Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Special Issue "Strengthening Derivation Chains in Cognitive Neuroscience": Research Report

Ignoring the alternatives: The N400 is sensitive to stimulus preactivation alone

James A. Michaelov* and Benjamin K. Bergen

Department of Cognitive Science, University of California San Diego, La Jolla, CA, USA

ARTICLE INFO

Article history:
Received 31 December 2022
Reviewed 7 March 2023
Revised 12 May 2023
Accepted 3 August 2023
Action editor Dan Mirman
Published online 14 August 2023

Keywords:
Human language comprehension
Cognitive modeling
Event-related brain potentials
Natural language processing

ABSTRACT

The N400 component of the event-related brain potential is a neural signal of processing difficulty. In the language domain, it is widely believed to be sensitive to the degree to which a given word or its semantic features have been preactivated in the brain based on the preceding context. However, it has also been shown that the brain often preactivates many words in parallel. It is currently unknown whether the N400 is also affected by the preactivations of alternative words other than the stimulus that is actually presented. This leaves a weak link in the derivation chain—how can we use the N400 to understand the mechanisms of preactivation if we do not know what it indexes? This study directly addresses this gap. We estimate the extent to which all words in a lexicon are preactivated in a given context using the predictions of contemporary large language models. We then directly compare two competing possibilities: that the amplitude of the N400 is sensitive only to the extent to which the stimulus is preactivated, and that it is also sensitive to the preactivation states of the alternatives. We find evidence of the former. This result allows for better grounded inferences about the mechanisms underlying the N400, lexical preactivation in the brain, and language processing more generally.

© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Perhaps the best studied neural signal of language comprehension, the N400 is a negative component of the event-related brain potential peaking roughly 400 msec after the presentation of a stimulus (Kutas & Federmeier, 2011; Kutas & Hillyard, 1980, 1984). Studying the amplitude of the N400 has provided key evidence about language processing—most notably that words and their meanings are preactivated in the brain before they are encountered during online language comprehension, and that this preactivation is correlated with the extent to which the words are contextually predictable (Federmeier, 2021; Kuperberg et al., 2020; Kutas et al., 2011; Kutas & Federmeier, 2011; Kutas & Hillyard, 1984; Van Petten & Luka, 2012). Specifically, the amplitude of the N400 response is large (more negative) by default, and is reduced in proportion to the extent that the word is predictable (Dambacher et al.,

* Corresponding author.
E-mail addresses: j1michae@ucsd.edu (J.A. Michaelov), bkbergen@ucsd.edu (B.K. Bergen).

https://doi.org/10.1016/j.cortex.2023.08.001

0010-9452/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

2006; Federmeier, 2021; Payne et al., 2015; Van Petten, 1993; Van Petten & Kutas, 1990, 1991; Van Petten & Luka, 2012). The predictability effect has been replicated numerous times when predictability is operationalized as *cloze probability* (Kutas & Federmeier, 2011; Kutas & Hillyard, 1984), the proportion of participants in a norming study to fill in a gap in a sentence with a specific word (Taylor, 1953, 1957). More recently, this has also been found to be the case when predictability is operationalized using the predictions of *language models* (Aurnhammer & Frank, 2019; Frank et al., 2015; Merkx & Frank, 2021; Michaelov et al., 2022, 2023; Szewczyk & Federmeier, 2022; Yan & Jaeger, 2020), computational systems designed to predict the probability of a word in context based on the statistics of language (Jurafsky & Martin, 2023).

However, while it is by now widely accepted that the amplitude of the N400 response to a word reflects its preactivation, there is a weak link in the derivation chain—exactly how the N400 indexes this preactivation is not clear. The current general consensus is that the amplitude of the N400 response to a word only reflects the extent to which the word or its semantic content were preactivated before the word was encountered (DeLong et al., 2014; DeLong & Kutas, 2020; Federmeier, 2021; Federmeier et al., 2007; Kuperberg et al., 2020; Kutas et al., 2011; Thornhill & Van Petten, 2012; Van Petten & Luka, 2012). We refer to this as the *stimulus-dependent* account.

The main kind of evidence supporting this idea comes from the N400's resilience to variability. A key line of research in this area involves looking at the effect of sentence constraint on the N400. The term *sentence constraint* in this context refers to the cloze probability of the highest-cloze continuation of a sentence—if the highest-cloze continuation has a high cloze probability, the sentence has a high constraint, while if it has a low probability, the sentence has a low constraint. The central finding is that with cloze probability as a metric of contextual predictability, sentence constraint does not impact N400 amplitude at all; only the cloze probability of the stimulus word itself does (Federmeier et al., 2002, 2007; Federmeier, 2007, Otten & Berkum, 2008; Van Petten et al., 1999; Vissers et al., 2006; Wlotko & Federmeier, 2007; for review see Federmeier, 2021; Kuperberg et al., 2020; Van Petten & Luka, 2012). For example, Federmeier et al. (2007) find that if a word such as *look* has a low cloze probability, it elicits a large N400 response no matter whether the preceding context is strongly constraining, such as in *the children went outside to* ***look*** (highest-cloze completion: *play*), or only weakly constraining, such as in *Joy was too frightened to* ***look*** (highest-cloze completion: *move*). The reliability of the effect across contexts with different degrees of constraint suggests that only the contextual predictability of the stimulus that is presented, and not the predictability of the most likely alternate word, impacts N400 amplitude.

However, this kind of finding still does not rule out the possibility that preactivation of other words can impact N400 amplitude. The aforementioned experiments only consider the extent to which two words (the highest-cloze continuation and the stimulus word) are preactivated. But many candidate words are typically possible in any position. Lexical prediction has been theorized to involve the graded preactivation of a more than two words, ranging from a few candidates, as proposed by Brothers and Kuperberg (2021) to 'large portions of [the] lexicon', as proposed by Smith and Levy (2013). If the N400 truly does index processing difficulty, this processing difficulty might include not only the effort required to activate neural representations associated with the actual stimulus, but also inhibition of the neural representations associated with other possible stimuli, as some researchers have argued (Debruille, 2007; Fitz & Chang, 2019; Hale, 2001; Hoeks et al., 2004). We refer to this as the *distribution-dependent* account in line with the idea that the N400 reflects the full distribution of stimulus preactivation across possible next words.

One approach to evaluating whether a larger cohort of predicted words affects the N400 is to create an aggregate metric derived from the cloze probabilities of all completions generated in the cloze task such as entropy (as in Stone et al., 2022). However, cloze has its limitations. For example, it is well-established that words with cloze probabilities of zero can vary in their degree of preactivation (see, e.g., DeLong et al., 2019; Ito et al., 2016; Metusalem et al., 2012). An alternative approach is to include information about potential preactivation across the entire lexicon (and thus provide a more complete assessment of alternate word predictability) by modeling preactivation with language models, which, given any context, can provide a probability distribution over all words in their vocabulary (Jurafsky & Martin, 2023).

While language models have been successfully used to predict N400 amplitudes recorded from experimental participants, thus far this has only involved stimulus-dependent metrics—namely, surprisal and probability (Aurnhammer & Frank, 2019; Frank et al., 2015; Merkx & Frank, 2021; Michaelov et al., 2022, 2023; Szewczyk & Federmeier, 2022; Yan & Jaeger, 2020). To the best of our knowledge, no study has thus far attempted to directly test whether N400 amplitude can be predicted by the probability assigned to any word other than the stimulus itself by a language model, and only one (Frank et al., 2015) has tested a metric even in part derived from the whole probability distribution. Because language models are currently the only way to calculate the contextual probability of all words in the lexicon, it is thus the case that the question of whether the amplitude of the N400 is affected by the extent to which words other than the stimulus itself were predicted has not been sufficiently investigated for any conclusions to be drawn. This severely limits the inferences we can draw from the N400 effect. Namely, we do not know whether the N400 indexes the preactivation of the stimulus alone, or also its alternatives.

This presents a problem for theoretical advancement. Making progress on neural mechanisms of language comprehension relies on reliable and sensitive signals such as the N400. Researchers hope to draw inferences from effects like the N400 about, for instance, what is preactivated during comprehension. But to do this requires a precise account of what affects those signals. In addition to presenting an obstacle to our understanding of language comprehension more generally—for example, whether language processing fits into our general understanding of predictive processing in the brain—the weak derivation link presents a challenge for investigating how certain linguistic features impact preactivation. The majority of contemporary work on the N400 investigates how the context preceding a stimulus impacts the extent to which the stimulus is preactivated in the brain (for review, see, e.g., Federmeier, 2021; Kuperberg et al., 2020), but uncertainty about whether the N400 reflects only the preactivation of the stimulus drastically reduces the scope of what we can hope to understand. This issue is especially important in a field where noise and small effect sizes can often lead to inconsistent findings across studies (for a recent discussion, see Nicenboim et al., 2020).

The aim of this study, therefore, is to test whether, to the extent that this can be evaluated using current methods, the amplitude of the N400 response solely reflects the preactivation of the stimulus presented, or whether it in some way also reflects the inhibition of alternatives. To do this, we use state-of-the-art large language models. This is because, as previously stated, the conventional cloze approach fails to capture preactivation that varies systematically between different words with a cloze probability of zero (e.g., DeLong et al., 2019; Ito et al., 2016; Metusalem et al., 2012). This may not just be a methodological issue; as discussed in subsection 2.1, it is likely that the task itself (which asks for the best completion of a sentence) may preclude more anomalous words being filled in. But even if the issue is purely methodological, human vocabularies are very large, on the order of tens of thousands of words (Brysbaert et al., 2016), making it impractical to collect judgments from enough participants for every possible word. There is also reason to believe that the probabilities derived from language models are actually more informative than cloze. In addition to being more clearly interpretable from an information-processing perspective—-they reflect the contextual probabilities of words based on the statistics of language alone—recent work has shown that the predictions of contemporary language models can outperform cloze probability as predictors of N400 amplitude (Michaelov et al., 2022). Thus, even if it were possible to collect and calculate cloze values for all words in the vocabulary, it might still be preferable to use language models.

2. Past approaches

2.1. *Constraint*

Since early work on the N400 (Kutas & Hillyard, 1984), cloze probability has been used to operationalize the extent to which words are preactivated such that their preactivation impacts N400 amplitude. Most subsequent work explicitly or implicitly assumes that the amplitude of the N400 is only (or at least, most importantly) correlated with the extent to which the stimulus itself is preactivated.

However, more recently, there have been attempts to consider the how the broader, distributed 'landscape of activation' (Federmeier, 2021, p. 1) impacts N400 amplitude. An exemplary case is the study carried out by Federmeier et al. (2007), who test whether sentence constraint—the cloze probability of the most probable word in context—impacts N400 amplitude. The idea is that if inhibition does impact N400 amplitude, one should expect to see it most clearly with low-probability stimuli in high-constraint sentences. Under an distribution-dependent account, the high-probability completion is preactivated to a large extent, and thus, when this prediction is violated, we should expect a strong inhibition response. But as discussed, Federmeier et al. (2007) did not find any effect of constraint, leading them, and many other researchers (DeLong & Kutas, 2020; Federmeier, 2007, 2021; Federmeier et al., 2002; Kuperberg et al., 2020; Kutas et al., 2011; Otten & Berkum, 2008; Thornhill & Van Petten, 2012; Van Petten et al., 1999; Van Petten & Luka, 2012; Vissers et al., 2006; Wlotko & Federmeier, 2007) to argue that N400 amplitude does not reflect inhibition. Under these accounts, N400 amplitude only reflects new activation elicited by the stimulus—that is, the activation of neural representations that were not already preactivated by the context.

