prediction-Based Learning and Processing of Event Knowledge. *Topics in Cognitive Science*, 1–18. https://doi.org/10.1111/tops.12482

Medler, D., & Binder, J. (2005). MCWord: An On-Line Orthographic Database of the English Language. http://www.neuro.mcw.edu/mcword/

Merkx, D., & Frank, S. L. (2021). Human Sentence Processing: Recurrence or Attention? *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 12–22. https://doi.org/10.18653/v1/2021.cmcl-1.2

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*(4), 545–567. https://doi.org/10.1016/j.jml.2012.01.001

Michaelov, J. A., & Bergen, B. K. (2020). How well does surprisal explain N400 amplitude under different experimental conditions? *Proceedings of the 24th Conference on Computational Natural Language Learning*, 652–663. https://doi.org/10.18653/v1/2020.conll-1.53

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2022). So Cloze yet so Far: N400 Amplitude is Better Predicted by Distributional Information than Human Predictability Judgements. *IEEE Transactions on Cognitive and Developmental Systems*. https://doi.org/10.1109/TCDS.2022.3176783

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. Retrieved November 27, 2019, from http://arxiv.org/abs/1301.3781

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in Pre-Training Distributed Word Representations. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Retrieved May 26, 2022, from https://aclanthology.org/L18-1008

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 3111–3119). Curran Associates, Inc. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Mohebbi, H., Modarressi, A., & Pilehvar, M. T. (2021). Exploring the Role of BERT Token Representations to Explain Sentence Probing Results. *arXiv:2104.01477 [cs]*. Retrieved April 20, 2021, from http://arxiv.org/abs/2104.01477

Nicenboim, B., Vasishth, S., & Rösler, F. (2020). Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia*, *142*, 107427. https://doi.org/10.1016/j.neuropsychologia.2020.107427

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., Matthew Husband, E., Ito, A., Kazanina, N., Kogan, V., Kohút, Z., Kulakova, E., Mézière, D., Politzer-Ahles, S., Rousselet, G., … Von Grebmer Zu Wolfsthurn, S. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20180522. https://doi.org/10.1098/rstb.2018.0522

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., … Huettig, F. (2018). Additional discussion of Yan, Kuperberg & Jaeger (2017). *Open Science Framework*. Retrieved July 28, 2021, from https://osf.io/mb2ud/
https://osf.io/mb2ud/

Nieuwland, M. S., & Van Berkum, J. J. A. (2005). Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, *24*(3), 691–701. https://doi.org/10.1016/j.cogbrainres.2005.04.003

OpenAI. (2021). OpenAI API. Retrieved August 18, 2021, from https://beta.openai.com

Otten, M., Nieuwland, M. S., & Van Berkum, J. J. (2007). Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience*, *8*(1), 89. https://doi.org/10.1186/1471-2202-8-89

Paczynski, M., & Kuperberg, G. R. (2012). Multiple influences of semantic memory on sentence processing: Distinct effects of semantic relatedness on violations of real-world event/state knowledge and animacy selection restrictions. *Journal of Memory and Language*, *67*(4), 426–448. https://doi.org/10.1016/j.jml.2012.07.003

Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2011). Using Language Models and Latent Semantic Analysis to Characterise the N400m Neural Response. *Proceedings of the Australasian Language Technology Association Workshop 2011*, 38–46. Retrieved October 6, 2020, from https://www.aclweb.org/anthology/U11-1007

Payne, B. R., Lee, C.-L., & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology*, *52*(11), 1456–1469. https://doi.org/10.1111/psyp.12515

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Piantadosi, S., & Hill, F. (2022). Meaning without reference in large language models. *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*. Retrieved January 12, 2023, from https://openreview.net/forum?id=nRkJEwmZnM

R Core Team. (2022). *R: A language and environment for statistical computing*. Manual. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rabovsky, M. (2020). Change in a probabilistic representation of meaning can account for N400 effects on articles: A neural network model. *Neuropsychologia*, *143*, 107466. https://doi.org/10.1016/j.neuropsychologia.2020.107466

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580

RStudio Team. (2020). *RStudio: Integrated development environment for r.* Manual. RStudio, PBC. Boston, MA. http://www.rstudio.com/

Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high-and low-frequency words. *Memory & Cognition*, *18*(4), 367–379. https://doi.org/10.3758/BF03197126

Saysani, A., Corballis, M. C., & Corballis, P. M. (2018). Colour envisioned: Concepts of colour in the blind and sighted. *Visual Cognition*, *26*(5), 382–392. https://doi.org/10.1080/13506285.2018.1465148

Shipp, S., Adams, R. A., & Friston, K. J. (2013). Reflections on agranular architecture: Predictive coding in the motor cortex. *Trends in Neurosciences*, *36*(12), 706–716. https://doi.org/10.1016/j.tins.2013.09.004

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Speer, R., Chin, J., Lin, A., Jewett, S., & Nathan, L. (2018). LuminosoInsight/wordfreq: V2.2. https://doi.org/10.5281/zenodo.1443582

Szewczyk, J. M., & Federmeier, K. D. (2022). Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language*, *123*, 104311. https://doi.org/10.1016/j.jml.2021.104311

