UNIVERSITY OF CALIFORNIA SAN DIEGO

Understanding the role of statistics in the predictive processing of language

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Cognitive Science with a Specialization in Anthropogeny

by

James A. Michaelov

Committee in charge:

Professor Benjamin K. Bergen, Chair
Professor Seana Coulson
Professor Victor Ferreira
Professor Marta Kutas
Professor Federico Rossano

2024

The dissertation of James A. Michaelov is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I'd like to thank Ben. None of this work would have been possible without his unmatched ability to turn a vague idea into an empirically rigorous and theoretically interesting research question over the course of a single meeting, or his seemingly infinite knowledge on any topic. But perhaps even more importantly, I would like to thank Ben for his generosity and attentiveness as an advisor, as well as for supporting me in following my research interests, even the more unexpected ones. And for his efforts against the thesis gnomes.

I would also like to thank the rest of my committee. Discussions with Seana have deeply shaped the direction of my research, both that included in this thesis and beyond. A big thank you also to Marta for sharing her wisdom and knowledge about the N400, and science and academia in general, as well as her ability to immediately identify the most important aspects of any experiment. I would like to thank Federico for making sure I remember what language is used for, and Vic for his experimental expertise and theoretical knowledge. Thank you all for your continued advice and support.

I would also like to thank the other members of the Language and Cognition Lab: Sean, Cameron, Tyler, Sam, Yoonwon, and Pam. One of the most valuable parts of the PhD has been the many academic discussions I've had with all of you, both scheduled and unscheduled—even (and maybe especially) those that haven't ended up as research projects. And I'm grateful for being able to call you friends outside the lab.

Thank you also to my collaborators beyond the lab, in particular Megan and Cyma, who contributed to the work in Chapter 8. Thank you also to Hannah and Jenny, who helped me take my first steps in psycholinguistics research.

Joining the Anthropogeny Specialization was one of the best decisions of my PhD, and I would like to thank CARTA for the opportunity. I would especially like to thank

Pascal for reigniting my interest in human origins, as well as the graduate students on the Anthropogeny Specialization and the many CARTA members from around the world for making this a truly unique program. Thanks also to Jesse, Matt, Nico, Katie, Meghan, and everyone else for making the Field Course an unforgettable one! Finally, I would like to acknowledge the financial support provided through the CARTA Fellowship and the CARTA Annette Merle-Smith Fellowship.

I would like to deeply thank the many friends I've made for being a second family here in San Diego. Thanks especially to Will and Phil(ip), who have been there from the start, to the other members of my cohort—Andrew, Vijay, Srishti, Alyssa, Yifan, Chaolan, Dalin, and Sam—and to Stephan, Mia, Sean K., Akshay, Quirine, Emily, Victor, Emilia, Ollie, Dillan, Harshada, Tania, Sean H., Davis, Zoe, Parla, Ana, Felix, Oisín, Shuai, Homero, Michael, Nojan, Camryn, Katja, Tiffany, Olivia, Andy, Ryan, Angusina, Ben, Nico, Laura, Anthony, Mark, Nina, and everyone else who I've met along the way. This journey wouldn't have been the same without you. Thanks also to friends back in the UK, and especially to Jess, Dotty, and David for making the journey out here.

Thank you to my family for their unwavering support, their encouragement to pursue my interests and passions, and their willingness to accept me moving halfway around the world to do so. I couldn't have done this without you.

Finally, I would like to thank Cat more than I can express here, for being by my side through all the ups and downs of the last six years, for sharing her adventurous spirit with me, and for making me laugh when I've needed it most.

Chapter 2, in full, is a reprint of the material as it appears in Michaelov, J. A., Coulson, S., & Bergen, B. K., "So Cloze yet so Far: N400 Amplitude is Better Predicted by Distributional Information than Human Predictability Judgements", *IEEE Transactions on Cognitive and Developmental Systems*, 2022. The dissertation author was the primary

investigator and author of this paper.

Chapter 3, in full, is a reprint of a manuscript currently under review for publication as Michaelov, J. A., & Bergen, B. K., "Better language models better model the N400". The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in Michaelov, J. A. & Bergen, B. K., "How well does surprisal explain N400 amplitude under different experimental conditions?", *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL2020)*, 2020. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in Michaelov, J. A. & Bergen, B. K. "Collateral facilitation in humans and language models", *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL 2022)*, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 6, in full, is a reprint of the material as it appears in Michaelov, J. A. & Bergen, B. K., "*Rarely* a problem? Language models exhibit inverse scaling in their predictions following *few*-type quantifiers" *Findings of the Association for Computational Linguistics: ACL 2023*, 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 7, in full, is a reprint of the material as it appears in Michaelov, J. A., Coulson, S., & Bergen, B. K., "Can Peanuts Fall in Love with Distributional Semantics?", *Proceedings of the Annual Meeting of the Cognitive Science Society, 45*, 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 8, in full, is a reprint of the material as it appears in Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S., "Strong Prediction: Language model surprisal explains multiple N400 effects", *Neurobiology of Language*, 2024.

The dissertation author was the primary investigator and author of this paper.

Chapter 9, in full, is a reprint of the material as it appears in Michaelov, J. A. & Bergen, B. K., "Ignoring the alternatives: The N400 is sensitive to stimulus preactivation alone", *Cortex*, 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 10, in full, is a reprint of the material as it appears in Michaelov, J. A. & Bergen, B. K., "On the Mathematical Relationship Between Contextual Probability and N400 Amplitude", *Open Mind*, 2024. The dissertation author was the primary investigator and author of this paper.

VITA

2017	MA (Hons) in Philosophy and Linguistics, University of Edinburgh
2018	MSc in Cognitive Science, University of Edinburgh
2024	PhD in Cognitive Science with a Specialization in Anthropogeny, University of California San Diego

PUBLICATIONS

Michaelov, J. A. & Bergen, B. K., "On the Mathematical Relationship Between Contextual Probability and N400 Amplitude", *Open Mind*, 2024

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S., "Strong Prediction: Language model surprisal explains multiple N400 effects", *Neurobiology of Language*, 2024

Michaelov, J. A.*, Arnett, C.*, Chang, T. A., & Bergen, B. K., "Structural priming demonstrates abstract grammatical representations in multilingual language models", *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, 2023

Michaelov, J. A. & Bergen, B. K., "Emergent inabilities? Inverse scaling over the course of pretraining", *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023

Michaelov, J. A. & Bergen, B. K., "Ignoring the alternatives: The N400 is sensitive to stimulus preactivation alone", *Cortex*, 2023

Michaelov, J. A. & Bergen, B. K., "*Rarely* a problem? Language models exhibit inverse scaling in their predictions following *few*-type quantifiers" *Findings of the Association for Computational Linguistics: ACL 2023*, 2023

Rezaii, N., Michaelov, J. A., Josephy-Hernandez, S., Ren, B., Hochberg, D., Quimby, M., & Dickerson, B. C., "Measuring Sentence Information via Surprisal: Theoretical and Clinical Implications in Nonfluent Aphasia", *Annals of Neurology*, 2023

Trott, S.*, Jones, C.*, Chang, T., Michaelov, J., & Bergen, B., "Do Large Language Models know what humans know?", *Cognitive Science*, 47(7), 2023

Michaelov, J. A. & Bergen, B. K. "Collateral facilitation in humans and language models", *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL 2022)*, 2022

Michaelov, J. A. & Bergen, B. K., "Do language models make human-like predictions about the coreferents of Italian anaphoric zero pronouns?", *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2022)*, 2022

Michaelov, J. A., Coulson, S., & Bergen, B. K., "So Cloze yet so Far: N400 Amplitude is Better Predicted by Distributional Information than Human Predictability Judgements", *IEEE Transactions on Cognitive and Developmental Systems*, 2022

Michaelov, J. A. & Bergen, B. K., "How well does surprisal explain N400 amplitude under different experimental conditions?", *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL2020)*, 2020

ABSTRACT OF THE DISSERTATION

Understanding the role of statistics in the predictive processing of language

by

James A. Michaelov

Doctor of Philosophy in Cognitive Science with a Specialization in Anthropogeny

University of California San Diego, 2024

Professor Benjamin K. Bergen, Chair

In recent years, converging evidence has suggested that prediction plays a role in language comprehension, as it appears to do in information processing in a range of cognitive domains. Much of the evidence for this comes from the N400, a neural index of the processing of meaningful stimuli which has been argued to index the extent to which a word was predicted before it was encountered. The main aim of this thesis is to investigate the extent to which this prediction can be explained as arising from the statistics of the linguistic inputs we receive over the course of our lives, in line with predictive processing in other cognitive domains. To do this, I turn to language models—computational systems

that can calculate the probability of a word given its context based on the statistics of language—and investigate how well their predictions correlate with the N400. The results show that probabilities calculated using language models are highly correlated with N400 amplitude, in many cases better than human-derived metrics such as cloze probability and plausibility, previously the best predictors of the N400. I also show that language model probabilities are able to qualitatively model a wide range of effects, showing significant differences based on the same experimental manipulations that lead to significant differences in N400 amplitude. In addition, the results show that language models that are better able to predict the next word in a sequence are better able to model N400 amplitude in both of these ways, showing both a closer fit to the data and more of the qualitative effects. Taken together, these results show a high degree of correlation between the N400 and predictions based on the statistics of language, consistent with the idea that the predictions indexed by the N400 are at least partly based on language statistics.

Chapter 1

Introduction

The idea that we proactively form expectations about upcoming stimuli has a long history (see, e.g., von Helmholtz, 1867; Bruner, 1951; Postman, 1951; Sanders, 1966). However, it is only in the last few decades that the evidence has mounted up in favor of the idea that prediction plays a role in the processing of stimuli in a wide range of cognitive domains, in particular vision (e.g., Rao and Ballard, 1999; Lee and Mumford, 2003; Summerfield et al., 2008; Alink et al., 2010; Egner et al., 2010; Girshick et al., 2011; Wyart et al., 2012; Kok et al., 2013) and audition (e.g. Wacongne et al., 2011; Rubin et al., 2016; Parras et al., 2017; Wacongne et al., 2011). A key element of many of these contemporary accounts of predictive processing is their convergence upon the idea that perception is the act of matching hierarchically-organized predictions about the world with our sensory inputs, and that we are constantly updating our representations to align better with these inputs, a theory known as *predictive coding* (for reviews, see, e.g., Rao and Ballard, 1999; Friston, 2005; Huang and Rao, 2011; Clark, 2013; de Lange et al., 2018).

A core component of the predictive coding account, therefore, is that our predictions are based on our previous experiences, and thus reflect statistical regularities—both

those of the recent past and over our lifetimes. Predictions based on previous experience over our lifetime are generally observed in the domain of vision through optical illusions and biases. For example, it is often argued that because the reason why the majority of gradient-filled circles in Figure 1.1 are generally perceived as a convex 'bump' and the central circle as a concave 'dimple' is that light tends to fall from above rather than below, and thus, if the gradient-filled circles were indeed three-dimensional, 'bumps' and 'dimples' would appear in this way (for review, see, e.g., Seriès and Seitz, 2013; de Lange et al., 2018). Similarly, it has been shown that the degree to which we over-estimate the orientations of shapes to be cardinal (i.e., horizontal or vertical) is correlated with the (high) proportion of cardinal orientations in the real world compared to other orientations (Girshick et al., 2011).

Prediction based on regularities in the recent past can also observed in experimental contexts. In experiments where participants are trained on sequences of sounds, images, or a combination of the two, researchers observe reduced neural activity in cases where an expected stimulus occurs, and in cases where such an stimulus does not occur, observe activity resembling that which would be elicited by the expected stimulus (Egner et al., 2010; Todorovic and de Lange, 2012; Kok et al., 2012; Ekman et al., 2017; Wacongne et al., 2011).

The aim of this thesis is to investigate the extent to which prediction based on the statistics of past experience can explain processing in a different cognitive domain—language. Stated directly, the thesis addresses the following question: **to what extent does the evidence support the idea that prediction based on the statistics of language occurs as part of the process of language comprehension?**

The general question of whether prediction occurs as part of language comprehension has been debated for decades (for reviews, see, e.g., Van Petten and Luka, 2012;

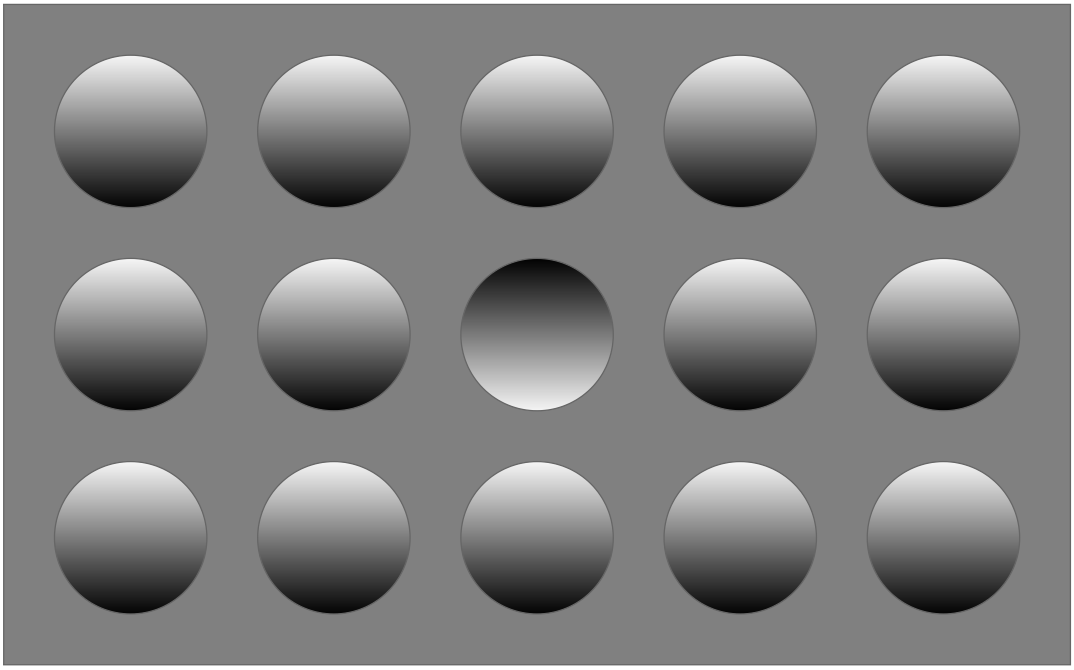

Figure 1.1: An illustration of how light-to-dark and dark-to-light gradients can lead to the illusion of a concave or convex surface. All figures in this chapter were created using *draw.io* (JGraph, 2024).

DeLong et al., 2014b; Luke and Christianson, 2016). One traditional argument opposing the idea observes that there are infinite possible continuations of a given sentence, which intuitively would make prediction (at least of a single candidate) needlessly costly (Forster, 1981; Jackendoff, 2002). However, it has long been known that words that are more congruous with their context or that have a higher contextual probability—generally operationalized as *cloze* probability, the proportion of people to fill in a gap in a sentence with a specific word (Taylor, 1953, 1957)—are read more quickly and recognized more easily than words that are not (see, e.g., Tulving and Gold, 1963; Miller and Isard, 1963; Fischler

and Bloom, 1979; Kutas and Hillyard, 1984). While strictly speaking, such results only directly indicate that words that are more contextually predictable are easier to process, the effect and its reliability have been argued to indicate that words and their meanings are predicted by the language comprehension system (Tulving and Gold, 1963; Luke and Christianson, 2016; Brothers and Kuperberg, 2021).

Stronger evidence for prediction occurring as part of the process of language comprehension comes from research on the N400, a neural index of processing. The N400 is a negative-going component of the event-related brain potential (ERP) that peaks roughly 400ms after the presentation of a meaningful stimulus such as a word. The N400 is described by Kutas and Federmeier (2011) in the following way:

> "The N400 window [...] provides a temporally delimited electrical snapshot of the intersection of a feedforward flow of stimulus-driven activity with a state of the distributed, dynamically active neural landscape that is semantic memory [...] a broad, multimodal neural network, whose current states have been shaped by recent and long-term experience of a wide range of types (e.g., based on world experience, long-standing and recent linguistic and nonlinguistic inputs, attentional states, and affect/mood)."

Mechanistically, then, the N400 can be thought of as reflecting the stimulus-driven activation of representations in long-term memory, which is reduced to the extent that these representations are already activated at the time at which the stimulus was encountered (Kutas and Federmeier, 2011; Van Petten and Luka, 2012; Federmeier, 2021). And generally, this *preactivation* of representations is considered to arise due the context preceding the stimulus (Kutas and Federmeier, 2011; Van Petten and Luka, 2012; DeLong et al., 2014b; Federmeier, 2021; DeLong and Kutas, 2020; Kuperberg et al., 2020). While this has not always been the case (see, e.g., Brown and Hagoort, 1993; Holcomb, 1993; van den Brink and Hagoort, 2004, for an example of an alternative theory), this mechanistic description of the N400 is compatible with the majority of contemporary accounts of the

N400, whether they specifically focus on the retrieval of access of semantic information in long-term memory (e.g., Kutas and Federmeier, 2000; Brouwer et al., 2012; Federmeier, 2021) or on the divergence between predictions and the actual stimulus encountered (Kuperberg et al., 2020, i.e., prediction error; see, e.g.,). Crucially, then, whether or not the N400 should be considered signal of prediction error (as in, e.g., Fitz and Chang, 2019; Hodapp and Rabovsky, 2021; Kuperberg et al., 2020), the fact that it is sensitive to the preactivation of a stimulus means that if preactivation at least partly arises due to prediction, the effects of prediction should be detectable in the N400. In other words, the N400 is a reliable index of prediction, whether or not this what it primarily indexes.

Under this framing of the N400, predictability effects—the fact that words with a higher cloze probability elicit smaller N400s than lower-cloze words—become more direct evidence for prediction in language comprehension than they do in the case of reading time. This is because, with the N400, such results suggest that words are preactivated in a way that is highly correlated with their predictability, and thus, unless there are confounds (for detailed discussion of this, see Chapter 8), this points to prediction. Indeed, there is evidence both specifically for prediction of the semantic content of words as well as specific words and their linguistic features. One such piece of evidence is the related anomaly effect, where words that are semantically related to the highest-probability continuation to a sentence elicit smaller N400s than unrelated words, even when they are not more appropriate sentence continuations (Federmeier and Kutas, 1999; Amsel et al., 2015; Ito et al., 2016; DeLong et al., 2019). For example, DeLong et al. (2019) find that following a context such as *the bartender chilled the champagne over some*, where the highest-cloze continuation is *ice*, words related to this continuation such as *hockey* elicit smaller N400s than unrelated words like *tricks*.

There are also several examples of the prediction of linguistic features. Perhaps

the best-known effect of this kind is the finding that articles elicit a smaller N400 response if they are congruent with a more predictable noun than an unpredictable one. Specifically, DeLong et al. (2005) find that for sentences such as *the day was breezy so the boy went outside to fly* where there is a very high-cloze continuation—in this case *a kite*—the indefinite article *a* elicits a significantly smaller N400 than *an*, with the opposite pattern for sentences where the most predictable noun begins with a vowel. Because in English the only factor that determines which version of the indefinite article will be used is whether the noun begins with a consonant or vowel, this finding suggests that the phonological form of the word *kite* is predicted as well as its meaning. Related effects have also been reported suggesting that the grammatical gender of words can be predicted in a similar way (Wicha et al., 2003a; Foucart et al., 2014; Martin et al., 2018). While such results have in the past been controversial (Martin et al., 2013; Ito et al., 2017a; DeLong et al., 2017; Ito et al., 2017b; Nieuwland et al., 2018b) and in some cases mixed (Wicha et al., 2003b, 2004; Kochari and Flecken, 2019), a recent meta-analysis by Nicenboim et al. (2020) suggests that such effects of the prediction of linguistic features tend to be relatively small—which makes them hard to detect given the level of noise in N400 data—but reliable, and more experiments testing this effect have supported this conclusion (Urbach et al., 2020; Nicenboim et al., 2020; Fleur et al., 2020).

Thus, overall, work on the N400 has provided a substantial amount of evidence for prediction in language comprehension, and that both semantic and formal features of words are predicted. It is perhaps then unsurprising that research on the N400, other later ERP components that appear to reflect other types of prediction (Van Petten and Luka, 2012; DeLong and Kutas, 2020; Kuperberg et al., 2020), and the aforementioned work on prediction based on behavioral metrics of language processing, a number of predictive coding accounts of language comprehension have arisen (Lewis and Bastiaansen, 2015;

Bornkessel-Schlesewsky and Schlesewsky, 2019; Kuperberg et al., 2020; Heilbron et al., 2022).

Returning to the main research question, we have seen that the research does indeed suggest that prediction occurs during the process of language comprehension, and that the N400 specifically indexes such prediction at the level of words and their meanings. But to what extent could these predictions be driven by the statistics of language? A number of studies have demonstrated that the N400 is sensitive to language statistics in that we see that more frequent words elicit smaller N400s (Van Petten and Kutas, 1990; Van Petten, 1993; Dambacher et al., 2006; Rugg, 1990; Fischer-Baum et al., 2014; Shain, 2024). However, in terms of contextual probability, at the time when the first paper (chronologically) of this thesis was being written (see Chapter 4), only four previously-published papers existed on this topic (Parviz et al., 2011; Frank et al., 2013, 2015; Frank and Willems, 2017), and one of these (Parviz et al., 2011) relates to the N400m, the magnetoencephalographic equivalent of the N400, and another (Frank et al., 2015) is an extended version of one of the others (Frank et al., 2013).

The aim of this thesis, therefore, is to evaluate in depth the extent to which prediction based on language statistics can provide an adequate explanation for the the effects we see on the N400. The approach taken to address this questions is outlined in the remainder of this chapter.

1.1 A Theoretical Framework For The Computational Study of the N400

This thesis investigates the extent to which contextual probabilities derived from the statistics of language can be used to model the N400. In this section, I present a

preliminary cognitive model of the N400, the neurocognitive processes that underlie it, and the extent to which statistical probabilities can shed light on these. This section then turns to how the model addresses several key questions in the field.

1.1.1 A cognitive model of the N400

In addition to the previously-discussed mechanistic account of the N400 provided by Kutas and Federmeier (2011), the starting point for the cognitive model of the N400 to be used in this thesis is the predictive coding account of language comprehension provided by Kuperberg et al. (2020), a simplified illustration of which is presented in Figure 1.2.

Under this account, the language system engages in prediction in hierarchical fashion. We make predictions about upcoming content at multiple levels of representation, from the phonological or orthographic to the discourse-level. These predictions are based on the preceding context at the same level and on predictions made at higher levels. When we perceive a linguistic stimulus, differences between the prediction and the true stimulus at each level lead to neural activity that passes 'up' to higher levels of representation—this can be thought of as prediction error.

Under this account, lexical semantic representations (i.e., 'semantic features' in Kuperberg et al., 2020) are activated based on currently-active lexical semantic representations and top-down predictions based on higher-level representations such as knowledge of events—for example, a context such as *at the homestead the farmer penalized the...* (Paczynski and Kuperberg, 2011) might lead to the activation of semantic features relating to an animate entity (almost certainly human) that could be involved in some way with farming. The N400 under this account reflects the amount of new lexical semantic representations that are activated upon encountering a new word—returning to the preceding example, we would expect a smaller N400 for something like *laborer*, which does indeed

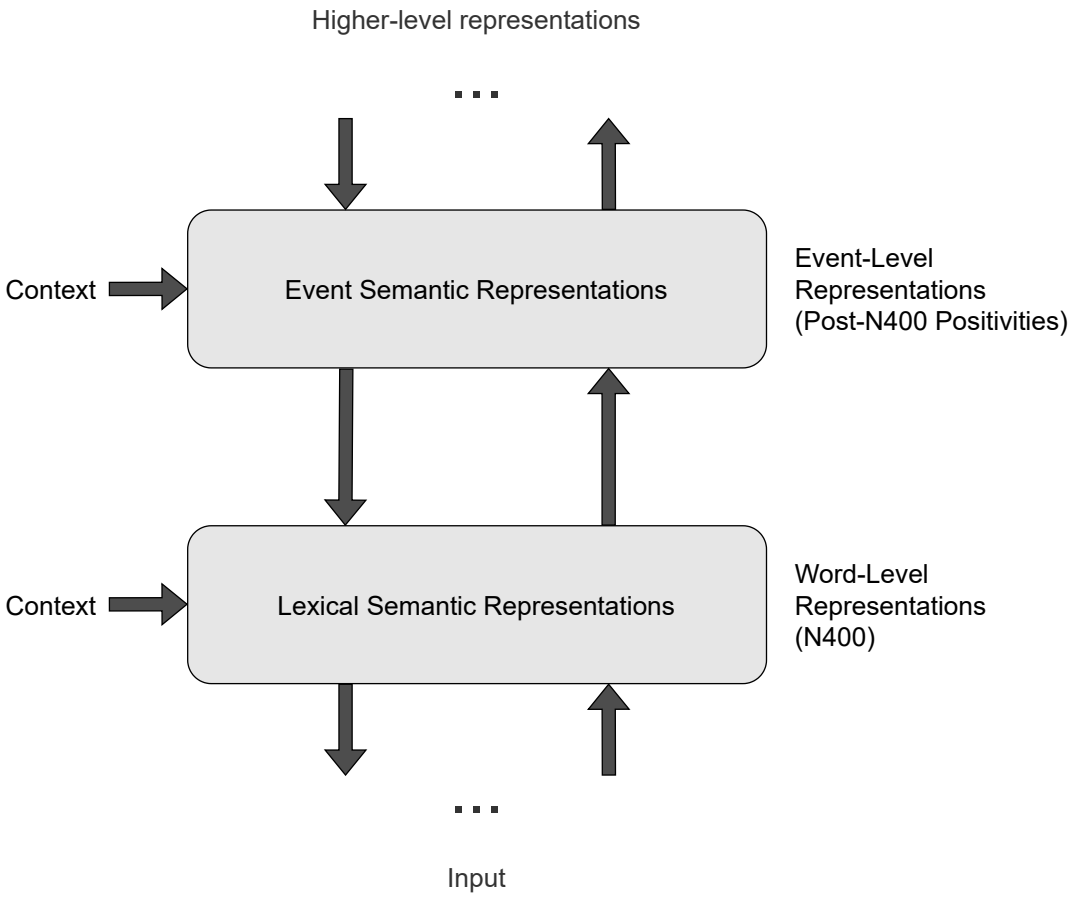

Figure 1.2: A high-level representation of Kuperberg et al.'s account of the N400 and related ERP components.

have the semantic features of being human and possibly being involved in farming; and we would expect a larger N400 for a word like *meadow*, which only shares a few of the features. In fact, this is precisely what Paczynski and Kuperberg (2011) do find. In this way, under the account presented by Kuperberg et al. (2020), the N400 reflects the difference between the predicted semantic features at the word level (i.e., the lexical semantic representations

in Figure 1.2) and those that are newly activated due to encountering the new stimulus.

Kuperberg et al.'s (2020) model also accounts for a range of other results showing that the N400 is relatively insensitive to certain types of semantic violation. Specifically, in cases where a verb that is semantically appropriate given the preceding context takes a inanimate subject when it requires an animate one, no N400 effect is observed. Examples of such sentences are provided in Equation (1) from Kuperberg et al. (2003) and Equation (2) from Kim and Osterhout (2005), where the critical word at which the N400 is measured is bolded.

(1) (a) Every morning at breakfast the *boys* would **eat**...

(b) Every morning at breakfast the *eggs* would **eat**...

(2) (a) The hearty meal was **devoured**...

(b) The hearty meal was **devouring**...

In these cases, under the Kuperberg et al. (2020) account, this incongruity only becomes apparent at a higher level of representation, that of event structure, with prediction error at this level being indexed by later positive components of the event-related brain potential referred to as 'post-N400' (DeLong and Kutas, 2020) or 'late' (Kuperberg et al., 2020) positivities (and previously collectively known as the P600; for discussion see Van Petten and Luka, 2012; DeLong and Kutas, 2020; Kuperberg et al., 2020). Again, this is illustrated in Figure 1.2—at the word level, the semantic features of words such as *eat* and *devour* are predicted based on the *breakfast* and *meal* contexts (respectively) and thus there is no clear difference in the N400. On the other hand, once these representations filter 'up' another level of representation to the level of event semantics, it becomes clear that there is a semantic violation, and thus, we see larger (more positive) post-N400 positivities in such cases.

While this account forms the basis for the model of the N400 used in this thesis, there are several aspects lacking that are important to the question at hand. The first is that it is important to note that semantic knowledge can arise from multiple sources, including both sensorimotor experience (Barsalou, 1999, see, e.g.,), and linguistic input (Marmor, 1978; Saysani et al., 2018; see Section 1.1.2 for further discussion). The second is that it lacks linguistic features. As previously discussed, there is evidence that linguistic features of words are predicted, for example in the case of their phonological form or gender (for review and meta-analysis, see Nicenboim et al., 2020). Words that are orthographically or phonologically similar to highly likely sentence continuations also elicit smaller N400s than words than are not, all else being equal (DeLong et al., 2019; Ryskin et al., 2021). Additionally, in contexts where a rhyme is expected, words that rhyme elicit a smaller N400 response than words that do not (Mantegna et al., 2019). Taken together, such findings suggest not only that linguistic features of words (such as form or grammatical features) are predicted during language comprehension, but that the N400 is sensitive to them.

Thus, we need to update the model proposed by Kuperberg et al. (2020) to also include semantic information derived solely from linguistic input, and to include word-level linguistic features that can be predicted and that therefore impact the N400—specifically, information about the form of the word (i.e., phonology or orthography) and its grammatical features. This updated model is illustrated in Figure 1.3, and forms the basis of the questions asked and addressed in this thesis.

1.1.2 Modeling the N400 computationally

As stated, for the purposes of this thesis, the N400 is taken to index the activation of the word-level representations of a stimulus driven by encountering the stimulus, reduced to the extent that these representations were already activated when the stimulus was

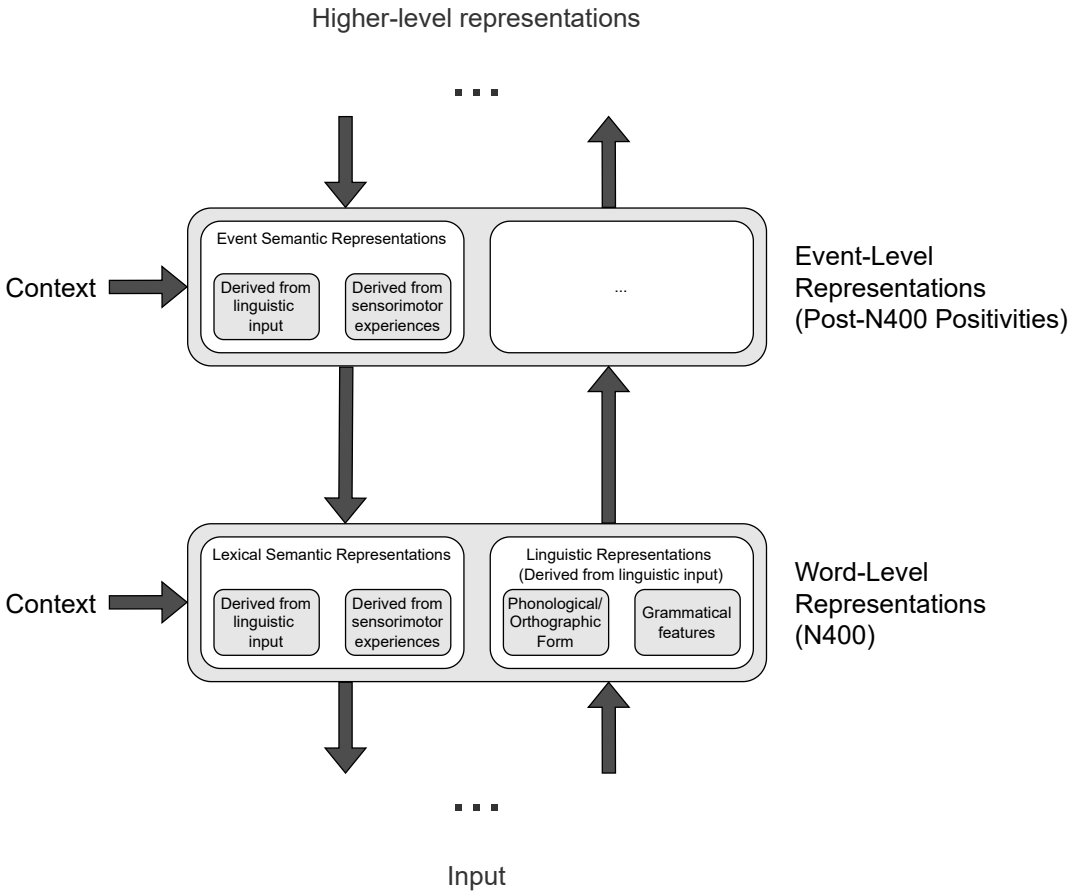

Figure 1.3: A high-level representation of the proposed cognitive model of the flow of information in the neurocognitive systems underlying the N400.

encountered. When considering the idea of this preactivation arising due to prediction, the main way to operationalize this effect is to consider the extent to which a given word is predictable, which is generally operationalized as cloze probability.

To address the main question of this thesis, however, we need to operationalize predictability in a way that is based on the statistics of the language input that an indi-

vidual receives over the course of their life. The most straightforward way to do this is to use a language model. A language model is a computational system that can predict the probability of a word given its context (Jurafsky and Martin, 2024a). Historically, these have taken the form of n-gram models that calculate the probability of a given sequence based on the number of times it occurs in a corpus, and thus, by calculating the probability of the sequence excluding the last word can be used to calculate the contextual probability of a word. More recently, language models have taken the form of neural networks that are trained to predict the probability of a word given its preceding context, with the most prominent architectures being recurrent neural networks (Jordan, 1986; Elman, 1990; Hochreiter and Schmidhuber, 1997) and more recently, transformers (Vaswani et al., 2017; Devlin et al., 2019; Radford et al., 2018).

Despite the fact that such models are trained on corpora—increasingly, vast amounts of text found online—that do not necessarily align with the linguistic experiences of humans, one might worry that this may lead to them being poor models of the language statistics learned by humans, and thus be unsuitable for the tasks at hand. However, in practice, this does not appear to be the case—the predictions of language models of all of these architectures have been shown to have a significant correlation with cloze probability (Smith and Levy, 2011), various metrics of reading time (McDonald and Shillcock, 2003a,b; Levy, 2008; Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2020; Hao et al., 2020; Wilcox et al., 2023a; Shain et al., 2024), and the N400 (Frank et al., 2013, 2015; Frank and Willems, 2017; Aurnhammer and Frank, 2019a,b; Yan and Jaeger, 2020; Merkx and Frank, 2021; Szewczyk and Federmeier, 2022; see also Parviz et al., 2011).

It is also worth noting that while the interpretation of the source of the predictions of language models is straightforward—namely, we know for certain that the predictions arise based on the statistics of text inputs—this is not to say that their predictions are

always transparent. While n-grams explicitly reflect the frequency of specific sequences in a corpus (with possible smoothing, see, e.g., Jurafsky and Martin, 2024a) and thus only learn the very surface-level statistics of language, language models with more advanced architectures—in particular, transformers—are able to learn a wide range of high-level and complex semantic (Wang et al., 2019b,a; Zellers et al., 2019; Bisk et al., 2020; Sakaguchi et al., 2019) and syntactic (Marvin and Linzen, 2018; Warstadt et al., 2019, 2020; Gauthier et al., 2020; Sinclair et al., 2022) properties of words and sequences of words, which may be drawn on in prediction. For example, in one striking study, Abdou et al. (2021) found that language models' representations of how similar colors are to other colors align well with such judgements in humans. Nonetheless, it is still the case that we know the learning of such representations and predictions derived from them are solely based on the statistics of language, and thus clearly interpretable in this way. It is also worth noting that humans can learn similar information from linguistic input alone—for example, a number of researchers (Marmor, 1978; Saysani et al., 2018) have found that congenitally blind participants' judgements of color similarities align with those of sighted participants. Thus, it is not as far-fetched as it may at first appear if predictability effects that hinge on such relationships occur in humans due to prediction based on language statistics rather than, for example, sensorimotor experience of the world (even if such experiences exist).

The overall approach taken in this thesis is to take these corpus-derived probabilities as calculated using a range of language models (specifically, recurrent neural networks and transformers) and to use these to model the N400. Across the thesis, two main approaches are taken—either using these probabilities to directly predict N400 amplitude (following Frank et al., 2013, 2015; Frank and Willems, 2017; and as has become increasingly popular, see, e.g., Aurnhammer and Frank, 2019a,b; Yan and Jaeger, 2020; Merkx and Frank, 2021; Szewczyk and Federmeier, 2022), or investigating whether surprisal quali-

tatively shows the same patterns as the N400—that is, investigating whether experimental manipulations that lead to significant differences in N400 also lead to significant differences in language model probabilities in the same direction (in a similar vein to Ettinger et al., 2016.

Returning to the cognitive model of prediction in language comprehension, Figure 1.4 shows the specific elements targeted in this thesis. Specifically, the approach taken in this thesis can be described as the modeling of the extent to which the activation of representations that are derived from past linguistic inputs (highlighted in blue) can explain the N400. The question of whether or not language models learn event structure or higher-order information from language statistics is currently a debated topic (see, e.g., Li et al., 2021), but it is highlighted on the basis that it is at least in principle possible for this to be the case.

It is also important to highlight that several key assumptions have been made by using the approach taken in this thesis. The first and perhaps strongest assumption made here is that the degree of preactivation of a word's form and meaning scales with the preactivation of the word, or at least can be modeled as such. This is far from a novel assumption. In fact, some version of this is implicit in any work that compares the relative degree to which words are preactivated (with a single value corresponding to the degree of preactivation of each word) and allows for the fact that words can share semantic or grammatical features—i.e., that does not explicitly interpret words as being represented as individual and self-contained units in the brain. Thus, virtually any recent work which operationalizes preactivation using a single value for contextual probability (e.g., Kutas and Hillyard, 1984; Frank et al., 2013, 2015; Aurnhammer and Frank, 2019a,b; Merkx and Frank, 2021; Szewczyk and Federmeier, 2021) or semantic association (e.g., Ettinger et al., 2016; Uchida et al., 2021) makes this assumption. However, it is still important to note

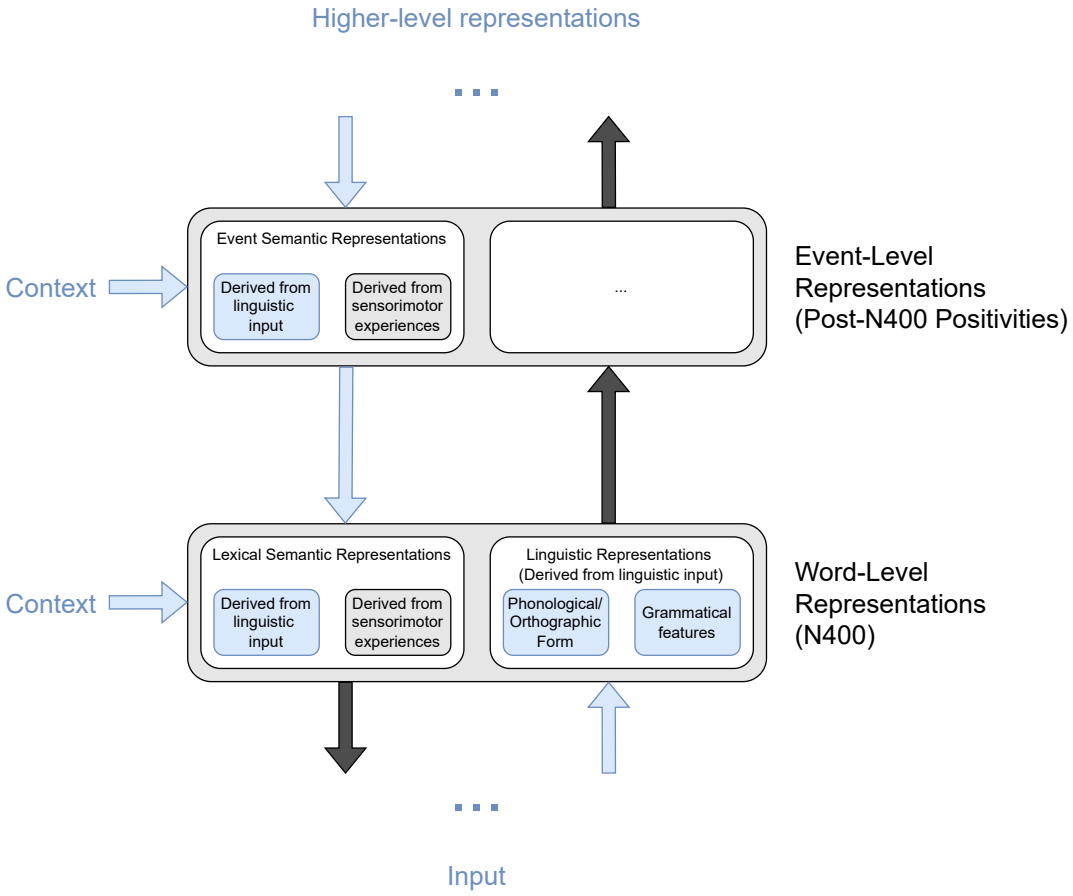

Figure 1.4: The high-level representation of the proposed cognitive model, with the elements under investigation highlighted.

that while such an approach provides a single value for a word based on both its form and meaning, the two are separable experimentally. Research shows, for example, that many of the high-level patterns in the N400 are relatively consistent no matter whether the last word in a sentence is presented as a word or as a picture of its referent (Ganis et al., 1996; Federmeier and Kutas, 2001; for reviews see Kutas and Federmeier, 2000; Federmeier et al.,

2016; Federmeier, 2021)—crucially, in the latter case, stimuli do not contain any linguistic representations.