However, as argued earlier, this approach does not speak to failed predictions for words other than the best completion, since it only takes into account the activation of the highest-probability item. Moreover, word prediction might not linearly impact N400 amplitude—it might or might not be ten times harder to inhibit a word with a probability of 50% than a word with a probability of 5%. And finally, this approach assumes that cloze probability actually reflects the proportion of activation given to a specific candidate word (as argued by Brothers & Kuperberg, 2021; Staub et al., 2015). While it may intuitively seem a given that cloze probability should be directly proportional to the relative activation level of each word, this is not necessarily the case, especially given that the cloze task may have specific deforming effects on the probability distribution. One possible example of this can be illustrated by looking at the related anomaly effect, where an anomalous word that is semantically related to the best (highest-cloze) completion of a sentence elicits a smaller N400 response than an anomalous word that is not (for review, see Amsel et al., 2015; DeLong et al., 2019; Federmeier & Kutas, 1999; Ito et al., 2016; Kutas & Hillyard, 1984; Metusalem et al., 2012). In such cases, while both semantically related and unrelated anomalous words have a cloze probability of zero (or almost zero) but elicit N400 responses of different amplitudes, when we look at language model predictions, we see that the semantically related words have a higher probability (Michaelov & Bergen, 2022a). This suggests that such semantically related anomalous words are in fact more likely than their unrelated counterparts, but this is not detectable by looking at cloze probability. In this case, it is likely that the cloze task discourages participants from filling in anomalous words, even if they are more likely in the context, and thus more strongly preactivated (for related discussion, see Michaelov et al., 2022; Smith & Levy, 2011).

2.2. *Surprisal*

One attempt to consider the full distribution of prediction is that of Levy (2008). Levy (2008) frames lexical processing difficulty as involving the effort required to reallocate neurocognitive resources upon encountering a stimulus, based on altering the entire predicted probability distribution. To do this Levy (2008) proposes that the relevant metric should be the Kullback–Leibler divergence (Kullback & Leibler, 1951) between the probability distribution of predictions and the 'true' probability distribution—a distribution where the actual next word (i.e., the stimulus word) has a probability of 1, and all other words have a probability of 0. It should be noted that while Levy's (2008) account is based on considering reading times as an index of lexical processing difficulty, it may in fact be even more applicable to the N400. As discussed, the N400 is frequently thought to reflect the extent to which encountering a stimulus shapes the activation of neurocognitive representations, or more specifically, indexes the processing difficulty associated with updating the activation states of the brain to bring the total landscape of activation in the brain in line with the new stimulus.

The Kullback–Leibler divergence thus appears to reflect both the extent to which the true stimulus was predicted and the extent to which other words were predicted. The problem, however, is that Levy (2008) finds that the Kullback–Leibler divergence between the probability distribution that is the output of language models and the true probability distribution is mathematically equivalent to the surprisal S of the stimulus itself, that is, the negative logarithm of the probability p of a word w_i given its preceding context, shown in Equation (1).

$$S = -\log(p(w_i)) \tag{1}$$

Thus, while under an information-theoretic account, surprisal may be a good characterization of processing difficulty envisioned as the updating of activation states in the brain—and indeed, Hale (2001) proposes surprisal as a metric of lexical processing difficulty that reflects the difficulty of disconfirming alternatives—it is critically determined solely by the predicted probability of the stimulus word. From a theoretical perspective, this is not a problem. The fact that the Kullback–Leibler divergence between the true and predicted probability distributions is equivalent to surprisal may actually help to explain the finding that the N400 does not appear to be sensitive to constraint—if the brain reflects information-theoretic principles, the effort required to update our probability distribution might indeed only be determined by the probability of the stimulus (with a logarithmic linking function). Empirically, surprisal has also been incredibly successful in the prediction and modeling of the N400 (Aurnhammer & Frank, 2019; Frank et al., 2015; Merkx & Frank, 2021; Michaelov & Bergen, 2020; Parviz et al., 2011; Szewczyk & Federmeier, 2022), with one recent study even finding the surprisal of the GPT-3 language model (Brown et al., 2020) to be the best predictor of the N400 measured thus far, beating other language models and even cloze probability, the canonical metric of word probability (Michaelov et al., 2022). Nonetheless, because surprisal is not affected at all by the extent to which other words are preactivated, it cannot be used to investigate whether the preactivation of non-stimulus words impacts N400 amplitude.

2.3. L^1 distance

Another metric that ostensibly includes information about the preactivation states of non-stimuli is developed by Fitz and Chang (2019). Fitz and Chang (2019) propose that rather than simply indexing prediction error of some kind, the N400 has a functional significance in itself as a learning signal used to update our neurocognitive representations of the statistics of language for use in production (for related accounts, see, e.g., Federmeier, 2021; Fitz & Chang, 2019; Kuperberg et al., 2020; MacDonald, 2013; Pickering & Garrod, 2013). For this reason, Fitz and Chang (2019) take the true and predicted probabilities for each word in their model's vocabulary, and then model N400 amplitude as the sum of absolute error for each word—that is, the sum of the difference between the true and predicted probability of each word. This is equivalent to the Manhattan distance or L^1 norm between the predicted and true probability distributions. However, like surprisal, this metric is in fact only dependent on the probability of the stimulus, as we show in Appendix A. Specifically, L^1 distance is has relationship to $p(w_i)$ shown in Equation (2).

$$L^1 = 2 - 2p(w_i) \tag{2}$$

Like surprisal, L^1 distance is a metric based on the distance between the true and predicted probability distributions, and like surprisal, it is in fact only dependent on the predicted probability of the stimulus. Again, this is a theoretically meaningful result. If we take the idea of proportional preactivation—that is, the idea that words are preactivated in proportion to probability—seriously, and expect the processing difficulty indexed by the N400 to reflect the sum of the absolute error between the true and predicted probabilities of words, then this mathematical result suggests that we only need to calculate the probability of the stimulus itself in order to understand the N400 response. Indeed, Fitz and Chang (2019) are successful in using L^1 distance to model N400 amplitude, though it should be noted that Fitz and Chang's (2019) main model is not a language model in the strict sense because it is trained using structured semantic information (though its output is still a probability distribution over words).

However, as is the case with Kullback–Leibler divergence, this means that L^1 distance cannot be used to investigate the question of whether the possible inhibition of preactivated stimuli impacts the processing difficulty indexed by the N400. But by the same token, what it does tell us is that if the distribution-dependent account of the N400 is true, the mathematical relationship between the true and predicted probability distributions cannot be L^1 distance. The same is true for Kullback–Leibler divergence. However, this does not rule out the possibility that other difference metrics between the true and predicted probability distribution could capture the effect—even including other L^k distance metrics. For example, it could be that the L^1 distance metric under-estimates the difficulty of inhibiting high-probability items relative to low-probability items, something which might be detectable using the L^2 (Euclidean) distance as the relevant metric. On the other hand, it might be that using L^1 distance under-estimates the difficulty in inhibiting low-probability items relative to high-probability items, something that could be addressed by using the $L^{0.5}$ distance as a metric.

2.4. Entropy

A final metric that has been used to predict N400 amplitude (Stone et al., 2022), but which does in fact take into account the full probability distribution of preactivation is entropy (Shannon, 1948). The equation for entropy H is given in Equation (3), where $\widehat{p}(w_i)$ is the predicted probability of w_i in context.

$$H = -\sum_i \widehat{p}(w_i)\log \widehat{p}(w_i) \tag{3}$$

Entropy reflects uncertainty—given a probability distribution over words, the distribution with the highest possible entropy would be a uniform distribution, and the lowest-entropy distribution is one where one word has a probability of 1 and the remaining words have a probability of 0. A theoretical account of how entropy should influence N400 amplitude is not necessarily intuitive. In line with work on constraint, one might expect that in cases with low-probability stimuli, a low-entropy distribution might lead to the most processing difficulty, as this would result from a probability distribution where one very high-probability word is greatly preactivated. On the other hand, Stone et al. (2022) hypothesize that we might be less likely to make predictions in situations with higher entropy—where there are a larger number of possible continuations of a sentence—and thus, higher entropy should be associated with larger N400 responses. In this way, either a positive or negative relationship between entropy and N400 amplitude is plausible based on previous work.

Of course, the fact that previous work on the N400 and language comprehension more generally can lead to multiple predictions is not in itself an issue—this is something that could be resolved empirically, if indeed it is the case that entropy impacts N400 amplitude. But there does remain a fundamental problem with entropy as a metric of processing difficulty: it does not take into account the actual stimulus. Specifically, it only reflects the activation state before the word is encountered. Thus, if stimulus preactivation itself impacts processing difficulty, entropy alone cannot be used to model it. In the one study the one study that directly tests the effect of entropy on N400 amplitude, Stone et al. (2022) do not find it to be a significant predictor, either as a main effect or in interaction with word probability. However, it is worth noting that Stone et al. (2022) calculate their entropy based on cloze probabilities, and thus only a limited number of possible preactivations are considered—the maximum number of different responses to filling in the blank in the cloze task in their study is 8 (Stone et al., 2021). If there are differences in levels of preactivation based on contextual probability beyond that reflected by cloze, as previously discussed, then this approach does not take into account the full distribution of preactivation. Thus, despite the aforementioned theoretical problems with entropy, it is still valuable to directly test how well entropy calculated from the full distribution of predictions—for example, by using probabilities derived from a language model—can predict N400 amplitude, which we do in the present work. This is especially so given the recent findings that entropy appears to correlate with some of the neural activity that occurs during language comprehension when measured using magnetoencephalography (Brodbeck et al., 2022; Huizeling et al., 2022).

One metric that at least at first glance would appear to be better suited to testing whether N400 amplitude is sensitive to the probability of words other than the stimulus is *cross-entropy*. Cross-entropy is a measure of the difference between two distributions that is often used as a loss function (Goodfellow et al., 2016; Jurafsky & Martin, 2023), and thus is in line with some theories of the N400 (e.g., Fitz & Chang, 2019). However, cross-entropy is the sum of the Kullback–Leibler divergence between the true and predicted probability distributions and the entropy of the true probability distribution (Goodfellow et al., 2016, p. 73). Given that the entropy of the true probability distribution is zero, this means that, at least for language models, the cross-entropy is equivalent to Kullback–Leibler divergence, and thus, surprisal. And so this metric is also only dependent of the probability of the stimulus.

There are also several other related metrics that bear mentioning. Frank et al. (2015) and Aurnhammer and Frank (2019) test how well *next-word entropy, the difference between entropy and next-word entropy,* and two forms of what they refer to as *Lookahead Information Gain* predict N400 amplitude as well as reading time. However, next-word entropy in this case refers to the entropy of the probability distribution of the predictions for the word *after* the stimulus, and thus does not take into account the preactivation at the time that the stimulus is encountered, or the actual stimulus itself. The two Lookahead Information Gain metrics are also both based on this probability distribution for the following word. Finally, it should also be noted that none of these four metrics were found to be good at modeling the N400 (Aurnhammer and Frank, 2019; Frank et al., 2015).

3. Language models and the N400

Using the predictions of language models rather than a human-derived metric such as cloze probability can evoke skepticism. As articulated above, language models allow us to test hypotheses about how the full distribution of preactivation may impact N400 amplitude, but this is naturally only a viable strategy if language model predictions bear a clear relationship to this preactivation. Intuitively it may seem problematic to use the predictions derived from systems trained only on text data with no grounding in sensorimotor experience of the world or explicit propositional knowledge to model the kinds of predictions that humans may make during language comprehension. However, as discussed, recent work has shown that the predictions of language models can model N400 amplitude incredibly successfully (Aurnhammer & Frank, 2019; Frank et al., 2015; Merkx & Frank, 2021; Michaelov & Bergen, 2020, 2022a; Michaelov et al., 2021, 2023; Szewczyk & Federmeier, 2022).