Tannenbaum, P. H., Williams, F., & Hillier, C. S. (1965). Word predictability in the environments of hesitations. *Journal of Verbal Learning and Verbal Behavior*, *4*(2), 134–140. https://doi.org/10.1016/S0022-5371(65)80097-4

Taylor, W. L. (1953). “Cloze Procedure”: A New Tool for Measuring Readability. *Journalism Quarterly*, *30*(4), 415–433. https://doi.org/10.1177/107769905303000401

Taylor, W. L. (1957). “Cloze” readability scores as indices of individual differences in comprehension and aptitude. *Journal of Applied Psychology*, *41*(1), 19–26. https://doi.org/10.1037/h0040591

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology*, *83*(3), 382–392. https://doi.org/10.1016/j.ijpsycho.2011.12.007

Uchida, T., Lair, N., Ishiguro, H., & Dominey, P. F. (2021). A Model of Online Temporal-Spatial Integration for Immediacy and Overrule in Discourse Comprehension. *Neurobiology of Language*, *2*(1), 83–105. https://doi.org/10.1162/nol_a_00026

Urbach, T. P., DeLong, K. A., Chan, W.-H., & Kutas, M. (2020). An exploratory data analysis of word form prediction during word-by-word reading. *Proceedings of the National Academy of Sciences*, *117*(34), 20483–20494. https://doi.org/10.1073/pnas.1922028117

Urbach, T. P., & Kutas, M. (2010). Quantifiers more or less quantify on-line: ERP evidence for partial incremental interpretation. *Journal of Memory and Language*, *63*(2), 158–179. https://doi.org/10.1016/j.jml.2010.03.008

Van Berkum, J. J. A. (2009). The Neuropragmatics of ’Simple’ Utterance Comprehension: An ERP Review. In R. Breheny, U. Sauerland, & K. Yatsushiro (Eds.), *Semantics and pragmatics: From experiment to theory* (pp. 276–316). Palgrave Macmillan.

Van Berkum, J. J. A. (2010). The brain is a prediction machine that cares about good and bad - Any implications for neuropragmatics? *Italian Journal of Linguistics*, *22*, 181–208. https://doi.org/10/component/file_539546/vanberkum-iljpap2010-definitive.pdf

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating Upcoming Words in Discourse: Evidence From ERPs and Reading Times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467. https://doi.org/10.1037/0278-7393.31.3.443

van den Brink, D., & Hagoort, P. (2004). The Influence of Semantic and Syntactic Context Constraints on Lexical Selection and Integration in Spoken-Word Comprehension as Revealed by ERPs. *Journal of Cognitive Neuroscience*, *16*(6), 1068–1084. https://doi.org/10.1162/0898929041502670

Van Petten, C. (1993). A comparison of lexical and sentence-level context effects in event-related potentials. *Language and Cognitive Processes*, *8*(4), 485–531. https://doi.org/10.1080/01690969308407586

Van Petten, C. (2014). Examining the N400 semantic context effect item-by-item: Relationship to corpus-based measures of word co-occurrence. *International Journal of Psychophysiology*, *94*(3), 407–419. https://doi.org/10.1016/j.ijpsycho.2014.10.012

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brainpotentials. *Memory & Cognition*, *18*(4), 380–393. https://doi.org/10.3758/BF03197127

Van Petten, C., & Kutas, M. (1991). Influences of semantic and syntactic context on open- and closed-class words. *Memory & Cognition*, *19*(1), 95–112. https://doi.org/10.3758/BF03198500

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

Van Petten, C., Weckerly, J., McIsaac, H. K., & Kutas, M. (1997). Working Memory Capacity Dissociates Lexical and Sentential Context Effects. *Psychological Science*, *8*(3), 238–242. https://doi.org/10.1111/j.1467-9280.1997.tb00418.x

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilke, C. O. (2020). *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. Manual. https://CRAN.R-project.org/package=cowplot

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception, & Psychophysics*, *72*(8), 2031–2046. https://doi.org/10.3758/BF03196680

Wu, Y. C., & Coulson, S. (2011). Are depictive gestures like pictures? Commonalities and differences in semantic processing. *Brain and Language*, *119*(3), 184–195. https://doi.org/10.1016/j.bandl.2011.07.002

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., & Matsumoto, Y. (2020). Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 23–30. https://doi.org/10.18653/v1/2020.emnlp-demos.4

Yan, S., & Jaeger, T. F. (2020). (Early) context effects on event-related potentials over natural inputs. *Language, Cognition and Neuroscience*, *35*(5), 658–679. https://doi.org/10.1080/23273798.2019.1597979

Yehezkel, S., & Pinter, Y. (2023). Incorporating Context into Subword Vocabularies. https://doi.org/10.48550/arXiv.2210.07095

Zeileis, A., Fisher, J. C., Hornik, K., Ihaka, R., McWhite, C. D., Murrell, P., Stauffer, R., & Wilke, C. O. (2020). Colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes. *Journal of Statistical Software*, *96*, 1–49. https://doi.org/10.18637/jss.v096.i01

Zeileis, A., Hornik, K., & Murrell, P. (2009). Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, *53*(9), 3259–3270. https://doi.org/10.1016/j.csda.2008.11.033