The second assumption is that we can model the relative preactivation of words as a probability distribution. Some views, such as the *proportional preactivation* account of Brothers and Kuperberg (2021), make this explicit. Brothers and Kuperberg (2021) explain preactivation as occurring when a limited supply of metabolic resources is distributed across possible candidates in proportion to their probability. Under other accounts such as that of Federmeier (2021), this is explicitly not the case: Federmeier (2021) argues that the relative degree to which words are preactivated is independent. In practice, the difference between these accounts is not usually important to the research question at hand, as the focus is generally on the difference between experimental conditions for each experimental item, and this can often be controlled for in statistical analyses by adding a random slope or intercept for each item. However, if preactivation is not competitive, it may be that differences between specific items are under-estimated both by modeling relative preactivation as probability and by controlling for item. Investigating the limits of modeling preactivation using a probability distribution is likely to be an important avenue of research, but given that language models are trained to output probabilities, is beyond the scope of this thesis. Additionally, stimulus item is controlled for in all studies, and so this issue is not likely to be a confound.

These two key assumptions bring the high-level approach to computational modeling in this thesis in line with the majority of current approaches to investigating how preactivation (and by extension, prediction) impacts reading time and the N400, even if these assumptions are not generally stated as such. Specifically, ever since it was established that cloze probability is correlated with behavioral (Fischler and Bloom, 1979) and neural (Kutas and Hillyard, 1984) indices of language processing, using probabilities to

operationalize predictability or the extent to which words or their meanings are preactivated has been widespread (for reviews of such work, see, e.g., Kutas and Van Petten, 1994; Van Petten and Luka, 2012; Luke and Christianson, 2016; Kutas and Federmeier, 2011; DeLong et al., 2014b; Kuperberg and Jaeger, 2016; Brothers and Kuperberg, 2021; Kuperberg et al., 2020).

1.2 Thesis outline

This thesis is divided into three main parts. In the first, the focus is on investigating how closely the predictions of language models of different types correlate with N400 amplitude. In Chapter 2, I present a study investigating how the extent to which the predictions of language models of three different architectures—recurrent neural networks with long short-term memory (Gulordava et al., 2018; Jozefowicz et al., 2016), autoregressive language models (Dai et al., 2019; Radford et al., 2019; Brown et al., 2020), and masked language models (Devlin et al., 2019; Lan et al., 2020; Liu et al., 2019)—each predict N400 amplitude, comparing both architecture and model size (in terms of number of parameters and training tokens). Chapter 3 takes this latter question further, teasing apart how number of parameters and training tokens affect the correlation between language model predictions and N400 amplitude using the Pythia (Biderman et al., 2023b) suite of language models, as well as trying to better understand the cause of this by also comparing the performance of each model at next-word prediction (Jelinek et al., 1977; Merity et al., 2017) and 5 natural language processing benchmarks (Warstadt et al., 2020; Paperno et al., 2016; Zellers et al., 2019; Bisk et al., 2020; Sakaguchi et al., 2019).

While Part 1 of this thesis looks at the overall correlation between language model predictions and the N400, Part 2 investigates whether the patterns in language model predictions and the N400 are the same—that is, whether statistically significant

patterns in the N400 response to stimuli correspond to statistically significant patterns in language model surprisal to the same stimuli. A range of N400 effects are investigated in this way in Chapters Chapter 4–Chapter 7. Chapter 8 then takes this a step further, investigating whether the language model predictions can explain variance N400 amplitude previously attributed to other factors such as plausibility or contextual similarity; thereby evaluating the extent to which the statistics of language can provide an explanation for previously-known N400 effects.

Parts 1 and 2 follow previous work (Frank et al., 2013, 2015; Frank and Willems, 2017) in transforming the statistical (i.e., language-model-based) probabilities into surprisal (negative log-probability) before using them to predict N400 amplitude. Part 3 investigates whether this is truly the best way to operationalize the relationship between the probabilities calculated by language models and the N400. First, Chapter 9 asks the question of whether predicted (statistical) probabilities of words other than the actual stimulus have an impact on N400 amplitude, which is found not to be the case. Next, Chapter 10 investigates the specific mathematical relationship between statistical probability and N400 amplitude, comparing un-transformed probability, surprisal, and surprisal to the power of a range of numbers between zero and two, following similar analyses in the reading time literature (Meister et al., 2021; Shain et al., 2024).

Finally, Chapter 11 provides a brief discussion and conclusion of the results of Chapter 2–Chapter 10, highlighting key takeaways and avenues for future research.

Part I

How well can language models model the N400?

Chapter 2

So Cloze yet so Far: N400 Amplitude is Better Predicted by Distributional Information than Human Predictability Judgements

Abstract

More predictable words are easier to process—they are read faster and elicit smaller neural signals associated with processing difficulty, most notably, the N400 component of the event-related brain potential. Thus, it has been argued that prediction of upcoming words is a key component of language comprehension, and that studying the amplitude of the N400 is a valuable way to investigate the predictions we make. In this study, we investigate whether the linguistic predictions of computational language models or humans better reflect the way in which natural language stimuli modulate the ampli-

tude of the N400. One important difference in the linguistic predictions of humans versus computational language models is that while language models base their predictions exclusively on the preceding linguistic context, humans may rely on other factors. We find that the predictions of three top-of-the-line contemporary language models—GPT-3, RoBERTa, and ALBERT—match the N400 more closely than human predictions. This suggests that the predictive processes underlying the N400 may be more sensitive to the statistics of language than previously thought.

2.1 Introduction

While it is widely accepted that predictable words are easier to process than unpredictable ones, the role of predictive processes in language comprehension has long been an issue of contentious debate (for reviews, see (Kutas et al., 2011; Van Petten and Luka, 2012; Luke and Christianson, 2016; Kuperberg and Jaeger, 2016)). One prominent position is that the language processor does not waste resources on predictive processing (Forster, 1981). Under such an account, because there are an infinite number of possible continuations for any given natural language string, linguistic predictions would be wrong far more often than they would be right. Thus, given the limited value of linguistic prediction, the language processor simply does not engage in it (Jackendoff, 2002). Advocates of this position have attributed observed predictability effects on language processing to the demands of integrating the meaning of a word into its preceding context (Schwanenflugel and Shoben, 1985; Traxler and Foss, 2000), some form of automatic spreading activation in the lexicon (West and Stanovich, 1982; Collins and Loftus, 1975), or both.

However, there is growing evidence in support of prediction as a component of language comprehension. Much of this research comes from looking at neural signals of processing difficulty, especially the N400, a negative-going component of the event-related

brain potential (ERP) that peaks roughly 400ms after the presentation of a meaningful stimulus (Kutas and Hillyard, 1980; Kutas and Federmeier, 2011). With linguistic stimuli, the size of the N400 is sensitive to semantic congruity—N400 amplitude is large by default, and is reduced if the word is facilitated by the preceding context (Van Petten and Luka, 2012; DeLong and Kutas, 2020; Kuperberg et al., 2020). In recent years, a range of studies have found that N400 amplitude modulations appear to reflect lexical properties of specific nouns that are semantically predictable; thus, researchers have argued that N400 predictability effects do not simply reflect ease of integration or spreading activation, and—at least some of the time—provide evidence for predictive processes in language comprehension (DeLong et al., 2005; Van Berkum et al., 2005; Otten et al., 2007; Kwon et al., 2017; Kuperberg et al., 2020; Nicenboim et al., 2020; Urbach et al., 2020; Fleur et al., 2020).

What are these predictions based on? Since the early days of N400 research, cloze probability (Taylor, 1953) has served as the chief metric of contextual word predictability (Kutas and Hillyard, 1984; Van Petten and Luka, 2012; Brothers and Kuperberg, 2021). The cloze probability of a given word is defined as the proportion of people who fill a gap in a sentence with that specific word (Taylor, 1953), and thus, provides a measure of how predictable a word is in a specific sentence context. It is well-established that words with a higher cloze probability elicit a smaller N400 response compared to words with lower cloze probabilities (Kutas and Hillyard, 1984; Kutas and Federmeier, 2011; Kuperberg et al., 2020), as well as being read faster and recognized faster (Brothers and Kuperberg, 2021)—in fact, some work has shown that cloze probability and N400 amplitude are inversely correlated at a level of over 90% (Kutas and Van Petten, 1994). A more recent operationalization of predictability is derived from language models (LMs), computational systems designed to predict a word in context. Unlike humans, these LMs are only trained

on text data as input, and consequently base their predictions solely on the statistics of language (Jurafsky and Martin, 2021). Thus, while linguistic predictions in humans may utilize a range of knowledge both linguistic and extra-linguistic, LMs learn the actual distributional probability of a word in context in the corpus on which they are trained (Smith and Levy, 2011; Brothers and Kuperberg, 2021).

Understanding the relationship between N400 amplitude and the statistics of language is vital to understanding the N400 (Michaelov and Bergen, 2020). Given the evidence that N400 amplitude is affected by linguistic input over the lifespan (Kutas and Federmeier, 2011), and the fact that they are models trained purely on linguistic input, LMs give us a precise way to model the extent to which linguistic input alone can predict the N400 response. On the other hand, there is no way to tell which sources of information and neurocognitive processes are involved when experimental participants complete the cloze task. Thus, even if cloze probability were to correlate more closely with N400 amplitude than LM predictions, it is less informative in terms of illuminating the basis of prediction in language comprehension.

However, recent work suggests that this trade-off between accuracy and explainability may be nearing an end. The statistics of language—as operationalized by LM predictions—can not only successfully predict single-trial N400 amplitudes (Frank et al., 2015; Aurnhammer and Frank, 2019b; Merkx and Frank, 2021; Michaelov et al., 2021) and the significant differences in N400 amplitude elicited by a range of experimental manipulations (Michaelov and Bergen, 2020), but at least for some stimuli may be better at this than cloze probability (Michaelov and Bergen, 2020; Michaelov et al., 2021). However, the two studies in which LM predictions outperform cloze have either looked at the effects without direct comparison to the N400 data (Michaelov and Bergen, 2020) or targeted data from an experiment intended to show the N400 responds to factors other than cloze (Michaelov

et al., 2021).

The goal of the present study is to test whether the amplitude of the N400 to words in sentence contexts can be better predicted by the statistics of language than by cloze probability—even under conditions that are maximally favorable to cloze. Using ERP data from a large-scale multiple-laboratory experiment (Nieuwland et al., 2018b), we used linear mixed effects regression models to examine how well the amplitude of the N400 elicited by experimental stimuli was predicted by the cloze probabilities gathered in the original experiment (Nieuwland et al., 2018b), and compared its performance to that of several pretrained neural network LMs (Gulordava et al., 2018; Jozefowicz et al., 2016; Devlin et al., 2019; Liu et al., 2019; Radford et al., 2019; Dai et al., 2019; Lan et al., 2020; Brown et al., 2020). Language models are the best way to capture prediction based on language statistics at present. If any contemporary models predict N400 amplitude better than cloze probability does, that would constitute compelling evidence that prediction, as measured by the N400, can be driven by language statistics.

2.2 Background

2.2.1 Cloze probability

Cloze probability has long been used to assess a word's predictability in context (Van Petten and Luka, 2012; DeLong et al., 2014b; Luke and Christianson, 2016; Brothers and Kuperberg, 2021). In addition to its use in understanding the N400 (Kutas and Hillyard, 1984; Kutas and Federmeier, 2011), it has been shown to predict behavioural correlates of processing difficulty, such as word reading time (Brothers and Kuperberg, 2021). In fact, when directly compared, cloze probability has previously been found to be better at predicting such behavioural metrics than LMs (Brothers and Kuperberg, 2021).

However, while cloze probability is a metric grounded in human judgements, it may not be as helpful in understanding online human comprehension as might appear at first glance. As discussed, predictability effects are thought to arise from individuals' graded predictions about upcoming words, whereas cloze probability is an aggregate measure over a sample of individuals based exclusively on their top predictions. In addition to the question of whether we should expect these two distributions to be equivalent, there is also a practical issue of sample size—less likely continuations require a larger sample of individuals in order for even a single experimental participant to produce. Indeed, as a language production task, its relevance for comprehension is unclear in view of disagreement regarding the extent of overlap between the production and comprehension systems (see Meyer et al., 2016; Hendriks, 2014 for review and discussion), it is not necessarily the case that the next-word probability of a word will be the same for both the production and comprehension system.

Beyond these concerns, and even if cloze is a good predictor of processing difficulty due to predictability overall (e.g. as measured by reading time), when investigating the N400, the temporal dimension must also be considered. Cloze probability is based on responses produced by experimental participants after reading a sentence with a gap that must be filled in. Given the substantial evidence that there are neurocognitive processes involved in human language comprehension that occur after the N400 (DeLong and Kutas, 2020; Kuperberg et al., 2020), even if it is the case that the N400 and cloze probability both reflect individuals' graded predictions, and that cloze responses are influenced by the predictions that underlie the N400 response, it should not be taken as a given that these predictions are the same. Thus, there is no *a priori* reason to assume that cloze probability is the best possible operationalization of the predictions that underlie the N400.

2.2.2 Language model predictions

LMs are trained to predict the probability of a word based only on the linguistic context. Given that such models do not explicitly learn meanings of words, and that the N400 response to a word is thought to be largely or wholly determined by meaning (Kutas and Federmeier, 2011; Kuperberg et al., 2020), intuitively, we may expect them to perform poorly at predicting the amplitudes of N400 responses to words. However, previous research has shown that LMs can learn semantic relationships between words (Rogers et al., 2021). Thus, the extent to which LMs can acquire semantic knowledge, and specifically, knowledge about the semantic relations between words, may be greater than would be expected *prima facie*. Whether or not humans can learn quite so much based on only linguistic input is an open question, but there is evidence that we may learn semantic relations between referents of words with which we have no direct experience (Marmor, 1978).

An additional benefit of using LM predictions to operationalize word predictability is that researchers know exactly what sources of information are used by these models—they are trained on specific data, and thus researchers can form hypotheses about how the specific kinds of information in these data may be used to predict upcoming linguistic input, and by which system. This is especially important given that, as discussed, we might expect the predictions underlying the N400 to also impact cloze probability. If factors beyond linguistic input such as world knowledge have an effect on N400 amplitude, as has been proposed (Kutas and Federmeier, 2011), then they are also likely to have an effect on cloze probability. For this reason, when using cloze probability to predict N400 amplitude, it may be impossible to disentangle the effect of each source of information, and thus limiting the extent to which we can understand the basis upon which the predictions underlying the N400 are made. Using metrics based on the statistics of language (for example, LM predictions) may therefore be one of the only ways to successfully isolate the

specific effect of linguistic input on N400 amplitude.

2.2.3 Language model surprisal

When LM predictions are used to investigate predictability effects on language comprehension, predictability is usually not operationalized as the raw probability of words as calculated by these models, but rather, their surprisal. The surprisal S of a word w_i is the negative logarithm of its probability given its preceding context $w_1...w_{i-1}$, as shown in (2.1).

$$S(w_i) = -\log P(w_i|w_1...w_{i-1}) \tag{2.1}$$

In addition to theoretical claims behind surprisal theory as an explanation of predictability effects in language comprehension (Hale, 2001; Levy, 2008; Smith and Levy, 2013), there is also an array of evidence showing that LM surprisal correlates with behavioural metrics of processing difficulty such as reading time (Boston et al., 2008; Demberg and Keller, 2008; Roark et al., 2009; Mitchell et al., 2010; Smith and Levy, 2011; Monsalve et al., 2012; Willems et al., 2016). A further body of research has found that LM surprisal is a significant predictor of N400 amplitude, with the surprisal of generally better-performing and more advanced LMs showing a better fit to the N400 data (Frank et al., 2015; Aurnhammer and Frank, 2019b; Merkx and Frank, 2021; Michaelov et al., 2021). Additionally, when LMs are given the same experimental stimuli as humans in neurolinguistic experiments, significant differences in surprisal often match significant differences in N400 as a function of experimental condition—again, with generally better-performing and more advanced models matching the human responses better (Michaelov and Bergen, 2020; Michaelov et al., 2021).

In previous work, operationalizing predictability as cloze probability generally appears to yield better results for human behavioural data than LM surprisal (Brothers

and Kuperberg, 2021); however, this has not been well-explored for the N400. To the best of our knowledge, only one published paper has directly compared how well cloze probability and LM surprisal predict N400 amplitude, finding that LM surprisal performs better (Michaelov et al., 2021). However, the comparison between cloze probability and LM prediction was not an aim of that previous study, and thus there are several caveats to be noted about this result. Firstly, the study investigated the N400 response to words with the same cloze probability but which were either related or unrelated to the highest-cloze completion—there is a well-established effect showing that the former elicit lower-amplitude N400s than the latter (Kutas and Hillyard, 1984; Kutas, 1993; Federmeier and Kutas, 1999; Thornhill and Van Petten, 2012; Ito et al., 2016). Thus, cloze is inherently at a disadvantage in prediction, given that the two conditions are controlled for cloze. The study also involved a condition where all stimuli had a cloze of zero; thus, none of the variance in N400 amplitude within this condition could be explained by cloze. Finally, the study compared raw cloze probability to LM surprisal—given that the surprisal calculated from cloze probability has been found to correlate with behavioural predictability effects (Smith and Levy, 2011; Lowder et al., 2018), a fair comparison would also involve cloze surprisal. The finding that surprisal can differ between words that are matched for cloze but either related or unrelated to the highest-cloze continuation of a sentence is also found in another study (Michaelov and Bergen, 2020), but this study only compares significant differences in surprisal to the significant differences reported in the original papers—there is no direct comparison made between the surprisal and N400 data.

2.2.4 The present study

In the present study, we aim to provide just such a fair comparison using modern LMs and openly available data from a large N400 study (n = 334) (Nieuwland et al.,

2018b). First, we use data from a study that was specifically designed to investigate the effect of cloze probability on N400 amplitude; thus, there are none of the aforementioned cases where experimental conditions are matched by cloze and differ in another way (that may be reflected in LM predictions, see (Michaelov and Bergen, 2020; Michaelov et al., 2021)). Additionally, we remove the data from all stimuli with a cloze probability of zero. Given that previous work has shown that there is variability in N400 amplitude between experimental conditions where all items had a cloze probability of zero (Metusalem et al., 2012; Ito et al., 2016), and some of these studies have been successfully modeled using LM predictions (Michaelov and Bergen, 2020), there is a chance that including these would give the LMs an unfair advantage. Finally, we compare both raw cloze probability and cloze surprisal to ensure that the log-transformation of LM probability is not a confound, as previous work has suggested that there may be a logarithmic linking function between human-derived metrics of word probability and processing difficulty (Smith and Levy, 2011; Lowder et al., 2018; Delaney-Busch et al., 2019).

2.3 Method

2.3.1 Original study and data

We use EEG data from a large-scale experiment by Nieuwland and colleagues (Nieuwland et al., 2018b). In this experiment, participants read sentences one word at a time, with ERPs time-locked to previously-determined target words. In the data provided, the N400 is operationalized as the mean amplitude voltage recorded from the centro-parietal region of interest (electrodes Cz, C3, C4, Pz, P3, and P4) 200–500ms after the presentation of the target word. We use the data provided for target nouns, which replicate the well-established finding that higher-cloze nouns elicit smaller (less negative) N400 responses

than lower-cloze nouns (Nieuwland et al., 2018b; Kutas and Hillyard, 1984; Kutas and Federmeier, 2011).

To calculate the cloze probability of items in the original study, each stimulus sentence was truncated before the target word (Nieuwland et al., 2018b). Thus, participants in the cloze task were presented with the preceding linguistic context for the target word and asked to complete the sentence. The cloze probabilities were then calculated on the basis of the responses from two sets of 30 participants, each of which completed the cloze task for half of the total stimulus sentences. The authors provide both the cloze and ERP data online (at https://osf.io/eyzaq/).

The electrophysiological experiment was carried out at 9 laboratories in the United Kingdom and comprises data from 334 participants, reaching a total of 25,849 trials. We removed all items with a cloze probability of zero for fair comparison with LM surprisal, as previously discussed. Finally, we used the cloze data to calculate cloze surprisal for each remaining item. Because all zero-cloze items were removed, this also removed the need for smoothing zero-probabilities, as has been done in previous related work (Lowder et al., 2018).

2.3.2 Language models

We operationalize corpus-based probability of a word in context as the probability calculated by a neural network LM. There are many different architectures for neural network LMs, some of which have been used to model behavioural and neural correlates of human language processing. Here we focus on the two most prolific and successful types of LM in recent years—RNNs and transformers.

RNNs

Until the development of transformer LMs (Vaswani et al., 2017), recurrent neural network (RNN) language models long dominated the field. With their memory bottleneck and their incremental processing of words (Keller, 2010; Merkx and Frank, 2021), RNNs have often been used as cognitive models of human language processing (Elman, 1990), including prior efforts to model the N400 (Frank et al., 2015; Aurnhammer and Frank, 2019b; Michaelov and Bergen, 2020; Merkx and Frank, 2021; Michaelov et al., 2021). In the present study, we use two RNN LMs referred to in the literature (see, e.g., (Futrell et al., 2019)) as GRNN (Gulordava et al., 2018) and JRNN (Jozefowicz et al., 2016). Previous research has found JRNN surprisal to more closely resemble N400 amplitude than does GRNN surprisal (Michaelov and Bergen, 2020). GRNN and JRNN surprisal were calculated using the code accompanying Michaelov and Bergen (Michaelov and Bergen, 2020).

Transformers

Transformer language models are a neural network LM architecture (Vaswani et al., 2017) that has been found to outperform RNNs at the standard language modeling task (predicting words from context, see (Dai et al., 2019) for review), as well as a range of other tasks (Devlin et al., 2019; Radford et al., 2019). Transformer LMs have also been shown to do better than RNNs at predicting N400 amplitude (Merkx and Frank, 2021; Michaelov et al., 2021). The present study includes two varieties of transformer LMs—*autoregressive language models* trained on the traditional task of predicting words based on their preceding linguistic context, and *masked language models*, trained to fill a gap in a sentence, and that thus can use words that appear both before and after in its prediction of the target word. We include the probabilities from three autoregressive LMs in our analysis—Transformer-XL (Dai et al., 2019), GPT-2 (Radford et al., 2019), and GPT-3

(Brown et al., 2020). The three masked LMs that we use to calculate word probability are BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020). For all transformer LMs except for GPT-3, we use the implementation of each model made available through the *transformers* (Wolf et al., 2020) package to calculate surprisal. GPT-3 predictions were accessed via the OpenAI API (OpenAI, 2021).

Table 2.1: Summary of language models used

Model	**Parameters**[1]	**Corpus size**[2]	**Ref.**
GRNN	71.8M	90M	(Gulordava et al., 2018)
JRNN	1.04B	1B	(Jozefowicz et al., 2016)
Transformer-XL[3]	285M	103M	(Dai et al., 2019)
GPT-2 (XL)	1.56B	∼8B	(Radford et al., 2019)
GPT-3 (Davinci)	175B	∼300B	(Brown et al., 2020)
BERT (L, C, WWM[4])	334M	3.3B	(Devlin et al., 2019)
RoBERTa (large)	355M	∼33B	(Liu et al., 2019)
ALBERT (XXL v2)[5]	206M	3.3B	(Lan et al., 2020)

[1] The number of free parameters for the *transformers* (Wolf et al., 2020) implementations of Transformer-XL, GPT-2, BERT, RoBERTa, and ALBERT were calculated using *pytorch* (Paszke et al., 2019). For JRNN and GPT-3, we utilized the models directly provided by the authors of the paper, and so use the number of parameters reported in the cited paper or its supplementary materials (Jozefowicz et al., 2016; Brown et al., 2020). While we use the author-provided GRNN, no estimate of model parameters is given in the original paper (Gulordava et al., 2018), so we calculated this with *pytorch* (Paszke et al., 2019).

[2] Number of words in training corpus is reported in the original papers (Gulordava et al., 2018; Jozefowicz et al., 2016; Dai et al., 2019; Devlin et al., 2019), or estimated (denoted by '∼'). ALBERT is trained on the same data as BERT (Lan et al., 2020). Training data for GPT-2 and RoBERTa are estimated based on a comparison of file size with the dataset used for BERT. GPT-3 is trained on 300 billion tokens; however, given that it uses byte-pair encoding for tokenization (Brown et al., 2020; Radford et al., 2019; Sennrich et al., 2016), the actual number of words is lower.

[3] We use the *transformers* (Wolf et al., 2020) implementation of Transformer-XL; some models reported in the original paper (Dai et al., 2019) have a higher number of parameters.

[4] Large, cased, whole-word masking, (see Devlin et al., 2019; Google Research, 2019).

[5] Note that while ALBERT has fewer free parameters than either BERT or RoBERTa, it shares parameters between layers, and so is actually a much larger model than either BERT or RoBERTa (Lan et al., 2020).

2.3.3 Language model predictions

The aforementioned LMs were thus used to predict the probability of the target nouns from the original study (Nieuwland et al., 2018b). Each stimulus sentence was truncated before the target word and the predicted probabilities generated by the models for each of the target words were recorded. Thus, all the models, including the masked LMs, were required to base their predictions on the preceding context. This procedure was intended to match the cloze task, where sentences were truncated in the same way, as well as the ERP experiment, where experimental participants had read only the preceding context when they reached the target word. These probabilities were then transformed into surprisals using the formula in (2.1). We used a logarithm of base 2 so that surprisal can be measured in bits (Futrell et al., 2019). For fair comparison, only words appearing in all models' vocabularies were included in the analysis.

2.3.4 Predicting the N400

The LM surprisal values, original cloze values, cloze surprisal values, and by-trial N400 amplitudes were all z-transformed before running statistical analyses. These z-transformed LM surprisals, cloze surprisals, and cloze probabilities were then used to predict the z-transformed by-trial N400 amplitudes. After the removal of data for all target words that either did not appear in all LMs' vocabularies or that had a cloze probability of zero, our final dataset consisted of N400 data from 15,551 trials, elicited by 94 different sentences.

Statistical analysis and data manipulation were carried out in *R* (R Core Team, 2020) using *Rstudio* (RStudio Team, 2020) and the *tidyverse* (Wickham et al., 2019), *lme4* (Bates et al., 2015), and *ggh4x* (van den Brand, 2021) packages, and the code provided by Nicenboim et al. (Nicenboim et al., 2020) for preparing the data (Nieuwland et al., 2018b).

To reduce the risk of Type I errors, all p-values in our analyses are corrected for multiple comparisons based on false discovery rate (Benjamini and Yekutieli, 2001).

2.4 Results

2.4.1 Preliminary analysis with cloze probability

First, we test whether the original finding, that higher-cloze nouns elicit smaller N400s than lower-cloze nouns, still holds for our subset of the data. We did this by following the original statistical methods as closely as possible (Nieuwland et al., 2018b). For this reason, we used linear mixed-effects regression models with the same covariates as in the original analyses; and in order to test the significance of variables, we use likelihood ratio tests on nested regressions.

After running all regressions (including those described in the following subsections), we found that including the original random effect structure of random slopes for experimental participant and item resulted in singular fits in several cases; so these were reduced to random intercepts in all models. Following the original analysis, we also included the laboratory in which the experiment was run as a fixed effect.

As in the original study, we found no interaction between cloze probability and laboratory ($\chi^2(8) = 7.357, p = 1$). However, unlike the original study, we found a significant effect of laboratory even when controlling for cloze probability ($\chi^2(8) = 36.280, p < 0.001$). This may be due to the difference in sample or in random effects structure. Crucially, we found a significant effect of cloze probability even when controlling for laboratory ($\chi^2(1) = 27.937, p < 0.001$). Thus, we replicated the noun predictability effect on our subset of the data.

2.4.2 Cloze surprisal and N400 amplitude

Running the same tests with cloze surprisal (i.e. negative log-transformed cloze probability) replacing cloze probability leads to the same results (Cloze surprisal x lab: $\chi^2(8) = 3.596, p = 1$; cloze surprisal: $\chi^2(1) = 29.403, p < 0.001$; lab: $\chi^2(8) = 36.241, p < 0.001$). Thus, we included laboratory as a covariate for our remaining analyses.

To compare cloze probability and cloze surprisal as predictors of N400, we compared the two best regressions including each as a main effect—namely, those also including laboratory as a main effect but not the interaction between the two. Since the two regressions are not nested, we employed Akaike's Information Criterion (AIC) (Akaike, 1973) to compare them. We found that the regression with cloze surprisal as a fixed effect has a slightly lower AIC (AIC = 113227.2) than the regression with cloze probability as a fixed effect (AIC = 113228.7).

These AIC values can be used to calculate evidence ratios based on Akaike weights (see (Wagenmakers and Farrell, 2004)). Based on this approach, we find that with an evidence ratio of 2.08, the cloze surprisal regression is 2.08 times more likely than the cloze probability regression to be the best model of the N400 data.

However, when comparing AIC values, a general rule of thumb is that when there is an absolute difference in AIC of 2 or less between two statistical models, they have similar levels of support, while a difference of 4 or more means that the model with a lower AIC has 'considerably' more evidential support (Burnham and Anderson, 2004). In this case, the cloze surprisal regression has an AIC which is 1.47 less than the cloze probability regression. Thus, despite the evidence ratio of 2.08, the two regressions should be considered to have similar levels of support, and so it is still not clear whether cloze probability or cloze surprisal is a better predictor of N400 amplitude.

In order to investigate this further, we ran additional analyses, finding that that

the two explain the same variance in N400 amplitude: adding cloze surprisal to the best cloze probability regression does not improve model fit ($\chi^2(1) = 1.638, p = 0.965$); and neither does adding probability to the best cloze surprisal regression ($\chi^2(1) = 0.171, p = 1$). However, given the lower (i.e., better) AIC of the cloze surprisal regression, we take cloze surprisal as the most explanatory representation of cloze for the remainder of our analyses.

2.4.3 Language model surprisal and N400 amplitude

We calculated the probability of each target word based on the predictions of GRNN (mean = 0.087; standard deviation = 0.190), JRNN (0.211 ± 0.291), Transformer-XL (0.092 ± 0.192), GPT-2 (0.382 ± 0.358), GPT-3 (0.526 ± 0.371), BERT (0.317 ± 0.355), RoBERTa (0.495 ± 0.374), and ALBERT (0.298 ± 0.316) for comparison with cloze (0.631 ± 0.348). These probabilities were then transformed into surprisal.

We tested whether the surprisal calculated from each LM is a significant predictor of N400 amplitude. To do this, we compared regressions with a main effect of laboratory and random intercepts for subject and item to those also including a main effect of the relevant LM's surprisal. In this way, the analysis matches those investigating the main effect of cloze probability and cloze surprisal. The results of these analyses are shown in Table 2.2. As can be seen, main effects of surprisal calculated using JRNN, Transformer-XL, GPT-2, GPT-3, BERT, RoBERTa, and ALBERT are all significant in their respective regressions, but the main effect of GRNN surprisal is only marginally significant.

Table 2.2: Significant predictors of N400 amplitude

Predictor	$\boldsymbol{\chi^2(df = 1)}$	**p**
GRNN surprisal	6.356	0.072
JRNN surprisal	**17.330**	**<0.001**
Tranformer-XL surprisal	**19.158**	**<0.001**
GPT-2 surprisal	**26.313**	**<0.001**
GPT-3 surprisal	**40.817**	**<0.001**
BERT surprisal	**30.760**	**<0.001**
RoBERTa surprisal	**37.848**	**<0.001**
ALBERT surprisal	**35.918**	**<0.001**

2.4.4 Comparison of model fit

We next compared the AICs of each linear mixed-effects regression model including LM surprisal with one that instead used cloze surprisal. These comparisons are presented in Figure 2.1, which shows the AIC of each LM surprisal regression with the AIC of the cloze surprisal regression subtracted. This allows for easier comparison of regression AIC, and has a clear interpretation—any regression with a relative AIC of less than zero has a better fit than the cloze surprisal regression.

As can be seen in Figure 2.1, the regressions based on the surprisals calculated from four LMs have lower AICs than cloze surprisal (AIC = 113227.2): GPT-3 (AIC = 113215.8; evidence ratio with cloze surprisal = 300.89), BERT (AIC = 113225.9; evidence ratio = 1.97), RoBERTa (AIC = 113218.8; evidence ratio = 68.18), and ALBERT (AIC = 3113220.7 ; evidence ratio = 25.98). The AIC of the remaining models is higher than that of cloze surprisal. It should be noted that in all but one case, the difference in AIC between the cloze surprisal and all other regressions is greater than 4, suggesting a meaningful difference in this respect (Burnham and Anderson, 2004). The one exception is the BERT regression (ΔAIC = 1.36)—thus, while the BERT regression is 1.97 times more likely than

the cloze surprisal regression to provide the best fit to the N400 data, we rely on the tests in the rest of this section to determine whether BERT surprisal is in fact a better predictor of N400 amplitude than cloze surprisal.

In sum, regressions based on the surprisals derived from GPT-3, RoBERTa, and ALBERT more closely fit the N400 data than the regression based on cloze surprisal, and this may also be the case for the BERT surprisal regression.

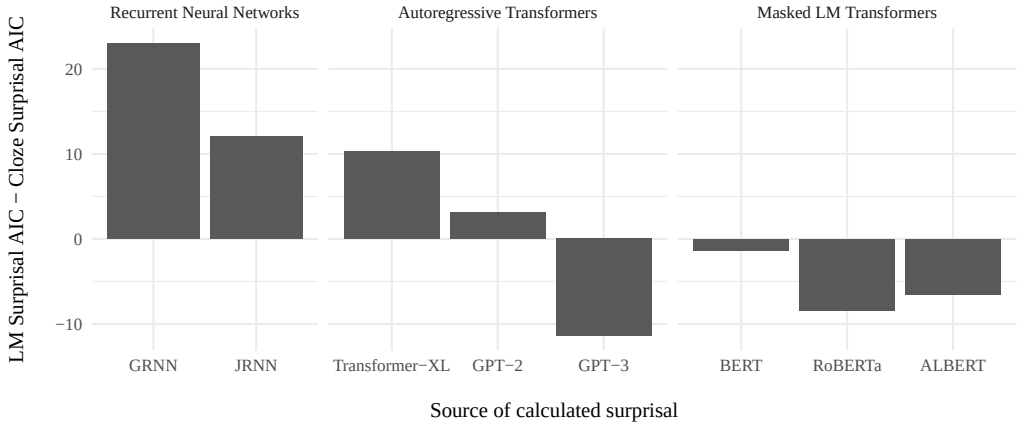

Figure 2.1: AICs of all regressions including fixed effects of the denoted surprisal and laboratory, as well as random intercepts for each item and experimental participants. For easier comparison, AIC is scaled by subtracting the AIC of the regression including cloze surprisal, laboratory, and the aforementioned random intercepts. Lower AICs indicate better model fit (Akaike, 1973).

2.4.5 Does language model surprisal improve fit of regressions based on human cloze data?

In addition to comparing the AICs of the models, following Brothers and Kuperberg (Brothers and Kuperberg, 2021), we compared how well cloze and LM surprisal predict N400 amplitude by constructing additional regressions with both variables and comparing them to regressions with only one. First, we compared the effect of adding the surprisal

calculated from each LM to a regression already including cloze surprisal. Thus, we tested whether each LM surprisal explains variance in N400 amplitude above and beyond that which is already explained by cloze surprisal. The results are shown in Table 2.4.5.

Table 2.3: Results of LRTs testing whether adding LM surprisal as a main effect improves the fit of regressions that already include cloze surprisal as main effect

Predictor	$\boldsymbol{\chi^2(df = 1)}$	**p**
GRNN surprisal	0.056	1
JRNN surprisal	3.982	0.260
Tranformer-XL surprisal	3.031	0.424
GPT-2 surprisal	5.088	0.142
GPT-3 surprisal	**12.168**	**0.004**
BERT surprisal	**9.639**	**0.015**
RoBERTa surprisal	**11.720**	**0.005**
ALBERT surprisal	**8.450**	**0.026**

As can be seen in Table 2.4.5, adding GPT-3, BERT, RoBERTa, or ALBERT surprisal to regressions already including cloze surprisal significantly improves their fit, while adding the surprisal of other LMs does not.

2.4.6 Does human cloze data improve fit of regressions based on language model surprisal?

We also ran the reverse analysis, investigating the effect of adding cloze surprisal to a regression that already includes one LM surprisal as a fixed effect. Thus, we test whether cloze surprisal explains variance in N400 amplitude not explained by each LM surprisal. The results are shown in Table 2.4.

Table 2.4: Results of LRTs testing whether adding cloze surprisal as a main effect improves the fit of regressions that already include LM surprisal as main effect

Predictor	χ^2**(df = 1)**	**p**
GRNN surprisal	**23.103**	**<0.001**
JRNN surprisal	**16.056**	**0.001**
Tranformer-XL surprisal	**13.277**	**0.002**
GPT-2 surprisal	**8.178**	**0.028**
GPT-3 surprisal	0.754	1
BERT surprisal	**8.282**	**0.027**
RoBERTa surprisal	3.276	0.380
ALBERT surprisal	1.935	0.820

As can be seen in Table 2.4, adding cloze surprisal to a regression already including GRNN, JRNN, Transformer-XL, GPT-2, or BERT surprisal improves their fit. By contrast, human cloze surprisal does not improve regressions already including surprisals from GPT-3, RoBERTa, or ALBERT.

In sum, surprisal calculated using GPT-3, RoBERTa, or ALBERT provides a better fit to N400 data than human cloze surprisals based on analyses in both directions, and BERT surprisal explains some variance in N400 amplitude not explained by human cloze surprisals.

2.5 General Discussion

In this study, we investigated whether linguistic predictions from language models or from human participants better predict the amplitude of the N400, a neural index of processing difficulty. We find that, across the board, the surprisal of three transformer LMs, GPT-3, RoBERTa, and ALBERT, are better predictors of N400 amplitude than cloze. This is consistent with prior work showing the correlation between LM surprisal and N400 amplitude (Frank et al., 2015; Aurnhammer and Frank, 2019b; Michaelov and

Bergen, 2020; Michaelov et al., 2021; Merkx and Frank, 2021). However, to the best of our knowledge, the present study provides the most convincing evidence to date that LM surprisal can outperform cloze as a predictor of N400 amplitude.

In contrast to a recent large-scale experiment and meta-analysis by Brothers and Kuperberg (Brothers and Kuperberg, 2021), our results do not show that raw cloze probability is a better predictor of language processing difficulty amplitude than cloze surprisal—in fact, if anything, cloze surprisal is the better predictor. Whether this is because there is a difference in how the N400 and the behavioral metrics analyzed by Brothers and Kuperberg (Brothers and Kuperberg, 2021) relate to word predictability or because of some other difference between the studies is a question for further research.

The skeptical reader might question whether there was some feature of our stimuli that offers an unfair advantage to the LMs over cloze measures. We find this unlikely, given that we have endeavoured to provide a 'level playing field'. First, unlike previous work that showed LM surprisal values provide a good account of N400 elicited by different kinds of semantic stimuli equated for cloze probability (Michaelov et al., 2021), the present study involved the experimental manipulation of the predictability of the words. There were no experimental conditions that were matched for cloze but that differed in some other systematic way. Thus, N400 amplitude variance in this study is almost exclusively due to differences in predictability. Second, all zero-cloze items were removed, meaning that any variation between items in terms of predictability was captured by both cloze and LM surprisal. Finally, we included both cloze probability and cloze surprisal as possible predictors to account for the possibility that one might be a better predictor than the other. In summary, the conditions of this study were maximally favorable towards cloze; and yet we see that even so, distributional information can better predict N400 amplitude.