Thus, at worst, language models appear to make predictions in line with the preactivation that underlies the N400 response. This in itself would not necessarily be surprising. The language we use encodes information about the world and our understanding of it to such an extent that its statistics can be used to calculate the semantic similarity of words (Landauer et al., 1998), identify structured semantic relations between words (Mikolov, Sutskever, et al., 2013), and even identify cultural biases (Bolukbasi et al., 2016). Thus, it may be that the statistics of language are able to approximate the statistics of the world—we are more likely to talk about more likely things. Therefore, even if the preactivation that occurs during online language comprehension is in fact largely based on our knowledge of the world (direct or indirect), this may be approximated well enough by the statistics of language that those statistics may be informative about neurocognitive systems underlying language comprehension.

However, there is a stronger alternative possibility: humans may actually be using the statistics of language in preactivation as part of language comprehension. Given the amount of information contained in the statistics of language (contemporary language models continue to improve performance at increasingly impressive tasks, see, e.g., Nie et al., 2020; Srivastava et al., 2022; Wang, Pruksachatkun, et al., 2019; Wang, Singh, et al., 2019), it would not in principle be surprising if the human language comprehension system took advantage of this. In fact, this would bring language processing in line with evidence for predictive coding in other domains, in which statistical learning is thought to play a key role. For example, in visual processing, there is evidence that environmental statistics are relevant from the level of neurons in the primary visual cortex to the overall encoding of scenes (de Lange et al., 2018; Rao & Ballard, 1999; Sherman & Turk-Browne, 2020).

In the domain of language specifically, learning from statistical information has been argued to be vital in acquisition, production, and comprehension (e.g., Ambridge et al., 2014; de Marneffe et al., 2012; Elman, 2009; Gerken, 2006, 2007; Gómez & Gerken, 2000; MacDonald, 2013; Newport and Aslin, 2004; Pickering & Garrod, 2007, 2013; Romberg & Saffran, 2010; Saffran et al., 1996; Seidenberg, 1997; Sherman et al., 2020). Indeed, there is already substantial evidence that the N400 is sensitive to factors that clearly relate to the statistics of language rather than just the organization of our semantic representations. Most notably, the N400 is sensitive to word frequency—words that are more frequent tend to 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) and their magnetoencephalographic equivalent (Halgren et al., 2002). Thus, rather than simply operationalizing predictability, language models may actually function as (computational-level) cognitive models of the neurocognitive system underlying lexical preactivation in the brain—a system engaging in lexical prediction at least in part based on the statistics of language.

4. The present study

The aim of the present study is to explore whether the amplitude of the N400 response is impacted not only by the extent to which a given stimulus was preactivated by its preceding context, but also by the extent to which other possible stimuli were preactivated. Most contemporary theoretical accounts of the N400, and by extension, the neurocognitive processes underlying language comprehension, assume that solely the stimulus word matters. But this has not yet been convincingly demonstrated.

To investigate this, we use language models to calculate several distribution-dependent metrics—that is, metrics that operationalize the difference between the true and predicted probability distribution—specifically, $L^{0.5}$ distance, L^2 distance, Hellinger distance, χ^2 distance, and cosine distance, as well as the previously-investigated constraint and entropy metrics (the equations for all metrics are presented in Table 2). We then test whether any of these can account for variance in N400 amplitude above and beyond that explained by predictability alone. We test this on the large N400 dataset made available by Szewczyk and Federmeier (2022), comprised of data from four published studies (Federmeier et al., 2007; Hubbard et al., 2019; Szewczyk et al., 2022; Wlotko & Federmeier, 2012) and one previously-unpublished ERP study.

We divide our study into two experiments. In the first, we test how well the predictability metrics calculated using seven contemporary language models predict N400 amplitude. Because our study tests whether metrics operationalizing the whole landscape of word preactivation predict N400 amplitude above and beyond the predictability of the stimulus itself, our first task is to find the best operationalization of predictability to compare these to. Previous work shows that surprisal is overall a better predictor of N400 amplitude than probability is (Szewczyk & Federmeier, 2022; Yan & Jaeger, 2020), especially for the best-performing models (Michaelov & Bergen, 2022b). However, Szewczyk and Federmeier (2022), analyzing the same dataset that we analyze, found that untransformed probability can also explain additional variance in N400 amplitude, especially for higher-probability items. As a result, we use both metrics as predictors in our linear mixed-effects regressions assessing how well different language models predict N400 amplitude.

In Experiment 2, we run our tests on the predictions of the best-performing language model: GPT-J (Wang & Komatsuzaki, 2021). First, we test whether any of the distribution-dependent metrics, entropy, or constraint out-perform predictability as predictors of N400 amplitude on their own, using the overall fit of linear mixed-effects regressions. We then test whether adding any of these to regressions already including the stimulus-only predictability variables improves model fit. If so, this would suggest that they explain variance not explained by predictability, and thus would provide evidence that the amplitude of the N400 response is impacted by the effort required to inhibit the activation of words other than the eliciting stimulus itself. If not, this would add to the evidence from research on sentence constraint suggesting that only the probability of the stimulus itself impacts N400 amplitude. The collection of metrics of each type that we use has, to the best of our knowledge, not been used previously to model N400 amplitude.

5. Experiment 1

5.1. *Introduction*

The overall purpose of the current study is to model the full landscape of neural preactivation using the probability of language models, and to use these probability distributions to investigate whether the amplitude of the N400 response to a stimulus is sensitive not only the extent to which it is preactivated, but also the extent to which alternatives are preactivated. To do this, in Experiment 1, we first select a language model that makes predictions that are highly correlated with word preactivation.

Previous work shows that surprisal from transformers—the current state-of-the-art language model architecture—correlate most closely with N400 amplitude compared with other models architectures (Merkx & Frank, 2021; Michaelov et al., 2022). In fact, the surprisals calculated using some of the most powerful models tested—ALBERT, RoBERTa, and GPT-3—have been found to out-perform cloze probability as predictors of N400 amplitude on one dataset (Michaelov et al., 2022). Given that the full probability distribution of GPT-3 is not directly accessible, it is not suitable for the present study. However, in recent work by Michaelov and Bergen (2022b), a much larger selection of contemporary transformer language models—including ALBERT, RoBERTa, and a number of models released after Michaelov et al. (2022)—are evaluated in terms of how well their probability and surprisal predicts N400 amplitude. Because surprisal appears to be a better predictor than probability overall, for the present study, we also include the two monolingual (i.e., trained only on English) transformer language models that generate surprisals which Michaelov and Bergen (2022b) find to be better correlated with N400 amplitude than ALBERT and RoBERTa—namely, GPT-J and OPT 6.7B. Since the publication of Michaelov and Bergen (2022b), a number of new language models have been released, and thus, we include 3 additional language models with a similar number of parameters as GPT-J and OPT 6.7B that have also been trained on datasets of the same order of magnitude: Pythia 6.9B (Biderman et al., 2023), Cerebras-GPT 6.7B (Dey et al., 2023), and StableLM-Base-Alpha 7B (Stability AI, 2023).

One thing that should be noted is that the set of models used comprises both autoregressive language models, those trained to predict a word based on only the preceding context; and masked language models, those trained to also predict based on the following context. In the present study, all models are only presented with the preceding context as humans were in the original N400 experiments, but it is unclear whether the fact that masked language models are also trained to ‘postdict’ (Huettig, 2015) makes them more or less human-like. While it would be impossible for us to use such postdictions during online comprehension, it is possible that we might still learn these reverse probabilities. Thus, in addition to the more practical question of which language model is best able to make predictions that correlate with the preactivation of neural representations during online language comprehension, the results of the present study may also shed light on what kinds of language statistics may be learned by humans.

5.2. *Method*

5.2.1. *Dataset*

The experimental stimuli and N400 data used in the present study come from a large dataset recently made available online by Szewczyk and Federmeier (2022) at https://osf.io/urvax/. This dataset is comprised of data from five experimental studies, which are described in more detail in this section. Four of the five experiments are from previously published papers (Federmeier et al., 2007; Hubbard et al., 2019; Szewczyk et al., 2022; Wlotko & Federmeier, 2012). We selected this dataset due to the fact that it covers a large number of stimuli, contains data from a large number of experimental participants, and is preprocessed in a consistent way across studies, so analyses can be run on all the data together. Furthermore, this dataset is well-suited to answer our main research question (addressed in Experiment 2) because, as will be discussed, all stimuli are based on those from the Federmeier et al. (2007) study—that is, the previously-discussed study that tested the effect of sentence constraint. For this reason, the stimuli were designed such that they included sentences with both high and low constraints, and thus vary in the shape of the probability distributions of possible continuations. While the stimuli were selected based on constraint calculated using cloze probability, and thus, we expect some variation between this and constraint as calculated using our language models (as well as between models), this allows us our analyses to account for a wide range of possible differences between true and predicted probability distributions.

In order to calculate probability and surprisal based on the original stimuli presented to the experimental participants, we truncated the stimuli such that they included the entire preceding context, using this as input to the language models. We then used the language models to calculate the probability of the critical words in the original stimuli, which we also negative log-transformed into surprisal. In our analysis, we only include words that are represented as a single token in all language models (i.e., are words in all language models' vocabularies). We only look at single-token critical words for each model because the other metrics that we calculate in Experiment 2 are only well-defined for such stimuli, and we only look at words that are single tokens for *all* language models so that we can compare performance across models. This exclusion criterion was decided before the analyses were carried out.

The dataset provided by Szewczyk and Federmeier (2022) provides single-trial N400 data. In it, the amplitude of the N400 response on a given trial is operationalized as the voltage amplitude at four centro-parietal electrodes (MiCe, MiPa, LMCe, RMCe) over the 300–500 msec time window. These N400 amplitudes are not baseline-corrected; instead, a baseline—the mean amplitude in the 100 msec before the presentation of the stimulus—is included as a variable, and in the original analysis is included as a covariate (Szewczyk & Federmeier, 2022).

As discussed, the data from five experiments are included in the dataset. Federmeier et al.'s (2007) is perhaps the best known of the studies, testing the effect of constraint on N400 amplitude. This study was built around a 2 × 2 design: sentences either had a high or low constraint, and for each sentence both the best (highest-cloze) completion and a low-cloze completion were used as critical words. This data subset included 7856 trials, collected for 564 stimuli from 32 experimental participants.

The second experimental study included in the dataset was conducted by Wlotko and Federmeier (2012). Stimuli in this experiment, which were selected from two previous studies (Federmeier et al., 2007; Wlotko & Federmeier, 2007) were selected to be plausible and vary ‘continuously through the full range of cloze probability’ (Wlotko & Federmeier, 2012, p. 359). This experiment contributed data from 4440 trials (300 stimuli; 16 experimental participants) to the dataset.

Third is a dataset from a study carried out by Hubbard et al. (2019). The stimuli in this study were 192 sentences selected from the Federmeier et al. (2007) experiment with the same 2 × 2 design: half of the sentences were high-constraint and half were low constraint; and each sentence had either the best completion or a low-cloze completion as the critical word. The data from this experiment included 5705 trials (32 experimental participants).

The final previously-published study included in the dataset is that of Szewczyk et al. (2022). The stimuli in this study were based on 168 sentence frames from previously-published studies including Federmeier et al. (2007), with high and low-cloze completions for each sentence frame. Stimuli were then expanded by adding an adjective before the completion that either increased the cloze probability of the low-cloze completion or further increased the cloze probability of the high-cloze completion. Thus there were four experimental conditions for each item, totaling 672 stimuli. Data from 4939 trials (32 experimental participants) were included from this study.

As previously discussed, the dataset also includes data from an unpublished study. The stimulus selection procedure is not mentioned in the paper (Szewczyk & Federmeier, 2022); however, looking at the data, we can see that all stimuli are present in one of the other four previously-published studies, and that the stimuli are comprised of a higher-cloze (mean = 57%) and lower-cloze (mean = 1%) critical word for each sentence frame. This study contributed 4822 trials (600 stimuli; 26 experimental participants) to the dataset.

Thus, the total dataset provided by Szewczyk and Federmeier (2022) was made up of 27,762 trials (138 experimental participants). Because of the overlap in stimuli between the different experiments, the total number of unique experimental stimuli was 1330. After removing data for stimuli where critical words are not tokens in all models' vocabulary, our analysis includes data from 25,506 trials (1238 stimuli; 138 experimental participants).

5.2.2. *Models*

The details of the seven models tested are provided in Table 1. All models are pretrained transformer language models, four of which are autoregressive—trained to predict the next word given the preceding context—and two of which are masked language models—trained to predict a word given the previous and following context. Note that in this study, we present all language models with only the preceding context. We used the *PyTorch* (Paszke et al., 2019) versions of all models made available through the *transformers* (Wolf et al., 2020) *Python* (Van Rossum & Drake, 2009) package.

5.2.3. *Statistical analysis*

All data manipulation, statistical analyses, and graphs were carried out and produced in R (R Core Team, 2020) using *Rstudio* (RStudio Team, 2020) and the *tidyverse* (Wickham et al., 2019) and *lme4* (Bates et al., 2015) packages. In this paper, we report how we determined all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, and all measures in the study. The sample size and all experimental manipulations were decided by the researchers who ran the original studies comprising the dataset (Federmeier et al., 2007; Hubbard et al., 2019; Szewczyk et al., 2022; Szewczyk & Federmeier, 2022; Wlotko & Federmeier, 2012). No part of the study procedures and no part of the analyses were pre-registered prior to the research being conducted. All data, code, and statistical analyses are available at https://osf.io/jrsgh.

Table 1 – Details of all the models used in the present study. Note that the ALBERT model uses shared parameters, and so the model is larger than the parameter counts suggest. The number of tokens for RoBERTa is estimated based on the fact that the dataset is 10 times larger than that on which ALBERT was trained.

Model Name	Parameters	Training data (tokens)
ALBERT XXL	.24B	3.3B
Cerebras-GPT 6.7B	6.7B	133B
GPT-J	6.1B	300B
OPT 6.7B	6.7B	180B
Pythia 6.9B	6.9B	300B
RoBERTa Large	.36B	33B
StableLM-Base-Alpha 7B	7.9B	800B

5.3. *Results*

We ran each of the preprocessed stimulus contexts through the seven language models, calculating the probability and surprisal for each critical word. We then combined this data with the single trial ERP data provided by the original authors, using linear mixed-effects regressions to predict N400 amplitude, with each regression including the probability and surprisal calculated using each language model as predictors. Following Szewczyk and Federmeier (2022), regressions also included baseline voltage, word frequency (log-transformed), concreteness, and orthographic neighborhood distance (OLD20), all of which were provided by Szewczyk and Federmeier (2022) as covariates. We also included random intercepts for each subject and sentence frame (each sentence frame in each experiment was treated as a separate sentence frame), as well as random slopes of the covariates (baseline voltage, word frequency, and orthographic neighborhood distance) for each of these. Following Michaelov et al. (2022), all variables were z-scored. In order to evaluate the performance of each metric, we compared each regression's Akaike Information Criterion (AIC) (Akaike, 1973), a metric of regression fit, where a lower AIC indicates a better fit.

Results are presented in Fig. 1, where AICs are are shown relative to the AIC of a baseline null model with the same predictors as the other regressions except without surprisal or probability. As can be seen, the best-performing model is GPT-J (AIC = 58549.22), followed by Pythia 6.9B (AIC = 58567.19), OPT 6.7B (AIC = 58568.82), Cerebras-GPT 6.7B (AIC = 58590.77), RoBERTa Large (AIC = 58627.10), StableLM-Base-Alpha 7B (AIC = 58708.78), and finally, ALBERT XXL (AIC = 58761.13). A

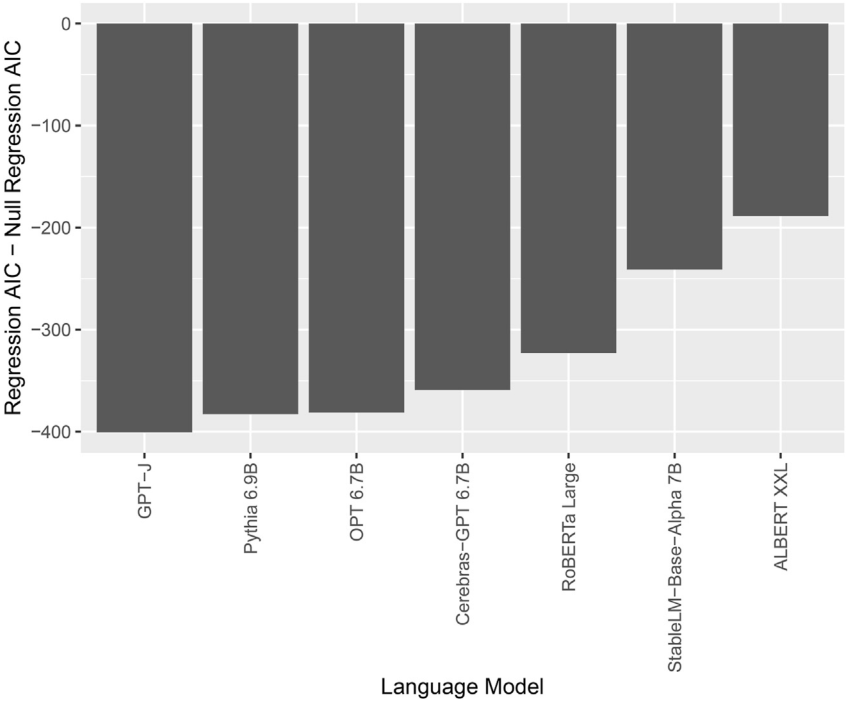

Fig. 1 – AICs of regressions including the probability and surprisal calculated from the indicated model as predictors. A lower AIC indicates a better fit.

difference in AIC of 4 or more is generally considered to indicate that the lower-AIC regression has 'considerably' more evidential support (Burnham & Anderson, 2004). Thus, the regression including GPT-J surprisal and probability is clearly the best-performing regression.

5.4. *Discussion*

The language model that best predicts N400 amplitude for this dataset is GPT-J, suggesting that its probability distributions most closely correlate with the preactivation state underlying the N400 response. We thus use metrics calculated from GPT-J for the remainder of our analyses.

The results of this experiment differ from the single-token results of Michaelov and Bergen (2022b) in that all but one of the autoregressive models tested here (StableLM-Base-Alpha 7B) performed better than the masked language models. It should be noted, however, that this result is in line with Michaelov and Bergen's (2022b) findings when analyzing the performance of language models at predicting N400 amplitude for stimuli including those made up of more than one token. Given this and the far larger number of experimental stimuli in the present study (1238 stimuli with single-token critical words compared to 37 single-token critical words and even 160 total critical words in Michaelov & Bergen, 2022b), it is likely that the results of the present study are more representative of the performance of the models at predicting N400 amplitude. Whether this is because the autoregressive architecture is more human-like or because the autoregressive models were trained on far more data than the masked language models is a question for future research.

6. Experiment 2

Equipped with a best-performing language model, we can now address the main research question, namely, whether the preactivation of possible stimuli other than the stimulus that elicits the N400 response can impact the amplitude of the response. To do this, we select a number of metrics that reflect the difference between the true and predicted probability distributions—that is, distribution-dependent metrics—as calculated using GPT-J. Many metrics relating the predicted and observed probability distributions across words were unsuitable for our analysis. Some, as discussed earlier, are linearly related to a metric of stimulus-dependent predictability. For example, total variation distance (as given in Gibbs & Su, 2002) is equivalent to half of the L^1 distance between the two distributions and thus is linearly related to probability. Similarly, because they involve element-wise multiplication between the distributions, Rényi divergence (as given in van Erven & Harremos, 2014) and Bhattacharyya distance (as given in Jain, 1976) simplify such that they become the logarithm of the stimulus probability multiplied by a constant, and thus, are directly proportional to surprisal. Other metrics are incalculable because in the true probability distribution, all words have a probability of zero with the exception of the true stimulus, which has a probability of 1. Because the zeros in the true distribution are meaningful, we do not use smoothing, and thus, we do not use any metrics that would involve dividing by or taking the logarithm of zero, e.g., Kullback–Leibler divergence in the opposite direction or information radius (as given in Manning & Schutze, 1999). We therefore selected two metrics that were both calculable and not linearly related to predictability: χ^2 distance and Hellinger distance. Beyond the aforementioned restrictions on suitable metrics, these specific metrics were not in themselves chosen for any theoretical reason beyond reflecting a difference between the true and predicted probability distributions. As discussed, the aim of the study is to test whether there is an effect of the full probability distribution on N400 amplitude at all rather than necessarily to precisely characterize such an effect. If either χ^2 and Hellinger distance successfully operationalize the difficulty inhibiting false predictions, then we should expect a negative correlation between the metric and N400 amplitude, indicating a stronger N400 response when there is a greater difference between the true and predicted probability distributions.

Other metrics were selected based on the theoretical perspective presented by Fitz and Chang (2019), which considers the probability distributions generated by predictive models to reflect the relative differences in preactivation between candidate stimuli, but also considers that these need not be meaningful as probabilities in themselves. Fitz and Chang (2019) operationalize the difference in the activation across all words before and after encountering a stimulus as L^1 distance; but as discussed, this is only dependent on the probability of the true stimulus itself. However, this is not the case for other L^k distances metrics. It may be the case, for example, that L^1 distance underestimates the extent to which lower-probability false predictions impact N400 amplitude, something which could be tested using a fractional L^k distance (in fact, fractional L^k norms are generally argued to be preferable for high-dimensional data; see Aggarwal et al., 2001). Conversely, if it is relatively more difficult to inhibit higher-probability false predictions than is operationalized by L^1 distance, it may be that a L^k distance with $k > 1$ is a more suitable way to operationalize this. In the present study, we test one of each of these: $L^{0.5}$ and L^2 distance. In addition to L^k distance, we also choose another distance metric that has had a large degree of success as a metric of the distance between two vectors in computational linguistics and psycholinguistics (Chwilla & Kolk, 2005; Dumais et al., 1988; Deerwester et al., 1990; Ettinger et al., 2016; Landauer et al., 1998; Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013; Parviz et al., 2011; Van Petten, 2014): cosine distance. As with other distribution-dependent metrics, if L^k or cosine distance successfully models the effect of inhibition on N400 amplitude, we should expect a negative correlation between the two; with a greater distance between the true and predicted probability distribution resulting a stronger N400 response.

We also compare these metrics (that to the best of our knowledge have not previously been used to predict N400 amplitude), with both constraint and entropy, also calculated from GPT-J. For constraint, we record the probability of the highest-probability continuation in a given context, analogous to the Best Completion calculated with cloze probabilities. To account for the possibility of a logarithmic linking function between constraint and the N400 (as there appears to be for predictability), we also convert these probabilities into surprisal, and test both metrics.