2.5.1 Theoretical implications

Our main result is that overall, GPT-3 surprisal, RoBERTa surprisal, and ALBERT surprisal were each found to be better predictors of N400 amplitude than cloze surprisal values gathered from human participants. While it is striking that cloze probability and surprisal values from a mere 30 participants provide a better fit to N400 data than do surprisal values from GRNN, JRNN, Transformer-XL, and GPT-2, we find that they do not explain any variance in N400 amplitude above and beyond that explained by GPT-3, RoBERTa, and ALBERT surprisal. Furthermore, the surprisal of these LMs, as well as BERT, explain variance in N400 amplitude not captured by cloze. When comparing LMs of the same type, our results also provide new evidence that supports the idea that LMs of higher quality perform better at modeling the N400 and other measures of online human sentence processing difficulty (Frank et al., 2015; Goodkind and Bicknell, 2018; Aurnhammer and Frank, 2019b; Merkx and Frank, 2021). When compared by perplexity, a common evaluation metric for autoregressive transformer LMs, GPT-3 outperforms Transformer-XL and GPT-2 (Dai et al., 2019; Radford et al., 2019; Brown et al., 2020). Similarly, ALBERT and RoBERTa each out-perform BERT at the GLUE benchmark (Wang et al., 2019b), which covers a wide range of natural language understanding tasks. Finally, all transformer LMs included in this analysis outperform the RNNs (GRNN and JRNN), replicating previous work that transformer LMs are better predictors of N400 amplitude than RNNs (Merkx and Frank, 2021; Michaelov et al., 2021).

This finding may offer additional insight into why our results diverge from previous behavioral studies showing that cloze probability (Brothers and Kuperberg, 2021) and cloze surprisal (Smith and Levy, 2011) are better predictors of processing difficulty than LM surprisal beyond the fact that the N400 and behavioral metrics of processing difficulty are not necessarily always comparable. The most sophisticated LM used in these studies

is the JRNN (in (Brothers and Kuperberg, 2021)), with n-grams also being used (Smith and Levy, 2011; Brothers and Kuperberg, 2021). Thus, our results are actually in line with such findings—in the present study, cloze probability and surprisal out-perform JRNN surprisal at predicting N400 amplitude. Our key finding is that more sophisticated, higher-quality LMs out-perform cloze—as LMs continue to advance and improve, their predictions appear to more closely match those of humans. Thus, our current best operationalizations of predictability based on the statistics of language are the best operationalizations of the predictions underlying the N400 response, and based on the present study, they may continue to get closer.

Until the present study, cloze has been the gold-standard method of operationalizing predictability, and, when tested, the best correlate of behavioural predictability effects (Smith and Levy, 2011; Brothers and Kuperberg, 2021). Thus, because the N400 is sensitive to manipulations that cannot be operationalized by cloze probability, it has been argued that it may be more productive to think of the N400 as reflecting 'preactivation' (Kuperberg et al., 2020), or the 'neural activation state at the time a physical stimulus is encountered' (DeLong and Kutas, 2020) rather than prediction *per se*. For example, besides its high degree of sensitivity to cloze probability, the amplitude of the N400 is also sensitive to factors ostensibly related to the organization of semantic memory. Consider the following set of stimuli from Ito et al. (Ito et al., 2016):

> Jack studied medicine in a university and works as a **doctor/patient/tenant** now.

Here, *doctor* is the highest-cloze continuation of the sentence, while both *patient* and *tenant* have a cloze probability of zero. However, despite the fact that *patient* and *tenant* are equally unpredictable and equally implausible continuations of the sentence (as judged by participants in their study), *patient* elicits a smaller (less negative) N400 than *tenant*. This is one example of a range of studies where words that are semantically related

to the preceding context (i.e. *medicine*) or to the most expected continuation of a sentence (i.e. *doctor*) elicit smaller N400 responses than semantically unrelated words, even when matched for cloze (Ito et al., 2016; Thornhill and Van Petten, 2012; Metusalem et al., 2012). Based on such experiments, it has been proposed that implausible continuations like *patient* are 'collaterally facilitated' by the preceding context (DeLong and Kutas, 2020), or, alternatively, that their preactivation is caused by a separate associative system (Frank and Willems, 2017).

However, recent work shows that the difference in N400 amplitude reported in Ito et al.'s (Ito et al., 2016) study can be successfully predicted based on GRNN and JRNN surprisal (Michaelov and Bergen, 2020). This suggests that manipulations thought to be separate or dissociable from predictability—in this case, semantic relatedness to the highest-cloze continuation—may be reducible to an appropriate measure of predictability. That is, *patient* and *tenant* are not in fact equally predictable, and the belief that they are is an artifact of cloze task. If even the GRNN and JRNN, which are among the worst-performing models in the present study, are able to successfully differentiate the predictability of *patient* and *tenant* (Michaelov and Bergen, 2020) without semantics learned explicitly or through experience of the world, this suggests that humans may not need to rely on such information for prediction either, at least within the N400 window.

The results of the present study may help to illuminate the functional significance of the N400 component by providing evidence for a unified explanation for its sensitivity to what seem to be disparate sources of contextual information. In previous work, we see that semantic relatedness, previously thought to be dissociable from predictability, can successfully be operationalized with LM surprisal (Michaelov and Bergen, 2020; Michaelov et al., 2021). In the present study, we see that predictability, previously thought to be best operationalized with cloze probability, can be operationalized with LM surprisal, with the

highest-quality LMs providing a better operationalization than cloze probability or cloze surprisal. Together, these results suggest that there may be something about the surprisal of high-quality LMs that makes them so well-suited to capturing the predictions of the neurocognitive system underlying the N400 response. LMs are systems trained to predict a word given its context based on the statistics of language. Their degree of success at predicting N400 amplitude relative to other approaches suggests that we should seriously consider that as part of language comprehension, humans may be doing the same.

2.5.2 Methodological implications

Our finding of the relationship between N400 amplitude and surprisal values from GPT-3, RoBERTa, and ALBERT has clear methodological implications. In future work, it may be advantageous for ERP language researchers who want to measure or control the predictability of their stimuli to use surprisal values from these LMs in addition to, or even instead of, cloze probability. As previously discussed, using cloze probability has several theoretical issues, but there are also practical reasons for favoring LM surprisal. For example, it is is easy to gather surprisal values for large stimulus sets (e.g. for every word in a collection of multiple sentences), while this may not be feasible for cloze. Additionally, the precision of cloze probability is limited by the number of participants used for the cloze task—with a limited number of participants, small differences in predictability may not be reflected in cloze, and further, this means that even with a large number of participants, variation in the predictability of zero-cloze items may not be detected. LM surprisal, by contrast, allows the researcher to differentiate between items even with a very low probability, making it possible to control for predictability over a wider range than does cloze probability.

However, in addition to these already-known reasons for preferring LM surprisal

to cloze, the results of the present study provide another, stronger argument for using LM surprisal over cloze. Even for stimuli that vary in measurable ways in terms of cloze, the surprisals calculated from GPT-3, RoBERTa, and ALBERT's predictions provide a better fit to the N400 data, suggesting that they may better operationalize the predictability underlying the variance in the N400 response to stimuli. Indeed, as discussed, given that these are the highest-quality models tested, we might expect that LM surprisal's ability to capture predictability may continue to improve. ERP language researchers already use other measures derived from linguistic corpora to control their language materials. For example, since the report that corpus-derived metrics of word similarity are correlated with N400 amplitude (Chwilla and Kolk, 2005; Parviz et al., 2011; Van Petten, 2014; Ettinger et al., 2016), many researchers have constructed their stimuli such that they are either matched in terms of these metrics, or include similarity metrics as covariates in their statistical analyses (Chwilla et al., 2007; Kuperberg et al., 2020; Nieuwland et al., 2020). The present study suggests that surprisals derived from high-quality LMs should be used analogously in ERP investigations of language processing.

2.6 Conclusion

Previous work has shown that LM predictions correlate with N400 amplitude when cloze does not (Michaelov and Bergen, 2020; Michaelov et al., 2021). The present study has shown that even in conditions maximally preferable for cloze, LM predictions correlate better with N400 amplitude. Thus, at least in terms of relative strength, the kinds of predictions made by LMs resemble the kinds of predictions made by humans as part of online language comprehension. Thus, the language comprehension system, or at least, the neurocognitive system underlying the N400 response, appears to be more finely attuned to the regularities in the statistics of language than previously thought.

2.7 Acknowledgements

The authors would like to thank Mante Nieuwland and collaborators for making their stimuli and data available online. The authors would also like to thank the anonymous reviewers for their helpful comments.

References

Chapter 2, in full, is a reprint of the material as it appears in Michaelov, J. A., Coulson, S., & Bergen, B. K., "So Cloze yet so Far: N400 Amplitude is Better Predicted by Distributional Information than Human Predictability Judgements", *IEEE Transactions on Cognitive and Developmental Systems*, 2022. The dissertation author was the primary investigator and author of this paper. Minor edits have been made to bring the formatting in line with the dissertation template.

Chapter 3

Better language models better model the N400

Abstract

The probability of a word in context, as captured by Large Language Models, is predictive of both behavioral and neural measures of human language processing. Intuitively, language models that are better at next word prediction might better model the effect of predictability on human language comprehension. Yet recent work suggests that language models can become too good at next-word prediction to model reading time, suggesting that the aspects of human comprehension indexed by reading time do not track perfectly with predictability from language statistics alone. However, it is unknown whether this decoupling is true of reading time only, or is intrinsic to measures of comprehension more generally. To address this question, we turn to another robust and well studied measure of online processing, the N400. We investigate how number of training tokens and performance at natural language benchmarks correlate with a language model's

ability to predict N400 amplitude. We find that across the board, models that are trained on more data and that perform better at next-word prediction and other more complex natural language tasks are better able to predict N400 amplitude. We interpret this difference between the N400 and reading time measures as potentially revealing the comparative importance of semantic prediction in the magnitude of the N400.

3.1 Introduction

The statistics of language have proven effective for modelling certain aspects of online human language comprehension. In the last few decades, for example, language models—computational models that predict the probability of a word in a given linguistic context—have been found to generate predictions that correlate strongly with behavioral (McDonald and Shillcock, 2003a,b) and neural (Parviz et al., 2011; Frank et al., 2015) indices of language processing. For many years, a consistent and perhaps unsurprising pattern has been reported whereby larger and more powerful n-gram (Goodkind and Bicknell, 2018; Wilcox et al., 2020), recurrent neural network (Aurnhammer and Frank, 2019a; Aurnhammer et al., 2021; Wilcox et al., 2020; Merkx and Frank, 2021), and transformer language models (Wilcox et al., 2020; Merkx and Frank, 2021) make predictions that better correlate with metrics of online human language comprehension.

However, in more recent years, language models have continued to advance, not only quantitatively in scale and next-word-prediction capabilities (see, e.g., Rae et al., 2022), but also qualitatively in that they are able to accomplish a wide range of tasks (Brown et al., 2020; Srivastava et al., 2022; Wei et al., 2022a; Hoffmann et al., 2022), to the point where text generated by them has become difficult to tell apart from that generated by humans (Brown et al., 2020; Köbis and Mossink, 2021; Clark et al., 2021; Jannai et al., 2023; Jones and Bergen, 2024a,b). Beyond being able to simply represent

n-gram relationships, they appear to be able to learn the semantic relationships between words and make predictions on this basis (for discussion, see, e.g. Michaelov and Bergen, 2022a).

Intuitively, one might expect this to lead to even better modeling of human language processing. Yet surprisingly, evidence from the reading time literature suggests the opposite (Oh et al., 2022; Oh and Schuler, 2023a,b; Oh et al., 2024; Shain et al., 2024; for related findings see Eisape et al., 2020; Kuribayashi et al., 2021). Oh et al. (2022), for example, find that larger GPT-2 models with a greater number of parameters predict reading time worse than GPT-2 models with fewer parameters (based on both self-paced reading and eye tracking experiments) . Oh and Schuler (2023a) observe a similar pattern when analyzing the effect of training data on the Pythia suite of transformer language models (Biderman et al., 2023b), finding that for all the models tested (from 70 million to 12 billion parameters), performance at predicting reading time data improves until the models are trained on 2 billion tokens, at which point performance begins to degrade again.

To the best of our knowledge, no equivalent results have been reported in the literature on neural indices of language processing, like the N400. The N400 is a neural index of language processing that, like reading time, is known to be highly correlated with the contextual probability of a word (Kutas and Federmeier, 2011; Van Petten and Luka, 2012; DeLong et al., 2014b; Kuperberg et al., 2020). While there have been no systematic attempts to quantify the exact relationship, previous studies using language models to predict N400 amplitude display an imperfect but relatively consistent relationship where, for a given set of models trained on the same dataset, models with more parameters (i.e., larger models) generally tend to predict N400 amplitude better than smaller models (Michaelov and Bergen, 2022b, 2023). Yet it is currently unknown how N400 prediction capability correlates with language model training data or overall performance at natural

language tasks, particularly at the scale where reading time prediction begins to falter. Aurnhammer and Frank (2019b), for example, report that models trained on more data perform better at predicting N400 amplitude; and Aurnhammer and Frank (2019a) and Merkx and Frank (2021) report that models that are better at next-word prediction (which they term 'higher-quality') also perform better at predicting the N400. However, crucially, the models in these studies are trained on a maximum of 95 million tokens; and as previously noted, decreased performance at predicting reading time has only been observed on models trained on over a billion tokens. Thus, as yet, we do not know whether better, more extensively trained language models predict N400 amplitude better or whether, as with reading time, word probability eventually diverges from the N400.

Our analysis therefore has several parts. In a first experiment, we follow Oh and Schuler (2023a) in using the Pythia suite of language models to tease apart how number of parameters and number of training tokens each contribute to performance at modeling the N400. We also follow previous work in looking at how well language modeling performance correlates with this. Like several previous studies (Goodkind and Bicknell, 2018; Oh and Schuler, 2023a,b; Oh et al., 2024), we operationalize basic language model quality using perplexity (a measure of how good language models are at predicting the next word in a sequence; Jelinek et al., 1977). But in a second experiment we further ask how performance at other natural language tasks (Warstadt et al., 2020; Paperno et al., 2016; Zellers et al., 2019; Bisk et al., 2020; Sakaguchi et al., 2020) correlates with N400 prediction capability. This extension aims to determine several things: whether generally better language models are better for modeling the N400, whether specific natural language model capabilities are more correlated with N400 prediction performance, and whether it is possible to use natural language modeling tasks to identify which models are best able to predict N400 amplitude even in the absence of information about number of model parameters or training tokens.

Finally, we run our analyses on two reading time datasets to compare with the N400 effects. We use the Provo Corpus (Luke and Christianson, 2018), on which Oh et al. (2024) demonstrate the decrease in performance for larger and higher-quality language models, as well as the data from a study by Smith and Levy (2013), which, to the best of our knowledge, has not been tested in this way.

3.1.1 Computational modeling of human language processing

The fact that indices of human language processing such as reading time and N400 amplitude are correlated with the contextual probabilities of words has been known for around four decades (Fischler and Bloom, 1979; Kutas and Hillyard, 1984), but it is only more recently that statistical models have been able to capture such patterns in ways that are meaningful for human processing research, from n-grams (McDonald and Shillcock, 2003a; Parviz et al., 2011) to transformers (Wilcox et al., 2020; Merkx and Frank, 2021). While different transformations of contextual probability have been used to predict both reading time and N400 amplitude, the current research suggests that a logarithmic transformation provides the best linking function for language-model-derived probabilities to both reading time (Smith and Levy, 2013; Shain et al., 2024) and the N400 (Yan and Jaeger, 2020; Szewczyk and Federmeier, 2022). Thus, the most common way to operationalize the effect of contextual probability on human language processing is to calculate the negative log-probability or *surprisal* of a word in context, the equation for which is given in Equation (3.1).

$$S = -\log P(w_i|w_1...w_{i-1}) \tag{3.1}$$

3.1.2 Scaling in Language Models

A number of studies have attempted to investigate whether there are formal scaling laws (or at least general patterns) governing the performance of language models based on number of parameters and training tokens for a given computational budget (for examples and discussion, see, e.g., Kaplan et al., 2020; Henighan et al., 2020; Hoffmann et al., 2022; Alabdulmohsin et al., 2022; Le Scao et al., 2022; Touvron et al., 2023; Tay et al., 2023; Sardana and Frankle, 2023; Muennighoff et al., 2023; Biderman et al., 2023b; Ruan et al., 2024). With the exception of certain conditions that induce so-called "inverse scaling" (see, e.g. Lin et al., 2022; McKenzie et al., 2023), models with more trainable parameters generally perform better at natural language processing tasks than those with fewer parameters, models trained on more tokens perform better than those trained on fewer, and models with more parameters can improve their performance with more training data to a greater extent than those with fewer parameters. Simply put: models larger in either dimension perform better than smaller models, and the two interact in that the number of parameters sets the ceiling of performance. Given this pattern, we follow Oh and Schuler (2023a,b) in henceforth referring to the number of parameters of a model as its *capacity*.

While the majority of research on scaling focuses on natural language tasks, a number of studies have either implicitly or explicitly investigated the extent to which model scale impacts performance at modeling online human language comprehension. As previously discussed, work on reading time has historically (with pre-transformer architectures, training data sizes of less than 2 billion tokens, or both) found that models trained on more data generally perform better at predicting reading time than models trained on less data (Aurnhammer and Frank, 2019b; Wilcox et al., 2020; Merkx and Frank, 2021; Wilcox et al., 2023a). As noted, however, more recent work with larger models has suggested that there is a limit to this, and that past a certain point, this effect is reversed. Specifically, a

recent set of studies (Oh et al., 2022; Oh and Schuler, 2023b,a; Oh et al., 2024) has shown that on models trained on over 2 billion tokens, both language model capacity and training tokens have a negative effect on language models' capability to predict reading time. For example, using two metrics of reading time from two separate datasets—self-paced reading response time from the Natural Stories Corpus (Futrell et al., 2021) and go-past duration from the Dundee Corpus (Kennedy et al., 2003)—Oh et al. (2022) and Oh and Schuler (2023b) find that smaller-capacity variants of the GPT-2, GPT-Neo, and OPT language models (all trained on the same datasets) predict the reading time metrics better than larger models of the same family. Thus, the findings show systematic evidence that higher-capacity models can actually perform worse than lower-capacity models. In addition, Oh and Schuler (2023a) compare the Pythia (Biderman et al., 2023b) models over the course of their training, finding that for all the models tested (of various capacities), performance at predicting the two reading time metrics improves until the thousandth step of training (when the models have been trained on roughly 2 billion tokens), before beginning to decrease again, with higher-capacity models showing a greater decrease in performance.

What could explain this pattern? As noted earlier in this section, the general finding is that larger models are generally better-performing at natural language tasks; and one possibility is that these larger models are *too good* at next-word prediction relative to humans. Specifically, the results show that when using language models to predict reading time, they systematically under-predict reading time for words with low contextual probabilities (Oh and Schuler, 2023b) and low overall (i.e., unigram) probabilities (Oh et al., 2024), suggesting that they are better able to predict such words than humans (Oh and Schuler, 2023b; Oh et al., 2024). Such an explanation is further strengthened by Shlegeris et al.'s (2022) finding that even the 350-million parameter GPT-3 model (which is trained on 300 billion text tokens; see Brown et al., 2020) can predict the next word in a sequence

better than any of the 60 humans they test, with larger models performing even better.

With the N400, on the other hand, all evidence thus far suggests that larger models are better predictors. Aurnhammer and Frank (2019b), for example, find that recurrent neural networks trained on more data are better able to predict N400 amplitude; while Michaelov et al. (2022) find that for recurrent neural networks, autoregressive transformers, and masked transformers, higher-capacity models and those trained on more data predict N400 amplitude better. In fact, while not discussed in the paper, the results presented in Michaelov and Bergen (2022b) show a clear pattern whereby the 7.5-billion-parameter (7.5B) XGLM (Lin et al., 2021) performs better than the 4.5B model, which in turn performs better than the 2.9B model, and so on with the 1.7B and 564-million parameter (564M) models; as well as similar (though not all perfect) patterns with the OPT (Zhang et al., 2022), GPT-Neo (Black et al., 2021), and GPT-2 (Radford et al., 2019) models. All these models are trained on over 2 billion tokens, and thus, this result shows that even at the scales at which negative scaling (as a function of capacity) has been found to emerge in reading time, such effects have not been found when modeling the N400.

Given Oh and Schuler's finding that the decrease in the extent to which they can predict reading time is driven by training data and the possible downstream effect of being *too good* at next-word prediction compared to humans, we explicitly target this in the present study. Thus, while we include models of varying capacities, our main aim is to investigate whether there is a point during training at which continuing to train a model leads to a model that is too good at predicting the next word to make predictions that correlate with N400 amplitude. For this reason, in addition to revealing how different metrics of language model quality correlate with models' ability to predict N400 amplitude, the evaluations of language model quality below also differentiate whether decreased performance with more training data is due to a model getting generally worse (which is possible

for very small models if they are 'overtrained'; see Hoffmann et al., 2022; Biderman et al., 2023b), or whether it is because they are in fact getting better at next-word prediction in a un-human-like manner.

3.1.3 Language Model Quality

If a language model has learned the statistical regularities of a language, it should calculate natural language to have a higher probability, than, for example, nonsensical strings. This is the logic behind the traditional way to measure language model performance, namely, by calculating language model *perplexity* (Jelinek et al., 1977) on a given corpus. The perplexity of a given sequence of words or tokens $w_1...w_N$ is given by Equation (3.2), where $P(w_i|w_1...w_{i-1})$ is the probability of a word w_i given the preceding words in the sequence $w_1...w_{i-1}$.

$$\text{PPL}(w_1...w_N) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_1...w_{i-1})}} \tag{3.2}$$

Thus, perplexity gives a measure of the average probability of a word in a text. For many years, perplexity on a held-out dataset was the primary way to measure language model performance, with a lower perplexity indicating a better model. Thus model perplexity's relationship to predicting metrics of human language comprehension has also been investigated. From a theoretical perspective, this question has more explanatory power than investigating scaling in that it is more directly interpretable. If we consider a language model's ability to predict the next word in a given context to be a measure of the language model's *quality* (Goodkind and Bicknell, 2018), then if lower-perplexity models better predict reading time or N400 amplitude, this suggests that the extent to which a model's predictions correlate with the human metric is tied to the model's quality. It is worth noting that while perplexity is the canonical and most widely-used metric of model

quality in studies of human language processing (Goodkind and Bicknell, 2018; Hao et al., 2020; Wilcox et al., 2020; Kuribayashi et al., 2021; Oh and Schuler, 2023a,b; Oh et al., 2024), other researchers have instead opted to use average log-probability (Aurnhammer and Frank, 2019a; Merkx and Frank, 2021) or average cross-entropy (Wilcox et al., 2023a). These are respectively equivalent to average negated surprisal (see Equation (3.1)) and surprisal (see Aurnhammer and Frank, 2019b; Michaelov and Bergen, 2023), respectively, and so are both monotonically related to perplexity, as can be seen from Equation (3.3).

$$\text{PPL}(w_1...w_N) = \exp\left\{-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_1...w_{i-1})\right\} \tag{3.3}$$

The results of analyses looking at model quality, however operationalized, are the same as those for model scale. N-grams, recurrent neural networks, and transformers with a lower perplexity (Goodkind and Bicknell, 2018; Hao et al., 2020; Wilcox et al., 2020), a higher average log-probability (Aurnhammer and Frank, 2019a; Merkx and Frank, 2021), or a lower average cross-entropy (Wilcox et al., 2023a) tend to perform better at predicting reading time than models of a lower quality up to a point, after which the correlation appears to decrease (Oh and Schuler, 2023a,b; Oh et al., 2024). Meanwhile, for the N400, we again only see evidence that higher-quality models perform better (Merkx and Frank, 2021). However, the most recent study to directly compare language model quality and goodness of fit to the N400 data is that of Merkx and Frank (2021), and as was previously noted, the fact that they do not find any negative scaling for reading time may suggest that the models are too small (at least in terms of training data) to observe this. Thus, in the present study, we run a similar analysis with contemporary models.

It is also important to note that perplexity as a metric of language model quality was developed at a time when language models were far less advanced that they are in the present. Today, language models have been developed that can generate not only

fully grammatical sequences, but seemingly coherent text (Stiennon et al., 2020; Nakano et al., 2022; Bai et al., 2022; though see, e.g., Bender et al., 2021; Raji et al., 2022, for a discussion of the risks of this). Thus, it makes sense to evaluate more specific tendencies in language model predictions—and indeed, contemporary benchmarks evaluate a wide range of possible patterns of 'capabilities' and 'behaviors' in the predictions of language models (see, e.g., Wang et al., 2019b,a; Srivastava et al., 2023; Bommasani et al., 2023b). We focus on five such tasks, which we describe in more detail in Section 3.3.

3.2 Experiment 1: The effect of scale on the N400

3.2.1 Introduction

In this paper, our main research question is how language models' performance at predicting N400 amplitude is impacted by how many tokens of text it is trained on as well as language model capacity. To do this, we draw on the approach used by Oh and Schuler (2023a) to ask the same question for reading time. Specifically, we use the Pythia suite (Biderman et al., 2023a), a collection of language models with different capacities, with checkpoints provided over the course of training. This allows us to investigate the extent to which the models' next-word predictions can be used to model N400 amplitudes from 9 previous studies (Federmeier et al., 2007; Nieuwland et al., 2018b; Wlotko and Federmeier, 2012; Hubbard et al., 2019; Lago et al., 2019; Ryskin et al., 2021; Szewczyk et al., 2022; Szewczyk and Federmeier, 2022; Michaelov et al., 2024).

3.2.2 Data Availability

All code, data, and statistical analysis scripts are provided at `https://osf.io/qbekt/?view_only=0f4ba6296eda442aaf8e49109eac145d`.

3.2.3 Method

General approach

Our general approach follows that of previous work using language models to predict metrics of reading time and the N400 (Oh et al., 2022; Oh and Schuler, 2023a,b; Szewczyk and Federmeier, 2022; Michaelov et al., 2022; Michaelov and Bergen, 2023). Specifically, we use each language model to calculate the surprisal of each critical word (i.e., each word for which the N400 was measured) given its preceding context. In cases where critical words are not in a language model's vocabulary, we take a sum of the surprisal of each token given their context (including previous tokens of the same), which is equivalent to taking the product of the probabilities and which leads to a well-defined surprisal for each such word.

To evaluate how closely these surprisals correlate with the N400, we also follow the aforementioned studies and use linear mixed-effects regression models to predict N400 amplitude, matching the structure of the regression (in terms of covariates and the random effects structure) in the original paper as closely as possible. In order to compare the quality of these regressions, we then calculate the Akaike Information Criterion (AIC; Akaike, 1973) of each regression (following, e.g., Michaelov et al., 2022; Michaelov and Bergen, 2023; Michaelov et al., 2024), and compare this for each checkpoint of each language model.

N400 datasets

We investigate how well the surprisal calculated using different language models predicts the amplitudes of N400 responses from 8 previously-published studies (Federmeier et al., 2007; Nieuwland et al., 2018b; Wlotko and Federmeier, 2012; Hubbard et al., 2019; Lago et al., 2019; Ryskin et al., 2021; Szewczyk et al., 2022; Michaelov et al., 2024) and

Table 3.1: A description of each of the N400 datasets, including the total number of experimental stimuli, experimental participants, and the number of trials.

Dataset	**Stimuli**	**Participants**	**Trials**
Federmeier et al. (2007)	564	32	7,856
Hubbard et al. (2019)	192	32	5,705
Szewczyk and Federmeier (2022)	600	26	4,822
Szewczyk et al. (2022)	672	32	4,939
Wlotko and Federmeier (2012)	300	16	4,440
Nieuwland et al. (2018b)	160	334	25,978
Lago et al. (2019)	856	104	9,892
Ryskin et al. (2021)	640	24	3,241
Michaelov et al. (2024)	500	50	5,526

1 dataset released as part of Szewczyk and Federmeier's (2022) study. We describe the details of each dataset below, and provide information about the number of stimuli, trials, and experiment participants in Table 3.1.

The first 5 datasets (Federmeier et al., 2007; Wlotko and Federmeier, 2012; Hubbard et al., 2019; Szewczyk et al., 2022; Szewczyk and Federmeier, 2022) are all variants of the original study by Federmeier et al. (2007), who measured N400 responses to low-cloze and high-cloze words in low-constraint and high-constraint contexts, where constraint refers to the cloze probability of the highest cloze sentence continuation. The stimuli in the other studies differ in specific ways—Hubbard et al. (2019) and Szewczyk and Federmeier (2022; previously unpublished) use a subset of the stimuli from Federmeier et al. (2007), Wlotko and Federmeier (2012) add additional stimuli to cover a wider range cloze probabilities, and Szewczyk et al. (2022) add adjectives that make the critical nouns either more or less likely.

We use the data from these studies as preprocessed by Szewczyk and Federmeier

(2022), who operationalize N400 amplitude as the mean un-baselined amplitude over the 300-500ms time window at four centro-parietal electrodes (MiCe, MiPa, LMCe, RMCe). Our statistical approach followed that in Szewczyk and Federmeier (2022) as much as possible (except to avoid regressions that would not converge or had singular fits), using a linear mixed-effects regression to predict N400 amplitude with fixed effects of surprisal, baseline amplitude, log-transformed frequency, the position of the word in the sentence, orthographic neighborhood distance, and concreteness, and including random slopes of baseline amplitude for each subject and item, random slopes of baseline amplitude for each subject, and random intercepts of subject and item.

The Nieuwland et al. (2018b) dataset is made up of the amplitudes of the N400 responses elicited by nouns in a large-scale study carried out by the authors. For each item, there were two sentence continuations: the highest-cloze continuation, and a low-cloze alternative. We use the data as preprocessed by Nieuwland et al. (2018b), where N400 is operationalized as the mean baselined voltage at 6 centro-parietal electrodes (Cz, C3, C4, Pz, P3, and P4) over the 200-500ms time window after stimulus presentation. Following the original study and Michaelov et al. (2022), we use linear mixed-effects regressions to predict N400 amplitude with laboratory as a fixed effect (the study was carried out over multiple laboratories), as well as random intercepts for each subject and item.

The Lago et al. (2019) dataset is made up of 5 experiments carried out by Lago et al. (2019) to investigate whether the semantics of an antecedent impact facilitation during coreference, specifically whether words related to the referent are primed by a possessive pronoun used to refer to it. The critical words in this study were nouns that were either semantically related or unrelated to the antecedent, and the antecedent is either repeated or referred to using a pronoun, giving four conditions; to which an additional control was added. Lago et al. (2019) operationalized N400 amplitude as the mean N400 amplitude

over the 300-500ms time period at each of 9 (P3, Pz, P4, CP3, CPz, CP4, C3, Cz, C4) centro-parietal electrodes. Lago et al. (2019) use Bayesian hierarchical linear models; but for consistency with our other analyses, we use frequentist linear mixed-effects regressions following the same structure as much as possible, which included surprisal as a fixed effect and random slopes of surprisal for each participant and item, as well as random intercepts for participant, item, electrode, and experiment.

The Ryskin et al. (2021) dataset is made up of data from a study investigating a noisy-channel account of human language comprehension. Items had four possible critical word continuations: a plausible one, a semantically incongruous one, a morphosyntactically incongruous one, and a semantically incongruous one similar enough in form to the plausible continuation that it could be interpreted as a recoverable mistake. N400 amplitude was operationalized as the mean baselined amplitude over the 300-500ms time window at 8 centro-parietal electrodes (C3, Cz, C4, CP1, CP2, P3, Pz, and P4). The original study again used a Bayesian regression, the structure of which we replicated as closely as possible with our frequentist linear mixed-effects models. Our regressions each included surprisal as a fixed effect with random slopes for each subject and item, as well as random intercepts for each of these as well as electrode.

The Michaelov et al. (2024) dataset has sentences with four possible endings: the highest-cloze continuation, a low-cloze but plausible continuation that is semantically related to this highest-cloze continuation, an equally low-cloze but unrelated continuation, and an implausible continuation, with the two low-cloze (but not implausible) continuations matched for cloze probability and plausibility. Stimuli varied in constraint but were not matched for this. The N400 was operationalized as the baselined voltage at 9 centro-parietal electrodes (C3, Cz, C4, CP3, CPz, CP4, P3, Pz, and P4) over the 300-500ms time window. Fit to N400 amplitude was calculated in the same way as in Michaelov

Table 3.2: Details of the Pythia models used in Experiment 1. The name provides the number of paramaters in millions (M) or billions (B) of parameters. Columns provide the number of layers, dimensions, and attention heads in each model, as well as the starting learning rate used during training (the learning rate decayed over the course of training as described in Biderman et al., 2023b).

Name	**Layers**	**Dimensions**	**Attn. Heads**	**LR**
Pythia 14M	6	128	4	1.0×10^{-3}
Pythia 70M	6	512	8	1.0×10^{-3}
Pythia 160M	12	768	12	6.0×10^{-4}
Pythia 410M	24	1024	16	3.0×10^{-4}
Pythia 1.4B	24	2048	16	2.0×10^{-4}
Pythia 2.8B	32	2560	32	1.6×10^{-4}
Pythia 6.9B	32	4096	32	1.2×10^{-4}

et al. (2024), using linear mixed-effects regressions predicting N400 with fixed effects of surprisal, log-transformed word frequency, and orthographic neighborhood distance, and random intercepts of context sentence, critical word, subject, and electrode.

Language Models

Following the analysis carried out by Oh and Schuler (2023a), we use the Pythia suite of language models to carry out our analyses. Pythia models are a set of autoregressive language models that have different capacities and that are trained on exactly the same text corpus, known as The Pile (Gao et al., 2020). In this paper, we analyze 7 of these models, the details of which are presented in Table 3.2.

Another aspect of the Pythia suite of models is that checkpoints are provided at a large number of stages over the course of training. This allows us to tease apart the effects of model size in terms of capacity and number of training tokens. In line with Oh and Schuler (2023a), we calculate the surprisal for all the stimuli using each Pythia model

at various stages over the course of training. We provide details of the training checkpoints and the corresponding number of total training tokens in Table 3.3.

Table 3.3: The training checkpoints analyzed in the present study and the corresponding number of total tokens the model has been trained on at that checkpoint.

Training Step	**Total Training Tokens**
Step 0	0
Step 1	2,097,152
Step 2	4,194,304
Step 4	8,388,608
Step 8	16,777,216
Step 16	33,554,432
Step 32	67,108,864
Step 64	134,217,728
Step 128	268,435,456
Step 256	536,870,912
Step 512	1,073,741,824
Step 1,000	2,097,152,000
Step 2,000	4,194,304,000
Step 4,000	8,388,608,000
Step 8,000	16,777,216,000
Step 16,000	33,554,432,000
Step 32,000	67,108,864,000
Step 64,000	134,217,728,000
Step 128,000	268,435,456,000
Step 143,000	299,892,736,000

Because of a known issue where the lowest-capacity Pythia models are sometimes unstable at the default 16-bit precision (see Schoelkopf, 2024), we run all models at 32-bit precision.

3.2.4 Results

Figure 3.1 shows two clear patterns: performance tends to improve over the course of training, and in the later stages where performance is better, larger models tend to do better than smaller models.

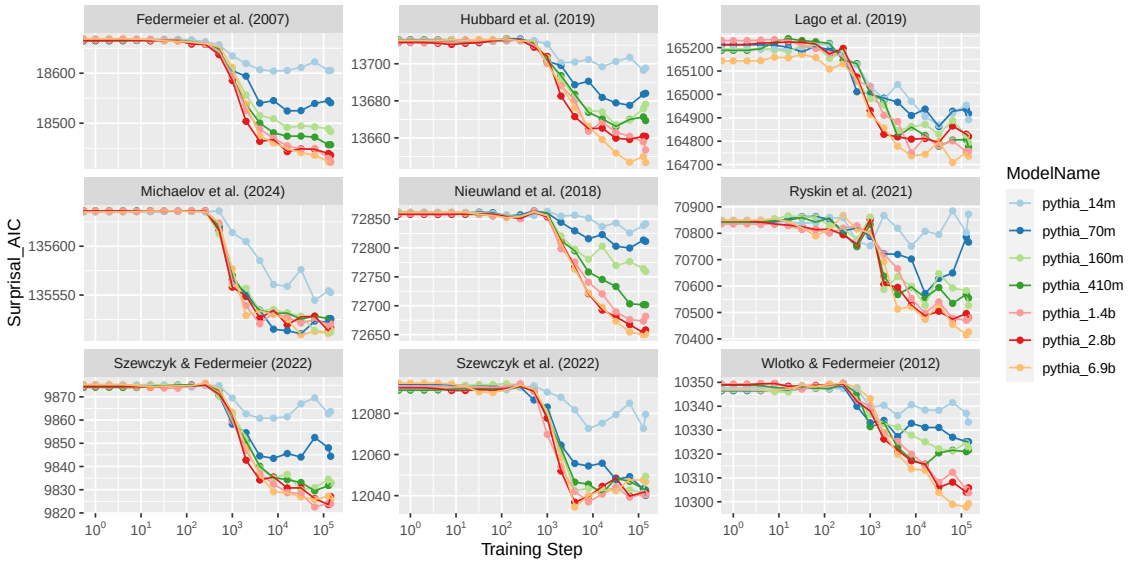

Figure 3.1: How language model performance at predicting N400 amplitude varies by model and over the course of training. A lower AIC indicates a better fit to the N400 data.

Specifically, on most datasets, all models begin to improve at around step 512, when the models are trained on around 1 billion tokens. In general, the models then continue to improve until the end of training (step 143,000), at which point the models are trained on roughly 300 billion tokens. During this period of training, the main differentiation in model performance is a function of number of parameters—larger models begin to predict N400 amplitude better earlier than smaller models, and the rate at which the larger models improve is faster. Finally, there appears to be a slow-down in the increase of performance, though the point at which this occurs appears to vary more widely—in fact, on one dataset (Szewczyk et al., 2022), there appears to be a slight decrease for the last part of training.

Because the performance patterns are highly nonlinear and idiosyncratic, we use

the Mann-Kendall Test (Mann, 1945; Kendall, 1948) to quantify the extent to which there is a significant negative trend in AIC (i.e., improved performance) over the course of training for each language model on each dataset from step 1000 until the end of training. As can be seen in Table 3.4, the results show a negative trend for nearly all models over nearly all datasets, and while there are a few individual exceptions, these are generally significant, even when correcting for multiple comparisons across all statistical tests carried out in this study (using the false discovery rate approach of Benjamini and Yekutieli, 2001).

3.2.5 Discussion

While there are some specific idiosyncracies, the main result of the experiment is clear: larger language models—both in terms of capacity and training data—perform better at predicting N400 amplitude than smaller models.

Furthermore, we see that unlike reading time where modeling performance appears to decrease after step 1000 when models are trained on about 2 billion tokens (Oh and Schuler, 2023a), the reverse is true for the N400: the majority of the improvement on most datasets occurs *after* step 1000, i.e., when the models are trained beyond 2 billion tokens.

3.3 Experiment 2: Language model quality

3.3.1 Introduction

What explains these scaling effects? The most straightforward explanation is that language models trained on more data, and which have more capacity, tend to be better at language modeling overall (Rae et al., 2022), and that better word predictions better align with human processing metrics. Where the effects of scale and quality have been

Table 3.4: The results of Mann-Kendall tests looking at the overall trend of AIC from step 1,000 to the end of training (step 143,000). FWOK07 refers to Federmeier et al. (2007), HRJF19 to Hubbard et al. (2019), LNJL19 to Lago et al. (2019), MBVBC24 to Michaelov et al. (2024), N18 to Nieuwland et al. (2018b), RSBEFG21 to Ryskin et al. (2021), SF22 to Szewczyk and Federmeier (2022), SMF22 to Szewczyk et al. (2022), and WF12 to Wlotko and Federmeier (2012). All p-values are corrected for multiple comparisons (Benjamini and Yekutieli, 2001).