6.1. *Method*

6.1.1. *Data*

For this experiment, we used experimental data from all stimuli in the dataset that have critical words that are in the vocabulary of the GPT-J language model (i.e., the data from all single-token critical words). Because we include constraint as a metric in our analysis, we also restrict our analysis to stimuli that are not the best completions in their context, following Federmeier et al. (2007)—that is, we exclude cases where the surprisal variant of the constraint metric is identical to stimulus surprisal. Our analysis thus includes data from 17,892 trials (873 stimuli; 138 experimental participants). Note that these exclusion criteria were decided before the analyses were carried out.

6.1.2. *Metrics*

All metrics used in this analysis are defined in Table 2. The correlations between all metrics is shown in Fig. 2.

6.2. *Results*

First, we compared how well each of the metrics performs compared to surprisal, probability, and an overall

Table 2 – The names of the metrics used in the present study and the equations used to calculate them. All equations are based on the version given in the citation, but have been adapted for consistency. $\hat{p}$ refers to the predicted probability, p to the true probability (i.e., 0 or 1), w_i to the critical word, and w_{BC} to the best completion (i.e., the word with the highest probability in a given context).

Metric Name	Equation	Citation
Surprisal	$-\log(\hat{p}(w_i))$	Levy (2008)
L^k Distance	$\sum_i(\lvert\hat{p}(w_i)-p(w_i)\rvert^k)^{\frac{1}{k}}$	Aggarwal et al. (2001)
χ^2 Distance	$\sum_i(\frac{(p(w_i)-p(w_i))^2}{\hat{p}(w_i)})$	Gibbs and Su (2002)
Hellinger Distance	$\left[\sum_i\left(\sqrt{p(w_i)}-\sqrt{\hat{p}(w_i)}\right)^2\right]^{\frac{1}{2}}$	Gibbs and Su (2002)
Cosine Distance	$1-\frac{\sum_i\hat{p}(w_i)p(w_i)}{\sqrt{\sum_i\hat{p}(w_i)^2}\sqrt{\sum_i p(w_i)^2}}$	Jurafsky and Martin (2023)
Entropy	$-\sum_i\hat{p}(w_i)\log\hat{p}(w_i)$	Jurafsky and Martin (2023)
Constraint (p)	$\hat{p}(w_{BC})$	–
Constraint (S)	$-\log(\hat{p}(w_{BC}))$	–

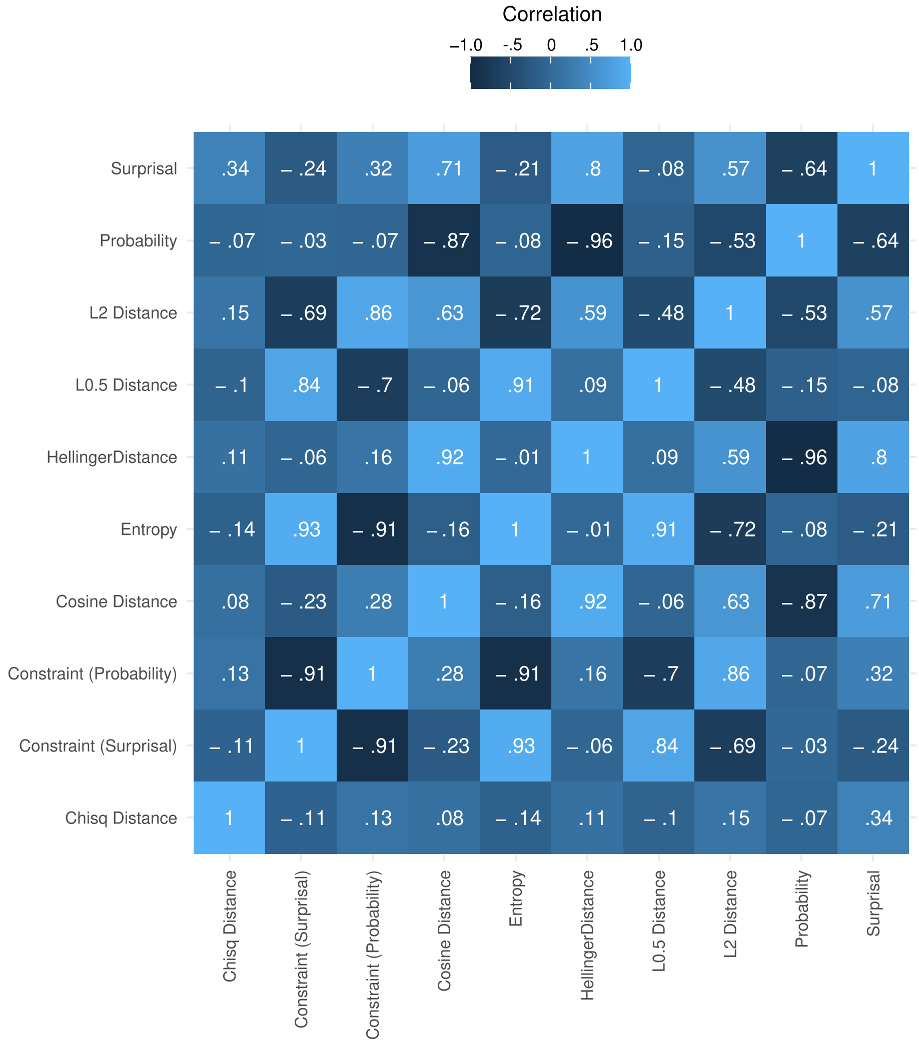

Fig. 2 – The Pearson Correlation *r* between all variables of interest in our study for all critical words that were single tokens for GPT-J.

predictability regression that includes both variables. We compared the AIC of linear mixed-effects models with each metric as a predictor and with the same covariates and random effects structure as those in Experiment 1, where, as in Experiment 1, all variables were z-scored. The results can be seen in Fig. 3, which shows that the aggregate predictability regression best fits the N400 data, followed by (in order of increasing AIC, and thus, decreasing fit) surprisal, Hellinger distance, probability, cosine distance, L^2 distance, constraint operationalized as probability, and χ^2 distance. On their own, constraint operationalized as surprisal, entropy, and $L^{0.5}$ distance appear to reduce model fit, compared to a model including just the covariates and random effects structure.

This result demonstrates that no distribution-dependent metric is a better predictor of N400 amplitude than a combination of surprisal and probability, or even surprisal alone. However, the question we seek to address is whether these variables can explain *any* variance in N400 amplitude not explained by predictability alone. Thus, in the final, critical step, we test whether adding any of the distribution-dependent metrics to the predictability regression improves fit. The results are shown in Fig. 4. As can be seen, the only metric that improves model fit numerically if added to the regression is cosine distance; the rest decrease model fit. However, as discussed in Experiment 1, generally only a difference in AIC of 4 or more is considered to reflect a substantial difference in model fit (Burnham & Anderson, 2004), suggesting that the improvement due to cosine distance is not meaningful.

In order to test directly and to verify whether there is indeed a lack of improvement from adding the other metrics, we run likelihood ratio tests comparing the predictability regression with regressions also including each distribution-dependent variable. We find that cosine distance does not improve model fit [$\chi^2(1) = 3.0969$, $p = .0784$], and neither does χ^2 distance [$\chi^2(1) = 1.8036$, $p = .1793$], entropy [$\chi^2(1) = .5557$, $p = .4560$], $L^{0.5}$ distance [$\chi^2(1) = .4025$, $p = .5258$], Hellinger

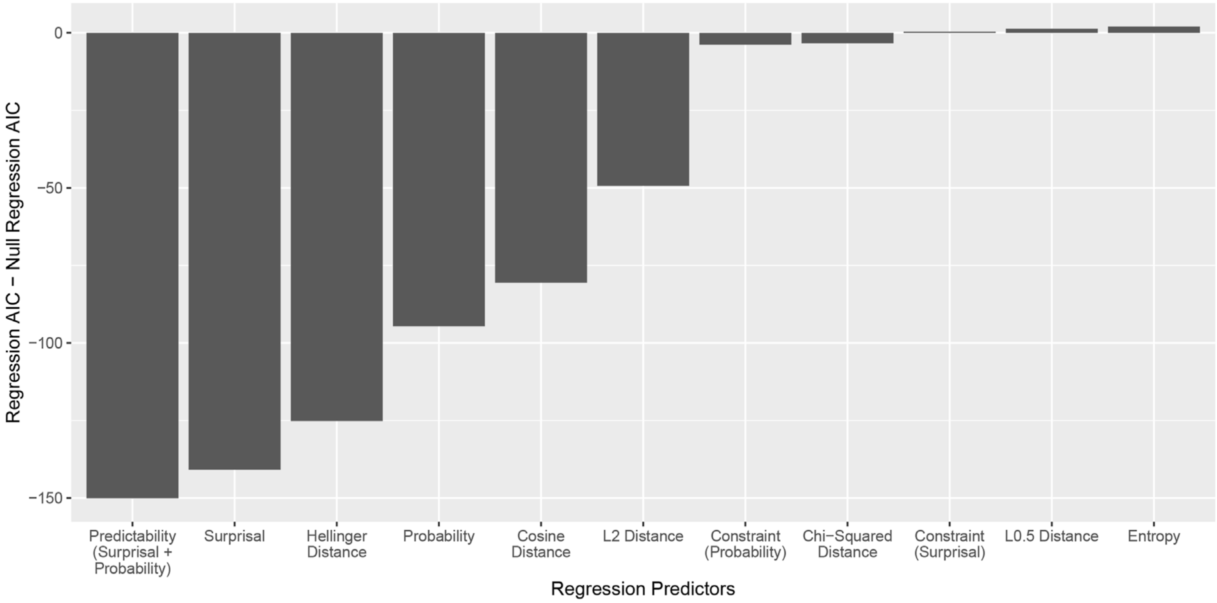

Fig. 3 – The AICs of all regressions including a single metric of interest as a predictor, as well as one including both predictability metrics (probability and surprisal).

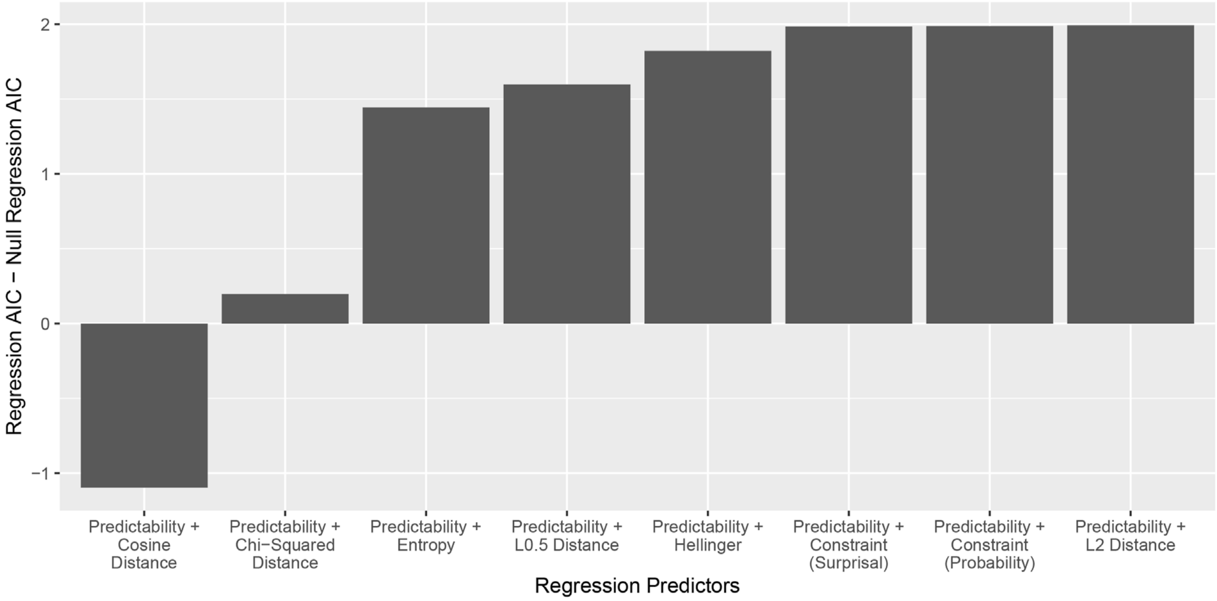

Fig. 4 – The AICs of all regressions including a single metric of interest as a predictor, as well as one including both predictability metrics (probability and surprisal).

distance [$\chi^2(1) = .1774$, $p = .6737$], either constraint metric [probability: $\chi^2(1) = .0113$, $p = .9153$; surprisal: $\chi^2(1) = .0145$, $p = .9042$], or L^2 distance [$\chi^2(1) = .0072$, $p = .9324$]. Thus, no distribution-dependent metric explains any variance in N400 amplitude above and beyond that explained by predictability.