	FWOK07		**HRJF19**		**LNJL19**	
Parameters	τ	p	τ	p	τ	p
14M	-0.317	0.0072	-0.317	0.0072	-0.529	< 0.0001
70M	-0.317	0.0072	-0.529	< 0.0001	-0.529	< 0.0001
160M	-0.846	< 0.0001	-0.423	0.0001	-0.582	< 0.0001
410M	-0.846	< 0.0001	-0.687	< 0.0001	-0.687	< 0.0001
1.4B	-0.899	< 0.0001	-0.899	< 0.0001	-0.582	< 0.0001
2.8B	-0.793	< 0.0001	-0.687	< 0.0001	-0.159	0.5971
6.9B	-0.952	< 0.0001	-0.899	< 0.0001	-0.529	< 0.0001

	MBVBC24		**N18**		**RSBEFG21**	
Parameters	τ	p	τ	p	τ	p
14M	-0.740	< 0.0001	-0.582	< 0.0001	0.211	0.1783
70M	-0.370	0.0010	-0.687	< 0.0001	-0.159	0.5971
160M	-0.846	< 0.0001	-0.793	< 0.0001	-0.423	0.0001
410M	-0.793	< 0.0001	-0.952	< 0.0001	-0.582	< 0.0001
1.4B	-0.634	< 0.0001	-0.846	< 0.0001	-0.793	< 0.0001
2.8B	-0.582	< 0.0001	-0.899	< 0.0001	-0.74	< 0.0001
6.9B	-0.582	< 0.0001	-0.899	< 0.0001	-0.687	< 0.0001

	SF22		**SMF22**		**WF12**	
Parameters	τ	p	τ	p	τ	p
14M	0.211	0.1783	-0.37	0.001	-0.264	0.0410
70M	-0.264	0.041	-0.793	< 0.0001	-0.74	0.1783
160M	-0.634	< 0.0001	-0.37	0.001	-0.74	0.1783
410M	-0.793	< 0.0001	-0.529	< 0.0001	-0.211	< 0.0001
1.4B	-0.846	< 0.0001	-0.423	0.0001	-0.899	< 0.0001
2.8B	-0.74	< 0.0001	-0.159	0.5971	-0.846	< 0.0001
6.9B	-0.793	< 0.0001	-0.053	1	-0.846	< 0.0001

investigated together, they've moved in lock-step. Oh and Schuler (2023a), for example, find that performance at modeling reading time decreases after models surpass a certain size, and when perplexity improves to below a value somewhere in the range of $2^7 - 2^{10}$ (i.e., average log_2-based surprisal falls between 7 and 10). Thus, in this experiment, we investigate whether the scaling effects observed in Experiment 1 can be explained by language model quality. Specifically, we ask whether better (i.e., higher-quality) language models predict N400 amplitude better.

As noted in Section 3.1, improvements in natural language technologies have led to the development of benchmarks aimed to test specific language model capabilities. In this experiment, we try to tease apart dimensions of model quality by selecting 5 such benchmarks—one designed to evaluate language models' knowledge of linguistic structure (BLiMP; Warstadt et al., 2020), and four more semantic benchmarks, designed to evaluate whether language models can predict what comes next in a text in human-like way based on contextual information and world knowledge (LAMBADA: Paperno et al., 2016; HellaSwag: Zellers et al., 2019; PiQA: Bisk et al., 2020; WinoGrande: Sakaguchi et al., 2020). Thus, we test how well the general ability to predict the next word in a sequence, grammmatical knowledge, and the ability to use world knowledge (or at least, to make predictions in line with world knowledge) correlate with a language model's ability to predict N400 amplitude.

3.3.2 Data Availability

All code, data, and statistical analysis scripts are provided at `https://osf.io/qbekt/?view_only=0f4ba6296eda442aaf8e49109eac145d`.

3.3.3 Method

General

We use the same language models and N400 datasets, and follow the same procedure for evaluating language model performance as in Experiment 1. The main difference is that rather than comparing models by scale (capacity and training tokens), we instead look at different metrics of model quality.

Metrics of Language Model Quality

We select 6 benchmarks designed to test language model quality in different ways, with each described below. We use the Language Model Evaluation Harness (Gao et al., 2021) to calculate each metric.

WikiText Test Set Perplexity The WikiText (Merity et al., 2017) test set is a text corpus that was explicitly designed to be used to evaluate language models. It is comprised of 60 Wikipedia articles with a total of 245,569 tokens. We calculate each model's word-level perplexity on this dataset.

BLIMP Accuracy The The Benchmark of Linguistic Minimal Pairs (BLiMP; Warstadt et al., 2020) is a benchmark designed to test language models' grammatical knowledge. BLiMP comprises 67 subsets of 1,000 sentence pairs. Each pair includes one grammatical version of a sentence and one ungrammatical version that differs by one word—for example, a grammatical sentence might be *The cats annoy Tim* and the equivalent ungrammatical sentence *The cats annoys Tim* (Warstadt et al., 2020). Each of the 67 subsets is designed to target a specific grammatical phenomenon, for example, the previous pair differ in whether they show the correct form of verb agreement for the noun subject (the plural word *cats*

agrees with *annoy* rather than *annoys*).

LAMBADA (OpenAI) Accuracy The LAnguage Modeling Broadened to Account for Discourse Aspects (LAMBADA; Paperno et al., 2016) benchmark is explicitly designed to test a language model's capability to predict a word based on a long context. Items were chosen such that human participants were generally able to guess the last word of a passage if they had read the whole passage up until the word to be predicted but not if they had only read the last sentence (Paperno et al., 2016). Thus, the task evaluates both language models' world knowledge and their ability to make predictions based on not just the preceding sentence. A language model 'answers' correctly if it successfully predicts the last word of the passage as the word with the highest contextual probability. LAMBADA has a test set made up of 5,153 items. The original LAMBADA dataset (Paperno et al., 2016) was preprocessed in a specific way to align with contemporary models at the time, such as removing capitalization. Given advances in language models since then, it has become increasingly common to use the un-preprocessed version of the benchmark developed by OpenAI (Radford et al., 2019) instead (see, e.g., Biderman et al., 2023b).

HellaSwag Accuracy HellaSwag (Zellers et al., 2019) is a task designed to test how well a language model can make predictions about how a text should continue, a task referred to by the authors as *commonsense natural language inference*. Specifically, the HellaSwag dataset is made up of captions (Krishna et al., 2017) of ActivityNet videos (a dataset of humans engaging in a wide range of activities; Heilbron et al., 2015) as well as articles from WikiHow, a website which Zellers et al. (2019) refer to as 'an online how-to manual'. The task for the language model is to identify which of a set of four candidate continuations is the most likely—the language model is considered to have 'answered' correctly if the

true continuation has the highest probability. The test set of the benchmark is made up of 10,000 such items (3,500 from ActivityNet captions and 6,500 from WikiHow).

PiQA Accuracy The 'Physical Interaction: Question Answering' (PiQA; Bisk et al., 2020) benchmark is designed to evaluate language models' world knowledge, or more specifically, the extent to which they can make predictions that align with physical knowledge about the world. It is made up of items that include a goal (stated as a question or statement) that involves accomplishing some form of physical task, and the language model has to predict which of two possible solutions is a more suitable answer or response. If the language model predicts the correct one to have a higher probability, this is considered a successful response. The PiQA dataset is made up of 3,084 such items.

WinoGrande Accuracy The WinoGrande (Sakaguchi et al., 2020) dataset is designed to be a difficult task based around Winograd schemata (Winograd, 1972; see also Levesque et al., 2012). The WinoGrande dataset is built around sentences such as *The trophy doesn't fit into the brown suitcase because* ***it****'s too large*, where the aim is to determine whether *it* in this case refers to *trophy* or *suitcase*. A language model would be correct in this case if it predicts the word *trophy* to be more likely than *suitcase* (for exact details of how this is implemented in practice, see Sakaguchi et al., 2019). An additional feature of WinoGrande that is designed to make it more difficult is that sentences were excluded where one candidate is more associated with the context than another—for example Sakaguchi et al. (2019) reject the possible item *The lions ate the zebras because* ***they*** *are predators* because *lions* are associated with being *predators* (and the words are more likely to co-occur). WinoGrande is made up of 43,972 items.

3.3.4 Results

First, we investigate how performance at each task correlates with model scale, which we present in Figure 3.2. We see the expected scaling pattern—language models with a higher number of parameters and those trained on more data perform best, with the best performance attained by the models that are largest on both of these axes. This pattern is clear for all datasets except for WinoGrande, on which the pattern also holds, but it is only the very largest models that appear to perform substantially better than chance.

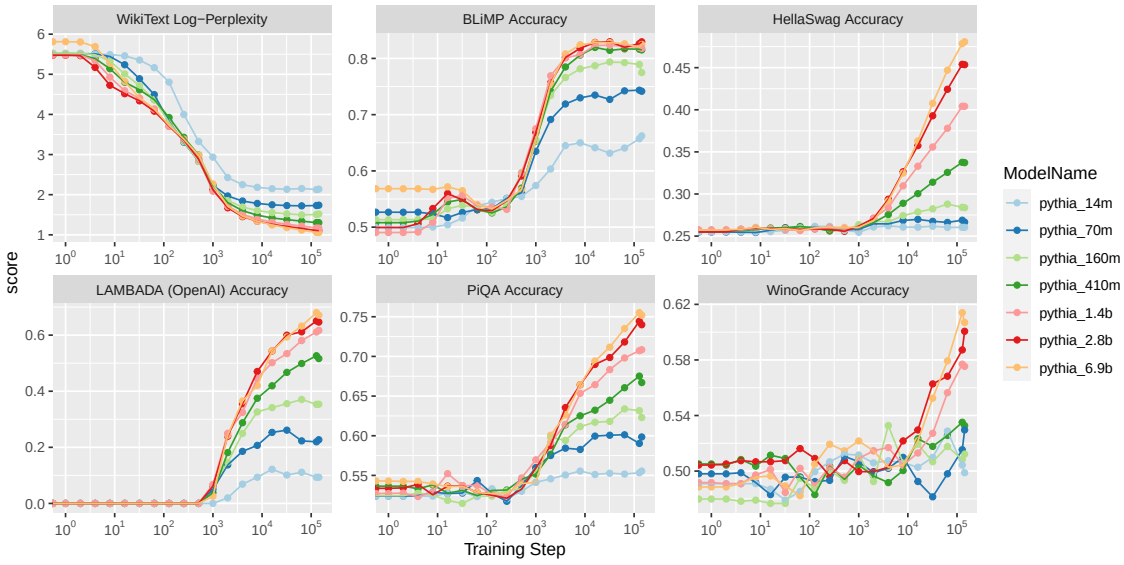

Figure 3.2: How performance at each benchmark examined varies by model and over the course of training. For WikiText Log-Perplexity, a lower score is better. For all the other benchmarks, a higher accuracy is better.

As with the scaling experiments, while there are some idiosyncracies, we see a clear pattern where higher-quality language models perform better at predicting N400 amplitude. Specifically, we see that there is a better fit to the N400 for language models with a lower perplexity (Figure 3.3), as well as those with a higher accuracy at BLiMP

(Figure 3.4), LAMBADA (Figure 3.5), HellaSwag (Figure 3.7), and PiQA (Figure 3.6). The one exception to this pattern is the WinoGrande benchmark (Figure 3.8)—while the models that best predict N400 amplitude do tend to be the models that perform best on WinoGrande, much of the variation in N400 prediction performance is not accounted for by WinoGrande performance.

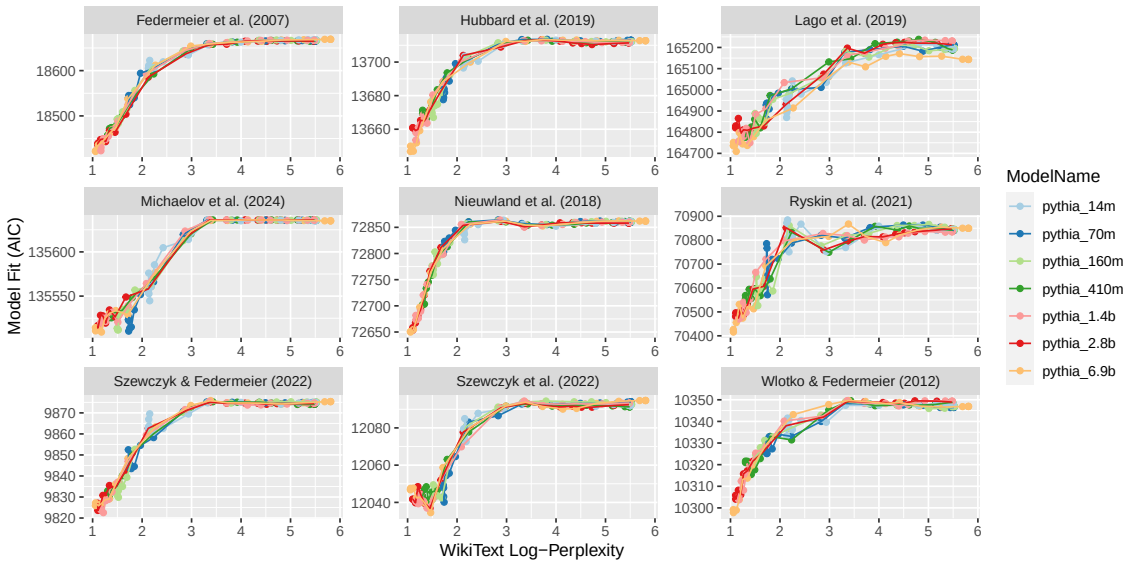

Figure 3.3: How language model performance at predicting N400 amplitude varies by each model's Log-Perplexity on the WikiText test set. A lower AIC indicates a better fit to the N400 data.

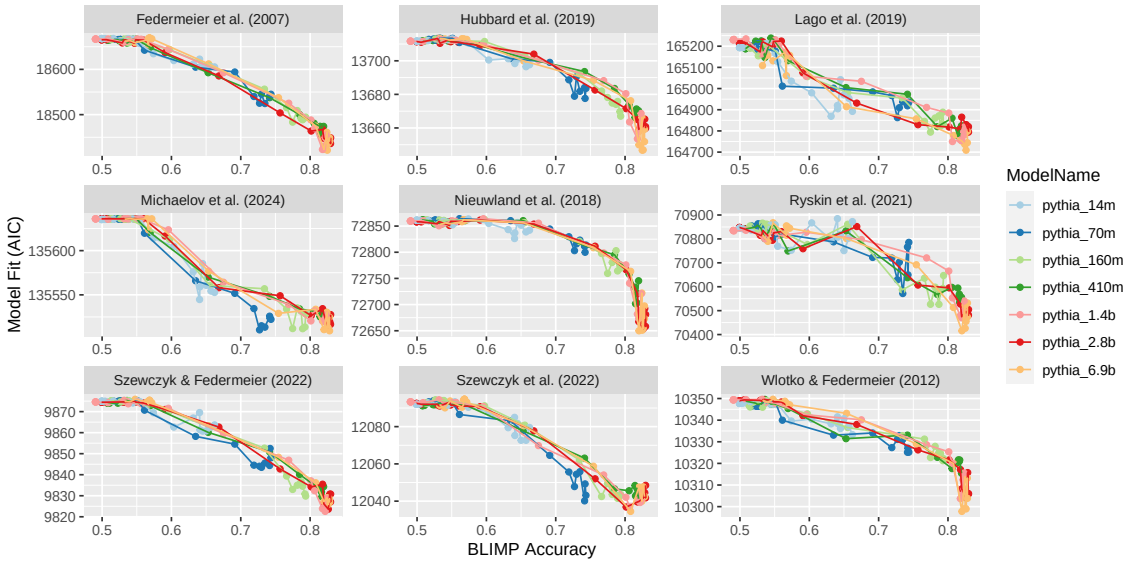

Figure 3.4: How language model performance at predicting N400 amplitude varies by each model's BLiMP accuracy. A lower AIC indicates a better fit to the N400 data.

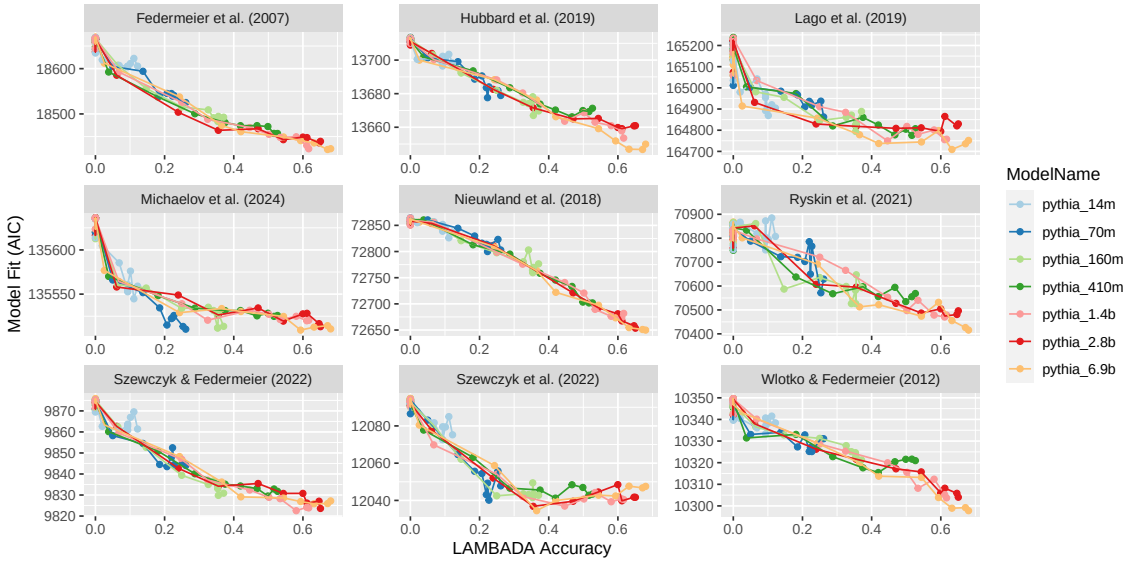

Figure 3.5: How language model performance at predicting N400 amplitude varies by each model's accuracy at the OpenAI version of the LAMBADA task. A lower AIC indicates a better fit to the N400 data.

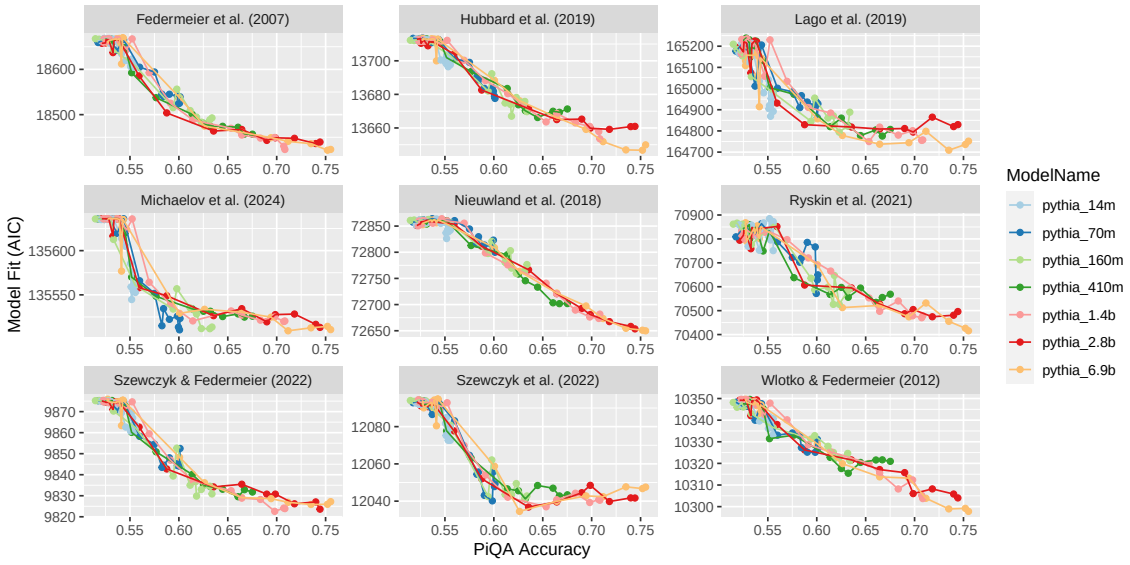

Figure 3.6: How language model performance at predicting N400 amplitude varies by each model's accuracy at the PiQA benchmark. A lower AIC indicates a better fit to the N400 data.

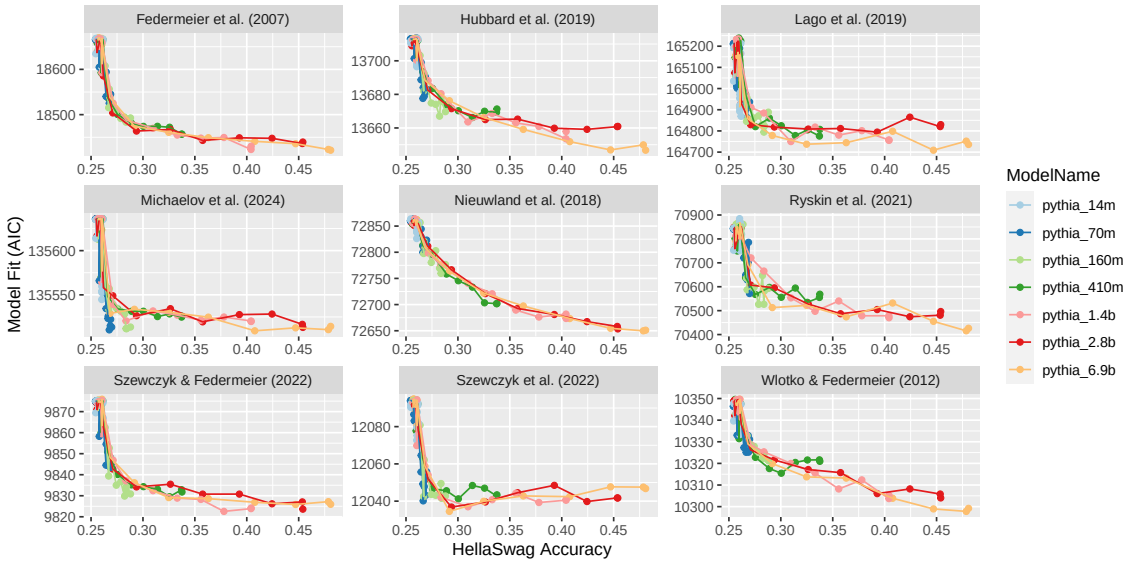

Figure 3.7: How language model performance at predicting N400 amplitude varies by each model's accuracy at the HellaSwag benchmark. A lower AIC indicates a better fit to the N400 data.

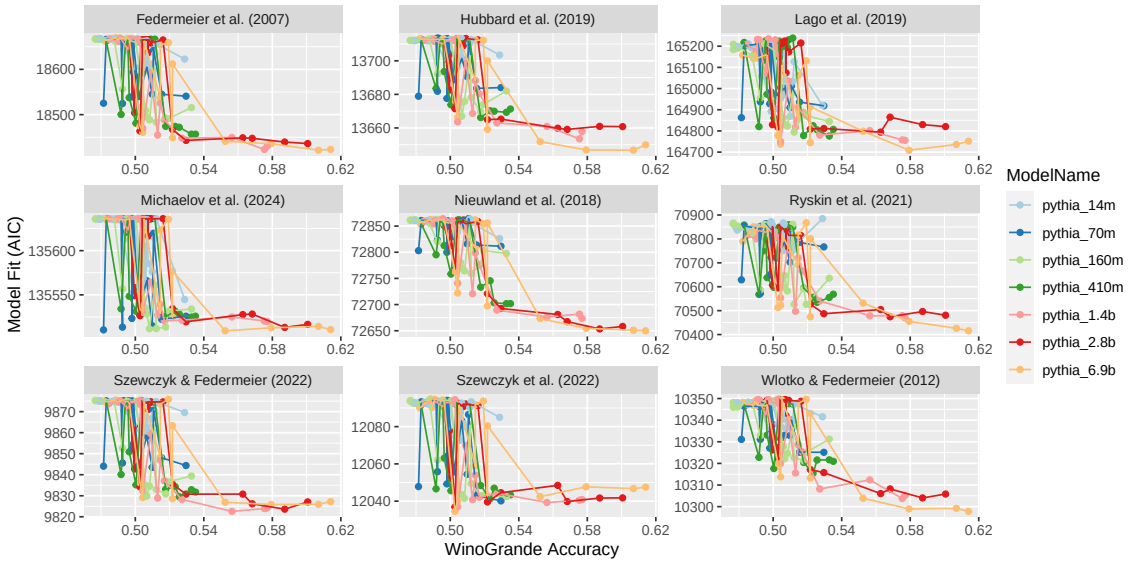

Figure 3.8: How language model performance at predicting N400 amplitude varies by each model's accuracy at the WinoGrande benchmark. A lower AIC indicates a better fit to the N400 data.

We again run Mann-Kendall tests on the same model steps (step 1000 to step 143,000—the end of training), but instead look at the trend in fit to the N400 data (AIC) ordered by the models' performance at each natural language benchark. We provide the results in Table 3.5. As can be seen, the results are even more consistent than for model scale—in all cases, better performance at the task correlates with better prediction of N400 amplitude. One possible reason for the difference in these statistical tests is that we use the data from all models for each test, increasing its statistical power.

3.3.5 Discussion

Overall, the results are clear. Higher-quality language models—whether evaluated based on the more traditional perplexity metric or on more recent benchmarks—generally perform better than lower-quality models at predicting N400 amplitude. As with scaling,

Table 3.5: The results of Mann-Kendall tests looking at the overall trend of AIC from step 1,000 to the end of training (step 143,000), ordered by performance at each benchmark. For WikiText Log-Perplexity, a lower value indicates a better score, while for all other metrics, a higher value indicates a better score. FWOK07 refers to Federmeier et al. (2007), HRJF19 to Hubbard et al. (2019), LNJL19 to Lago et al. (2019), MBVBC24 to Michaelov et al. (2024), N18 to Nieuwland et al. (2018b), RSBEFG21 to Ryskin et al. (2021), SF22 to Szewczyk and Federmeier (2022), SMF22 to Szewczyk et al. (2022), and WF12 to Wlotko and Federmeier (2012). All p-values are corrected for multiple comparisons (Benjamini and Yekutieli, 2001).

	FWOK07		**HRJF19**		**LNJL19**	
Task	τ	p	τ	p	τ	p
WikiText	0.906	< 0.0001	0.836	< 0.0001	0.687	< 0.0001
BLiMP	-0.833	< 0.0001	-0.784	< 0.0001	-0.634	< 0.0001
HellaSwag	-0.864	< 0.0001	-0.813	< 0.0001	-0.649	< 0.0001
LAM. (OAI)	-0.869	< 0.0001	-0.85	< 0.0001	-0.67	< 0.0001
PiQA	-0.887	< 0.0001	-0.836	< 0.0001	-0.642	< 0.0001
WinoGrande	-0.428	0.0029	-0.423	< 0.0001	-0.348	0.0004

	MBVBC24		**N18**		**RSBEFG21**	
Task	τ	p	τ	p	τ	p
WikiText	0.599	< 0.0001	0.907	< 0.0001	0.764	< 0.0001
BLiMP	-0.538	< 0.0001	-0.82	< 0.0001	-0.708	< 0.0001
HellaSwag	-0.571	< 0.0001	-0.873	< 0.0001	-0.755	< 0.0001
LAM. (OAI)	-0.605	< 0.0001	-0.913	< 0.0001	-0.76	< 0.0001
PiQA	-0.607	< 0.0001	-0.874	< 0.0001	-0.773	< 0.0001
WinoGrande	-0.303	0.0029	-0.478	< 0.0001	-0.376	0.0001

	SF22		**SMF22**		**WF12**	
Task	τ	p	τ	p	τ	p
WikiText	0.814	< 0.0001	0.599	< 0.0001	0.829	< 0.0001
BLiMP	-0.763	< 0.0001	-0.587	< 0.0001	-0.793	< 0.0001
HellaSwag	-0.808	< 0.0001	-0.559	< 0.0001	-0.82	< 0.0001
LAM. (OAI)	-0.818	< 0.0001	-0.572	< 0.0001	-0.821	< 0.0001
PiQA	-0.822	< 0.0001	-0.578	< 0.0001	-0.843	< 0.0001
WinoGrande	-0.425	< 0.0001	-0.241	0.0316	-0.422	< 0.0001

this is in contrast to recent work on reading time (Oh et al., 2022; Oh and Schuler, 2023a,b; Oh et al., 2024; Shain et al., 2024), suggesting that higher-quality language models are worse predictors of metrics of human language processing.

Specifically, we see that general ability to predict the next word in a sequence (i.e., perplexity) is highly correlated with the extent to which language model predictions are correlated with N400 amplitude, suggesting that, when modeling the N400, there may not be an issue of language models being able to predict language *too well* (compare Oh and Schuler, 2023a; Oh et al., 2024). This is especially highlighted by the fact that even at the logarithmic scale shown in Figure 3.3, the extent to which N400 prediction improves accelerates as perplexity lowers.

Next, we see that performance at BLiMP, LAMBADA, PiQA, and HellaSwag are correlated with the extent to which language model predictions match N400 amplitude. It is not possible to establish a causal relationship using this approach. But these results do demonstrate that language models that can better accomplish such tasks—i.e., being able to distinguish grammatical from ungrammatical sentences, and to predict text continuations that align with world knowledge and commonsense reasoning over continuations that do not—are better able to predict N400 amplitude. There's an intriguing contrast here with reading time. These results may suggest that peak prediction of N400 amplitude requires learning more complex relationships between words than does predicting reading time.

Finally, with WinoGrande, we see that while the models that perform best at the task (the largest models trained on the most data) also tend to be those that generate predictions that most closely align with N400 amplitude, for the majority of models, the two are not correlated. This suggests that being able to perform well at WinoGrande is not a prerequisite for language models to be able to predict N400 amplitude well. In other words, achieving a good WinoGrande score requires a higher-quality model than is

needed to predict N400 amplitude well, at least in our sample of language models. It does appear to be the case for several datasets, however—most clearly Nieuwland et al. (2018b) and Wlotko and Federmeier (2012)—that for the high-quality models that do begin to perform better at WinoGrande, this performance is correlated with how well the models predict N400 amplitude. Thus, while this suggests that much of the improvement in N400 improvement can be achieved by models that perform poorly at WinoGrande, being able to learn the regularities in language that lead to better WinoGrande performance may nonetheless improve N400 modeling.

3.4 Experiment 3: Negative scaling with reading time

3.4.1 Introduction

In Experiments 1 and 2, we found that unlike previous work on reading time, language models' performance at predicting N400 amplitude increases as models are trained on more text data, and that higher-quality language models predict N400 amplitude better than lower-quality models. To enable a comparison across studies, we designed Experiment 1 to match the corresponding study for reading time (Oh and Schuler, 2023a) as closely as possible, and likewise for our analysis of WikiText Perplexity in Experiment 2. However, these first two experiments also introduced several new elements, namely, the Mann-Kendall tests (Mann, 1945; Kendall, 1948) and the use of natural language benchmarks designed to probe specific language model capabilities.

To validate these choices, and to ensure that differences between N400 amplitude and reading time measures replicate with this approach, a third experiment runs the same analyses from Experiments 1 and 2 on two reading time datasets. As noted in Section 3.1, one of these (Luke and Christianson, 2018) has previously shown language

model performance decreases as the models are trained on more data and improve their next-word prediction performance (Oh et al., 2024). While the other (Smith and Levy, 2013) has been modeled computationally in previous work (Wilcox et al., 2020; Shain et al., 2024), it has not thus far been shown to have this property—in fact, unlike most of the other datasets, Shain et al. (2024) find that the larger (both in capacity and training data) GPT-3 predictions display a closer fit to the reading time data than GPT-2 predictions.

3.4.2 Data Availability

All code, data, and statistical analysis scripts are provided at `https://osf.io/qbekt/?view_only=0f4ba6296eda442aaf8e49109eac145d`.

3.4.3 Method

General Method

We follow the same general methods as in Experiments 1 and 2. We use the same language models and natural language benchmarks, and follow the same procedure for estimating fit to the data.

Reading Time Datasets

Luke and Christianson (2018) The Provo Corpus is an eye-tracking dataset collected from 470 participants reading passages collected from a range of sources such as online articles and public-domain fiction (Luke and Christianson, 2018). Following previous work (e.g. Oh et al., 2024), we look at go-past duration, a metric of the time between the first time a word is fixated and the first time gaze moves to the right of that word. We calculated fit to the data by constructing a linear mixed-effects regression model predicting log-transformed go-past duration using surprisal, word length, unigram surprisal, position

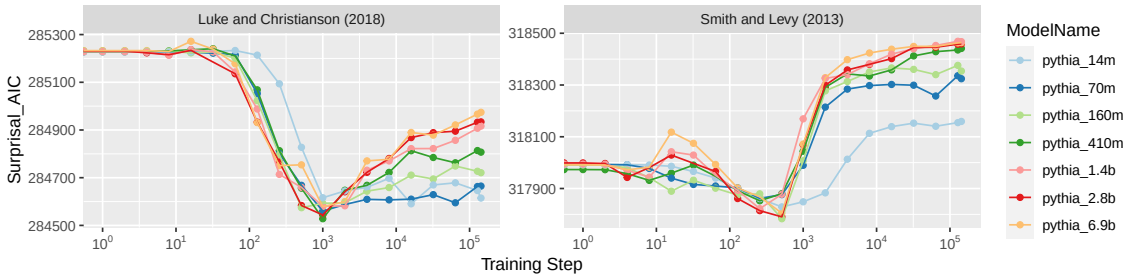

Figure 3.9: How language model performance at predicting reading time varies by model and over the course of training. A lower AIC indicates a better fit to the reading time data.

in the sentence, whether the previous word was fixated, and saccade length. We also included random slopes for each of these for each subject, as well as a random intercept for each subject and sentence.

Smith and Levy (2013) This dataset, sometimes known as the Brown SPR corpus (e.g., in Shain et al., 2024) is made up of self-paced reading times from 35 participants reading extracts from the Brown corpus of American English. Self-paced reading time was operationalized as the time between when a button was pressed to move onto the word and the next button press. We calculated fit to the data by constructing a linear mixed-effects regression model predicting log-transformed reading time with surprisal, word length, unigram surprisal, and the word's position in the sentence as fixed effects. We also included a random slope of each of these variables for each subject, as well as random intercepts for each subject and sentence.

Table 3.6: The results of Mann-Kendall tests looking at the overall trend of AIC on the reading time datasets from step 1,000 to the end of training (step 143,000), ordered by training step. All p-values are corrected for multiple comparisons (Benjamini and Yekutieli, 2001).

	Luke & Christianson (2018)		**Smith & Levy (2013)**	
Parameters	τ	p	τ	p
14M	0.053	1.0000	0.899	< 0.0001
70M	0.687	< 0.0001	0.634	< 0.0001
160M	0.74	< 0.0001	0.582	< 0.0001
410M	0.687	< 0.0001	0.899	< 0.0001
1.4B	0.952	< 0.0001	0.899	< 0.0001
2.8B	0.952	< 0.0001	0.952	< 0.0001
6.9B	0.899	< 0.0001	0.952	< 0.0001

3.4.4 Results

First, we look at how well the Pythia models predict reading time over the course of training as we did for the N400 in Experiment 1. The fit to the data for the predictions generated from each language model are shown in Figure 3.9. While different to each other, the results show the opposite pattern to those in Experiments 1 and 2 and are consistent with the findings of Oh and Schuler (2023a)—fit increases until step 1000, and then begins to decrease, with higher-capacity models showing a greater decrease. In fact, on the Smith and Levy (2013), we see that the predictions of the fully-trained language models lead to a worse fit than the untrained models.

These results are echoed in the Mann-Kendall tests (Table 3.6), where fit to the reading time data gets significantly worse over the course of training from step 1,000 to the end of training.

We also run the same analyses looking at model quality (Figure 3.10). Perplexity exhibits the same pattern as training steps—an improvement in the ability to predict

reading time, followed by a drop in performance. For the other benchmarks, how robust this is varies, but once accuracy on the benchmark improves by about 5-10% relative to the starting point, a clearer pattern emerges where better benchmark performance correlates with worse fit to reading time.

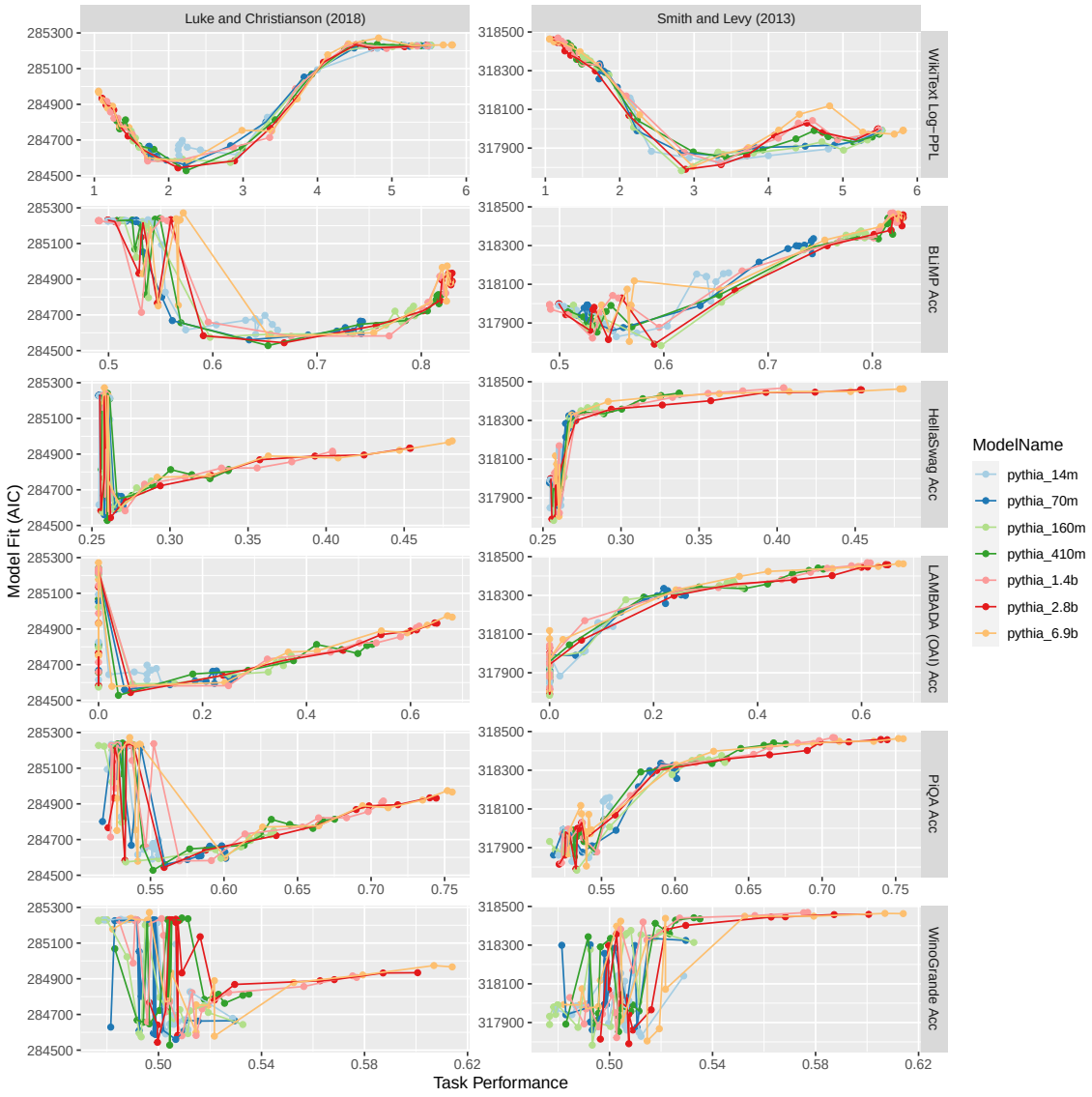

Figure 3.10: How language model performance at predicting reading time varies by each model's accuracy at the each benchmark. A lower AIC indicates a better fit to the reading time data.

This also shows up in the Mann-Kendall tests, as shown in Table 3.7. Across the board, better performance at a language model benchmark is significantly correlated with

Table 3.7: The results of Mann-Kendall tests looking at the overall trend of AIC on reading time from step 1,000 to the end of training (step 143,000), ordered by language model performance at each task. All p-values are corrected for multiple comparisons (Benjamini and Yekutieli, 2001).