6.3. *Discussion*

Our results replicate and extent several findings. First, as in previous work (Aurnhammer & Frank, 2019; Frank et al., 2015; Michaelov & Bergen, 2022b; Szewczyk & Federmeier, 2022), surprisal is the best single predictor of N400 amplitude overall. Second, like Szewczyk and Federmeier (2022), we find that including un-transformed probability as a predictor in addition to surprisal improves fit to the N400 data in this dataset. However, we extend this finding to also include GPT-J, a model that appears calculate probabilities that more closely correlate with N400 amplitude both when used directly and transformed into surprisal (Michaelov & Bergen, 2022b) compared to GPT-2 (Radford et al., 2019), the model used by Szewczyk and Federmeier (2022). Finally, as in previous work, neither constraint (Federmeier, 2007; Federmeier et al., 2002, 2007; Otten & Berkum, 2008; Van Petten et al., 1999; Vissers et al., 2006; Wlotko & Federmeier, 2007) nor entropy (Stone et al.,

2021) predict N400 amplitude above and beyond predictability. Crucially, our study extends these findings to probabilities derived from language models in addition to cloze probability.

In this experiment we set out to investigate whether the preactivation of stimuli other than the actually-occurring stimuli impact the amplitude of the N400 response using metrics operationalizing the difference between the true distribution for each critical word and the distribution predicted by GPT-J. We found that neither the variables that treat this difference as a difference between probability distributions (χ^2 distance and Hellinger distance) nor the metrics that treat it as the distance between two vectors (cosine distance, $L^{0.5}$ distance, and L^2 distance) explain any variance in N400 amplitude not explained by predictability alone, as operationalized by probability and surprisal.

7. General discussion

It has long been widely believed (with a few exceptions, e.g., Debruille, 2007; Fitz & Chang, 2019; Hoeks et al., 2004) that the N400 is only sensitive to the preactivation of the stimulus that it is elicited by, and not the rest of the landscape of activation elicited by its context. This premise forms the basis of the majority of contemporary accounts of the effect (e.g., Brouwer et al., 2012; Brouwer & Hoeks, 2013; Delogu et al., 2019; DeLong et al., 2014; Federmeier, 2021; Kuperberg et al., 2020; Kuperberg & Jaeger, 2016; Kutas et al., 2011; Van Petten & Luka, 2012). But, as discussed in section 1, this never been fully tested—previous work has looked at constraint (Federmeier, 2007; Federmeier et al., 2002, 2007; Otten & Berkum, 2008; Van Petten et al., 1999; Vissers et al., 2006; Wlotko & Federmeier, 2007), or in one more recent study, entropy based on the words generated by the cloze task (Stone et al., 2022). In both cases, the approaches only consider a small subset of the full landscape of preactivation at the time when the stimulus is encountered—in the case of constraint, only the extent to which the most predictable word is expected, and in the case of the cloze-derived entropy study (Stone et al., 2022), the degree to which at most 8 predictable words are expected.

Thus, prior to the current study, a key link in the derivation chain was weak. Do metrics that consider the full probability distribution predict variance in the amplitude of the N400 not captured by metrics that consider only the probability of the stimulus itself? Our results suggest that they do not—no distribution-dependent metric on its own predicts N400 amplitude better than surprisal, and like constraint and entropy, none of the distribution-dependent metrics explain a significant amount of the variance in N400 amplitude above and beyond that explained by predictability alone.

7.1. *What impacts N400 amplitude?*

In our experiments, no distribution-dependent metric significantly predicts N400 amplitude once predictability has been accounted for. In addition, no individual distribution-dependent is a better predictor of N400 amplitude than surprisal. These results are consistent with the account that the amplitude of the N400 response is dependent only on the extent to which the stimulus itself was preactivated.

The present study is the first to directly test whether the full distribution of preactivation can impact N400 amplitude. The finding that no distribution-dependent metric better correlates with N400 amplitude than surprisal (which only reflects the preactivation of the stimulus itself) suggests that the extent to which a word is preactivated is still the best predictor of N400 amplitude; and this is further strengthened by the fact that no distribution-dependent metric explains variance not explained by either surprisal or probability. Thus, the derivation chain is strengthened, and we can more confidently make inferences directly from N400 effects about the degree to which the neural representations associated with given stimuli are activated before they are encountered. It is therefore possible to investigate exactly which factors impact and modulate this—as one example, the line of research investigating whether the amplitude of the N400 response, and hence, preactivation, is sensitive to the animacy features of entities under discussion (Kim & Osterhout, 2005; Kuperberg, 2007; Kuperberg et al., 2003; Nieuwland et al., 2013; Nieuwland & Van Berkum, 2005; Paczynski & Kuperberg, 2011, 2012; Szewczyk & Schriefers, 2011, 2013; Vega-Mendoza et al., 2021; Wang et al., 2020).

7.2. *Surprisal and predictive coding*

The research carried out in the present study is compatible with most contemporary accounts of the N400. However, as noted in section 3, a strong interpretation of the study and results uses the predictive coding framework, under which the neurocognitive system responsible for the preactivation underlying the N400 response is a predictive system (Bornkessel-Schlesewsky & Schlesewsky, 2019; Kuperberg et al., 2020; Lewis & Bastiaansen, 2015). As shown in the current work, language models can serve as computational-level cognitive models of at least part of this proposed system. The results of the present study also provide evidence to support the predictive coding account of the N400.

Under a predictive coding account, the functional significance of neural metrics of processing difficulty is twofold: the new activation is information that allows the current stimulus to be correctly processed by the system; and the new activation is a learning signal (Clark, 2013; Huang & Rao, 2011; Rao & Ballard, 1999). In the language domain, this learning signal is thought to allow the neurocognitive system underlying language comprehension (and under some accounts also production, see, e.g., Fitz & Chang, 2019; Kuperberg et al., 2020) to learn and adapt, either long-term as part of continual language learning, or to a specific situation (Bornkessel-Schlesewsky & Schlesewsky, 2019; Hodapp & Rabovsky, 2021; Kuperberg et al., 2020).

While all metrics tested in the present study could conceivably fulfill both of these roles, it is striking that surprisal, the best-performing metric, also seems best suited to fulfilling the role of learning signal. As discussed, when comparing the true and predicted probabilities generated by language models, surprisal is equivalent to cross-entropy. This is interesting because cross-entropy is precisely the loss function used to train virtually all language models (Jurafsky & Martin, 2023). In other words, if we were to determine what would be the best loss function for a neurocognitive system engaging in lexical prediction to use, based on current research, it would be cross-entropy—and thus, surprisal. For this reason, the fact that surprisal is the metric most correlated with N400 amplitude is striking. In this way, our results provide indirect evidence to support the predictive coding account of the N400.

7.3. *Mechanistic implications*

Predictability alone explaining variance in N400 amplitude is consistent with two specific mechanistic accounts of how the preactivation that occurs as part of online language comprehension is indexed by the N400 response.

The first is that the processing difficulty indexed by the N400 is only due to the activation of the neural representations associated with the stimulus that were not already activated due to the preceding context. That is, the amplitude of the N400 response is not just stimulus-dependent, but also only reflects this stimulus-driven activation. This is in line with most contemporary accounts of the N400 (DeLong et al., 2014; DeLong & Kutas, 2020; Federmeier, 2021; Kuperberg et al., 2020; Kuperberg & Jaeger, 2016; Kutas et al., 2011; Kutas & Federmeier, 2011; Van Petten & Luka, 2012). So what happens to words that are preactivated but not encountered? One possibility is that the metabolic resources required for preactivation (see, e.g., Brothers & Kuperberg, 2021; Levy, 2008) are constantly required to be expended to maintain preactivation, and thus, simply stopping doing so is enough to suppress them. Alternatively, there may not be any active suppression or inhibition at all—the evidence suggests that highly probable words that are not presented as stimuli can remain activated over the course of an experiment (Rommers & Federmeier, 2018).

The other mechanistic account consistent with the results is that inhibition does indeed contribute to the processing difficulty indexed by the N400 response, but that the effort required to do this is dependent on the extent to which the stimulus was preactivated. Under such an account, it is simply the case that surprisal, or another metric that is only dependent on the preactivation state of the stimulus, mathematically expresses the combined processing difficulty of activating the representations associated with the stimulus and inhibiting others. Indeed, given the number of metrics of the difference between the true and predicted probability distributions that simplify to a stimulus-dependent metric—Kullback-Leibler divergence, Rényi divergence (a generalization of Kullback-Leibler divergence), Bhattacharyya distance, total variation distance, and L^1 distance—perhaps it would not be surprising if this were the case. This idea is in line with the account of Hale (2001), who envisions surprisal as reflecting the difficulty of disconfirming predictions, and perhaps implicitly in line with the account of Fitz and Chang (2019), who argue that N400 amplitude reflects the activation and inhibition effort and present L^1 distance as the metric to express this—which, as we show, is a stimulus-dependent metric. If this is the case, however, it does not diminish the importance of determining whether the amplitude of the N400 response is sensitive to the preactivation of the stimulus only or the to the whole distribution (i.e., the whole landscape of activation in long-term memory). The weak link in the derivation chain has still been strengthened—we can be more comfortable in using the N400 to understand exactly how much a given stimulus was preactivated under one experimental condition relative to another—but further work would need to be carried out to investigate exactly to what extent the activation and inhibition contribute to the final amplitude measured.

8. Conclusions

In this study, we used computational methods to investigate the question of whether the amplitude of the N400 response to a word is impacted only by the degree to which the word was preactivated or to the entire landscape of activation elicited by the preceding context. We found that across the data from the five experiments modeled, surprisal was the best single predictor of N400 amplitude. Furthermore, no metrics reflecting the extent to which words other than the stimulus were preactivated explained any variance in N400 amplitude beyond that explained by surprisal and probability. This result supports the idea that N400 amplitude is only sensitive to the degree to which the stimulus itself was preactivated at the point at which it was encountered. Based on this and another property of surprisal—its equivalence with cross-entropy for language model predictions—we argue that the results of the present study support a predictive coding account of the N400.

Human studies statement

No experiments involving human subjects were carried out by the authors for the study reported in this manuscript. The study involves data collected from experiments using computational language models and the reanalysis of published data.

Data availability

The data and code used in the present study have been made available at https://osf.io/jrsgh.

CRediT role statement

James Michaelov: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization.

Benjamin Bergen: Supervision, Writing - Review & Editing.