	Luke & Christianson (2018)		**Smith & Levy (2013)**	
Task	τ	p	τ	p
WikiText	-0.765	< 0.0001	-0.887	< 0.0001
BLiMP	0.691	< 0.0001	0.799	< 0.0001
HellaSwag	0.76	< 0.0001	0.843	< 0.0001
LAM. (OAI)	0.774	< 0.0001	0.873	< 0.0001
PiQA	0.745	< 0.0001	0.852	< 0.0001
WinoGrande	0.535	< 0.0001	0.47	< 0.0001

a worse fit to the reading time data.

3.4.5 Discussion

The results of the experiment are clear: beyond 2 billion tokens, models trained on more data are worse predictors of reading time, with higher-capacity models performing worse than lower-capacity models. Similarly, higher-quality models predict reading time less well than lower-quality models. These results provide further support to the hypothesis that language models can become too good at next-word prediction to be able to model human reading time (Oh and Schuler, 2023a,b; Oh et al., 2024). Because we arrived at them in the same way as Experiments 1 and 2, these results also indirectly provide further support for the conclusions based on those studies—the pattern observed for reading time does not appear to apply to N400 amplitude.

3.5 General Discussion

In this study, we carry out the first analysis systematically comparing how model capacity, number of training tokens, and multiple metrics of quality of language models correlate with how well a language model can predict N400 amplitude. In contrast to the recent work on reading time (Oh et al., 2022; Oh and Schuler, 2023a,b; Oh et al., 2024; Shain et al., 2024) as well as a reading time replication reported here in Experiment 3, we find no evidence of a decrease in performance for any of these with more training, more capacity, or better performance in the N400.

While similar indications exist for small recurrent neural networks (Aurnhammer and Frank, 2019b), the present study for the first time establishes a relationship between between training tokens and N400 prediction for transformers. We find that transformer language models trained on more tokens predict N400 amplitude better. These results are also consistent with previous findings that all else being equal, models with a larger capacity predict N400 amplitude better than smaller models (Michaelov and Bergen, 2022b).

We also scale up the finding that higher-quality transformers predict N400 amplitude better (Merkx and Frank, 2021) in two ways. First, we find that this result holds for models with a much wider range of perplexities (not including step 0, roughly 11–650,000, compared to roughly 90–5000 in Merkx and Frank, 2021). Current results also extend this line of research by showing that performance at other natural language benchmarks assessing both syntactic and semantic processing is correlated with the extent to which a language model can be used to predict N400 amplitude.

Overall, the results are clear: better language models are better at predicting N400 amplitude. The practical upshot of Experiment 2 on language quality is that perplexity and benchmark performance can be reliable indicators of how well a language model will predict N400 amplitude, at least for models trained only on autoregressive language

modeling (i.e., pretrained-only or 'foundation' models Bommasani et al., 2021). This is especially important if larger and higher-quality language models continue to prove to be better at predicting N400 amplitude, as training state-of-the-art models of this type requires resources beyond those available to most researchers studying human language comprehension. The results of this study suggest that even if relying on pretrained models that are not fully open-source (see discussion in Bommasani et al., 2023a, 2024; Lambert et al., 2023; White et al., 2024), it may still be possible to assess how suitable a language model is for predicting N400 amplitude using purely empirical measures.

3.5.1 Theoretical Implications and Further Discussion

The main question raised by what precedes is how to explain the difference between the reading time and N400 results—specifically, why does performance at predicting reading time begin to decrease after models are trained beyond 2B tokens, whereas performance at predicting N400 amplitude continues to increase? Finding an answer may require a deeper look at the nature of the N400 itself.

A relatively uncontroversial mechanistic account of the N400 consistent with the majority of current evidence is that the N400 reflects the activation of semantic representations in long-term memory driven by a stimulus, and that this activation is attenuated by the extent to which these representations were already activated at the time at which the stimulus is encountered (Kutas and Federmeier, 2011; Van Petten and Luka, 2012; Kuperberg et al., 2020; Federmeier, 2021). There are a variety of accounts of what drives this 'preactivation.' The first is prediction—of particular words, semantic features, or even of formal aspects of words such as phonology or grammatical features (for discussion, see, e.g., DeLong et al., 2005; Nicenboim et al., 2020). Others include semantic association between the stimulus and individual words or events in the context (Delogu et al., 2019;

Aurnhammer et al., 2021; Van Petten, 2014), an overlap in the semantic features shared by the word and previous words in the context (Federmeier, 2021), or a combination of these (Federmeier, 2021). In any case, however, a fundamental feature of the N400 is that it is highly sensitive to the semantics of an utterance and its context.

This is important in light of the fact that one of the key commonalities shared by four of the six benchmarks analyzed is the extent to which semantic knowledge is crucial to solving them. The clear, positive correlation between fit to the N400 data and performance at HellaSwag, PiQA, and LAMBADA is suggestive that it may be precisely the ability to make predictions based on the semantics of the context that leads to the improved performance at predicting N400 amplitude. The relatively muddier relationship to WinoGrande performance may also be further evidence of this. As mentioned earlier, WinoGrande was intentionally designed to avoid semantic associations between the context and the critical word. If both humans and language models use of this context in processing—respectively generating the N400 effect and word surprisal—then tasks in which it's impossible to use semantic associations with context may be less useful at predicting the model-human fit. Indeed, some accounts propose that associative mechanisms lead to preactivation in human language comprehension; in some cases, overlap in semantic features between the critical word and the words of the context (Federmeier, 2021) are sufficient to account for the N400. Computational approaches that use the distance between word vectors to model the degree of association between words (e.g., Van Petten, 2014; Frank and Willems, 2017) work similarly, and the results of Michaelov and Bergen (2022a) and Michaelov et al. (2024) show that a purely predictive language model can emergently predict more semantically associated words to be more likely (and this may be the case to a greater extent for better models, see, e.g., Michaelov et al., 2021). In any case, contextual association is a useful cue in predicting, and thus it is possible that predictive systems like language models rely

on this, and so they are better able to make predictions when the context is informative.

Overall, then, the benchmark results suggest that when it comes to predicting the N400, a model that is better able to deal with the semantics of a text will better fit electrophysiological measures, to the extent that this improvement may outweigh its better-than-human next-word prediction capabilities that are such a problem for predicting reading time. Reading time, by contrast, may not be as sensitive to semantic coherence with context, which could explain why better, more semantically sensitive language models trend downwards as they scale up.

3.5.2 Limitations and Future Research

There are other possible explanation for the difference between the N400 and reaction time. Perhaps the most important one lies with the differences in the stimuli themselves used in the various studies. While the N400 datasets we analyze use stimuli explicitly constructed for their respective experiments, the reading time datasets on which inverse scaling has been observed in previous work (e.g., by Oh et al., 2022; Oh and Schuler, 2023a,b; Oh et al., 2024; de Varda and Marelli, 2023; Shain et al., 2024) are naturalistic in that they are either already-existing texts or based on them.

There are several different ways in which this could lead to differences in the extent to which the predictions of language models match humans measures. One is that with constructed stimuli, critical words are generally of the same part of speech (often content words such as nouns or verbs), while with naturalistic reading corpora, a wide variety of parts of speech are included. Language models have been found to show different levels of performance at predicting reading time (Oh and Schuler, 2023a) and the N400 (Frank et al., 2015) depending on part of speech, which could account for some of the difference. Indeed, Oh and Schuler (2023a) note that a key issue with using higher-quality models to

predict reading time is that because they are able to predict low-probability words better than humans, when a regression is fit to the data, this regression is likely to systematically over-predict other more probable words such as function words.

The sourcing of the reading time datasets that show negative scaling effects either directly from or based on already-existing text may also present another problem in itself—that of 'data leakage' (Wilcox et al., 2023a). Language models have been shown to 'memorize' data, with larger models showing this to a greater extent (Tirumala et al., 2022; Carlini et al., 2022; Biderman et al., 2023a). Thus, if a language model is trained on data that appears one of the reading time studies, this may artificially increase how likely a language model predicts a certain word to be—and it is precisely the over-prediction of unlikely and low-frequency words that has been argued to be at least one of the driving forces behind the negative scaling effects on reading time (Oh and Schuler, 2023b; Oh et al., 2024). Thus far, the results of the one study investigating this possibility suggests that it may not be as much of a problem as it may sound (Wilcox et al., 2023a), but further work is needed before it can completely be discounted.

One way to tease apart data-based explanations like these from more theoretically interesting ones would be to use reading time and N400 data drawn from the same data sets. We leave that however for future research.

3.6 Conclusion

We investigated how the number of tokens on which a language model is trained impacts how well it can predict N400 amplitude, finding a positive relationship—overall, models trained on more data predict N400 amplitude better. We also find that higher-quality models—those that perform better on natural language benchmarks—consistently predict N400 amplitude better than lower-quality models. These results hold even for

language models that show worse performance at predicting reading time as they increase along each of these dimensions. This difference may allow for a better understanding of the factors that go into and differentiate human measures like reading time and the N400—including semantic associations with context—as well as into language model prediction.

3.7 Acknowledgements

The experiments with the language models were carried out using hardware provided by the NVIDIA Corporation as part of an NVIDIA Academic Hardware Grant.

Chapter 3, in full, is a reprint of a manuscript currently under review for publication as Michaelov, J. A., & Bergen, B. K., "Better language models better model the N400". The dissertation author was the primary investigator and author of this paper. Minor edits have been made to bring the formatting in line with the dissertation template.

Part II

Can language models be used to model N400 effects?

Chapter 4

How well does surprisal explain N400 amplitude under different experimental conditions?

Abstract

We investigate the extent to which word surprisal can be used to predict a neural measure of human language processing difficulty—the N400. To do this, we use recurrent neural networks to calculate the surprisal of stimuli from previously published neurolinguistic studies of the N400. We find that surprisal can predict N400 amplitude in a wide range of cases, and the cases where it cannot do so provide valuable insight into the neurocognitive processes underlying the response.

4.1 Introduction

The N400 component of the event-related brain potential is generally understood to be a neural signal of processing difficulty (Kutas and Federmeier, 2011). After over 1,000 articles published on the topic, we know that all else being equal, an upcoming word that is supported by the semantics of the context will elicit a lower-amplitude N400 than a word that is not (Kutas and Federmeier, 2011; Kuperberg et al., 2020). However, despite the great amount of experimental research on the topic, many aspects of the N400 are still not well understood.

In addition to 'long-standing and recent linguistic [...] inputs' (Kutas and Federmeier, 2011, p. 641), the context that impacts N400 amplitude is thought to include factors such as world experience, attentional state, and mood (Kutas and Federmeier, 2011). Over the last decade, there have been a number of attempts to use computational modeling to test hypotheses about the neurocognitive processes underlying the N400 and how the aforementioned factors may impact its amplitude (Parviz et al., 2011; Laszlo and Plaut, 2012; Laszlo and Armstrong, 2014; Rabovsky and McRae, 2014; Frank et al., 2015; Ettinger et al., 2016; Cheyette and Plaut, 2017; Brouwer et al., 2017; Delaney-Busch et al., 2017; Rabovsky et al., 2018; Venhuizen et al., 2019; Fitz and Chang, 2019).

As the majority of experimental research on the N400 involves manipulating the relationship between the stimulus and the preceding linguistic context (Kutas and Federmeier, 2011), a computational account of how linguistic inputs impact N400 amplitude is a logical starting point. Language models are inherently models of linguistic prediction based only on language input. Since N400 amplitude reflects how unexpected an upcoming word is based on context, the predictions of a language model can be used to model how expected a word is based on the linguistic input, and thereby investigate the extent to which N400 amplitude is explainable by linguistic input alone.

Recent research has shown that *surprisal*, a measure of how unlikely a language model predicts the next word in sequence to be, correlates overall with N400 amplitude (Frank et al., 2015; Aurnhammer and Frank, 2019b). Thus, to investigate the extent to which N400 amplitude is explained by linguistic input alone, we ask to what extent surprisal can explain the variance observed in N400 amplitude.

In order to investigate this, we run experimental stimuli from eleven experiments from six papers (Urbach and Kutas, 2010; Kutas, 1993; Ito et al., 2016; Osterhout and Mobley, 1995; Ainsworth-Darnell et al., 1998; Kim and Osterhout, 2005) through two recurrent neural network language models (Jozefowicz et al., 2016; Gulordava et al., 2018), systematically comparing the significant predictors of N400 amplitude and surprisal. We find that in the majority of cases, significant differences in surprisal predict significant differences in N400 amplitude, and discuss the implications of the cases where it does not.

4.2 Background

4.2.1 The N400

The N400 is a negative deflection in the event-related brain potential (ERP) that peaks roughly 400ms after the presentation of a stimulus (Kutas and Hillyard, 1980; Kutas and Federmeier, 2011). Most current accounts agree that N400 amplitude reflects processing difficulty for a specific lexical item, where a lower amplitude reflects prior activation of some of the semantic content associated with the word (Kutas and Federmeier, 2011; Kuperberg, 2016; Kuperberg et al., 2020).

Recent research has found that N400 amplitude '*decreases* with supportive context, but does not *increase* when predictions are violated' (DeLong and Kutas, 2020, p. 2, emphasis in original; see Kutas and Federmeier, 2011; Van Petten and Luka, 2012; Luke

and Christianson, 2016; Kuperberg et al., 2020, for discussion). Crucially, therefore, we should not think of N400 amplitude as a general measure of prediction error. It is not the case that the N400 elicited by a word increases when the word is more semantically anomalous or unexpected based on the preceding context; rather, it is the case that N400 amplitude is reduced when the word is semantically congruous or predictable because it is facilitated by the preceding context.

This facilitation can occur in a large number of ways. All else being equal, words that are more semantically congruous, typical, or plausible completions of a sentence elicit lower N400 amplitudes than words that are more semantically incongruous, atypical, and implausible completions, respectively (e.g. Kutas and Hillyard, 1980; Urbach and Kutas, 2010; Ito et al., 2016; Osterhout and Mobley, 1995; Ainsworth-Darnell et al., 1998; Kim and Osterhout, 2005; Kutas and Federmeier, 2011).

One well-known correlate of N400 amplitude is the cloze probability (Taylor, 1953; Bloom and Fischler, 1980) of a word—the probability that it will be offered to fill a specific gap in a sentence by a given sample of individuals in a norming study. All else being equal, higher-cloze completions elicit lower N400 amplitudes (Kutas and Hillyard, 1984; Kutas and Federmeier, 2011). Additionally, even when matched for cloze, words semantically related to the highest-cloze completion elicit lower-amplitude N400s than unrelated words (Kutas, 1993; Federmeier and Kutas, 1999; Ito et al., 2016).

4.2.2 Cognitive Plausibility of RNN-LMs in N400 modeling

To disentangle the effect of linguistic input from other factors affecting N400 amplitude, a valid model of such linguistic input is needed. Recurrent Neural Network Language Models (RNN-LMs) are, in many ways, perfect models of the 'long-standing and recent linguistic [...] inputs' (Kutas and Federmeier, 2011, p. 641) thought to impact N400

amplitude. Long-standing linguistic inputs in humans are made up of previous language experience, which is analogous to a model's training data; and recent linguistic input is the linguistic context that impacts how humans understand the current utterance, which is analogous to the word sequence preceding the word to be predicted in the model's test data.

Beyond being largely developed as models of human language comprehension (Elman, 1990), recurrent neural network language models (RNN-LMs) have certain properties that make them reasonable models of human cognition. Keller (2010) identifies five features of the human language processing system that he argues are vital for a language model to be cognitively plausible. Three of these are exemplified by unidirectional RNN-LMs—like humans, they can make *predictions* about upcoming words, have a distance-based *memory cost*, and process language word-by-word in order in an *incremental* fashion (unlike bidirectional RNN-LMs and most transformer networks). The two remaining features, *efficiency and robustness* and *broad coverage* are determined more by the model's specific architecture and training than general architecture.

4.2.3 Surprisal and N400 amplitude

As discussed in Section 4.2.1, the neurolinguistic evidence suggests that the N400 is a measure of lexical processing difficulty. Recent work, both theoretical and experimental (e.g. Hale, 2001; Levy, 2008; Boston et al., 2008; Demberg and Keller, 2008; Smith and Levy, 2008; Roark et al., 2009; Brouwer et al., 2010; Mitchell et al., 2010; Monsalve et al., 2012; Fossum and Levy, 2012; Frank and Thompson, 2012; Smith and Levy, 2013; Frank, 2014; Willems et al., 2016; Delaney-Busch et al., 2017), has argued that surprisal, the negative logarithm of the probability of a word w_i given its preceding context $w_1...w_{i-1}$, as shown in Equation (4.1), is a good predictor of lexical processing difficulty.

$$S(w_i) = -\log P(w_i|w_1...w_{i-1}) \tag{4.1}$$

Several researchers (Frank et al., 2015; Delaney-Busch et al., 2017; Aurnhammer and Frank, 2019b) have directly demonstrated that surprisal is correlated with N400 amplitude. In their study, Delaney-Busch et al. (2017) use a Bayesian approach to calculate the surprisal associated with a target word given a related or unrelated prime (using word association norms and word frequency), and find that this is correlated with N400 amplitude. Frank et al. (2015) and Aurnhammer and Frank (2019b) used a number of language models (including RNN-LMs) to calculate the surprisal of words in a natural language text, and compared this to the N400 elicited by these words in human participants, finding a statistically significant correlation.

Frank et al. (2015) and Aurnhammer and Frank (2019b) also find that surprisal is a better predictor of N400 amplitude than a number of RNN-LM-derived metrics based on the full probability distributions predicted by the model such as entropy. We suggest that this may be explained by the aforementioned finding that while the N400 amplitude for a word decreases when its semantic content has been pre-activated, it does not increase when a specific prediction is violated. In other words, N400 amplitude is a kind of positive prediction error—a measure of how not-predicted the target word was. This is what surprisal is by definition—it only takes into account how much the actual target word was predicted and is not affected by the rest of the probability distribution. The other metrics, on the other hand, also take into account the rest of the predicted probability distribution, which does not appear to be reflected in N400 amplitude. Thus, there is a theoretical reason for using surprisal to predict N400 amplitude based on previous neurolinguistics research.

4.2.4 Predicting N400 effects

An alternative approach, that taken by Ettinger et al. (2016), is to use a language-model-derived metric as an analogue of the N400 and investigate whether experimental manipulations in the stimuli that result in statistically significant differences in N400 amplitude also result in statistically significant differences in the chosen metric. This approach allows researchers to investigate whether the reason for the correlation between the metric and N400 amplitude is in fact the experimental manipulation or some other factor.

This is the general approach that we take in this study; however, rather than focusing on the cosine similarity between the word embedding of target word and the combined embeddings of the previous words in the sentence (Ettinger et al., 2016), we model N400 amplitude as surprisal (following Frank et al., 2015; Delaney-Busch et al., 2017; Aurnhammer and Frank, 2019b). Additionally, whereas Ettinger et al.'s (2016) proof-of-concept paper is based on 40 sample sentences from a single study investigating one phenomenon, we use stimuli from eleven experiments (with over 100 sentences each) covering a wide range of phenomena.

4.2.5 Other Models of N400 amplitude

While a number of other researchers have used neural networks to model specific N400 findings this way (Laszlo and Plaut, 2012; Laszlo and Armstrong, 2014; Rabovsky and McRae, 2014; Cheyette and Plaut, 2017; Brouwer et al., 2017; Rabovsky et al., 2018; Venhuizen et al., 2019; Fitz and Chang, 2019), these studies differ in that these models all have semantic representations as part of their input or are trained to learn to output some form of semantic representation. Thus, these models are also limited to the experiments for which they were trained.

For the same reason, these models can also not be used on their own to disentangle

the effects of linguistic input from the semantic knowledge provided to them—this can only be done by comparison to models without this. While two of the studies compare their models to simple recurrent networks (SRNs) trained on the same data (Rabovsky et al., 2018; Fitz and Chang, 2019), these SRNs are not representations of the extent of what is possible with lingusitic input alone—these models are simple (for example, they do not use long short-term memory), and much of the power of RNNs comes from large training datasets (see, e.g., the discussion in Chelba et al., 2014).

Finally, it should be noted that while all of the studies discussed in this section aim to model real N400 effects, only two (Laszlo and Armstrong, 2014; Rabovsky and McRae, 2014) use stimuli from real N400 experiments; in the remaining studies, stimuli are chosen to represent manipulations that studies have found to influence N400 amplitude. Given that the N400 is still not fully understood, it is important to verify that the experimental manipulations investigated actually do elicit the expected N400 effect. For this reason, we only use experimental stimuli provided for published N400 experiments, and compare the effect on surprisal directly to the reported effects on N400 amplitude.

4.3 Approach, Motivations, and Hypotheses

The aim of this study is to investigate the boundary conditions of using surprisal to model N400 amplitude. While there is evidence that surprisal and N400 amplitude are correlated overall (Frank et al., 2015; Aurnhammer and Frank, 2019b), it is unclear what variance in N400 amplitude is actually being explained by surprisal. While it is tempting to assume that surprisal is correlated with the N400 because the same factors that lead to reduced N400 amplitudes lead to reduced surprisal, this has thus far not been shown empirically.

This is the question that we investigate in this paper: which experimental manip-

ulations that elicit a difference in N400 amplitude elicit the same difference in surprisal, and which do not?

We do this by running the (English language) stimuli from previously published N400 studies through two neural networks that have been used extensively to model human language processing (e.g., in Wilcox et al., 2018; Futrell et al., 2019; Wilcox et al., 2019; An et al., 2019; Costa and Chaves, 2020). The two models used are the the best English LSTM from Gulordava et al. (2018) and BIG LSTM+CNN INPUTS from Jozefowicz et al. (2016), henceforth (following Futrell et al., 2019) GRNN and JRNN, respectively. These models are both LSTM-RNN-LMs, but differ most notably in size and training data: The JRNN has two hidden layers (8192 and 1024 units), a 793471-word vocabulary, and was trained on 1 billion tokens (Chelba et al., 2014); while the GRNN has two hidden layers (both 650 units), a 50000-word vocabulary, and was trained on 90 million tokens.

In addition to answering questions about the nature of the neurocognitive systems underlying the N400, the results of this study also serve as a baseline for future research—they represent the best that current cognitively plausible neural network language models can do at predicting N400 amplitude using surprisal. Thus, future research that argues for additional sources of information or neurocognitive processes being involved in the N400 on the basis of modeling success should demonstrate that the inclusion of such components in the model improves upon the results presented here.

This aim of establishing a useful baseline is another reason for our choice of models—both are provided pre-trained by the authors, allowing for our results to be replicated and expanded upon. We also only use sets of stimuli that have been made available in papers or their supplementary materials. The stimuli from these papers (Urbach and Kutas, 2010; Kutas, 1993; Ito et al., 2016; Osterhout and Mobley, 1995; Ainsworth-Darnell et al., 1998; Kim and Osterhout, 2005), which cover a range of experimental manipula-

tions that are discussed in Section 4.4, are included in text format in our supplementary materials[1].

4.4 Experiments

Figure 4.1 is a visualization of the findings of the original N400 studies and the results of the simulations. Given the differences in measurements, there is no scale—the heights of the bars indicate which conditions elicited higher or lower N400 amplitudes or surprisals relative to the others in the same experiment or simulation. All and only the significant differences between conditions for significant predictors of the N400 or surprisal are shown, not including significant interactions with recording locations on the scalp (which are beyond the scope of the present study). Black bars represent successful modeling of the differences in N400 amplitude, red bars represent unsuccessful or partially unsuccessful modeling, and purple bars indicate that the results are more complex than can be represented in this way. Only stimuli sets with over 100 stimulus sentences were run through the models (GRNN and JRNN); and while the models were not able to predict the surprisal of all target words (due to limited vocabularies or being unable to process certain characters in sentences), both models successfully calculated the surprisals of over 100 target words in each study. Stimuli, target word surprisals, and the code used to run the models are all included in our supplementary materials.

Where possible, the significant predictors of the surprisal of the GRNN and JRNN models were selected via backwards model selection using likelihood ratio tests of linear-mixed effects models (R Core Team, 2020; Bates et al., 2015) with and without the predictor under investigation as a main effect. When this was not possible, the significance of predictors were evaluated using a Type III ANOVA with Satterthwaite's method for estimating

[1] `https://github.com/jmichaelov/does-surprisal-explain-n400`

degrees of freedom (Kuznetsova et al., 2017). Significant differences between experimental conditions (i.e. between the levels of a predictor) were calculated via t-test based on the selected linear-mixed effects model, using Satterthwaite's method to estimate degrees of freedom (Kuznetsova et al., 2017). In this paper, significant predictors and significant differences between conditions are considered those where $p < 0.05$ in the relevant statistical test. All code for the statistical analyses is included in our supplementary materials.

The remainder of this section discusses the experiments (and the original N400 studies on which they are based) in more detail.

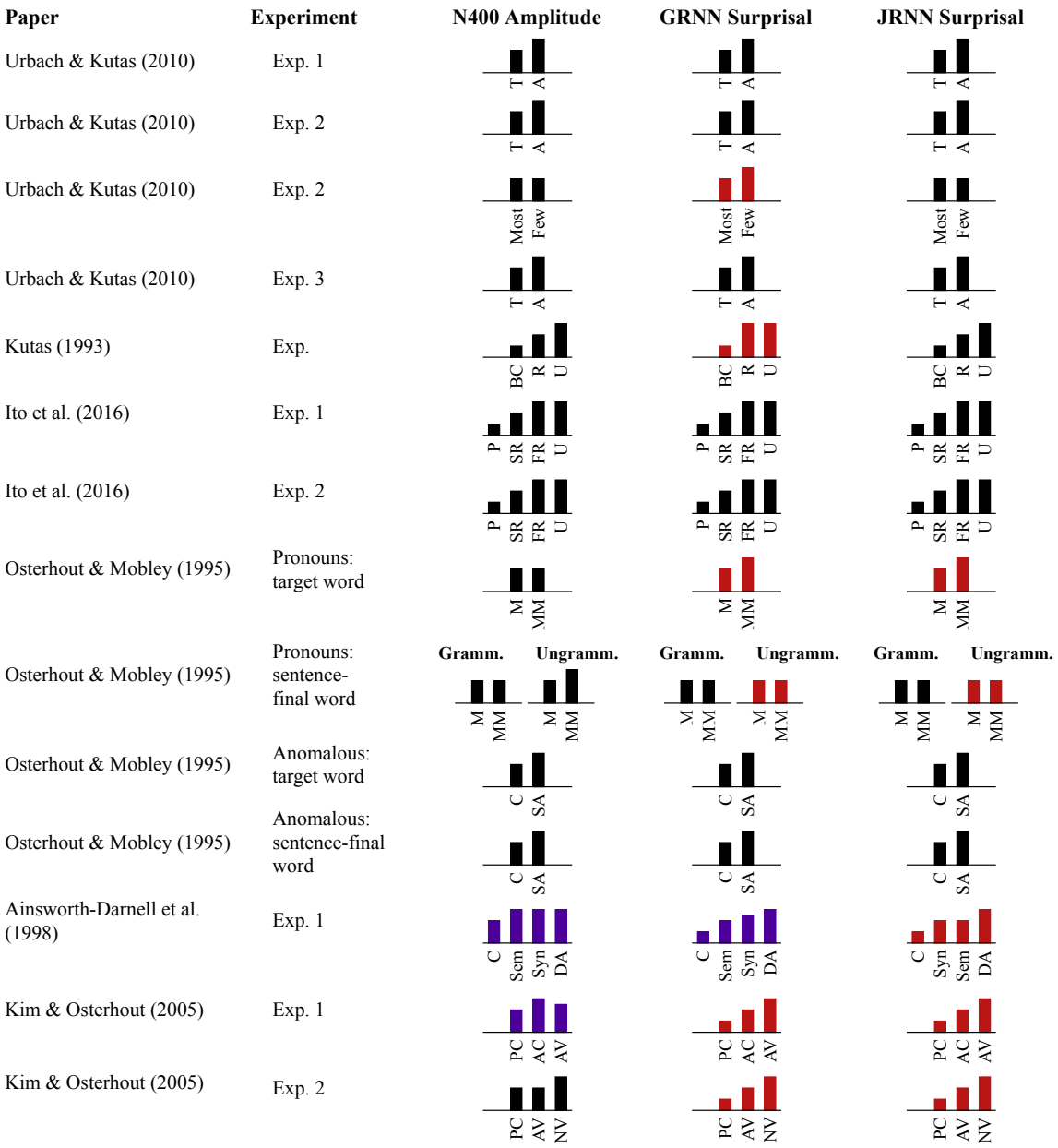

Figure 4.1: The significant differences between all conditions of significant predictors of N400 amplitude in the original studies and the surprisal of the GRNN and JRNN models. **Black** bars indicate successful modeling of the differences in N400 amplitude, **red** bars indicate unsuccessful or partially unsuccessful modeling, and **purple** bars indicate that the results are more complex than shown.

4.4.1 Urbach and Kutas (2010): Experiment 1

Experiment 1 of Urbach and Kutas (2010) investigates the N400's sensitivity to the typicality of a patient of a described event. There were two kinds of sentences in this experiment exemplified by the following stimulus pair: *prosecutors accuse* ***defendants*** (TYPICAL; T in Figure 4.1) / ***sheriffs*** (ATYPICAL; A) *of committing a crime.* As expected, the N400 elicited by TYPICAL object nouns is significantly lower in amplitude than that elicited by ATYPICAL object nouns.

Typicality was also a significant predictor of the surprisal of both the GRNN and JRNN models (GRNN: $p < 0.001$; JRNN: $p < 0.001$), with TYPICAL object nouns eliciting a lower surprisal than ATYPICAL ones (GRNN: $p < 0.001$; JRNN: $p < 0.001$).

4.4.2 Urbach and Kutas (2010): Experiment 2

Expanding on Experiment 1, Urbach and Kutas (2010) ask whether the results are affected by whether the sentences begin with the word *most* or *few* (or synonymous expressions), e.g. ***most*** *prosecutors accuse defendants.* The main effect of typicality remained. In addition, while the main effect of quantifier type was not significant overall (nor was there an interaction with typicality without an interacting electrode location variable), Urbach and Kutas (2010) found that FEW-type quantifiers reduced the N400 amplitude of ATYPICAL patients and reduced the extent to which N400 amplitude was lowered for TYPICAL patients, with this latter effect being found to be statistically significant via t-test.

Typicality predicted the surprisals of both RNNs in the same direction as in Experiemnt 1 ($p < 0.001$ for all statistical tests). The surprisal of the GRNN was also significantly predicted by quantifier type ($p < 0.001$), with FEW-type quantifiers eliciting significantly higher surprisals ($p < 0.001$). As this pattern is limited only to the GRNN (and the analogous main effect does not appear in Experiment 3 for either model), this

finding is not considered further. The t-test comparing the N400 of TYPICAL objects under the FEW and MOST quantifiers does not replicate with surprisal—there is no significant difference (GRNN: $p = 0.107$; JRNN: $p = 0.249$).

4.4.3 Urbach and Kutas (2010): Experiment 3

Experiment 3 of Urbach and Kutas (2010) is a variant of Experiment 2. Instead of MOST or FEW sentence beginnings, the words *often* or *rarely* appear after the subject (agent) noun, e.g. *prosecutors **often** accuse defendants of committing a crime.* The aim of this was to investigate whether proximity of the quantifier to the target noun had an effect. Urbach and Kutas (2010) again found the same result—only typicality was a significant predictor of N400 amplitude overall; and a t-test found that the N400 reduction for TYPICAL nouns was attenuated by the word *rarely.*

GRNN and JRNN surprisals were only significantly predicted by typicality, with typical nouns eliciting a lower surprisal than atypical nouns ($p < 0.001$ for all tests). The t-test comparing the N400 of TYPICAL objects under the FEW and MOST quantifiers does not replicate with surprisal—there is no significant difference (GRNN: $p = 0.367$; JRNN: $p = 0.283$).

4.4.4 Kutas (1993)

Kutas (1993) examines the effect of relatedness to the BEST COMPLETION (the highest-cloze completion). An example of a BEST COMPLETION (BC) and RELATED completion can be demonstrated by the following stimulus pair: *The pizza was too hot to **chew*** (RELATED; R) / ***eat*** (BC). An example of a BC and UNRELATED pair is the following sentence: *The paint turned out to be the wrong **consistency*** (UNRELATED; U)/ ***color*** (BC). BC nouns were found to elicit the lowest N400 amplitude, followed by RELATED nouns,

followed by UNRELATED nouns. Experimental condition is a significant predictor of both GRNN and JRNN surprisal. However, while the surprisals in the GRNN are different between the BC and other nouns ($p < 0.001$ for both RELATED and UNRELATED), there is no significant difference between RELATED and UNRELATED ($p = 0.820$). On the other hand, the surprisals of the JRNN are lowest for BC nouns, followed by RELATED nouns, followed by UNRELATED nouns ($p < 0.001$ for all pairwise comparisons).

4.4.5 Ito et al. (2016): Experiments 1 and 2

Ito et al. (2016) further investigate the relatedness effect by investigating whether a word that is related in form to the most PREDICTABLE word (i.e. the best completion) has a similar effect on N400 amplitude as being semantically related. The conditions can be illustrated with the following example sentence: *The student is going to the library to borrow a* ***book*** (PREDICTABLE; P)/ ***hook*** (FORM-RELATED; FR)/ ***page*** (SEMANTICALLY RELATED; SR)/ ***sofa*** (UNRELATED; U) *tomorrow.* In both Experiments 1 and 2, where the difference was in the amount of time that the stimuli were presented, Ito et al. (2016) found that experimental condition was a significant predictor, and specifically that PREDICTABLE words elicited the lowest N400 amplitude, followed by SEMANTICALLY RELATED words, followed by the FORM-RELATED and UNRELATED completions, which did not differ in N400 amplitude.

We found the same pattern in the surprisal of both models ($p < 0.001$ for condition as a predictor; $p < 0.001$ for all significant pairwise comparisons; FR vs. U with GRNN surprisal: $p = 0.080$; FR vs. U with JRNN surprisal: $p = 0.399$).

4.4.6 Osterhout and Mobley: Experiment 2

Pronoun Matching

Osterhout and Mobley (1995) investigate the effect on the amplitude of the N400 elicited by words in sentences where pronouns either do or do not match a preceding noun, as illustrated in the following example: *The aunt heard that* ***she*** (MATCH; M) / ***he*** (MISMATCH; MM) *had won the lottery.* The MISMATCH sentences can be interpreted as grammatical sentences where the pronoun refers to a different person than that denoted by the sentence subject; or ungrammatical sentences, where the pronoun refers back to the sentence subject with the wrong gender. Osterhout and Mobley (1995) ask whether there is a difference in N400 amplitude between the two conditions, and whether this is affected by which interpretation is taken by participants.

Target Words First, Osterhout and Mobley (1995) look at the N400 measured at the pronoun itself, finding no significant effect of condition.

For both RNN-LMs, however, experimental condition is a significant predictor of surprisal, with matched pronouns eliciting a significantly lower surprisal ($p < 0.001$ for all tests).

Sentence-Final Words The N400 was also measured at the last word in the sentence. Under this condition, it was found that there was a reduced N400 for matching compared to mismatching pronouns, but only for participants who interpreted mismatching sentences to be ungrammatical.

In both models, condition was not found to be a significant predictor of surprisal (GRNN:$p = 0.775$; JRNN: $p = 0.112$). However, whether this is a successful replication of the responses of the participants who found the sentence to be grammatical ('Gramm.' in

Figure 4.1) or a failure to replicate the results of those who found the sentence ungrammatical ('Ungramm.' in Figure 4.1) is unclear without further research, and thus this result is not discussed further in this paper.

Semantic Anomaly

In parallel to the pronoun stimuli, Osterhout and Mobley (1995) also compared N400 responses to sentences under the following experimental conditions: *The boat sailed down the river and* ***sank*** (CONTROL; C) / ***coughed*** (SEMANTICALLY ANOMALOUS; SA) *during the storm.*

Target Words N400 amplitude was significantly lower in response to the experimentally manipulated CONTROL words compared to SEMANTICALLY ANOMALOUS words. This effect was replicated in the surprisals of both models ($p < 0.001$ for all tests).

Sentence-Final Words The N400 and surprisals to sentence-final words followed the same pattern as target words ($p < 0.001$ for all tests).

4.4.7 Ainsworth-Darnell et al. (1998)

Ainsworth-Darnell et al. (1998) investigate the difference in N400 amplitude in response to syntactic and semantic anomaly, operationalized in the following way: *The chef entrusted the recipe* ***to relatives*** *before he left Italy* (CONTROL; C) / *The chef entrusted the recipe* ***to carrots*** *before he left Italy* (SEMANTIC ANOMALY; SEM) / *The chef entrusted the recipe* ***relatives*** *before he left Italy* (SYNTACTIC ANOMALY; SYN) / *The chef entrusted the recipe* ***carrots*** *before he left Italy* (DOUBLE ANOMALY; DA). While previous research argued that the N400 does not respond to SYNTACTIC ANOMALY, they found that the CONTROL nouns elicited lower N400 amplitudes than nouns in other conditions,

but they did not find a significant difference between the SYNTACTIC ANOMALY and SEMANTIC ANOMALY conditions or between the SEMANTIC ANOMALY and DOUBLE ANOMALY conditions. Ainsworth-Darnell et al. (1998) do not report a test comparing the SYNTACTIC ANOMALY and DOUBLE ANOMALY conditions, but it should be noted that SYNTACTIC ANOMALY has a lower amplitude (based on the graphs) than SEMANTIC ANOMALY, so an unreported significant difference between these should not be ruled out.

Experimental condition is a significant predictor of both GRNN and JRNN surprisal ($p < 0.001$). For both models, the surprisal is lower for words in the CONTROL condition compared to other conditions ($p < 0.001$ for all pairwise comparisons), and there is no significant difference between word in the SYNTACTIC ANOMALY and SEMANTIC ANOMALY conditions (GRNN: $p = 0.274$; JRNN: $p = 0.056$). The surprisals of the two models differ in that while DOUBLE ANOMALY words differ from SEMANTIC ANOMALY words in both models (GRNN: $p < 0.001$; JRNN: $p < 0.001$), they do not differ from the SYNTACTIC ANOMALY in GRNN surprisal but they do in JRNN surprisal (GRNN: $p = 0.059$; JRNN: $p < 0.001$). Based on these findings and inspection of the graphs in Ainsworth-Darnell et al. (1998), it appears that syntactic anomaly of this kind has a larger relative effect on surprisal than N400 amplitude.

4.4.8 Kim and Osterhout (2005): Experiment 1

Experiment 1 Kim and Osterhout (2005) investigate whether words that violate the event-structure of the described event are still facilitated if they are related to the event being described. The stimuli were of the following form: *The murder had been* ***witnessed*** *in the dark* (PASSIVE CONTROL; PC) / *The bystanders had been* ***witnessing*** *the crime* (ACTIVE CONTROL; AC) / *The murder had been* ***witnessing*** *by the three bystanders* (ATTRACTION VIOLATION; AV). General analysis found that condition only marginally pre-

dicted N400 amplitude, but pairwise comparison found one significant difference bwetween conditions: PC completions elicited lower-amplitude N400s than AC completions.

In both models, condition was a significant predictor of surprisal, and PCs elicited the lowest surprisals, followed ACs, followed by AVs ($p < 0.001$ for all tests).

4.4.9 Kim and Osterhout (2005): Experiment 2

Experiment 2 added the NO-ATTRACTION VIOLATION (NV) condition to the study, which is exemplified by the following sentence: *The unpleasant cough syrup was* ***witnessing*** *in the dark*. These were compared to results of the PC and AV conditions in Experiment 1. There was a significant main effect of condition, with PCs and AVs eliciting significantly lower-amplitude N400s than NVs.

Condition was a significant predictor the surprisals of both RNNs, with PCs eliciting a lower surprisal than AVs, followed by NVs with the highest surprisals ($p < 0.001$ for all tests).

4.5 General Discussion

We compared human N400 responses with surprisal in two RNN-LMs presented with the same stimuli, in the interest of determining the extent to which exposure to linguistic input alone can account for this particular component of human language processing. The results confirmed previous findings that surprisal is generally a good predictor of N400 amplitude, while also clearly demonstrating limitations of the models at capturing the human behavior.

4.5.1 Successful Predictions

The models effectively predicted certain kinds of contrast that the N400 is sensitive to.

Cloze The surprisals of both models for the Kutas (1993) and Ito et al. (2016) studies show that the surprisal of a language model is sensitive to cloze probability in the same direction as N400 amplitude—higher-cloze words elicit lower N400 amplitudes than lower-cloze words, and the same is true of surprisal.

Relatedness The results of the Kutas (1993) and Ito et al. (2016) experiments also show that surprisal matches N400 amplitude in that words that are related to the highest-cloze completion in terms of semantics, but not form, elicit a lower surprisal than semantically unrelated words, even controlling for these words' cloze.

Semantic typicality The surprisals of both models to the stimuli from Urbach and Kutas's (2010) three experiments demonstrate that the surprisal of a language model patterns in the same way as N400 amplitude in that more typical words (in a given context) elicit a lower surprisal than atypical words in the same context.

Semantic anomaly While the results are framed in the opposite direction in the original studies, the results from the Anomaly stimuli from Osterhout and Mobley (1995) and Experiment 1 of Ainsworth-Darnell et al. (1998) show that, all else being equal, completions that are not semantically anomalous (labeled 'controls' in these experiments) elicit a lower surprisal from language models than semantically anomalous completions, which is the result reported for N400 amplitude in the original studies.

Event structure violations The results for Experiment 2 of Kim and Osterhout (2005) show that both surprisal and N400 amplitude are reduced when a word is in line with event-structure norms, compared to a word that is not and is semantically unrelated to the preceding context.

4.5.2 Limitations and further directions

At the same time, there are areas where the predictive capabilities of the models are limited.

Quantifiers While the surprisal of the models matched the significant differences in Experiments 2 and 3 of Urbach and Kutas (2010) based on typicality overall, it did not replicate the finding that N400 amplitude was less reduced for TYPICAL nouns when they appeared with FEW or RARELY quantifiers. Thus, it may be the case that some more explicit (or at least more specific) representation of quantification is involved in the neurocognitive processes underlying the N400 than can be modeled by surprisal alone.

Event structure violations Overall, the surprisal of both models is more sensitive to morphosyntactic or event structure violations than N400 amplitude is (for a discussion on the extent to which these can be considered separate in the context of ERPs, see Kuperberg, 2016). For the stimuli from both Kim and Osterhout (2005) experiments, despite the ATTRACTION VIOLATION stimuli eliciting both a significantly reduced N400 amplitude and surprisal compared to the NO-ATTRACTION VIOLATION stimuli, surprisal remained significantly higher for ATTRACTION VIOLATION stimuli than either of the control stimuli, which is not the case with N400 amplitude. Thus, by contrast with the case of quantifiers discussed above (Urbach and Kutas, 2010), which seems to require a more detailed semantic representation, shallower or broader semantic representation might be

needed to capture responses to the kinds of stimuli presented in Kim and Osterhout (2005). If the goal is to improve the extent to which models capture human behavior, then there might be ways to accomplish this. Frank and Willems (2017), for example, use cosine distance between the sum of the vectors of all the preceding words in the sentence and the target word to predict the BOLD response (using fMRI) in N400 areas. Given the collateral facilitation of words semantically related to the highest-cloze completions of sentences, it is not unreasonable to assume that a similar process of spreading activation may occur for the preceding as well as the predicted upcoming word in the sentence. One way to implement this could be to weight the RNN model's predictions of the next word by each word's similarity to a general sentence-vector such as that used by Frank and Willems (2017) before the probabilities are transformed into surprisal[2].

Morphosyntactic Anomaly While there has been some discussion about the extent to which event structure violation and morphosyntactic anomalies can be considered separate in the context of ERPs (see, e.g. Kuperberg, 2016), there are clear cases where the surprisal of the language models appear to be more sensitive to morphosyntactic anomaly than N400 amplitude is. This can be seen in humans in the results of Experiment 1 of Ainsworth-Darnell et al. (1998), where words that exhibit either semantic or syntactic anomalies elicit equally reduced surprisal. By contrast, the models predict grammatical continuations to a sentence over ungrammatical ones. This leads to lower surprisals for semantically anomalous words that are syntactically acceptable than those that are both syntactically and semantically anomalous. This difference between humans and the models supports the idea that there needs to be some way to weight predictions by semantic relatedness to the preceding context.

[2]See Kuperberg's (2016) discussion on bag-of-word approaches to the N400.

4.6 Conclusions

Previous work has found that surprisal is a good predictor of N400 amplitude overall. Comparisons of surprisal in RNN-LMs to human N400 responses to the same input sentences showed for the first time that suprisal manages to account for a wide range of phenomena found in human N400 experiments. But at the same time, there are linguistic phenomena where it overpredicts, and others where it underpredicts a significant difference in the human N400 response. From the perspective of human language processing, this suggests that the activation of semantic and lexical features indexed by the N400 cannot be entirely captured by exposure to linguistic input alone. Specifically, quantification, aspects of event structure, and morphosyntactic anomalies seem to require some other learning architecture than the bottom-up statistical learning represented by standard recurrent neural networks. From the perspective of model-building, in order to improve a language-model based cognitive model of the N400, we need to allow for the addition of more shallow semantic processing (independent of syntax and event structure) such as an implementation of spreading activation.

4.7 Acknowledgements

Chapter 4, in full, is a reprint of the material as it appears in Michaelov, J. A. & Bergen, B. K., "How well does surprisal explain N400 amplitude under different experimental conditions?", *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL2020)*, 2020. The dissertation author was the primary investigator and author of this paper. Minor edits have been made to bring the formatting in line with the dissertation template.

Chapter 5

Collateral facilitation in humans and language models

Abstract

Are the predictions of humans and language models affected by similar things? Research suggests that while comprehending language, humans make predictions about upcoming words, with more predictable words being processed more easily. However, evidence also shows that humans display a similar processing advantage for highly anomalous words when these words are semantically related to the preceding context or to the most probable continuation. Using stimuli from 3 psycholinguistic experiments, we find that this is also almost always also the case for 8 contemporary transformer language models (BERT, ALBERT, RoBERTa, XLM-R, GPT-2, GPT-Neo, GPT-J, and XGLM). We then discuss the implications of this phenomenon for our understanding of both human language comprehension and the predictions made by language models.

5.1 Introduction

Humans process words more easily when they are more contextually predictable, whether predictability is determined by humans (Fischler and Bloom, 1979; Brothers and Kuperberg, 2021) or language models (McDonald and Shillcock, 2003a; Levy, 2008; Smith and Levy, 2013). Work on the N400, a neural signal of processing difficulty, has provided evidence that the neurocognitive system underlying human language comprehension preactivates words based on the extent to which they are predictable from the preceding context—thus, predictable words are easier to process because they or their features have already been activated before they are encountered (Kutas and Hillyard, 1984; Van Petten and Luka, 2012). This has led many to argue that we should consider the human language comprehension system to be engaging in prediction (DeLong et al., 2005; Kutas et al., 2011; Van Petten and Luka, 2012; Bornkessel-Schlesewsky and Schlesewsky, 2019; Kuperberg et al., 2020; DeLong and Kutas, 2020; Brothers and Kuperberg, 2021).

However, words that are either semantically related to the elements of the preceding context or to the most likely next word are also processed more easily, even if they are semantically implausible and ostensibly unpredictable. These are known as *related anomaly* effects. For an example of the former, consider the sentences in (1) that were used as experimental stimuli by Metusalem et al. (2012).

(1) My friend Mike went mountain biking recently. He lost control for a moment and ran right into a tree. It's a good thing he was wearing his _____.

(a) *helmet*

(b) *dirt*

(c) *table*

Helmet is the most predictable continuation of the sentence, as determined based

on cloze probability (Taylor, 1953, 1957)—the proportion of people to fill in a gap in a sentence with a specific word. Thus, unsurprisingly, *helmet* elicited the smallest N400 response, indicating that it is most easily processed. *Dirt* and *table* are both implausible continuations, and equally improbable based on human responses (both have a cloze probability of zero). Yet Metusalem et al. (2012) found that *dirt*, which is semantically related to the preceding context of *mountain biking*, elicits a smaller N400 response than *table*, which is not. This suggests that something about *dirt*'s relation to the *mountain biking* event causes it to be preactivated more than *table*, despite their seemingly equal implausibility and unpredictability.

The sentences in (2), used as experimental stimuli by Ito et al. (2016), provide an example of the other previously-discussed form of related anomaly—where a word semantically related to the most probable continuation (in this case, that with the highest cloze) is easier to process than one that is not. Even though *tail* and *tyre* are both implausible continuations with a cloze probability of zero, Ito et al. (2016) find that *tail*, which is semantically-related to the highest-cloze continuation *dog*, elicits a smaller N400 response than *tyre*, which is not.

(2) Meg will go to the park to walk her _____ tomorrow.

(a) *dog*

(b) *tail*

(c) *tyre*

In sum, words related to elements of the preceding context or to the most probable continuation of a sequence appear to be more preactivated in the brain than words that are not, even when both are highly anomalous. This effect has been replicated many times (Kutas and Hillyard, 1984; Kutas et al., 1984; Kutas, 1993; Federmeier and Kutas, 1999;

Metusalem et al., 2012; Rommers et al., 2013; Ito et al., 2016; DeLong et al., 2019; for review see DeLong et al., 2019).

The key question, therefore, is whether the same neurocognitive system underlying the predictability effects on the N400 also underlie related anomaly effects. Under one account (DeLong et al., 2019; DeLong and Kutas, 2020), the predictive system that underlies predictability effects also leads to these related anomalous words being 'collaterally facilitated' (DeLong and Kutas, 2020, p. 1045) due to their shared semantic features. Under this account, therefore, related anomaly effects can all be explained as by-products of our predictive system and the semantic organization of information in the brain. However, there is no direct evidence that this is the case—in fact, given the metabolic costs of preactivation (Brothers and Kuperberg, 2021), it may intuitively seem unlikely that an efficient predictive system would lead to implausible and otherwise anomalous words being preactivated. In fact, many researchers have argued that one or more associative mechanisms are required to explain related anomaly and other similar effects (Lau et al., 2013; Ito et al., 2016; Frank and Willems, 2017; Federmeier, 2021).

As systems designed specifically to predict the probability of a word given its context, language models offer a means to test the viability of the former hypothesis. If language models calculate that related but anomalous words are more predictable than unrelated anomalous words, this would demonstrate that related anomaly effects can be produced by a system engaged in prediction alone. This would show that it is possible that related anomalies can be 'collaterally facilitated' (DeLong and Kutas, 2020, p. 1045) by a predictive mechanism in human language comprehension. Thus, it would remove the need to posit additional associative mechanisms on the basis of related anomaly effects, which could greatly simplify our understanding of human language comprehension.

This is what we test in the present study. We run the stimuli from 3 psycholin-

guistic experiments carried out in English (Ito et al., 2016; DeLong et al., 2019; Metusalem et al., 2012) through 8 contemporary transformer language models (Devlin et al., 2019; Radford et al., 2019; Liu et al., 2019; Lan et al., 2020; Conneau et al., 2020; Black et al., 2021; Wang and Komatsuzaki, 2021; Lin et al., 2021), calculating the surprisal (negative log-probability) of each word for which the N400 was measured. We then compare whether, in line with the N400 response, anomalous words that are semantically related to the context have significantly lower surprisals than unrelated words.

5.2 Related work

There have been a wide range of attempts to computationally model the N400 (Parviz et al., 2011; Laszlo and Plaut, 2012; Laszlo and Armstrong, 2014; Rabovsky and McRae, 2014; Frank et al., 2015; Ettinger et al., 2016; Cheyette and Plaut, 2017; Brouwer et al., 2017; Rabovsky et al., 2018; Venhuizen et al., 2019; Fitz and Chang, 2019; Aurnhammer and Frank, 2019b; Michaelov and Bergen, 2020; Merkx and Frank, 2021; Uchida et al., 2021; Szewczyk and Federmeier, 2022; Michaelov et al., 2022). One of the most successful and influential approaches has been to model the N400 using the surprisal calculated from neural language models—surprisal has been found to be a significant predictor of single-trial N400 data (Frank et al., 2015; Aurnhammer and Frank, 2019b; Merkx and Frank, 2021; Michaelov et al., 2021; Szewczyk and Federmeier, 2022; Michaelov et al., 2022), and has been found to be similar to the N400 response in how it is affected by a range of experimental manipulations (Michaelov and Bergen, 2020; Michaelov et al., 2021). A key finding is that better-performing and more sophisticated language models perform better at predicting the N400 (Frank et al., 2015; Aurnhammer and Frank, 2019b; Michaelov and Bergen, 2020; Merkx and Frank, 2021; Michaelov et al., 2021, 2022). For this reason, we use contemporary transformer language models in the present study.

We use experimental stimuli from 3 experiments. Stimuli from one of these experiments (Ito et al., 2016) have been previously used in computational analyses of the N400. This is one of several sets that Michaelov and Bergen (2020) attempt to model using recurrent neural network (RNN) language models, finding that they can indeed calculate that words related to the highest-cloze continuation are more predictable than unrelated words. In the present study, we test whether this result can be replicated on a larger number of language models, and specifically, transformer language models.

There has also been work looking at how language models deal with semantic relatedness to the highest-cloze continuation based on stimuli from other N400 experiments. Michaelov and Bergen (2020), for example, find that in cases where the related and unrelated words are both plausible, the related continuations are more strongly predicted by RNNs (Gulordava et al., 2018; Jozefowicz et al., 2016), in line with the original N400 results (Kutas, 1993). Michaelov et al. (2021) conceptually replicate this finding on a different dataset (Bardolph et al., 2018) using one of the same RNNs (Jozefowicz et al., 2016) and GPT-2 (Radford et al., 2019). However, these prior efforts differ from the present study in that they investigate N400s and surprisal to words that are all plausible continuations of the sentence, and where they both have a low but generally non-zero cloze probability. In the stimuli analyzed in the present study, by contrast, both the related and unrelated words are anomalous—they have a cloze probability of zero, and are implausible continuations. Thus, their preactivation does, at least intuitively, appear to be more clearly 'collateral'.

We are only aware of one previous study that directly compares the predictions of transformers and the human N400 response on related anomaly stimuli. Ettinger (2020) evaluates BERT in terms of its similarity to cloze—because the predictions of a language model, being incremental, may show similar effects to those found in the N400 (see also Michaelov and Bergen, 2020 for discussion). For this reason, Ettinger (2020) tests how good

BERT is at predicting the highest-cloze (most probable) continuations in the stimuli over anomalous but semantically related continuations, but does not directly look at the related anomaly effect—whether the related anomalous continuations are more strongly predicted than the unrelated anomalous continuations. Thus, to the best of our knowledge, the present study is the first to investigate whether the predictions of transformer language models display related anomaly effects like humans do.

Finally, there has been some work investigating whether language models display priming effects (e.g. Prasad et al., 2019; Misra et al., 2020; Kassner and Schütze, 2020; Lin et al., 2021; Lindborg and Rabovsky, 2021). The effect found by Metusalem et al. (2012)—that words related to the events described in the context are preactivated more strongly than words that are not—is a form of semantic priming, as it results in the increased preactivation of a word based on the semantic content stimulus that has been recently encountered (i.e. the event described in the preceding linguistic context). Thus, our investigation of the patterns in the prediction of the the stimuli from Metusalem et al. (2012) is intended to further our knowledge of priming in language models—specifically, whether there are systematic ways in which context shapes the extent to which anomalous words are predicted.

5.3 General Method

In this study, we took the stimuli from a range of experiments (Ito et al., 2016; DeLong et al., 2019; Metusalem et al., 2012) and ran them through a number of transformer language models. We used the *transformers* (Wolf et al., 2020) implementations of the (largest and most up-to-date versions of each of the) following models: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), XLM-R Conneau et al. (2020), GPT-2 (Radford et al., 2019), GPT-Neo (Black et al., 2021), GPT-J (Wang

and Komatsuzaki, 2021), and XGLM (Lin et al., 2021). We chose these models to cover a number of both autoregressive (GPT-2, GPT-Neo, GPT-J, XGLM) and masked (BERT, RoBERTa, ALBERT, XLM-RoBERTa) language model architectures. Given the recent increase in popularity of multilingual language models, we also made sure to include one autoregressive (XGLM) and one masked (XLM-RoBERTa) multilingual language model, in case there is a difference based on the number of languages that a model is trained on.

All experimental stimuli used in the present study have been made available by the original authors of their respective papers as appendices or supplementary materials. In our analysis, we truncated all stimuli to be the preceding context of the critical word (the word for which the N400 was measured). We then used the language models to calculate the probability of the next word, and negative log-transformed (using a logarithm of base 2, following Futrell et al., 2019) these probabilities to calculate the surprisal of each word. For words not present in the vocabulary of each model, we tokenized the word, and then progressively calculated the surprisal of each sub-word token given the preceding context; with the sum of all the surprisals (equivalent to the the negative log-probability of the product of all the probabilities) being used as the total surprisal for the word. In this way, we calculated the surprisal of each critical word given its preceding context only.

All graphs and statistical analyses were created and run in *R* (R Core Team, 2020) using *Rstudio* (RStudio Team, 2020) and the *tidyverse* (Wickham et al., 2019), *lme4* (Bates et al., 2015), and *lmerTest* (Kuznetsova et al., 2017) packages. All reported p-values are corrected for multiple comparisons based on false discovery rate across all statistical tests carried out (Benjamini and Hochberg, 1995). Because of this correction procedure, if any models display related anomaly effects, this is evidence that prediction alone can account for them.

All of the code for running the experiments and carrying out the statistical anal-

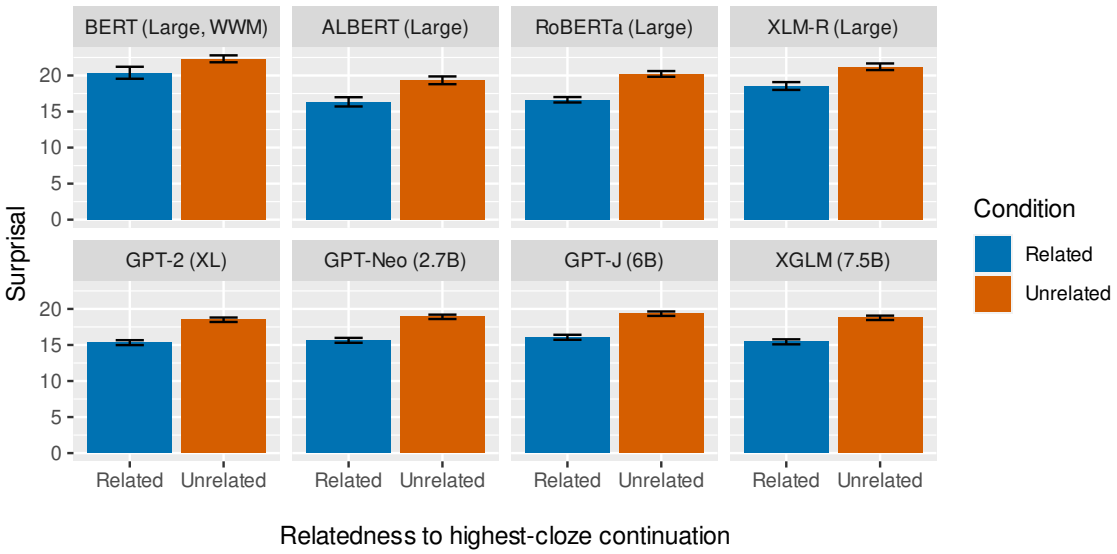

Figure 5.1: Mean surprisal elicited by each language model for the Ito et al. (2016) stimuli related and unrelated to the most probable (highest-cloze) continuation of each sentence. Error bars indicate standard error.

yses is provided at `https://github.com/jmichaelov/collateral-facilitation`.

5.4 Experiment 1: Ito et al. (2016)

5.4.1 Introduction

We begin with Ito et al. (2016), who investigated whether relatedness to the highest-cloze continuation of a given sentence impacts the amplitude of the N400 response. They presented human participants with experimental stimuli that included a word that was either the highest-cloze continuation of a sentence, semantically related to that highest-cloze continuation, similar to the highest-cloze continuation in terms of their form (e.g. *hook* and *book*), or unrelated. For the purposes of the present study, we are interested in

semantic relatedness and thus do not consider the formal relatedness condition. Thus, we look at the stimuli from the three experimental conditions exemplified in (3)—an example of Predictable, Related, and Unrelated continuations for one sentence frame.

(3) Lydia cannot eat anymore as she is so _____ now.

- *full* (Predictable)
- *half* (Related)
- *mild* (Unrelated)

Ito et al. (2016) find that related continuations elicit a smaller N400 response than unrelated continuations. As stated, this finding was successfully modeled using the surprisal of two RNN language models by Michaelov and Bergen (2020).

In the present study, we aim to investigate whether this can be replicated with contemporary transformer language models. Thus far, only one study (Merkx and Frank, 2021) has directly compared the N400 prediction capabilities of RNNs and transformers while matching number of parameters, training data, and language modeling performance, finding that transformers are better predictors of N400 amplitude overall. We might therefore expect that the transformers used in the present study should model the related anomaly effect found by Ito et al. (2016) at least as well as the RNNs used by Michaelov and Bergen (2020). However, a key feature of Merkx and Frank's (2021) study is that it uses naturalistic stimuli. This makes the experiment more ecologically valid, but as has been pointed out (Michaelov and Bergen, 2020; Brothers and Kuperberg, 2021), this means that we cannot tell whether the higher correlation between surprisal and N400 amplitude is due to any factors that we are interested in investigating—Merkx and Frank (2021) do not consider how relatedness to a previously-mentioned event or to most predictable continuation impacts surprisal and the N400. For this reason, it is in fact far from clear

Table 5.1: The results of a Type III ANOVA (using Satterthwaite's method for estimating degrees of freedom; Kuznetsova et al., 2017) on the Ito et al. (2016) stimuli, testing for which language models experimental condition (related or unrelated) is a significant predictor of their surprisal. This is the case for all language models.

Model	**Test Statistic**	**Corrected p**
BERT	$F(1, 120) = 7.15$	0.0093
ALBERT	$F(1, 92) = 20.6$	< 0.0001
RoBERTa	$F(1, 159) = 60.8$	< 0.0001
XLM-R	$F(1, 126) = 21.2$	< 0.0001
GPT-2	$F(1, 157) = 64.0$	< 0.0001
GPT-Neo	$F(1, 152) = 64.1$	< 0.0001
GPT-J	$F(1, 149) = 62.5$	< 0.0001
XGLM	$F(1, 146) = 72.6$	< 0.0001

that we should expect this specific related anomaly effect to be modeled as well by transformers as by RNNs. However, if it is, this would demonstrate the effect in two different language model architectures, further strengthening the idea that a predictive system alone can explain related anomaly effects.

Thus, in the present study, we investigate whether the results of Michaelov and Bergen (2020) replicate beyond the two RNNs tested, and crucially, whether the results replicate with transformer language models. Specifically, we test whether the surprisal elicited by implausible stimuli related to the highest-cloze continuation is lower than the surprisal elicited by implausible stimuli unrelated to the highest-cloze continuation.

5.4.2 Results

The results of the experiment are shown in Figure 5.1. As can be seen, numerically, related words elicit lower surprisals than unrelated words, indicating that they were more highly predicted by the language models. This in turn suggests that these models do in

fact collaterally predict the related continuations.

In order to test this more directly, we ran statistical analyses of the surprisals elicited by the language models. This was done by constructing linear mixed-effects regressions for each language model surprisal with experimental condition as a main effect, and the maximal random effects structure that would successfully converge for all models (see Barr et al., 2013). For all regressions except for that predicting RoBERTa surprisal, this random effects structure was a random intercept of sentence frame and of critical word. For the RoBERTa surprisal regression, the latter random intercept was removed due to it causing a singular fit. As creating null models with only the random effects structure resulted in singular fits for multiple regressions, we were unable to run likelihood ratio tests to test whether experimental condition—that is, whether the word was semantically related or unrelated to the highest-cloze continuation—was a significant predictor of surprisal. For this reason, we instead tested whether experimental condition was a significant predictor of surprisal by running a Type III ANOVA using Satterthwaite's method for estimating degrees of freedom (Kuznetsova et al., 2017) on the aforementioned linear mixed-effects models that included experimental condition as a fixed effect.

The results of the tests are shown in Table 5.1. As can be seen, condition is a significant predictor of the surprisal from every language model, confirming that language models predict related stimuli to be more likely than unrelated stimuli.

The results of this experiment demonstrate that all the language models tested—BERT, ALBERT, RoBERTa, XLM-R, GPT-2, GPT-Neo, GPT-J, and XGLM—display the related anomaly effect in response to the Ito et al. (2016) stimuli. All eight models predict implausible continuations that are related to the most probable continuations to be more likely those that are unrelated.

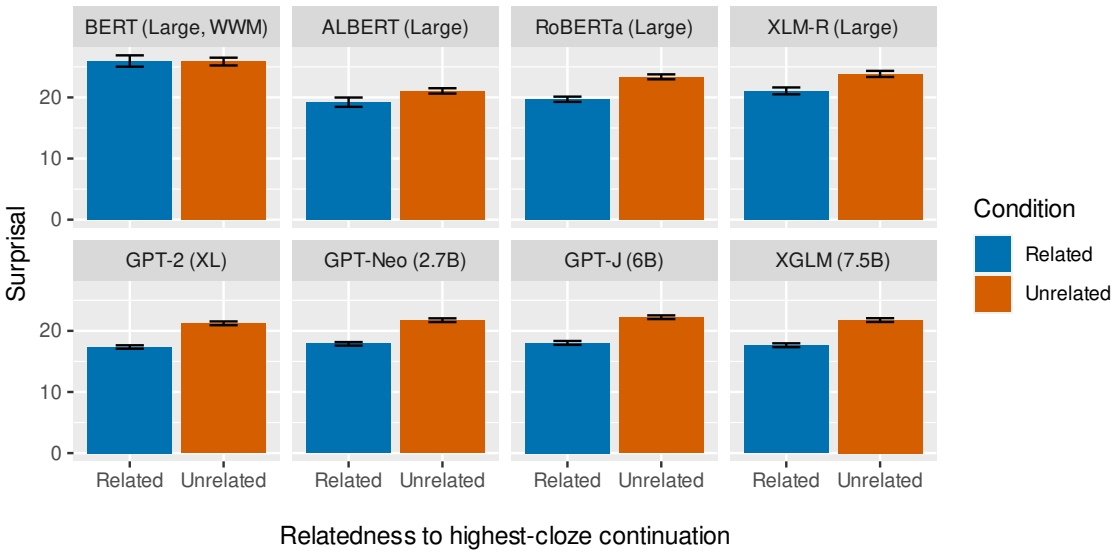

Figure 5.2: Mean surprisal elicited by each language model for the DeLong et al. (2019) stimuli related and unrelated to the most probable (highest-cloze) continuation of each sentence. Error bars indicate standard error.

5.5 Experiment 2: DeLong et al. (2019)

5.5.1 Introduction

DeLong et al. (2019) also investigated the difference between the N400 amplitude elicited by implausible words that are related or unrelated to the most predictable (highest-cloze) continuation. As in Ito et al. (2016), these stimuli were chosen such that both related and unrelated words were highly implausbile—in this case, 'unpredictable words were strategically chosen not to make sense in their given contexts' (DeLong et al., 2019, p. 4). These stimuli are exemplified by the set shown in (4).

(4) The commuter drove to work in her _____ after breakfast.

Table 5.2: The results of a Type III ANOVA (using Satterthwaite's method for estimating degrees of freedom; Kuznetsova et al., 2017) on the DeLong et al. (2019) stimuli, testing for which language models experimental condition (related or unrelated) is a significant predictor of their surprisal. This is the case for all language models except BERT.

Model	**Test Statistic**	**Corrected p**
BERT	$F(1, 159) =< 0.1$	0.9322
ALBERT	$F(1, 112) = 6.3$	0.0138
RoBERTa	$F(1, 159) = 50.7$	< 0.0001
XLM-R	$F(1, 132) = 18.2$	0.0001
GPT-2 XL	$F(1, 134) = 120.7$	< 0.0001
GPT-Neo	$F(1, 142) = 111.7$	< 0.0001
GPT-J	$F(1, 141) = 132.6$	< 0.0001
XGLM	$F(1, 159) = 122.4$	< 0.0001

- *car* (Predictable)
- *brakes* (Related)
- *poetry* (Unrelated)

Like Ito et al. (2016), DeLong et al. (2019) find that overall, related continuations elicit a smaller N400 response than unrelated continuations.

5.5.2 Results

As in Experiment 1, we ran the stimuli from the original experiment through the 8 language models and calculated the surprisal of each critical word. The results of the experiment are shown in Figure 5.2. In all models except BERT, related stimuli all elicit numerically lower surprisals than unrelated stimuli, indicating that they were more highly-predicted by the language models.

We again ran the same statistical test as in Experiment 1, testing whether exper-

imental condition (related or unrelated to the highest-cloze continuation) is a significant predictor of the surprisal elicited by the stimuli in each language model. The ALBERT, XLM-R, GPT-2, GPT-Neo, and GPT-J regressions had random intercepts of sentence frame and critical word, while the BERT, RoBERTa, and XGLM regressions had only random intercepts for sentence frame. The results of the Type III ANOVA are shown in Table 5.2. Condition is a significant predictor of the surprisal of every model except BERT—in these models, related stimuli are predicted to be more likely continuations of the sentence than unrelated stimuli. Thus, with the exception of BERT, we replicate the findings of Experiment 1.

5.6 Experiment 3: Metusalem et al. (2012)

5.6.1 Introduction

Metusalem et al. (2012) investigated the extent to which relatedness to the event described in the preceding context impacts the amplitude of the N400 response. Metusalem et al. (2012) presented human participants with experimental stimuli that included either the most probable (highest-cloze) continuation of a sentence, an implausible continuation that was related to the event described, or an implausible continuation that was unrelated to the event described. All of the implausible stimuli also had a cloze probability of zero. The stimuli are exemplified by the set for a single sentence frame shown in (5).

(5) We're lucky to live in a town with such a great art museum. Last week I went to see a special exhibit. I finally got in after waiting in a long _____.

- *line* (Predictable)
- *painting* (Related)
- *toothbrush* (Unrelated)

Metusalem et al. (2012) found that despite their implausibility and improbability (based on cloze), critical words related to the event described in the context preceding them elicited smaller N400 responses than words that were unrelated to the event, a clear example of a related anomaly effect.

5.6.2 Results

As in Experiments 1 and 2, we ran the stimuli from the original experiment through the 8 language models and calculated the surprisal of each critical word. The results of the experiment are shown in Figure 5.3. As in Experiment 1, numerically, in all models related stimuli elicit lower surprisals than unrelated surprisals, indicating that they were more highly predicted by the language models.

We again ran the same statistical analyses as in Experiments 1 and 2, constructing linear mixed-effects regression models, all of which had random intercepts of sentence frame and critical word. Using a Type III ANOVA, we tested whether experimental condition (related or unrelated to the event described in the preceding context) is a significant predictor of N400 amplitude. The results are shown in Table 5.3. As can be seen, experimental condition was a significant predictor of the surprisal of all models.

5.7 General Discussion

5.7.1 Summary of Results

In all but one specific case—BERT in Experiment 2—experimental condition significantly predicted language model surprisal in the same direction as human N400 responses. The results of Experiments 1 and 2, therefore demonstrate convincingly that, like humans, language models do tend to predict that anomalous words related to the

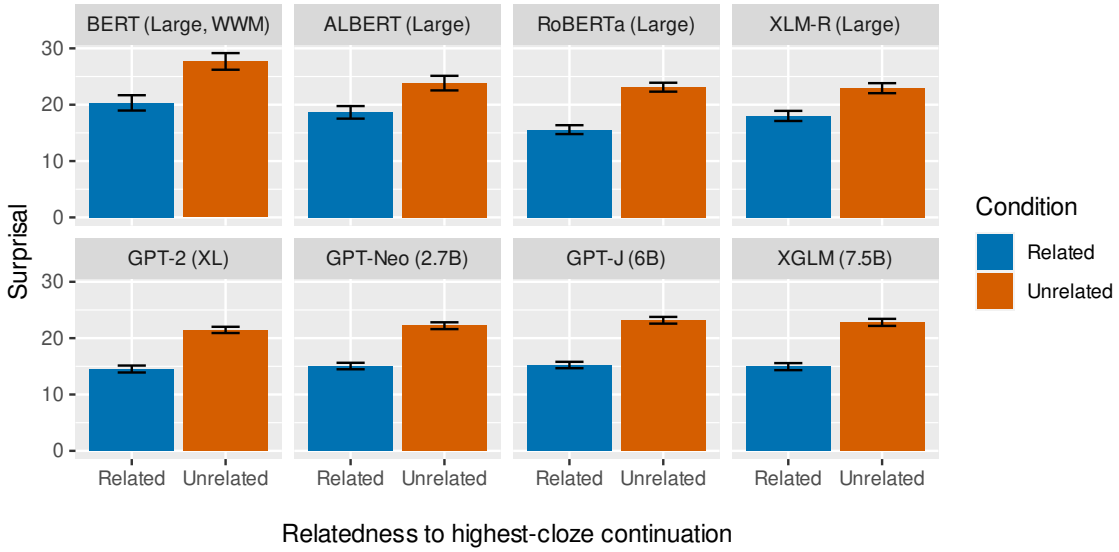

Figure 5.3: Mean surprisal elicited by each language model for the Metusalem et al. (2012) stimuli related and unrelated to the most probable (highest-cloze) continuation of each sentence. Error bars indicate standard error.

Table 5.3: The results of a Type III ANOVA (using Satterthwaite's method for estimating degrees of freedom; Kuznetsova et al., 2017) on the Metusalem et al. (2012) stimuli, testing for which language models experimental condition (related or unrelated) is a significant predictor of their surprisal. This is the case for all language models.

Model	**Test Statistic**	**Corrected p**
BERT	$F(1, 29) = 77.1$	< 0.0001
ALBERT	$F(1, 29) = 78.7$	< 0.0001
RoBERTa	$F(1, 28) = 188.1$	< 0.0001
XLM-R	$F(1, 34) = 83.4$	< 0.0001
GPT-2 XL	$F(1, 35) = 211.5$	< 0.0001
GPT-Neo	$F(1, 42) = 200.1$	< 0.0001
GPT-J	$F(1, 35) = 265.5$	< 0.0001
XGLM	$F(1, 33) = 222.5$	< 0.0001

most probable continuation are more probable than anomalous words that are not. The results of Experiments 3, analogously, demonstrate that like humans, language models tend to predict that anomalous words related to a relevant event described in the preceding context are more probable than anomalous words that are not. Thus, like the human language comprehension system, language models exhibit related anomaly effects.

5.7.2 Psycholinguistic implications

These results have clear implications for psycholinguistic research on the effects of related anomalies on human language processing. First, a predictive system can display the effects—in fact, there is only one set of stimuli for which not all models do. This demonstrates the sufficiency of a predictive system for preactivating related anomalous stimuli to a greater degree than unrelated anomalous stimuli. In other words, based on a parsimony criterion, there is no need to posit that related anomaly effects on human language processing require something beyond a predictive system such as an associative system, either instead of or in addition to a predictive one.

Second, both kinds of related anomaly effect explored—the reduction in N400 amplitude correlated with relatedness to the most probable continuation and that correlated with relatedness to the event in the preceding context—are explainable by a single mechanism. This may seem counterintuitive, given how intuitively different the effects may seem. Yet this finding is consistent with the idea in the literature that the two effects can be considered different variants of the same phenomenon (DeLong et al., 2019; DeLong and Kutas, 2020).

Given that this study is based on computational modeling, we should note that the results do not constitute direct proof of a neurocognitive predictive system or of the lack of the involvement of an additional associative mechanism. However, they are consistent with

such accounts, and open the door for future research, both computational and experimental. For example, it may be the case that other phenomena that have been argued to constitute evidence for a separate associative mechanism (see Federmeier, 2021, for review) may also be explainable on the basis of prediction. On the other hand, the approach we use here can also be used to design stimuli that do not differ in probability in order to further test whether prediction can explain all related anomaly effects.

5.7.3 Implications for NLP

The results of the present study demonstrate that related anomaly effects occur in contemporary transformer language models. Based on the present study, this does not appear to be impacted by whether the model is an autoregressive or masked language model; or by whether the model is monolingual or multilingual. In fact, the only model that does not show the effect every time is BERT, the least powerful model tested (all other models are either larger, trained on more data, or both). Thus, in line with previous research showing that higher-quality language models better predict human processing metrics (Merkx and Frank, 2021), the present results suggest that better language models are also more likely to display human-like patterns of prediction.

The results of this study also have several implications for understanding how the predictions of humans and language models relate. As has been previously discussed, some researchers have argued that we should evaluate the predictions of language models based on cloze probability (Ettinger, 2020). In fact, some have suggested training models on cloze probabilities (Eisape et al., 2020). However, the results of this study, along with others (Frank et al., 2015; Aurnhammer and Frank, 2019b; Michaelov and Bergen, 2020; Aurnhammer and Frank, 2019b; Merkx and Frank, 2021; Szewczyk and Federmeier, 2022; Michaelov et al., 2022), suggest that the predictions of language models are highly

correlated with N400 amplitude; and recent work has argued that that the activation states of transformers are highly correlated with activation in the brain during language comprehension more generally (Schrimpf et al., 2020). Thus, while it may be useful for certain tasks to have cloze-like predictions, it may be the case that we are generally more likely to get N400-like predictions from language models.

If so, this is a cause for both optimism and pessimism. Given that humans are the gold-standard in natural language tasks generally, if a language model can make predictions that closely match those that humans make as part of language comprehension, this may also suggest that the representations learned are at least in some ways functionally similar to those that humans use to generate the same predictions. On the other hand, by the same token, it may suggest a limit to the possibilities of language modeling alone—there is much more to language comprehension than the kinds of prediction that underlie the N400 response (see, e.g., Ferreira and Yang, 2019; DeLong and Kutas, 2020; Kuperberg et al., 2020).

5.8 Conclusion

In order to better understand related anomaly effects in humans, we investigated whether contemporary transformer language models display them. We found that in all but one case, they do, suggesting that related anomaly effects in both humans and language models may be driven by prediction alone.

5.9 Appendices

5.9.1 Limitations

As mentioned the discussion section, one limitation of the present study is that while it demonstrates that it is possible for related anomaly effects to emerge from a system engaged in prediction alone, it does not directly demonstrate that this is what is occurring in humans.

A further limitation is that we model the results of three related anomaly experiments out of the larger total number that have been carried out (for review, see DeLong et al., 2019). However, given how consistent related anomaly effects appear to be (DeLong et al., 2019), and how consistent our results are (after statistical correction for multiple comparisons, all three related anomaly effects are modeled by all but one transformer, which only fails to model one effect), we do not believe this presents a problem for our analysis.

Finally, the three experiments modeled were all carried out in English. Related anomaly effects have been reported in other languages (DeLong et al., 2019) such as Dutch (Rommers et al., 2013); and these are not modeled in our study. Thus, it is an open question whether our results generalize to related anomaly effects in languages other than English. However, we also note the evidence that higher-quality models are better at predicting N400 amplitude (Merkx and Frank, 2021). For this reason, given the overwhelming focus on English in computational linguistics (Bender, 2009, 2011; Tsarfaty et al., 2013; Munro, 2015; Mielke, 2016; Kim et al., 2016; Amram et al., 2018; Bender, 2019; Clark et al., 2022), current language model architectures are likely to be best suited to predicting English—indeed, current state-of-the-art models such as GPT-3 (Brown et al., 2020), OPT (Zhang et al., 2022), PaLM (Chowdhery et al., 2022), and LaMDA (Thoppilan et al., 2022) are

trained mostly or only on English data. Thus, while the focus on modeling English may be an issue for the field as a whole, in this case, focusing on experiments carried out in English may in fact give us the best possible chance to evaluate what the human predictive system *could* predict.

5.9.2 Models used

The details of the models used in this study are provided in Table 5.4.

Table 5.4: Transformer language models used in the present study. All were accessed using the transformers (Wolf et al., 2020) package. *Full name* refers to the model's full name on the Hugging Face Model Hub (Wolf et al., 2020).

Model Name	**Full Name**	**Reference**
BERT	`bert-large-cased-whole-word-masking`	Devlin et al. (2019)
ALBERT	`albert-xxlarge-v2`	Lan et al. (2020)
RoBERTa	`roberta-large`	Liu et al. (2019)
XLM-R	`xlm-roberta-large`	Conneau et al. (2020)
GPT-2 XL	`gpt2-xl`	Radford et al. (2019)
GPT-Neo	`EleutherAI/gpt-neo-2.7B`	Black et al. (2021)
GPT-J	`EleutherAI/gpt-j-6B`	Wang and Komatsuzaki (2021)
XGLM	`facebook/xglm-7.5B`	Lin et al. (2021)

5.10 Acknowledgements

We would like to thank the authors of the original N400 experiment papers—Wen-Hsuan Chan, Martin Corley, Katherine A. DeLong, Jeffrey L. Elman, Mary Hare, Aine Ito, Marta Kutas, Andrea E. Martin, Ken McRae, Ross Metusalem, Mante S. Nieuwland, Martin J. Pickering, and Thomas P. Urbach—for making their stimuli available. We would also like to thank the anonymous reviewers for their helpful comments, the other members of the Language and Cognition Lab at UCSD for their valuable discussion, and the San

Diego Social Sciences Computing Facility Team for their technical assistance. This work was partially supported by a 2021-2022 Center for Academic Research and Training in Anthropogeny Annette Merle-Smith Fellowship awarded to James A. Michaelov, and the RTX A5000 used for this research was donated by the NVIDIA Corporation.