(**Note:** The original N400 data were gathered and preprocessed by the original authors of the respective studies whose data are re-analyzed in the present study. The above CRediT roles apply only to the work carried out as part of the present study and not these previously-published experiments.)

Open practices

The study in this article earned Open Material badge for transparent practices. The material used in this study are available at: https://osf.io/jrsgh.

Declaration of competing interest

None.

Acknowledgments

This work was partially supported by the Center for Academic Research and Training in Anthropogeny [Annette Merle-Smith Fellowship].

Appendix A. The stimulus-dependence of L^1 distance

In this appendix, we show that the L^1 distance between the true and predicted probability distributions for a given word w_i is only dependent on the probability of the word $p(w_i)$ and not the probabilities of other words.

First, we note that the sum of the absolute error for each word is the sum of the absolute error E for the true next word w_i and the absolute error for all the words that are not the true next word (i.e., every $w_{\neg i}$):

$$L^1 = E(w_i) + \sum E(w_{\neg i}) \tag{A.1}$$

For the true next word, the absolute error is a positive prediction error, the difference between 1 and the predicted probability of the word p_{true}:

$$E(w_i) = 1 - p(w_i) \tag{A.2}$$

For all other words, the absolute error is a negative prediction error, the predicted probability of the false word $p(w_{\neg i})$ minus the true probability, 0:

$$E(w_{\neg i}) = p(w_{\neg i}) - 0 \tag{A.3}$$

This simplifies to:

$$E(w_{\neg i}) = p(w_{\neg i}) \tag{A.4}$$

Since the distribution is a probability distribution, all probabilities add up to 1, and thus:

$$p(w_i) + \sum p(w_{\neg i}) = 1 \tag{A.5}$$

This means that the following is also the case:

$$\sum p(w_{\neg i}) = 1 - p(w_i) \tag{A.6}$$

We can substitute Equation (A.2) and Equation (A.6) into the equation for total Manhattan distance Equation (A.1), getting:

$$L^1 = (1 - p(w_i)) + (1 - p(w_i)) \tag{A.7}$$

Which can be simplified to:

$$L^1 = 2 - 2p(w_i) \tag{A.8}$$

REFERENCES

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In J. Van den Bussche, & V. Vianu (Eds.), *Database theory — ICDT 2001* (pp. 420–434). Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-44503-X_27.

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). Budapest, Hungary: Akadémiai Kiadó. https://doi.org/10.1007/978-1-4612-1694-0_15.

Ambridge, B., Pine, J. M., & Lieven, E. V. M. (2014). Child language acquisition: Why universal grammar doesn't help. *Language, 90*, e53–e90. https://doi.org/10.1353/lan.2014.0051. URL: https://muse.jhu.edu/article/552064.

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. URL: http://www.sciencedirect.com/science/article/pii/S0749596X1500042X.

Aurnhammer, C., & Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia, 134*, Article 107198. https://doi.org/10.1016/j.neuropsychologia.2019.107198. URL: http://www.sciencedirect.com/science/article/pii/S0028393219302404.

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

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., & van der Wal, O. (2023). Pythia: A suite for analyzing large language models across training and scaling. URL: https://arxiv.org/abs/2304.01373v1. arXiv:2304.01373.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc.. URL: https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

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. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00298/full#B34.

Brodbeck, C., Bhattasali, S., Cruz Heredia, A. A., Resnik, P., Simon, J. Z., & Lau, E. (2022). Parallel processing in speech perception with local and global representations of linguistic context. *eLife, 11*, Article e72056. https://doi.org/10.7554/eLife.72056

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*, Article 104174. https://doi.org/10.1016/

j.jml.2020.104174. URL: http://www.sciencedirect.com/science/article/pii/S0749596X20300887.

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research, 1446*, 127–143. https://doi.org/10.1016/j.brainres.2012.01.055. URL: http://www.sciencedirect.com/science/article/pii/S0006899312001588.

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. URL: https://www.frontiersin.org/articles/10.3389/fnhum.2013.00758/full.

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. In *Advances in neural information processing systems* (pp. 1877–1901). Curran Associates, Inc.. URL: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology, 7*. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01116.

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

Chwilla, D. J., & Kolk, H. H. J. (2005). Accessing world knowledge: Evidence from N400 and reaction time priming. *Cognitive Brain Research, 25*, 589–606. https://doi.org/10.1016/j.cogbrainres.2005.08.011. URL: http://www.sciencedirect.com/science/article/pii/S0926641005002259.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 181–204. https://doi.org/10.1017/S0140525X12000477. URL: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/whatever-next-predictive-brains-situated-agents-and-the-future-of-cognitive-science/33542C736E17E3D1D44E8D03BE5F4CD9.

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research, 1084*, 89–103. https://doi.org/10.1016/j.brainres.2006.02.010. URL: https://www.sciencedirect.com/science/article/pii/S0006899306003854.

Debruille, J. B. (2007). The N400 potential could index a semantic inhibition. *Brain Research Reviews, 56*, 472–477. https://doi.org/10.1016/j.brainresrev.2007.10.001. URL: https://www.sciencedirect.com/science/article/pii/S0165017307002202.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9.

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences, 22*, 764–779. https://doi.org/10.1016/j.tics.2018.06.002. URL: http://www.sciencedirect.com/science/article/pii/S1364661318301396.

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*, Article 103569. https://doi.org/10.1016/j.bandc.2019.05.007. URL: https://www.sciencedirect.com/science/article/pii/S0278262618304299.

DeLong, K. A., Chan, W.h., & Kutas, M. (2019). Similar time courses for word form and meaning preactivation during sentence comprehension. *Psychophysiology, 56*, Article e13312. https://doi.org/10.1111/psyp.13312. URL: https://onlinelibrary.wiley.com/doi/abs/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*, 1044–1063. https://doi.org/10.1080/.23273798.2019.1708960

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*, 631–645. https://doi.org/10.1111/lnc3.12093. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12093.

de Marneffe, M. C., Grimm, S., Arnon, I., Kirby, S., & Bresnan, J. (2012). A statistical model of the grammatical choices in child production of dative sentences. *Language and Cognitive Processes, 27*, 25–61. https://doi.org/10.1080/01690965.2010.542651

Dey, N., Gosal, G., Chen, Z. C., Khachane, H., Marshall, W., Pathria, R., Tom, M., & Hestness, J. (2023). Cerebras-GPT: Open compute-optimal language models trained on the Cerebras Wafer-Scale Cluster. URL: https://arxiv.org/abs/2304.03208v1. arXiv:2304.03208.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '88* (pp. 281–285). Washington, D.C., United States: ACM Press. https://doi.org/10.1145/57167.57214. URL: http://portal.acm.org/citation.cfm?doid=57167.57214.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science, 33*, 547–582. https://doi.org/10.1111/j.1551-6709.2009.01023.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1551-6709.2009.01023.x.

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

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology, 44*, 491–505. https://doi.org/10.1111/j.1469-8986.2007.00531.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8986.2007.00531.x.

Federmeier, K. D. (2021). Connecting and considering: Electrophysiology provides insights into comprehension. *Psychophysiology n/a.* , Article e13940. https://doi.org/10.1111/psyp.13940. URL: https://onlinelibrary.wiley.com/doi/abs/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*, 469–495. https://doi.org/10.1006/jmla.1999.2660. URL: https://linkinghub.elsevier.com/retrieve/pii/S0749596X99926608.

Federmeier, K. D., McLENNAN, D. B., Ochoa, E. D., & Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology, 39*, 133–146. https://doi.org/10.1111/1469-8986.3920133. URL: https://www.cambridge.org/core/journals/psychophysiology/article/impact-of-semantic-memory-organization-and-sentence-context-information-on-spoken-language-processing-by-younger-and-older-adults-an-erp-study/AD2EC099F575F24AA9B4D4E0166EAEC9.

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research, 1146*, 75–84. https://doi.org/10.1016/j.brainres.2006.06.101. URL: https://linkinghub.elsevier.com/retrieve/pii/S0006899306019986.

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*, 1342–1355. https://doi.org/10.1080/23273798.2014.927067

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. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010028518300124.

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. URL: http://www.sciencedirect.com/science/article/pii/S0093934X14001515.

Gerken, L. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition, 98*, B67–B74. https://doi.org/10.1016/j.cognition.2005.03.003. URL: https://www.sciencedirect.com/science/article/pii/S0010027705000727.

Gerken, L. (2007). Acquiring linguistic structure. In *Blackwell handbook of language development* (pp. 173–190). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470757833.ch9 https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470757833.ch9.

Gibbs, A. L., & Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review, 70*, 419–435. https://doi.org/10.1111/j.1751-5823.2002.tb00178.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2002.tb00178.x.

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4*, 178–186. https://doi.org/10.1016/S1364-6613(00)01467-4. URL: https://www.sciencedirect.com/science/article/pii/S1364661300014674.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies 2001 - NAACL '01* (pp. 1–8). Pittsburgh, Pennsylvania: Association for Computational Linguistics. https://doi.org/10.3115/1073336.1073357. URL: http://portal.acm.org/citation.cfm?doid=1073336.1073357.

Halgren, E., Dhond, R. P., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J. D., & Dale, A. M. (2002). N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. *NeuroImage, 17*, 1101–1116. https://doi.org/10.1006/nimg.2002.1268. URL: https://linkinghub.elsevier.com/retrieve/pii/S1053811902912681.

Hodapp, A., & Rabovsky, M. (2021). The N400 ERP component reflects a learning signal during language comprehension. *bioRxiv*. https://doi.org/10.1101/2021.03.25.436922, 2021.03.25.436922. URL: https://www.biorxiv.org/content/10.1101/2021.03.25.436922v1.

Hoeks, J. C. J., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research, 19*, 59–73. https://doi.org/10.1016/j.cogbrainres.2003.10.022. URL: http://www.sciencedirect.com/science/article/pii/S0926641003002866.

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

Hubbard, R. J., Rommers, J., Jacobs, C. L., & Federmeier, K. D. (2019). Downstream behavioral and electrophysiological consequences of word prediction on recognition memory. *Frontiers in Human Neuroscience, 13*. URL: https://www.frontiersin.org/articles/10.3389/fnhum.2019.00291.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research, 1626*, 118–135. https://doi.org/10.1016/j.brainres.2015.02.014. URL: https://www.sciencedirect.com/science/article/pii/S0006899315001146.

Huizeling, E., Arana, S., Hagoort, P., & Schoffelen, J. M. (2022). Lexical frequency and sentence context influence the brain's response to single words. *Neurobiology of Language, 3*, 149–179. https://doi.org/10.1162/nol_a_00054

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. URL: http://www.sciencedirect.com/science/article/pii/S0749596X15001242.

Jain, A. K. (1976). On an estimate of the Bhattacharyya distance. *IEEE Transactions on Systems, Man, and Cybernetics SMC-, 6*, 763–766. https://doi.org/10.1109/TSMC.1976.4309450

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed.) [Online Draft]. URL: 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*, 205–225. https://doi.org/10.1016/j.jml.2004.10.002. URL: http://www.sciencedirect.com/science/article/pii/S0749596X04001159.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. URL: *The Annals of Mathematical Statistics, 22*, 79–86 https://www.jstor.org/stable/2236703. arXiv:2236703.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research, 1146*, 23–49. https://doi.org/10.1016/j.brainres.2006.12.063. URL: http://www.sciencedirect.com/science/article/pii/S0006899306036821.

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*, 12–35. URL: https://www.mitpressjournals.org/doi/abs/10.1162/jocn_a_01465.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*, 32–59. https://doi.org/10.1080/23273798.2015.1102299. URL: https://www.tandfonline.com/doi/full/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*, 117–129. https://doi.org/10.1016/S0926-6410(03)00086-7. URL: https://linkinghub.elsevier.com/retrieve/pii/S0926641003000867.

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). New York, NY, US: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195395518.003.0065.

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*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123. URL: http://www.annualreviews.org/doi/10.1146/annurev.psych.093008.131123.

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

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

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259–284. https://doi.org/10.1080/01638539809545028. URL: http://www.tandfonline.com/doi/abs/10.1080/01638539809545028.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*, 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006. URL: http://www.sciencedirect.com/science/article/pii/S0010027707001436.

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. URL: http://www.sciencedirect.com/science/article/pii/S0010945215000714.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00226. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00226/full.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Merkx, D., & Frank, S. L. (2021). Human sentence processing: Recurrence or attention?. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 12–22). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.cmcl-1.2. URL: https://aclanthology.org/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*, 545–567. https://doi.org/10.1016/j.jml.2012.01.001. URL: http://www.sciencedirect.com/science/article/pii/S0749596X12000034.

Michaelov, J. A., Bardolph, M. D., Coulson, S., & Bergen, B. K. (2021). Different kinds of cognitive plausibility: Why are transformers better than RNNs at predicting N400 amplitude?. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society* (pp. 300–306). Vienna, Austria: University of Vienna (Hybrid).

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2023). Strong prediction: Language model surprisal explains multiple N400 effects. *Neurobiology of Language*, 1–29. https://doi.org/10.1162/nol_a_00105. URL: https://direct.mit.edu/nol/article/doi/10.1162/nol_a_00105/115605/Strong-Prediction-Language-Model-Surprisal.

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

Michaelov, J. A., & Bergen, B. (2022a). Collateral facilitation in humans and language models. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)* (pp. 13–26). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics. URL: https://aclanthology.org/2022.conll-1.2.

Michaelov, J. A., & Bergen, B. K. (2022b). The more human-like the language model, the more surprisal is the best predictor of N400 amplitude. In *NeurIPS 2022 workshop on information-theoretic principles in cognitive systems*. URL: https://openreview.net/forum?id=uCgYvb8GNQZ.

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. (2013a). Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs] URL: http://arxiv.org/abs/1301.3781. arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). 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* (Vol. 26, pp. 3111–3119). Curran Associates, Inc.. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*, 127–162. https://doi.org/10.1016/S0010-0285(03)00128-2. URL: https://www.sciencedirect.com/science/article/pii/S0010028503001282.

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*, Article 107427. https://doi.org/10.1016/j.neuropsychologia.2020.107427. URL: https://linkinghub.elsevier.com/retrieve/pii/S0028393220300981.

Nieuwland, M. S., Martin, A. E., & Carreiras, M. (2013). Event-related brain potential evidence for animacy processing asymmetries during sentence comprehension. *Brain and Language, 126*, 151–158. https://doi.org/10.1016/j.bandl.2013.04.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0093934X13000898.

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*, 691–701. https://doi.org/10.1016/j.cogbrainres.2005.04.003. URL: https://www.sciencedirect.com/science/article/pii/S0926641005001102.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2020). Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4885–4901). Association for Computational Linguistics, Online. https://doi.org/10.18653/.v1/2020.acl-main.441. URL: https://www.aclweb.org/anthology/2020.acl-main.441.

Otten, M., & Berkum, J. J. A. V. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes, 45*, 464–496. https://doi.org/10.1080/01638530802356463

Paczynski, M., & Kuperberg, G. R. (2011). Electrophysiological evidence for use of the animacy hierarchy, but not thematic role assignment, during verb-argument processing. *Language and Cognitive Processes, 26*, 1402–1456. https://doi.org/10.1080/01690965.2011.580143

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*, 426–448. https://doi.org/10.1016/

j.jml.2012.07.003. URL: https://www.sciencedirect.com/science/article/pii/S0749596X12000745.

Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2011). Using language models and latent semantic analysis to characterise the N400m neural response. In *Proceedings of the Australasian Language Technology Association Workshop 2011, Canberra, Australia* (pp. 38–46). URL: https://www.aclweb.org/anthology/U11-1007.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*. Curran Associates, Inc.. URL: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

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, 1456–1469. https://doi.org/10.1111/psyp.12515. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/psyp.12515.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences, 11*, 105–110. https://doi.org/10.1016/j.tics.2006.12.002. URL: https://www.sciencedirect.com/science/article/pii/S1364661307000034.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*, 329–347. https://doi.org/10.1017/S0140525X12001495. URL: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/an-integrated-theory-of-language-production-and-comprehension/B8078F8F7AAEE99DE0579ACF32039B6A.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: https://www.R-project.org/.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners* (Vol. 24).

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*, 79–87. https://doi.org/10.1038/4580. URL: https://www.nature.com/articles/nn0199_79.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *WIREs Cognitive Science, 1*, 906–914. https://doi.org/10.1002/wcs.78. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcs.78.

Rommers, J., & Federmeier, K. D. (2018). Lingering expectations: A pseudo-repetition effect for words previously expected but not presented. *NeuroImage, 183*, 263–272. https://doi.org/10.1016/j.neuroimage.2018.08.023. URL: https://www.sciencedirect.com/science/article/pii/S1053811918307213.

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

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

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. URL: *Science, 274*, 1926–1928 https://www.jstor.org/stable/2891705. arXiv:2891705.

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science, 275*, 1599–1603. https://doi.org/10.1126/science.275.5306.1599. URL: https://www.science.org/doi/full/10.1126/science.275.5306.1599.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sherman, B. E., Graves, K. N., & Turk-Browne, N. B. (2020). The prevalence and importance of statistical learning in human cognition and behavior. *Current Opinion in Behavioral Sciences, 32*, 15–20. https://doi.org/10.1016/j.cobeha.2020.01.015. URL: http://www.sciencedirect.com/science/article/pii/S2352154620300152.

Sherman, B. E., & Turk-Browne, N. B. (2020). Statistical prediction of the future impairs episodic encoding of the present. *Proceedings of the National Academy of Sciences, 117*, 22760–22770. https://doi.org/10.1073/pnas.2013291117. URL: https://www.pnas.org/content/117/37/22760.

Smith, N. J., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. *Proceedings of the Annual Meeting of the Cognitive Science Society, 33, 7*.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*, 302–319. https://doi.org/10.1016/j.cognition.2013.02.013. URL: http://www.sciencedirect.com/science/article/pii/S0010027713000413.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. URL: http://arxiv.org/abs/2206.04615. arXiv:2206.04615.

Stability AI. (2023). StableLM-Base-Alpha 7B. URL: https://huggingface.co/stabilityai/stablelm-base-alpha-7b.

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language, 82*, 1–17. https://doi.org/10.1016/j.jml.2015.02.004. URL: https://www.sciencedirect.com/science/article/pii/S0749596X15000236.

Stone, K., Vasishth, S., & von der Malsburg, T. (2021). Does entropy modulate the prediction of German long-distance verb particles? Data and code. https://doi.org/10.17605/OSF.IO/h75jm.

Stone, K., Vasishth, S., & von der Malsburg, T. (2022). Does entropy modulate the prediction of German long-distance verb particles? *PLoS One, 17*, Article e0267813. https://doi.org/10.1371/journal.pone.0267813. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0267813.

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*, Article 104311. https://doi.org/10.1016/j.jml.2021.104311. URL: https://www.sciencedirect.com/science/article/pii/S0749596X21000942.

Szewczyk, J. M., Mech, E. N., & Federmeier, K. D. (2022). The power of "good": Can adjectives rapidly decrease as well as increase the availability of the upcoming noun? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*, 856–875. https://doi.org/10.1037/xlm0001091

Szewczyk, J. M., & Schriefers, H. (2011). Is animacy special?: ERP correlates of semantic violations and animacy violations in sentence processing. *Brain Research, 1368*, 208–221. https://doi.org/10.1016/j.brainres.2010.10.070. URL: https://www.sciencedirect.com/science/article/pii/S0006899310023395.

Szewczyk, J. M., & Schriefers, H. (2013). Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. *Journal of Memory and Language, 68*, 297–314. https://doi.org/10.1016/j.jml.2012.12.002. URL: http://www.sciencedirect.com/science/article/pii/S0749596X12001295.

Taylor, W. L. (1953). “Cloze procedure”: A new tool for measuring readability. *Journalism Quarterly, 30*, 415–433. https://doi.org/10.1177/107769905303000401. URL: http://journals.sagepub.com/doi/10.1177/107769905303000401.

Taylor, W. L. (1957). “Cloze” readability scores as indices of individual differences in comprehension and aptitude. *Journal of Applied Psychology, 41*, 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*, 382–392. https://doi.org/10.1016/j.ijpsycho.2011.12.007. URL: http://www.sciencedirect.com/science/article/pii/S0167876011003862.

van Erven, T., & Harremos, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory, 60*, 3797–3820. https://doi.org/10.1109/TIT.2014.2320500

Van Petten, C. (1993). A comparison of lexical and sentence-level context effects in event-related potentials. *Language and Cognitive Processes, 8*, 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*, 407–419. https://doi.org/10.1016/j.ijpsycho.2014.10.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167876014016377.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning Memory and Cognition, 25*, 394–417. https://doi.org/10.1037/0278-7393.25.2.394. URL: https://arizona.pure.elsevier.com/en/publications/time-course-of-word-identification-and-semantic-integration-in-sp.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brainpotentials. *Memory & Cognition, 18*, 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*, 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*, 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015. URL: http://www.sciencedirect.com/science/article/pii/S0167876011002819.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Vega-Mendoza, M., Pickering, M. J., & Nieuwland, M. S. (2021). Concurrent use of animacy and event-knowledge during comprehension: Evidence from event-related potentials. *Neuropsychologia, 152*, Article 107724. https://doi.org/10.1016/j.neuropsychologia.2020.107724. URL: https://www.sciencedirect.com/science/article/pii/S0028393220303961.

Vissers, C. T. W. M., Chwilla, D. J., & Kolk, H. H. J. (2006). Monitoring in language perception: The effect of misspellings of words in highly constrained sentences. *Brain Research, 1106*, 150–163. https://doi.org/10.1016/j.brainres.2006.05.012. URL: https://www.sciencedirect.com/science/article/pii/S0006899306013862.

Wang, B., & Komatsuzaki, A. (2021). GPT-J-6B: A 6 billion parameter autoregressive language model. URL: https://github.com/kingoflolz/mesh-transformer-jax.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019a). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32, pp. 3266–3280). Curran Associates, Inc.. URL: https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019b). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=rJ4km2R5t7.

Wang, L., Wlotko, E., Alexander, E., Schoot, L., Kim, M., Warnke, L., & Kuperberg, G. R. (2020). Neural evidence for the prediction of animacy features during language comprehension: Evidence from MEG and EEG representational similarity analysis. *The Journal of Neuroscience, 40*, 3278–3291. https://doi.org/10.1523/JNEUROSCI.1733-19.2020. URL: http://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.1733-19.2020.

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*, 1686. https://doi.org/10.21105/joss.01686

Wlotko, E. W., & Federmeier, K. D. (2007). Finding the right word: Hemispheric asymmetries in the use of sentence context information. *Neuropsychologia, 45*, 3001–3014. https://doi.org/10.1016/j.neuropsychologia.2007.05.013. URL: https://linkinghub.elsevier.com/retrieve/pii/S0028393207002126.

Wlotko, E. W., & Federmeier, K. D. (2012). So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *NeuroImage, 62*, 356–366. https://doi.org/10.1016/j.neuroimage.2012.04.054. URL: http://www.sciencedirect.com/science/article/pii/S1053811912004508.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

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