Chapter 5, in full, is a reprint of the material as it appears in Michaelov, J. A. & Bergen, B. K. "Collateral facilitation in humans and language models", *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL 2022)*, 2022. The dissertation author was the primary investigator and author of this paper. Minor edits have been made to bring the formatting in line with the dissertation template.

Chapter 6

Rarely a problem? Language models exhibit inverse scaling in their predictions following *few*-type quantifiers

Abstract

How well do language models deal with quantification? In this study, we focus on *few*-type quantifiers, as in *few children like toys*, which might pose a particular challenge for language models because the sentence components without the quantifier are likely to co-occur, and *few*-type quantifiers are rare. We present 960 English sentence stimuli from two human neurolinguistic experiments to 22 autoregressive transformer models of differing sizes. Not only do all the models perform poorly on *few*-type quantifiers, but overall the larger the model, the worse its performance. This inverse scaling is consistent

with previous work suggesting that larger models increasingly reflect online rather than offline human processing, and we argue that the decreasing performance of larger models may challenge uses of language models as the basis for natural language systems.

6.1 Introduction

Quantifiers can dramatically alter the meaning of an utterance. Consider the sentences in (1).

(1) (a) Most sharks are harmless.

(b) Most sharks are dangerous.

(c) Few sharks are harmless.

(d) Few sharks are dangerous.

Despite the fact that (a) and (c) have the same content words in the same syntactic arrangement, the statements have starkly different meanings. The same is true of (b) and (d). Being able to successfully comprehend these differences is useful, and in an example such as this one, vitally important[1].

Yet current work suggests that language models deal poorly with quantifiers—they struggle to predict which quantifier is used in a given context (Pezzelle et al., 2018; Talmor et al., 2020), and also perform poorly at generating appropriate continuations following logical quantifiers (Kalouli et al., 2022). This is especially concerning given the recent trend of using large language models (sometimes referred to as 'foundation models'; Bommasani et al., 2021) as general systems that can perform multiple tasks, including question answering, without specific training (Brown et al., 2020; Raffel et al.,

[1]Note that most sharks are in fact harmless to humans; see, e.g., `https://www.floridamuseum.ufl.edu/discover-fish/sharks/shark-attack-faq/`.

2020; Lin et al., 2021; Srivastava et al., 2022; Hoffmann et al., 2022; Rae et al., 2022; Zhang et al., 2022; Chowdhery et al., 2022). It is thus crucial that such systems be able to distinguish among sentences like those in (1) in human-like ways both during training and when generating responses.

The aim of the present study is to evaluate how well language models take into account the meaning of a quantifier when generating the text that follows it, and to investigate whether this scales with model size. We are particularly interested in the question of whether language models exhibit *inverse scaling*—that is, whether as model size increases, performance decreases rather than increases (Perez et al., 2022; McKenzie et al., 2022a). Inverse scaling is an issue of serious concern for developing and training new language models, since inverse scaling could indicate 'outer misalignment' (Perez et al., 2022)—that the training approach is leading to models that produce undesirable outputs, which may get worse as performance at training objectives increases. Inverse scaling is also a concern for models' ultimate use. As models increase in size and perform better at a wider range of benchmarks (for recent examples, see, e.g., Srivastava et al., 2022; Chowdhery et al., 2022), they may be increasingly assumed to be trustworthy and general-purpose, and thus able to perform well tasks on which they have not been tested (Raji et al., 2021). This could lead to a range of possible harms, from misidentifying whether something is dangerous or not (as in the opening example), to amplifying biases (Bender et al., 2021).

To test how well language models deal with quantifiers, we follow the approach of Ettinger (2020) in using sentences from a study on human language comprehension to inform our evaluation. Ettinger (2020) found that following a negation, the predictions of $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ in simple sentences expressing a proposition with or without negation (from Fischler et al., 1984) do not appear sensitive to negation—for example, $\text{BERT}_{\text{LARGE}}$ predicts the final word of *a robin is a* ***bird*** to be more likely than *a robin*

is a ***tree***, but also predicts that *a robin is not a* ***bird*** is more likely than *a robin is not a* ***tree***. In this way, the models' predictions more closely match those made by humans 'online'—that is, incrementally during the process of language comprehension—than our fully-formed 'offline' judgements: in their original study, Fischler et al. (1984) found that the word *bird* elicited an N400 response of smaller amplitude than *tree* in both contexts, indicating that it was more strongly predicted.

Similar effects have been reported (Kassner and Schütze, 2020; Kalouli et al., 2022) for other transformers such as Transformer-XL (Dai et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020), as well as ELMo (Peters et al., 2018). Worse, recent work suggests that as language models increase in size, their ability to deal with negation may degrade: an inverse scaling relationship has been reported for performance at a wide range of tasks when prompts include negation (McKenzie et al., 2022b; Jang et al., 2023), though it is possible that this may reverse at extremely large scales (Wei et al., 2022b).

Negation may be particularly challenging for statistical language models because its presence radically alters the meaning of a sentence, but negation occurs in only about 10% of sentences (Jiménez-Zafra et al., 2020). Quantifiers similarly impose radical modulations to meaning while also being relatively infrequent (see subsection 6.5.4). In the present study, we focus on quantifiers indicating typicality such as *most* and *few*. To the best of our knowledge, only one study has evaluated model predictions following any quantifiers (Kalouli et al., 2022), and it focused on words corresponding to logical quantifiers such as *all*, *every*, and *some*. The few studies involving the quantifiers we address either focus on predicting the quantifier itself (Pezzelle et al., 2018; Talmor et al., 2020), or use RNNs to investigate modeling significant effects on the N400 without any form of evaluation (Michaelov and Bergen, 2020). This study, therefore, represents the first attempt to

explicitly evaluate the predictions of language models following *most* and *few*-type quantifiers.

In the present study, we carry out two experiments. In the first, following Ettinger (2020), we use the stimuli from a previously published N400 study (Urbach and Kutas, 2010). In it, Urbach and Kutas (2010) found that while *most* and *few*-type quantifiers do impact N400 amplitude, it is not enough to reverse predictions—*few farmers grow* ***crops*** elicits a smaller N400 response than *few farmers grow* ***worms***, indicating that *crops* was more strongly predicted than *worms*, even though experimental participants judged it to be less plausible off-line. We test whether language models show the same pattern of insensitivity towards the quantifiers that humans do in online measures. In this way, we test how closely the predictions of language models correlate with those underlying the human N400 response.

In our second experiment, we extend our study further. Experiment 1 aims to replicate the original N400 results of Urbach and Kutas (2010); however, one thing that it does not account for is that while a given complete sentence (e.g., *few farmers grow* ***crops.***) can be highly unlikely and implausible, sentences beginning with the same words may not be (for example, in the plausible sentence *few farmers grow* ***crops*** *in the winter*). Experiment 1 does not distinguish between these possibilities, and while it is important to test the sensitivity of language models to *few*-type quantifiers, if they fail to show a difference for complete sentences including the final period (e.g., *few farmers grow* ***crops.***), this is more concerning. Thus, in Experiment 2, we run the same stimuli as Experiment 1, but including a period following the final word (e.g., *crops./worms.*).

6.2 Experiment 1: Replication of Urbach and Kutas (2010)

6.2.1 Materials

In this experiment, we use all the stimuli from two experiments carried out by Urbach and Kutas (2010). These are made up of 120 sentence frames with 8 different sentence types falling into 4 experimental conditions, for a total of 960 sentences. The 4 conditions had a 2x2 design—each stimulus was either typical (T) or atypical (A), and had either a *most*-type or *few*-type quantifier. An example of the 8 sentence types comprising one sentence frame is shown in (2).

(2) (a) *Most* squirrels gather **nuts**... (T, *most*)

(b) *Most* squirrels gather **nails**... (A, *most*)

(c) *Few* squirrels gather **nuts**... (T, *few*)

(d) *Few* squirrels gather **nails**... (A, *few*)

(e) Squirrels *often* gather **nuts**... (T, *most*)

(f) Squirrels *often* gather **nails**... (A, *most*)

(g) Squirrels *rarely* gather **nuts**... (T, *few*)

(h) Squirrels *rarely* gather **nails**... (A, *few*)

The quantifiers used in sentences (a)-(d) differed by sentence frame; see subsection 6.5.4 for a full list.

6.2.2 Language Models

To cover a range of language models with different training data and numbers of parameters, we run our analyses on the GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), GPT-Neo (Black et al., 2021; including GPT-J, Wang and Komatsuzaki, 2021), and

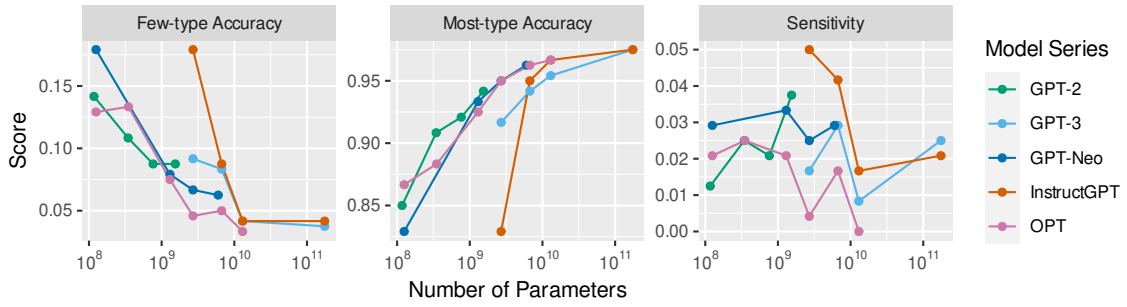

Figure 6.1: Accuracy and sensitivity of all models.

OPT (Zhang et al., 2022) language models. We also include an analysis of the first series of InstructGPT models (`text-davinci-001` etc.), which were finetuned on human-written and highly-rated model-generated responses (OpenAI, 2023).

6.2.3 Evaluation

For each stimulus sentence, we calculate the surprisal of the critical word, that is, the word for which the N400 response was measured in the original study. Because humans only encounter the context preceding the critical word when processing the word, and because the language models we analyze are all autoregressive, we only consider the surprisal of the critical word given its preceding context. To do this we truncated the sentence before the critical word, and then used the relevant language model to calculate the probability p of the target word given the preceding context, which was then converted to surprisal S following Equation 6.1.

$$S = -\log p(w_i|w_1...w_{i-1}) \tag{6.1}$$

In previous work of this type (e.g., Ettinger, 2020), only words that were single

tokens in the models' vocabularies were used. In this study, all models are autoregressive, so for multi-token words, consecutive sub-word tokens can be predicted, the product of which is a well-defined probability for the whole word. The surprisal of such words, then, is the sum of the surprisals of the sub-word tokens. Calculating surprisal this way allows us to compare the predictions of all the models for all the stimuli in the original experiment.

In order to evaluate how well each model takes into account the quantifier in its predictions, we compared which of the two possible critical words (typical or atypical) had a lower surprisal, i.e., was more strongly predicted by the model. To align with human plausibility judgements, following a *most*-type quantifier, the typical continuation was judged to be correct, and following a *few*-type quantifier, the atypical continuation was judged to be correct. Accuracy was calculated as the fraction of the stimulus pairs for which the model predicted the appropriate critical word—that is, predicted the correct continuation more strongly than the incorrect one. For example, the set of stimuli presented in (2) is made up of 4 pairs of stimuli, and for a model to achieve 100% accuracy (4/4), it would need to predict (a) over (b), (d) over (c), (e) over (f), and (h) over (g). This design intrinsically controls for any differences in unconditioned probability among the final words themselves.

Following Ettinger (2020), we also analyzed model sensitivity to the quantifiers. In the present study, this corresponds to the question of whether, for a given sentence frame, the model makes a different prediction following a *few*-type quantifier than it does following a *most*-type quantifier. We defined sensitivity as the proportion of stimuli for which the model correctly predicts the critical word following both the *most*-type and the *few*-type quantifier. Thus, the stimuli in each sentence frame provide 2 data points for sensitivity: in (2), sensitivity is calculated for (a)-(d) and for (e)-(h). For the (a)-(d) stimuli, a model would be considered sensitive to the quantifier if it correctly predicted (a)

over (b) *and* (d) over (c). Code and data are available at `https://osf.io/vjyw9`.

6.2.4 Results

Each model's accuracy at predicting the critical words following *most*- and *few*-type quantifiers is shown in Figure 6.1. All model series show the same general tendencies in accuracy: (1) they perform quite poorly for *few*-type quantifiers but relatively well for *most*-type quantifiers; and (2) as model size increases, word prediction following *most*-type quantifiers improves, but it degrades following *few*-type quantifiers. Figure 6.1 does show small exceptions to this pattern. From GPT-2 762M to 1542M and from InstructGPT 13B to 175B, while *most*-performance increases, *few*-performance does not decrease. Furthermore, from OPT 125M to 350M, and from OPT 2.7B to 6.7B, there is actually a slight improvement. Nonetheless, these differences are small compared to the overall decreases in performance, and the general trends are still clear—for example, no model performs better on *few*-type quantifiers than a model two or more sizes smaller.

With sensitivity, as shown in Figure 6.1, some models improve as they increase in size, and some get worse; however, even the greatest distance between the sensitivity of two models in the same series (InstructGPT 2.7B and 13B) is only 3.4%. Thus, other than the general fact that sensitivity is low across all models, there does not appear to be any clear pattern, suggesting that sensitivity does not drive the effects seen in accuracy. All accuracy and sensitivity scores can be found in subsection 6.5.3.

6.2.5 Discussion

These results show that contemporary autoregressive transformer models perform poorly on *few*-type quantifiers, and that as these models increase in size, they tend to improve at predicting words following *most*-type quantifiers but get worse at predicting

words following *few*-type quantifiers. In fact, we see that models that better predicted the more typical word after a *most*-type quantifier were also worse at predicting the less typical word following a *least*-type quantifier. The fact that models were evaluated on which of the two options they predicted to be more likely, combined with generally poor and largely invariant sensitivity (peaking at 5%), suggests that the larger models generally made predictions increasingly in accordance with typicality, overwhelming any sensitivity to quantifier type. This aligns with previous work on negation and logical quantifiers in language models (Ettinger, 2020; Kassner and Schütze, 2020; Kalouli et al., 2022), as well as the N400 results of the original study by Urbach and Kutas (2010).

6.3 Experiment 2: Sentence-final nouns

6.3.1 Method

The models and evaluation approach were identical to Experiment 1. The materials were identical to Experiment 1 with the single difference that all nouns were followed by a period, and the surprisal of this period was included when calculating the total surprisal of the critical word (e.g., *nuts.* or *nails.* for the example presented in (2)). Thus, surprisal reflected both the surprisal of the critical word in context and the surprisal of the word being followed by a period, i.e., being the last word in the sentence. For a discussion of modeling the probability of sentence-final words in this way, see Szewczyk and Federmeier (2022).

6.3.2 Results

Results are shown in Figure 6.2. As in Experiment 1, larger models perform worse overall. However, there is a small improvement in the very largest GPT-3 and InstructGPT

models relative to the second-largest models of the same type, both in *few*-type accuracy and sensitivity. Performance also increases on these metrics between OPT 2.7B and OPT 6.7B; however, this decreases with OPT 13B. All accuracy and sensitivity scores can be found in subsection 6.5.3.

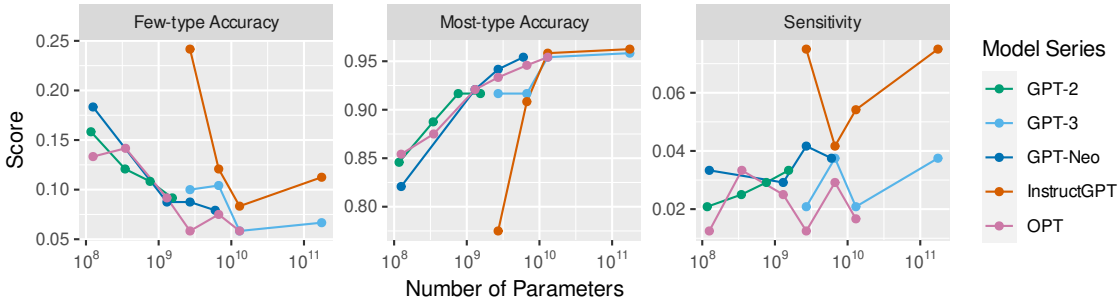

Figure 6.2: Accuracy and sensitivity of all models on stimuli with added periods (e.g., *Few squirrels gather* ***nuts.***).

6.3.3 Discusion

Overall, the results are similar to those of Experiment 1: Larger models of the same type perform worse than smaller models. Whether the small improvement of the largest GPT-3 and InstructGPT models relative to the second-largest models is a fluctuation like that seen for OPT or the beginnings of a U-shaped curve (see Wei et al., 2022b) is a question for further research.

6.4 General Discussion

In this study, we investigated whether language models show the same insensitivity towards *few*-type and *most*-type quantifiers observed in the predictions made by humans during language comprehension, as indexed by the N400 response. We find that

when tested on the same stimuli, they do, predicting the ostensibly implausible *few squirrels gather* ***nuts*** to be more likely than *few squirrels gather* ***nails***. Moreover, we find that as language models increase in size, they tend to show this effect to a greater extent, an example of inverse scaling. Based on our analysis of sensitivity and accuracy with *most*-type quantifiers, we hypothesize that these results are due to a low degree of sensitivity to quantifiers and an increase in sensitivity to typicality. In other words, language models appear to be increasingly sensitive to the fact that *squirrels gather* ***nuts*** is more plausible than *squirrels gather* ***nails***, but not to the effect on meaning that is caused by a preceding *most* or *few*.

It is often assumed that as models increase in size and are trained on more data, their performance on natural language tasks generally improves—indeed, evidence supports this (Brown et al., 2020; Raffel et al., 2020; Lin et al., 2021; Srivastava et al., 2022; Hoffmann et al., 2022; Rae et al., 2022; Zhang et al., 2022; Chowdhery et al., 2022). However, the predictions of larger models and those trained on more data also increasingly correlate with human incremental online predictions, in particular those indexed by N400 amplitude (Frank et al., 2015; Aurnhammer and Frank, 2019a,b; Michaelov and Bergen, 2020; Merkx and Frank, 2021; Michaelov et al., 2021, 2022). The two are often aligned—it is easier for humans to process well-formed sentences with plausible semantics (Frisch and Schlesewsky, 2005; Nieuwland et al., 2020). But in cases such as the present study, the two are not aligned, and we see instead that the predictions of larger models correlate better with human online predictions, even when these are contrary to offline judgements. Thus, the increased performance we see at tasks corresponding to offline human judgements—and note that virtually all manually-annotated tasks are based on offline human judgements—may in fact be a by-product of the models' predictions resembling the online predictions.

Fortunately, the literature boasts a wealth of psycholinguistic studies where met-

rics of online prediction such as the N400 appear to conflict with offline judgements. Future work could use these to identify phenomena where language models may struggle to make predictions in line with human judgements. Such cases are important to detect as use of LMs becomes more widespread. But by the same token, the present study shows that as language models increase in size, even when augmented by finetuning on desirable responses, they can make predictions that align less and less with explicit human judgements.

This may be a clear indication of an inherent 'outer misalignment' present in language models: while humans might like language models to generate plausible sentences, by their nature they can only generate the most statistically probable ones. Just as there is no guarantee of accuracy or coherence (Bender et al., 2021), there is no guarantee of plausibility. While it may be possible to tailor training to avoid specific known issues, this misalignment between probability and plausibility may pose a fundamental challenge with current approaches that aim to use language models as general-purpose natural language systems.

6.5 Appendices

6.5.1 Limitations

There are two main limitations to our study. The first is that the stimuli used were limited to those provided by Urbach and Kutas's (2010) study. This is because, as stated, we wanted to be able to compare the patterns in the language models' predictions to the patterns in the human N400 response. Thus, we do not look at logical quantifiers like Kalouli et al. (2022), or any others that have previously been studied (in, e.g., Pezzelle et al., 2018; Talmor et al., 2020).

The other (and perhaps more important) limitation is in the models we were able to use. Crucially, we were not able to access models larger than GPT-3 175B such as PaLM 540B (Chowdhery et al., 2022). This is important because recent work has shown that some inverse scaling patterns become U-shaped (i.e., as language model size increases, performance degrades and then improves again) with such larger models (Wei et al., 2022b).

6.5.2 Ethics Statement

Our work complies with the ACL Ethics Policy. Beyond this, we are not aware of any way in which the results of this study may be harmful—in fact, if anything, identifying the limitations of large language models is something that is likely to reduce possible harms by demonstrating cases where their use is not suitable.

From an environmental perspective, we did not train any models; we only used pretrained models for analysis, limiting energy consumption. With the exception of the GPT-3 and InstructGPT models and OPT 13B, all analyses were run on an NVIDIA RTX A6000 GPU, taking a total of 43 minutes. OPT 13B was too large to run on this GPU, and thus was run on an Intel Dual Xeon E7-4870 CPU for a total of 22 hours and 39 minutes.

Finally, the GPT-3 and the InstructGPT models were run using the OpenAI API, and thus we do not have access to information about the GPUs used.

6.5.3 Scores

The performance of each model on the original stimuli is presented in Table 6.1 and the performance for the stimuli with the added period at the end of the critical word inb Table 6.2.

Table 6.1: Accuracy and sensitivity scores for all models for original stimuli.

	Critical word		
	Accuracy		Sens.
Model	***most***	***few***	
GPT-2 117M (`gpt2`)	0.850	0.142	0.013
GPT-2 345M (`gpt2-medium`)	0.908	0.108	0.025
GPT-2 762M (`gpt2-large`)	0.921	0.088	0.021
GPT-2 1542M (`gpt2-xl`)	0.942	0.088	0.038
GPT-3 2.7B (`ada`)	0.917	0.092	0.017
GPT-3 6.7B (`babbage`)	0.942	0.083	0.029
GPT-3 13B (`curie`)	0.954	0.042	0.008
GPT-3 175B (`davinci`)	0.975	0.038	0.025
InstructGPT 2.7B (`text-ada-001`)	0.829	0.179	0.050
InstructGPT 6.7B (`text-babbage-001`)	0.950	0.088	0.042
InstructGPT 13B (`text-curie-001`)	0.967	0.042	0.017
InstructGPT 175B (`text-davinci-001`)	0.975	0.042	0.021
GPT-Neo 125M (`EleutherAI/gpt-neo-125m`)	0.829	0.179	0.029
GPT-Neo 1.3B (`EleutherAI/gpt-neo-1.3B`)	0.933	0.079	0.033
GPT-Neo 2.7B (`EleutherAI/gpt-neo-2.7B`)	0.950	0.067	0.025
GPT-J 6B (`EleutherAI/gpt-j-6b`)	0.963	0.062	0.029
OPT 125M (`facebook/opt-125m`)	0.867	0.129	0.021
OPT 350M (`facebook/opt-350m`)	0.883	0.133	0.025
OPT 1.3B (`facebook/opt-1.3b`)	0.925	0.075	0.021
OPT 2.7B (`facebook/opt-2.7b`)	0.950	0.046	0.004
OPT 6.7B (`facebook/opt-6.7b`)	0.963	0.050	0.017
OPT 13B (`facebook/opt-13b`)	0.967	0.033	0

Table 6.2: Accuracy and sensitivity scores for all models for stimuli with added period.

	Critical word + period		
	Accuracy		**Sens.**
Model	***most***	***few***	
GPT-2 117M (`gpt2`)	0.846	0.158	0.021
GPT-2 345M (`gpt2-medium`)	0.887	0.121	0.025
GPT-2 762M (`gpt2-large`)	0.917	0.108	0.029
GPT-2 1542M (`gpt2-xl`)	0.917	0.092	0.033
GPT-3 2.7B (`ada`)	0.917	0.1	0.021
GPT-3 6.7B (`babbage`)	0.917	0.104	0.038
GPT-3 13B (`curie`)	0.954	0.058	0.021
GPT-3 175B (`davinci`)	0.958	0.067	0.038
InstructGPT 2.7B (`text-ada-001`)	0.775	0.242	0.075
InstructGPT 6.7B (`text-babbage-001`)	0.908	0.121	0.042
InstructGPT 13B (`text-curie-001`)	0.958	0.083	0.054
InstructGPT 175B (`text-davinci-001`)	0.963	0.112	0.075
GPT-Neo 125M (`EleutherAI/gpt-neo-125m`)	0.821	0.183	0.033
GPT-Neo 1.3B (`EleutherAI/gpt-neo-1.3B`)	0.921	0.088	0.029
GPT-Neo 2.7B (`EleutherAI/gpt-neo-2.7B`)	0.942	0.088	0.042
GPT-J 6B (`EleutherAI/gpt-j-6b`)	0.954	0.079	0.038
OPT 125M (`facebook/opt-125m`)	0.854	0.133	0.013
OPT 350M (`facebook/opt-350m`)	0.875	0.142	0.033
OPT 1.3B (`facebook/opt-1.3b`)	0.921	0.092	0.025
OPT 2.7B (`facebook/opt-2.7b`)	0.933	0.058	0.013
OPT 6.7B (`facebook/opt-6.7b`)	0.946	0.075	0.029
OPT 13B (`facebook/opt-13b`)	0.954	0.058	0.017

6.5.4 Quantifiers

Table 6.3 lists all quantifiers used and the proportion of sentences in WikiText-103 that contain them.

	***Most*-type**		***Few*-type**	
	Quantifier	**Frequency**	**Quantifier**	**Frequency**
	most	0.025177	few	0.005870
	almost all	0.000305	almost no	0.000098
	practically all	0.000009	practically no	0.000008
	a large number of	0.000300	a small number of	0.000131
	nearly all	0.000170	rather few	0.000001
	lots of	0.000153	hardly any	0.000017
	a lot of	0.000745	a very few	0.000010
	many	0.015874	few	0.005870
	often	0.005766	rarely	0.000610
Total		0.046809		0.006717

Table 6.3: All quantifiers used by Urbach and Kutas (2010). In each sentence frame, *most* and *few*-type quantifiers were matched based on their meanings as length in number of words (Urbach and Kutas, 2010). Matched quantifiers are shown beside each other. As can be seen, *few* is matched to both *most* and *many*. The frequency of each quantifier is given in terms of the proportion of sentences in WikiText-103 (Merity et al., 2017) that contain it. The total frequencies are the number of sentences in WikiText-103 that contain at least one of either the *few*-type or *most*-type quantifiers; not the sum of the individual quantifier frequencies.

6.6 Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. We would also like to acknowledge the other members of the Language and Cognition Lab at UCSD for their valuable discussion, as well as Roger Levy and attendees of the MIT Computational Psycholinguistics Laboratory meeting. Finally, we would like to thank the

San Diego Social Sciences Computing Facility Team for the use of the Social Sciences Research and Development Environment (SSRDE) cluster. The RTX A6000 used for this research was donated by the NVIDIA Corporation.

Chapter 6, in full, is a reprint of the material as it appears in Michaelov, J. A. & Bergen, B. K., "*Rarely* a problem? Language models exhibit inverse scaling in their predictions following *few*-type quantifiers" *Findings of the Association for Computational Linguistics: ACL 2023*, 2023. The dissertation author was the primary investigator and author of this paper. Minor edits have been made to bring the formatting in line with the dissertation template.

Chapter 7

Can Peanuts Fall in Love with Distributional Semantics?

Abstract

Context changes expectations about upcoming words—following a story involving an anthropomorphic peanut, comprehenders expect the sentence *the peanut was in love* more than *the peanut was salted*, as indexed by N400 amplitude (Nieuwland and van Berkum, 2006). This updating of expectations has been explained using Situation Models—mental representations of a described event. However, recent work showing that N400 amplitude is predictable from distributional information alone raises the question whether situation models are necessary for these contextual effects. We model the results of Nieuwland and van Berkum (2006) using six computational language models and three sets of word vectors, none of which have explicit situation models or semantic grounding. We find that a subset of these can fully model the effect found by Nieuwland and van Berkum (2006). Thus, at least some processing effects normally explained through

situation models may not in fact require explicit situation models.

7.1 Introduction

It is widely believed that prediction plays a key role in language processing, with more predictable words being processed more easily (Fischler and Bloom, 1979; Kutas and Hillyard, 1984; Levy, 2008; Kutas et al., 2011; Van Petten and Luka, 2012; DeLong et al., 2014b; Luke and Christianson, 2016; Kuperberg et al., 2020). Perhaps the strongest evidence for this comes from the N400, a neural signal of processing difficulty that is highly correlated with lexical probability—contextually probable words elicit an N400 response of smaller (less negative) amplitude than contextually improbable words, whether predictability is determined based on human judgements (Kutas and Hillyard, 1984; for review see Van Petten and Luka, 2012) or a corpus (Parviz et al., 2011; Frank et al., 2015; Aurnhammer and Frank, 2019b; Merkx and Frank, 2021; Szewczyk and Federmeier, 2022; Michaelov et al., 2022, 2023).

A striking feature of the predictions indexed by the N400 is how flexible they can be. Under normal circumstances, a sentence such as *the peanut was in love* would be highly improbable, much more so than *the peanut was salted*. Following the short story in (1), however, this changes (Nieuwland and van Berkum, 2006).

(1) A woman saw a dancing peanut who had a big smile on his face. The peanut was singing about a girl he had just met. And judging from the song, the peanut was totally crazy about her. The woman thought it was really cute to see the peanut singing and dancing like that.

In fact, Nieuwland and van Berkum (2006), who tested this in Dutch, found that in the context of (1), the last word of *de pinda was* ***verliefd*** ('the peanut was **in love**')

elicited a smaller N400 than *de pinda was* ***gezouten*** ('the peanut was **salted**'). How does such a dramatic reversal occur?

One possibility put forward by Nieuwland and van Berkum (2006) is that while reading the context, the reader's mental representation of the peanut is altered such that it is treated as an animate entity. This, as Nieuwland and van Berkum (2006) note, is in line with theories of situation models, which argue that we track the entities under discussion, as well as their properties and relations. Such accounts generally involve explicit structures or schemata, grounding in world knowledge or experience, extraction of propositional information, or a combination of these (see, e.g., Bransford et al., 1972; Kintsch and van Dijk, 1978; Johnson-Laird, 1980; Garnham, 1981; Johnson-Laird, 1983; van Dijk and Kintsch, 1983; Kintsch, 1988; Zwaan et al., 1995a,b; Radvansky et al., 1998; Kintsch, 1998; Zwaan and Radvansky, 1998; Zwaan and Madden, 2004; Kintsch, 2005; Van Berkum et al., 2007; Kintsch and Mangalath, 2011; Butcher and Kintsch, 2012; Zwaan, 2014, 2016; Zacks and Ferstl, 2016; Kintsch, 2018; Hoeben Mannaert and Dijkstra, 2021). On a situation model account, the reader alters their semantic representation of the peanut to give it animate features in accordance with the information that it can sing, dance, and show emotions, thereby facilitating the processing of *in love*.

The hypothesis that structured or grounded situation models explain N400 effects such as those found by Nieuwland and van Berkum (2006) is generally accepted (e.g., Hagoort and van Berkum, 2007; Filik and Leuthold, 2008; Warren et al., 2008; Rosenbach, 2008; Ferguson and Sanford, 2008; Ferguson et al., 2008; Menenti et al., 2009; Bicknell et al., 2010; de Groot, 2011; Metusalem et al., 2012; Aravena et al., 2014; Zwaan, 2014; Xiang and Kuperberg, 2015; Kuperberg et al., 2020) and has been shown to be viable using computational models (Venhuizen et al., 2019). However, there are alternative explanations.

The present study asks whether the effect can instead be explained by lexical preactivation based on distributional linguistic knowledge, following the findings that the statistics of language can be used to model N400 effects (Ettinger et al., 2016; Michaelov and Bergen, 2020; Michaelov et al., 2021; Michaelov and Bergen, 2022a; Uchida et al., 2021; Michaelov et al., 2023) and predict single-trial N400 amplitude (Chwilla and Kolk, 2005; Parviz et al., 2011; Van Petten, 2014; Frank et al., 2015; Aurnhammer and Frank, 2019a,b; Merkx and Frank, 2021; Michaelov et al., 2021; Szewczyk and Federmeier, 2022; Michaelov et al., 2023).

Specifically, we look at two possible ways in which this might arise. One, which we refer to as *event-level priming*, refers to the idea that a word associated with a previously-discussed event may be more likely to be predicted by virtue of this. This is something that has been previously reported in the N400—Metusalem et al. (2012), for example, found that that merely being related to the event under discussion leads to a smaller N400 response to a word even when that word is inappropriate. Michaelov and Bergen (2022a) model this with transformer language models—systems trained to calculate the probability of a word given its context based on the statistics of language alone—showing that this effect is explainable with distributional information. Thus, it may be the case that the fact that *in love* is related to, for example, being *crazy about* someone that leads to it being predicted to be more likely than *salted*. Following Michaelov and Bergen (2022a), we investigate this using 6 Dutch transformer language models (Havinga, 2021, 2022c,a,b; de Vries et al., 2019; Delobelle et al., 2020), testing whether they show the same effect as humans—that is, whether they predict the canonical sentence the *peanut was salted* to be less likely than the noncanonical sentence *the peanut was in love*.

An alternative possibility is *lexical priming*. More simply than in the case of event-level priming, it may be the case that the preceding context involving words such as

dancing, *smile*, *singing*, *crazy*, and *cute* exerts a stronger pressure on prediction of *in love* than *peanut* does on *salted*. Intuitively, one might expect that a system (neurocognitive or computational) displaying event-level priming is likely to display lexical priming—indeed, lexical priming is a possible mechanism by which at least some part of event-level priming could be achieved. The fact that lexical priming is likely to occur in a system displaying event-level priming is also supported by the fact that language models show both (Kassner and Schütze, 2020; Misra et al., 2020; Michaelov and Bergen, 2022a). Thus, in the present study, we distinguish between two possible explanations of the effect found by Nieuwland and van Berkum (2006): lexical priming alone, and event-level priming that may include lexical priming.

As discussed, language models can be used to model the latter. To model the former, we turn to word vectors—representations of words derived from their co-occurrence statistics, either directly or based on word embeddings learned by neural networks (see, e.g., Dumais et al., 1988; Landauer et al., 1998; Mikolov et al., 2013b; Pennington et al., 2014; Mikolov et al., 2018; Tulkens et al., 2016; Grave et al., 2018). The cosine distance between the vector of each critical word (e.g. *in love* or *salted*) and the mean of the vectors of the words in the preceding context can therefore be used to test how similar the critical word is to the words preceding it (Ettinger et al., 2016; Uchida et al., 2021), and thereby model the effects of lexical priming alone. To do this this we used three sets of Dutch word vectors (from Tulkens et al., 2016; and Grave et al., 2018).

7.2 Background

A number of researchers have attempted to model the N400 computationally, including using language models (Parviz et al., 2011; Frank et al., 2015; Aurnhammer and Frank, 2019b; Michaelov and Bergen, 2020; Merkx and Frank, 2021; Michaelov et al., 2021,

2022; Szewczyk and Federmeier, 2022; Michaelov et al., 2023) and the distances between vector representations of words (Parviz et al., 2011; Van Petten, 2014; Ettinger et al., 2016; Uchida et al., 2021; Michaelov et al., 2023). There have also been several attempts to computationally model whether the amplitude of the N400 response is impacted by situation models (Uchida et al., 2021; Venhuizen et al., 2019) and thematic roles (Brouwer et al., 2017; Fitz and Chang, 2019; Rabovsky et al., 2018).

To our knowledge, only one previous study (Uchida et al., 2021) has directly attempted to model the discourse effect found by Nieuwland and van Berkum (2006), and it does not rely on purely distributional linguistic information. Uchida et al. (2021) base their model on Wikipedia2Vec (Yamada et al., 2020) vectors—while these include distributional information derived from the surface-level statistics of language, they also include information about hyperlinks between Wikipedia pages, and thus structured semantic relations based on human judgements of relevance and importance (Yamada et al., 2020). Additionally, Uchida et al. (2021) only look at the English-translated version of the single stimulus item presented in (1), and thus, it is unclear whether the results generalize to all the stimuli in the original study. The current study overcomes these inferential limitations by using the original Dutch stimuli and by using neural language models and word vectors trained only on natural language input.

7.3 The present study

We investigate the adequacy of distributional knowledge to explain the human N400 effect found by Nieuwland and van Berkum (2006) using predictions of neural network language models and the distance between the word vectors of the critical words and their context. Specifically, we ask this question for two possible variants of the effect found by Nieuwland and van Berkum (2006).

Nieuwland and van Berkum (2006) presented experimental participants with short stories such as those in (1) including "canonical" sentences like *the peanut was **salted*** or "noncanonical" ones like *the peanut was **in love***. One approach to whether language models and humans show the same prediction patterns (taken by Uchida et al., 2021) is to compare the statistical metrics the critical words elicit in the context of the full story versus in isolation. Without preceding context, these sentences should produce values that match the canonicality of the sentence, but the difference should attenuate or reverse following the story context.

Thus, we ran a statistical analysis testing for an interaction between stimulus length (full story or only the last sentence) and canonicality (canonical or noncanonical). Such an interaction would reveal a context-dependent difference in the effect of canonicality on our statistical metrics; and thus would replicate in neural language models the effect found by Nieuwland and van Berkum (2006).

However, an interaction between stimulus length and canonicality in this direction could result from either a reversal or a decrease in the magnitude of the canonicality effect. Canonical stimuli might elicit lower surprisals or smaller cosine distances in both context conditions, but of different magnitudes. For this reason, we label the effect measured by an interaction (in the expected direction) a **reduction effect**.

Nieuwland and van Berkum (2006) did not employ the 2 x 2 design that would allow them to detect an interaction—they compared the N400 in context only, finding that canonical stimuli actually elicited larger N400 responses than noncanonical stimuli. To replicate this finding, we test whether the canonical full-length stimuli elicit higher surprisals or greater cosine distances than the noncanonical full-length stimuli, a **reversal effect**.

If either language models or word vectors can successfully model the reversal ef-

fect, this would suggest that distributional information is sufficient to explain the data reported by Nieuwland and van Berkum (2006). Thus, while situation models and extralinguistic information may be involved in the neurocognitive system underlying the N400, additional evidence is required to prove this. If neither can model either effect, this would undermine the claim that distributional information is sufficient to explain the effect found by Nieuwland and van Berkum (2006). Finally, if either language models or word vectors can successfully model the reduction effect but not the full reversal effect, this may support the idea that distributional information could be used as part of the neurocognitive system underlying the N400 response, but that it is not sufficient to yield the dynamic contextual sensitivity humans display. Situation models and other sources of information might explain the remainder.

7.4 Method

7.4.1 Materials

Stimuli were used from the original experiment, and are provided online[1] by the authors (Nieuwland and van Berkum, 2006). We compared the effect of experimental condition on the N400 and on neural network surprisal (as in Michaelov and Bergen, 2020) and the cosine similarity between the word vector of the critical word and the mean of the word vectors in its context (as in Ettinger et al., 2016).

The stimuli use 60 full-length story frames, each of which has either a canonical or noncanonical predicate, for 120 unique stories. As the aim is to model human online comprehension processes, the models only used the text before the critical words (e.g., *in love* or *salted*) to predict the critical words, so stories were truncated after the critical word. For the critical sentence stimuli, we isolated the last sentence of these truncated stories,

[1] `https://www.researchgate.net/publication/268208198`

including and up to the critical word in each story (e.g., *The peanut was in love*). This produced 240 stimuli, as shown in Table 7.1.

Table 7.1: Experimental stimuli derived from Nieuwland and van Berkum (2006).

Predicate Type	**Stimulus Length**	**Count**
Canonical	Full-length	60
Canonical	critical sentence	60
Noncanonical	Full-length	60
Noncanonical	critical sentence	60

7.4.2 Statistical Analysis

Statistical analysis and data manipulation were carried out 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. Code, data, and statistical analyses are provided at `https://osf.io/wnj76`.

7.5 Experiment 1: Language Models

7.5.1 Language models

We used six pretrained models available through the *transformers* package (Wolf et al., 2020). These were all of the available monolingual Dutch language models using standard architectures and training procedures at the time of analysis. Four of these models—Dutch versions of the medium (Havinga, 2021) and large (Havinga, 2022c) GPT-2 models (Radford et al., 2019) and Dutch versions of the 125 million parameter (Havinga, 2022a) and 1.3 billion parameter (Havinga, 2022b) GPT-Neo models (Black et al., 2021)—were autoregressive, meaning that they are trained to predict a word based only on its preceding context. The remaining two models—BERTje (de Vries et al., 2019) and Rob-

BERT v2 (Delobelle et al., 2020), based on BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), respectively—are masked language models, meaning that they are also trained to predict a word based on the text following the critical word. However, as stated, in the present study, all models were only provided with the context preceding the critical words. We ran the stimuli through each language model, calculating the surprisal of each critical word that was in the model's vocabulary (we restricted our analyses to these items). To do this, we calculated the negative of the logarithm of the probabilities provided for each critical word by each of the language models. We then tested for the reduction and reversal effects with these surprisal values. The language models were run in *Python* (Van Rossum and Drake, 2009), using the *PyTorch* (Paszke et al., 2019) implementation of each model, as provided by the *transformers* package (Wolf et al., 2020).

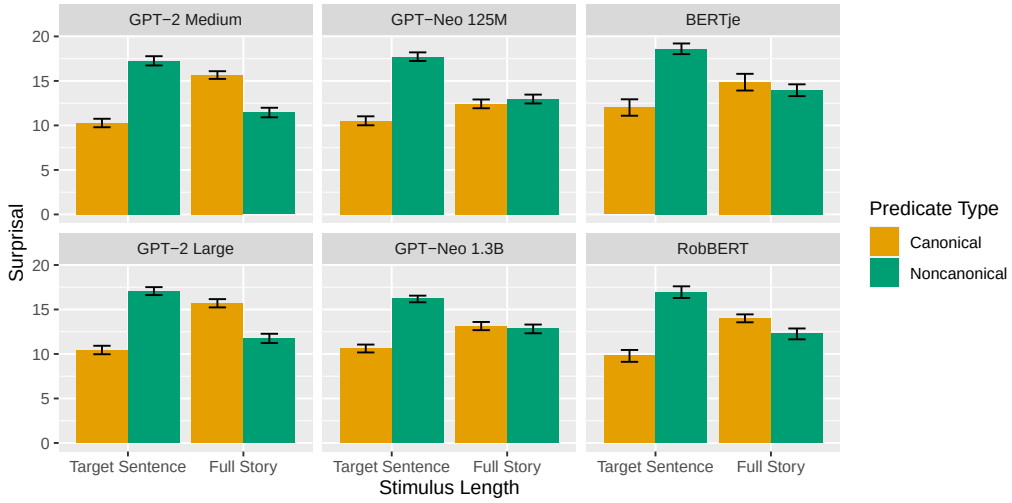

Figure 7.1: Surprisal elicited by critical words for each predicate type and stimulus length.

7.5.2 Reduction effect

In order to test the reduction effect, we constructed linear mixed-effects regression models, with the surprisal calculated from each language model as the dependent variable. In each model, predicate type (canonical or noncanonical) and stimulus length (full-length or critical sentence) were fixed effects and story frame (each of the 60) was a random intercept. For the regressions with the autoregressive models and BERTje surprisal as their dependent variables, we then constructed regressions also including an interaction between predicate type and stimulus length. Using likelihood ratio tests, we found that these regressions including the interaction fit the data significantly better than those without the interaction (GPT-2 Medium: $\chi^2(1) = 112.0, p < 0.001$; GPT-2 Large: $\chi^2(1) = 115.9, p < 0.001$; GPT-Neo 125M: $\chi^2(1) = 67.3, p < 0.001$; GPT-Neo 1.3B: $\chi^2(1) = 56.3, p < 0.001$; BERTje: $\chi^2(1) = 44.4, p < 0.001$), indicating a significant interaction between predicate type and stimulus length. The regression with RobBERT surprisal as its dependent variable and no interaction had a singular fit, but the regression with the interaction did not. Thus, instead of running a likelihood ratio test to investigate whether there was a significant interaction, we used a a Type III ANOVA with Satterthwaite's method for estimating degrees of freedom (Kuznetsova et al., 2017) on the regression with the interaction, finding it to be a significant predictor of RobBERT surprisal ($F(1, 71.2) = 81.1, p < 0.001$). Note that all reported p-values are corrected for multiple comparisons based on false discovery rate (Benjamini and Yekutieli, 2001).

For all language models, there was a significant interaction between predicate type and stimulus length. Further inspection of the regressions showed that in all cases, the interaction was in the expected direction. Thus, all models displayed the reduction effect. This can be seen visually in Figure 7.1—in all models, when only the critical sentence was presented, the mean surprisal for critical words in canonical sentences is

lower than for critical words in noncanonical sentences. Conversely, when the full-length story is presented to the language models, the critical words in the noncanonical sentences elicit a lower or roughly-equal surprisal than the critical words in the canonical sentences.

7.5.3 Reversal effect

To test for which models this latter finding was statistically significant, we initially attempted to fit linear mixed-effects regression models for each the full-length and critical sentence stimulus results for each language model; however, this led to several models with singular fits. Instead, we carried out pairwise two-tailed t-tests, comparing the surprisal of canonical and noncanonical stimuli for full-length and critical sentence stimuli for each language model.

First, we test whether the decontextualized canonical critical sentence stimuli elicit significantly lower surprisals than noncanonical critical sentence stimuli. After correction for multiple comparisons, they do so in all language models (GPT-2 Medium: $t(88.7) = -9.91$, $p < 0.001$; GPT-2 Large: $t(88.1) = -10.1$, $p < 0.001$; GPT-Neo 125M: $t(88.6) = -10.3$, $p < 0.001$; GPT-Neo 1.3B: $t(85.5) = -9.62$, $p < 0.001$; BERTje: $t(48.4) = -5.99$, $p < 0.001$; RobBERT: $t(55.1) = -7.67$, $p < 0.001$).

Next, in order to investigate the reversal effect, we test whether canonical full-length stimuli elicit lower surprisals than noncanonical full-length stimuli. After correction for multiple comparisons, only the Dutch GPT-2 models successfully model the reversal effect—they are the only models for which canonical full-length stimuli elicit significantly higher surprisals than noncanonical full-length stimuli (GPT-2 Medium: $t(86.3) = 6.11$, $p < 0.001$; GPT-2 Large: $t(88.4) = 5.65$, $p < 0.001$).

The difference in other models was not significant after correction for multiple comparisons (GPT-Neo 125M: $t(88.9) = -0.77$, $p = 1.000$; GPT-Neo 1.3B: $t(88.9) = 0.47$,

$p = 1.000$; BERTje: $t(51.5) = 0.79$, $p = 1.000$; RobBERT: $t(46.6) = 2.32$, $p = 0.120$).

However, it is worth noting that the contrast between the two sets of results (critical sentence only vs. full stimulus) means that significant canonicality effects for the critical sentence stimuli disappear in the full-length stimuli, underscoring the presence of a reduction effect in the Dutch GPT-Neo models, BERTje, and RobBERT.

7.5.4 Discussion

Nieuwland and van Berkum (2006) found that in a suitably supportive context, noncanonical stimuli like *de pinda was* ***verliefd*** ('the peanut was **in love**') elicit smaller N400 responses than canonical stimuli such as *de pinda was* ***gezouten*** ('the peanut was **salted**')—context not only mitigated but reversed the effect of animacy violation.

We find that two language models also display this reversal effect: Dutch GPT-2 Medium (Havinga, 2021) and Dutch GPT-2 Large (Havinga, 2022c). When these models are presented with the same contexts, the surprisal of critical words in the noncanonical condition is lower than that elicited by those in the canonical condition.

This is not the case for the remaining four language models: Dutch GPT-Neo 125M (Havinga, 2022a), Dutch GPT-Neo 1.3B (Havinga, 2022b), BERTje (de Vries et al., 2019), and RobBERT (Delobelle et al., 2020). However, these models do display the weaker reduction effect, and further, the absence of a significant difference between conditions for these models when presented with the full stories shows that the difference between canonical and noncanonical critical sentence stimuli is not just reduced, but disappears entirely.

It may be tempting to infer that the architecture of autoregressive transformers, and in particular, those based on the GPT-2 architecture, leads to success capturing the effect. However, it should be noted that before correction for multiple comparisons, Rob-

BERT also successfully displays the reversal effect. In addition, not all language models had the same vocabulary, and thus, a different number of items were analyzed across models[2]. For these reasons, and because these models are all of various sizes and trained on several different datasets, we believe it would be premature to draw conclusions about how language model architecture impacts whether a model displays the reversal effect.

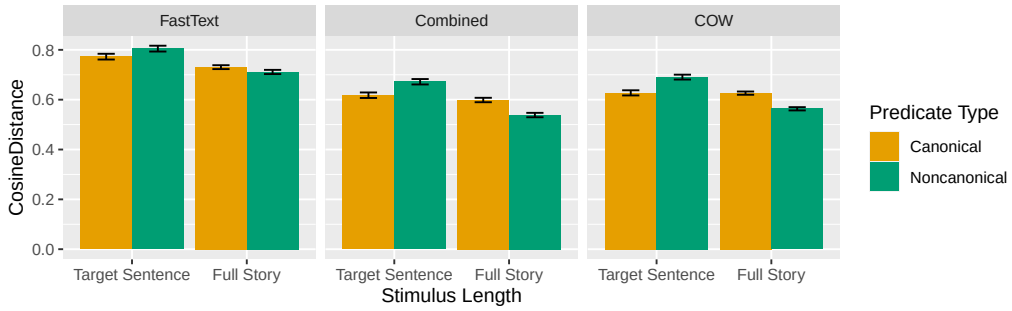

Figure 7.2: Cosine distance elicited by critical words for each predicate type and stimulus length.

7.6 Experiment 2: Word Vectors

7.6.1 Cosine Distance

In this study, we used 3 sets of pretrained word vectors: the 300-dimensional Dutch *fastText* vectors (Grave et al., 2018) trained on Dutch text from Wikipedia[3] and Common Crawl[4] and two 320-dimensional Dutch word vectors released by Tulkens et al. (2016)—one trained on *COW* (COrpora from the Web; Schäfer and Bildhauer, 2012) and one trained on a *Combined* corpus made up of the SoNaR corpus (Oostdijk et al., 2013)

[2]Though it should be noted that an alternate analysis including all critical words by operationalizing the suprisal of multi-token words as the sum of their tokens' surprisals (see Michaelov and Bergen, 2022b) shows the same qualitative results for all models except for BERTje—which performs worse.

[3]`https://nl.wikipedia.org/`

[4]`https://commoncrawl.org/`

and text from Wikipedia and Roularta[5]. Cosine distance was calculated (using *SciPy*; Virtanen et al., 2020) between the mean of the word vectors for all words in the preceding context and the word embedding for the critical word. All critical words were present in the vectors, so all experimental items were included in the analysis; it should be noted though that words in the context that were not present in the vectors were ignored when calculating cosine distance. The cosine distances for critical words in each condition are shown in Figure 7.2.

7.6.2 Reduction effect

As with language model surprisal, we constructed linear mixed-effects regressions with predicate type and stimulus length as fixed effects and story from as a random intercept. With these models, the cosine distance calculated using each set of word vectors was the dependent variable. The interaction between predicate type and stimulus length was significant for all vectors after correcting for multiple comparisons (fastText: $\chi^2(1) = 12.0, p = 0.003$; Combined: $\chi^2(1) = 40.8, p < 0.001$; COW: $\chi^2(1) = 66.4, p < 0.001$).

7.6.3 Reversal effect

When comparing the cosine distances calculated between the embedding of the critical words and the preceding words of the critical sentence using two-tailed t-tests as with surprisal, there was a significant difference between canonical and noncanonical critical words for Combined and COW vectors (Combined: $t(116.9) = -3.48$, $p = 0.004$; COW: $t(116.5) = -4.45$, $p < 0.001$), but not fastText vectors (fastText: $t(118.0) = -1.96$, $p = 0.237$).

Similarly, when comparing the cosine distances between the critical word and

[5] `https://www.roularta.be`

the preceding words of the full story, there was a significant difference between canonical and noncanonical critical words for Combined and COW vectors (Combined: $t(117.0) = 4.82$, $p < 0.001$; COW: $t(117.0) = 6.78$, $p < 0.001$), but not fastText vectors (fastText: $t(117.4) = 1.68$, $p = 0.418$).

7.6.4 Discussion

The cosine distances calculated from all three sets of word vectors displayed the reduction effect, and two out of three displayed the reversal effect. Thus, the results suggest that the N400 effect reported by Nieuwland and van Berkum (2006) can be explained by lexical priming based on distributional linguistic knowledge alone.

The present study corroborates the finding of Uchida et al. (2021), and expands upon it in several ways. First, we explicitly tested for the reversal effect—not just whether canonical and noncanonical stimuli differ depending on whether there is a preceding story or not, but also whether the noncanonical sentence is more expected than the canonical when the story is present. Second, we found that word vector cosine distance can model the effect for multiple stimuli, not just the *peanut was in love* example. Third, we found that the effect can be modeled in Dutch, the language in which the human study was carried out. And finally, we found that vectors derived from text data only (i.e., without additional information) are able to model the effect.

7.7 General Discussion

Human comprehenders use context to update expectations about upcoming words, making a sentence that would be highly unlikely on its own more predictable than a sentence that would be relatively likely on its own. More strikingly, humans do this even when the event described is implausible, violating the constraint, for instance, that only animate,

conscious entities can fall in love. The human comprehension system is quite flexible if it can update expectations about what peanuts, for example, can do, based only a story that indirectly implies the animacy of a fictional peanut.

It has often been assumed that this flexibility requires situation models that are explicitly structured (Venhuizen et al., 2019) or involve non-linguistic world knowledge (Uchida et al., 2021). However, the present findings show that it is possible for purely linguistic language models model with no direct experiential grounding to update their expectations based on the linguistic context and knowledge of the statistics of language. Thus, the dynamics of event-level priming based on the distributional statistics of language may in some implicit, unspecified way approximate the effects on language comprehension previously ascribed to situation models.

In fact, the results of the present study provide evidence for an even simpler explanation. Within final sentences alone, canonical critical words were more similar to their contexts than noncanonical words, but when we include the full story context, it is the noncanonical critical words that are more similar to their contexts. It is already well-established that the amplitude of the N400 to a given word is reduced when it is semantically related to a previously-seen word (Bentin et al., 1985; Rugg, 1985; Van Petten and Kutas, 1988; Kutas and Hillyard, 1989; Holcomb, 1988; Kutas, 1993; Lau et al., 2013). Overall, then, our results show that in principle, it is possible that the pattern in the N400 responses reported by Nieuwland and van Berkum (2006) may not rely on situation models or even event-level priming, but rather reflect some form of lexical priming.

It may still be the case that humans use structured or semantically-rich situation models in online language comprehension (see, e.g., Kuperberg et al., 2020). However, the results of the study carried out by Nieuwland and van Berkum (2006) appear to provide weaker evidence for this than previously believed. Language model predictions or even

lexical priming based on language statistics appear to be sufficient to explain the effect, at least qualitatively—a valuable line of future research would be to test whether these can fully account for the effect in single-trial N400 data.

7.8 Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. This work was partially supported by a 2021-2022 Center for Academic Research and Training in Anthropogeny Annette Merle-Smith Fellowship awarded to James A. Michaelov.

Chapter 7, in full, is a reprint of the material as it appears in Michaelov, J. A., Coulson, S., & Bergen, B. K., "Can Peanuts Fall in Love with Distributional Semantics?", *Proceedings of the Annual Meeting of the Cognitive Science Society, 45*, 2023. The dissertation author was the primary investigator and author of this paper. Minor edits have been made to bring the formatting in line with the dissertation template.

Chapter 8

Strong Prediction: Language model surprisal explains multiple N400 effects

Abstract

Theoretical accounts of the N400 are divided as to whether the amplitude of the N400 response to a stimulus reflects the extent to which the stimulus was predicted, the extent to which the stimulus is semantically similar to its preceding context, or both. We use state-of-the-art machine learning tools to investigate which of these three accounts is best supported by the evidence. GPT-3, a neural language model (LM) trained to compute the conditional probability of any word based on the words that precede it, was used to operationalize contextual predictability. In particular, we used an information theoretical construct known as surprisal (the negative logarithm of the conditional probability). Contextual semantic similarity was operationalized by using two high-quality co-occurrence-

derived vector-based meaning representations for words: GloVe and fastText. The cosine between the vector representation of the sentence frame and final word was used to derive Contextual Cosine Similarity (CCS) estimates. A series of regression models were constructed, where these variables, along with cloze probability and plausibility ratings, were used to predict single trial N400 amplitudes recorded from healthy adults as they read sentences whose final word varied in its predictability, plausibility, and semantic relationship to the likeliest sentence completion. Statistical model comparison indicated GPT-3 surprisal provided the best account of N400 amplitude and suggested that apparently disparate N400 effects of expectancy, plausibility and contextual semantic similarity can be reduced to variations in the predictability of words. The results are argued to support predictive coding in the human language network.

8.1 Introduction

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

likely to reflect activity in the very sort of interactive neural networks (Fedorenko and Thompson-Schill, 2014) propose.

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

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

One reason for the continued dispute on this issue is that advocates of each account have mostly focused on a subset of N400 effects, discounting the relevance of less amenable phenomena and arguing that they are potentially explicable given a suitable operationalization of either expectancy or semantic similarity. Whereas advocates of pre-

dictive processing focus on expectancy effects (DeLong et al., 2014b; Kuperberg and Jaeger, 2016; Bornkessel-Schlesewsky and Schlesewsky, 2019; Kuperberg et al., 2020), advocates of contextual similarity and combined accounts focus on the way that N400 amplitude is modulated by the presence of semantically related words in the immediate context (Lau et al., 2013; Ettinger et al., 2016; Uchida et al., 2021; Federmeier, 2021). By contrast, the present study examines manipulations of the expectancy, plausibility, and the relatedness of sentence final words to the words that precede them.

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

8.1.1 Predictive Preactivation Account

One prominent account of the N400 is that it reflects the activation of semantic features associated with the eliciting word (Kutas and Federmeier, 2011). According to this account, contextual congruity effects occur because elements of the prior context have

already activated some of these associated features. If relevant features associated with a word have been activated by the preceding context—whether these be semantic features (Kuperberg et al., 2020; Federmeier, 2021) or a combination of semantic, grammatical, and phonological features (as supported by the work of DeLong et al., 2005; Van Berkum et al., 2005; Otten et al., 2007; Nicenboim et al., 2020; Urbach et al., 2020; Fleur et al., 2020)—they need not be newly activated when the word is encountered, and thus the amplitude of the N400 is less than when words are encountered alone or in less supportive contexts.

The most obvious source of support for predictive preactivation lies in the close relationship between N400 amplitude and the expectancy metric known as *cloze probability* (the proportion of people to fill in the relevant gap in a sentence with a given word; Taylor, 1953, 1957). A higher-cloze continuation of a sentence elicits a smaller (i.e., more positive) N400 response, while a lower-cloze continuation elicits a larger (more negative) N400 (Kutas and Hillyard, 1984; Kutas and Federmeier, 2011). In fact, in previous work the two variables have been reported to have a Pearson correlation coefficient r of -0.9 or more (Kutas and Van Petten, 1994; Kutas and Federmeier, 2011). As the cloze task requires participants to predict an upcoming word, cloze probability has often been argued to reflect how predictable a word is in context (Tannenbaum et al., 1965; Fischler and Bloom, 1979; Kutas and Hillyard, 1984; Kutas et al., 2011; Van Petten and Luka, 2012; Luke and Christianson, 2016; Kuperberg et al., 2020; Brothers and Kuperberg, 2021). Moreover, the negative correlation between N400 amplitude and cloze probability tells us that N400 amplitude is not simply a categorical indicator of surprise, but reflects the predictability of the eliciting word in a more fine-grained way.

Beyond the graded predictability effect, the predicted preactivation account is supported by the way that N400 amplitude is modulated by sentence context. Research has shown that words elicit a large N400 when presented alone, a large N400 when presented

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

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

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

Aurnhammer and Frank, 2019b; Michaelov and Bergen, 2020; Ettinger, 2020; Merkx and Frank, 2021; Michaelov et al., 2022; Szewczyk and Federmeier, 2022).

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

8.1.2 Contextual Semantic Similarity

An alternative explanation of the neural activity underlying the N400 is contextual semantic similarity. Under this account, as we comprehend a sentence, the semantic features of each word are activated and briefly maintained, thereby reducing the neural activity required in response to words with overlapping features (Federmeier, 2021). While this feature-based account is compatible and indeed central to some prediction-based accounts of the N400 (e.g. Kuperberg et al., 2020), the key difference is that the activations are limited to semantic features of previously encountered words. That is, there is no additional

spreading activation to related words or semantic features, and, crucially, no prediction. Some investigators have suggested that contextual semantic similarity accounts for all variation in N400 amplitude (Ettinger et al., 2016; Uchida et al., 2021), while others suggest semantic similarity acts in concert with a prediction mechanism (see, e.g. Federmeier, 2021; Lau et al., 2013; Frank and Willems, 2017).

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

Outside the realm of negation and quantification, initial studies showed that the presence of a strongly related word within either a meaningful sentence (e.g., *When the* ***moon*** *is full, it is hard to see many* ***stars*** *or the Milky Way*) or a grammatically legal but meaningless word string (e.g., *When the* ***moon*** *is rusted, it is available to buy many* ***stars*** *or the Santa Ana*) leads to a smaller N400 to **stars** than if the prior context does not include a related word (Van Petten, 1993; Van Petten et al., 1997). However, other studies

indicate that N400 is not driven solely by an automatic semantic comparison process during sentence comprehension. Coulson and colleagues found much smaller N400s to the second words of related (*tin/**aluminum***) than unrelated (*tin/**disposal***) word pairs when the pairs were presented by themselves (Coulson et al., 2005). The word pairs were then embedded in sentences that were compatible or incompatible with the word-pair relationship, like the quartet below.

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

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

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

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

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

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

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

8.1.3 Multiple Systems Accounts

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

According to other accounts, both systems are constantly active but implemented in different brain circuits. In one fMRI experiment, (Frank and Willems, 2017) found

that contextual semantic similarity was correlated with activations in the anterior middle temporal sulcus, the precuneus, and bilateral angular gyri, whereas predictability was correlated with activations in the left inferior temporal sulcus, left posterior fusiform gyri, bilateral superior temporal gyri, and bilateral amygdalae. In view of the limited temporal resolution of fMRI, however, it is also possible that these findings reflect a disparate impact of contextual similarity and predictability at distinct stages of language processing.

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

8.1.4 The Present Study

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

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

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

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

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

We then use state-of-the-art language models to calculate the predictability and contextual similarity of our stimuli and investigate how well these metrics predict the single-trial N400 amplitudes elicited by the stimuli. To operationalize predictability, we used the transformer neural network language model, GPT-3. Research has shown that in general, larger language models trained on more data provide the best fits to human data, and that transformer neural networks are the architecture best suited to predicting N400 data

(Merkx and Frank, 2021). However, rather than using the conditional probabilities assigned by GPT-3 to our stimuli, we instead utilize *surprisal* scores, the negative logarithm of the probability assigned by the language model to a given word in context. Previous work has shown that when directly compared, language model surprisal is a better predictor of N400 amplitude than raw probability (Szewczyk and Federmeier, 2022; Yan and Jaeger, 2020).

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

We expect that our experimental manipulation of predictability, plausibility, and relatedness to the best completion will replicate each of these well-documented effects on the N400, as would be evidenced by an effect of experimental condition. In particular, we expect the Best completions to elicit the least negative (most positive) N400, the Im-

plausible completions to elicit the most negative N400, and the Related and Unrelated completions to fall in between the two. Despite the fact that the Related and Unrelated completions are matched for cloze probability and plausibility, the Related completions are expected to elicit smaller N400 than Unrelated completions.

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

8.2 Materials and Methods

8.2.1 Participants

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

8.2.2 Materials

Our stimuli were based on the original stimuli of the experiment carried out by (Thornhill and Van Petten, 2012). These stimuli were of the form given in Table 8.1. For

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

Sentences were normed via online surveys using the same participant pool we used to recruit participants for the EEG study. First, cloze probability measures were collected from UCSD students such that each sentence frame was completed by at least 35 participants. In this survey, participants were provided with sentence frames and instructed to produce the last word of the sentence. Average cloze probability and standard deviation for each condition are shown in Table 8.1.

All sentences were also rated for plausibility by a separate group of at least 30 students. In this survey, participants read one sentence at a time and were asked to rate each on a scale from 1 (very plausible) to 5 (very implausible). Multiple stimulus lists were employed so that each participant viewed only one of the four versions of each sentence frame. Average plausibility ratings for each experimental condition are shown in Table 8.1. All sentences in the Implausible condition had ratings above 3.5, with an average rating of 4.3. By contrast, the other conditions all had ratings below 2, suggesting participants found them plausible.

These stimuli were initially constructed as part of a larger study. In order to use the computational tools required to test our hypotheses, we opted to analyze a subset of the data such that critical words of all sentence stimuli appeared as whole tokens in GPT-3, GloVe, and fastText—that is, critical words were present as whole words in the

vocabularies of these models. We then further selected stimuli such that, as in (Thornhill and Van Petten, 2012), there was no overall difference in cloze probability between the related and unrelated completions. We also additionally ensured that there was no overall difference in plausibility. Thus, the two conditions differed only in how related they were to the Best Completion for that sentence. This resulted in a final stimulus set of 125 sentence frames in 4 conditions, for a total of 500 items. The stimuli were presented along with 165 other sentences that were part of the larger study and thus similar in character to the experimental sentences. As for the experimental sentences, these additional stimuli were equally likely to end with the Best, Related, Unrelated, or Implausible completion for the sentence frame as each participant saw approximately 41 non-experimental stimuli in each condition—in addition to the approximately 31 experimental sentences in each condition.

Table 8.1: Descriptive Statistics for Sentences: Mean and standard deviation of cloze probabilities and plausibility ratings (1 = very plausible; 5 = very implausible) for each experimental condition.

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

8.2.3 Procedure

Testing consisted of a single experiment session, with words presented centrally using RSVP presentation. For each sentence, participants first saw a break screen, then pressed a key to display the sentence. A fixation character remained on the screen while words were presented for 300ms, followed by a 200ms blank screen. The final word was displayed for 1200ms. After some sentences, participants saw a question about the content of the previous sentence (e.g., "Was the previous sentence about banking?") and responded

Yes or No with a button press.

8.2.4 EEG Recording and Analysis

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

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

N400 amplitude was operationalized as the mean voltage 300-500ms post-onset recorded from nine centro-parietal electrodes: C3, Cz, C4, CP3, CPz, CP4, P3, Pz, and P4. All graphs and statistical analyses were run in *R* (R Core Team, 2022) using *Rstudio* (RStudio Team, 2020) and the *tidyverse* (Wickham et al., 2019), *lme4* (Bates et al., 2015), *lmerTest* (Kuznetsova et al., 2017), *corrr* (Kuhn et al., 2022), *colorspace* (Zeileis et al., 2020, 2009), *gridExtra* (Auguie, 2017), and *cowplot* (Wilke, 2020) packages. All figures use

colorblind-friendly palettes (Jackson, 2016; Zeileis et al., 2020; Chang, 2022). All reported p-values are corrected for multiple comparisons based on false discovery rate (Benjamini and Yekutieli, 2001).

8.2.5 Computational Metrics

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

GPT-3 Surprisal

The OpenAI API (OpenAI, 2021) was used to access the predictions of the largest original GPT-3 model (*Davinci*), which has 175 billion parameters (Brown et al., 2020). Each sentence stimulus was input into the API and GPT-3 was used to calculate the probability of the final word given its preceding context. This figure was then used to

calculate the log-probability of each critical word. Since these log-probabilities used the natural exponent as a base, they were converted to the logarithm of base two and multiplied by negative one. The resultant surprisal values are thus measured in bits (see, e.g.,Futrell et al., 2019, for discussion).

GloVe Cosine Similarity

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

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

fastText Cosine Similarity

To calculate fastText Contextual Cosine Similarity, we utilized the fastText (Bojanowski et al., 2017) vectors made available through the fastText website (`https://fasttext.cc/`)—specifically, the version with a 2 million word vocabulary, 300-dimensional vectors, and no sub-word information trained on 600 billion tokens from the Common Crawl

corpus. As with the GloVe vectors, we calculated the cosine similarity between the vector corresponding to the stimulus word and the mean vector of the preceding context. In addition to calculating fastText Contextual Cosine Similarity, we also calculate fastText Best Completion Cosine Similarity or fastText BCCS.

8.3 Results

Figure 8.1 shows grand average ERP waveforms for words in each of the four conditions (Best Completion, Related, Unrelated, and Implausible) along with topographic maps. By convention, negative voltage is plotted upwards making it apparent that, as predicted, the Implausible condition elicited the largest (most negative) N400, and the Best Completions elicited the smallest (most positive) N400. The Unrelated condition fell in between these two extremes, and, as predicted, elicited more negative ERPs than did the Related condition (which was virtually overlapping the Best Completion condition, despite the large difference in their average cloze probability). The topographic maps were formed by first calculating point-by-point difference waves obtained by subtracting the amplitude of ERPs recorded at each electrode in the Best Completion condition from their counterparts in the Related, Unrelated, and Implausible conditions, respectively. The mean amplitude 300-500ms was then measured on each difference wave and plotted on the scalp to visualize the relative pattern of positive and negative voltage. The posterior negativity apparent in all three plots is characteristic of N400 ERP effects reported in sentence reading paradigms like the one used here.

Figure 8.2 presents normalized (z-scored; and in the case of surprisal and plausibility, multiplied by -1) values in each experimental condition for the outcome variable (N400) and for each of our predictors. Note that the human derived metrics of cloze probability and plausibility reflect our experimental design. The Best Completions were intended

to be predictable, while the Related, Unrelated, and Implausible conditions were designed to be unexpected, with Related and Unrelated conditions equated for cloze probability. Similarly, Best Completions, Related, and Unrelated conditions were all intended to be plausible, whereas the Implausible condition was intended to be implausible. Figure 1 indicates that all of the computational metrics were associated with differences between Best Completions and Implausible endings. Related and Unrelated conditions were quite similar on some metrics—such as GloVe Contextual Cosine Similarity (CCS) and fastText CCS—and differed on others, such as GPT-3 surprisal and both measures of Best Completion Cosine Similarity (BCCS).

Figure 8.3 presents a heatmap of correlations between the various predictors used in the regression analyses below. Recall that Contextual Cosine Similartity (CCS) is the cosine of the angle between the vector for each word and the mean of the vectors for each of the words in the preceding sentence context and serves as an operationalization of contextual semantic similarity. Best Completion Cosine Similarity (BCCS) is the cosine of the angle between the vector for each word and the vector for the word that is the best completion for the sentence frame and is relevant for some multiple systems accounts. Although the two kinds of embeddings (GloVe and fastText) yielded virtually identical estimations of similarity between pairs of words—as reflected in the 0.98 correlation between GloVe BCCS and fastText BCCS—they differed somewhat in their estimates of contextual semantic similarity (CCS) as GloVe CCS and fastText CCS had a correlation coefficient of 0.66. Relative to GloVe CCS, fastText CCS was more associated with cloze probability (0.39 versus 0.32), GPT-3 surprisal (-0.61 versus -0.46), and plausibility (-0.56 versus -0.37). Relative to GloVe CCS, the fastText CCS measure also showed more sensitivity to the semantic relationship between each unexpected ending and the best completion, as evidenced by a greater correlation with fastText BCCS (0.52 versus 0.33) and even with

GloVe BCCS (0.54 versus the 0.4 correlation between GloVe CCS and Glove BCCS).

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

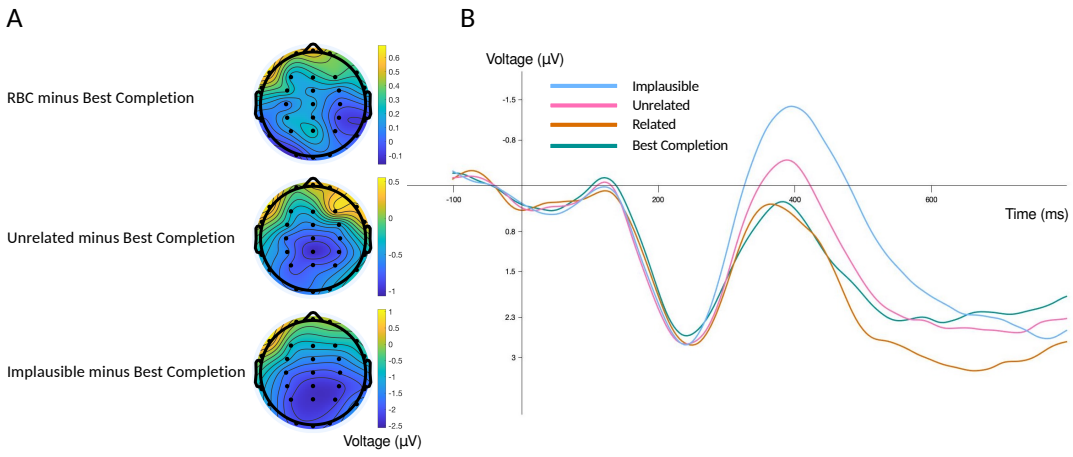

Figure 8.1: ERP scalp maps and waveforms. Panel A shows the topography of the mean amplitude 300-500ms of the difference wave for the RBC and Best Completion conditions (top), Unrelated and Best Completion (middle), and Implausible and Best Completion (bottom) using a spherical spline interpolation. Panel B shows the ERP waveforms for each condition (Best Completion, Related, Unrelated, Implausible) as measured at the Centroparietal Electrode Cluster used in the regression models.

Figure 8.2: Average values of all predictors under each experimental condition. For easier comparison across predictors, we plot negative surprisal and plausibility, and the values of all predictors were z-scored. For easier comparison to the N400 waveform, the y-axis is reversed, with negative values plotted upwards. Error bars show the standard error.

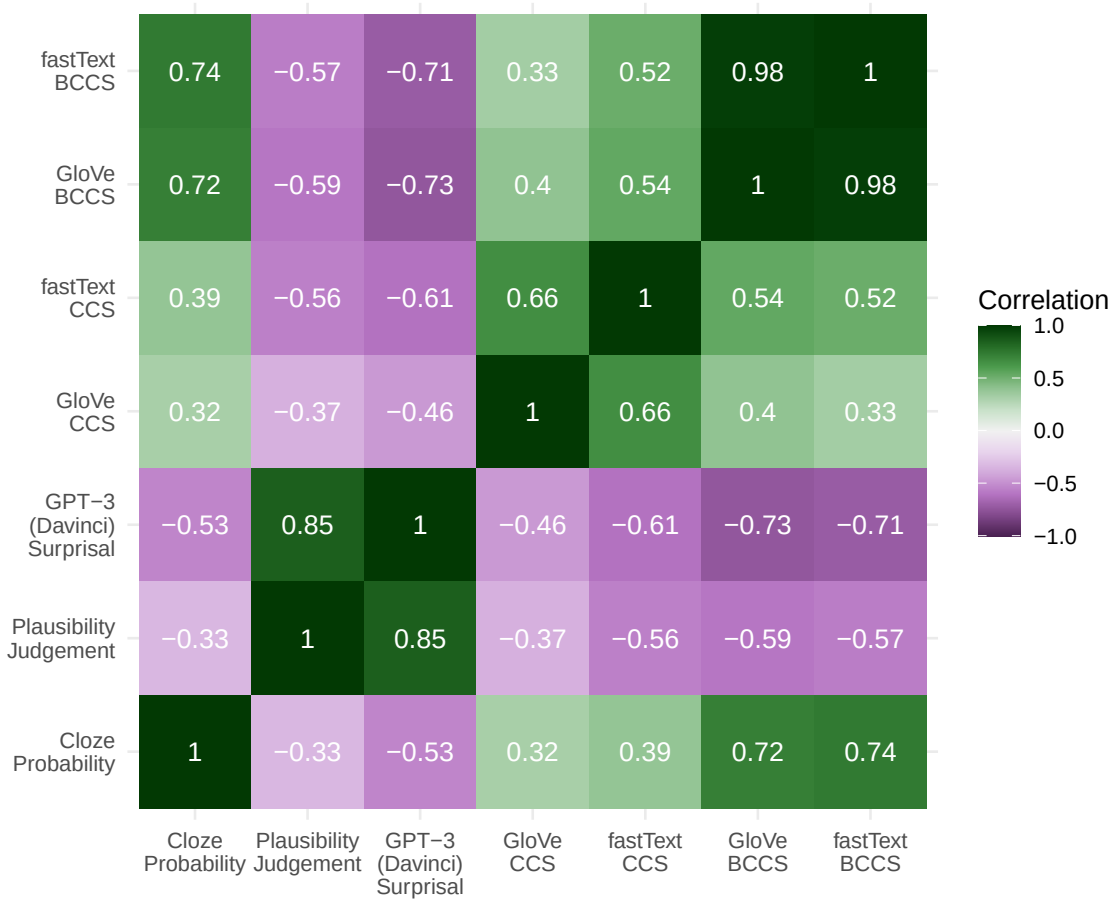

Figure 8.3: Heatmap of correlations between predictors

8.3.1 Single Factor Accounts

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

In order to compare these predictors, we constructed linear mixed-effects regres-

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

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

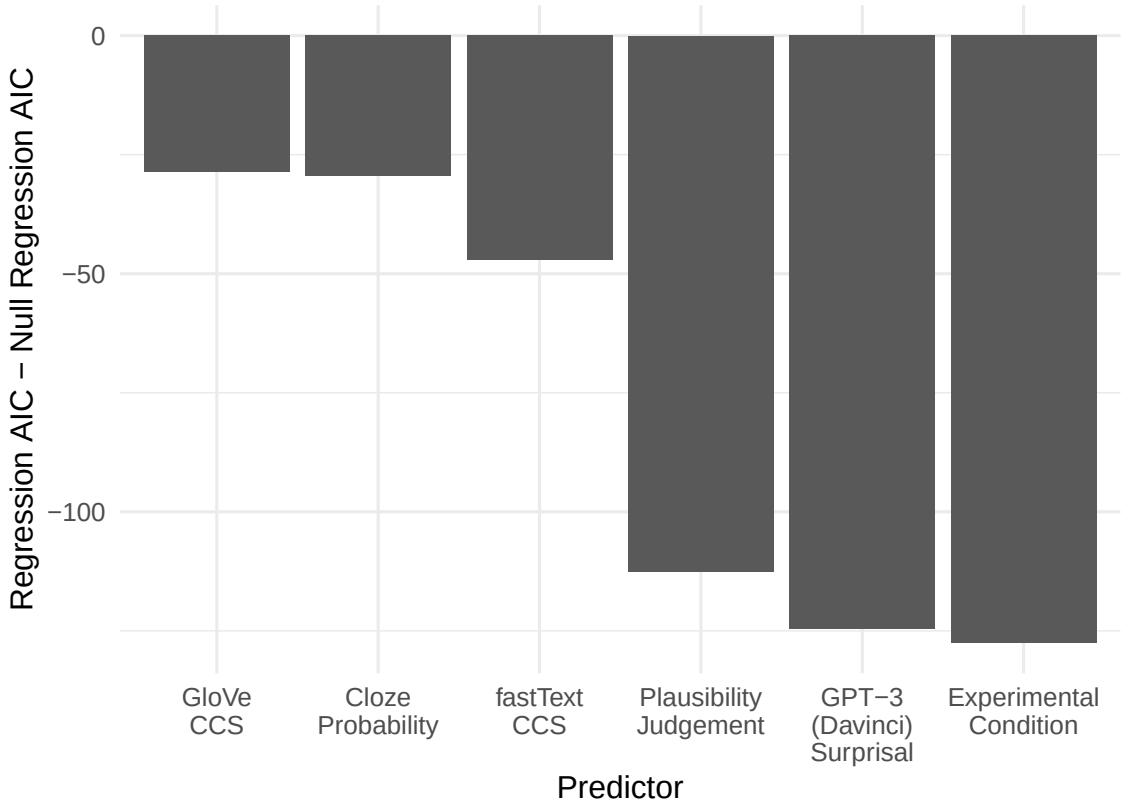

Figure 8.4: The AICs of the regressions resulting from the single factor analyses. CCS refers to Contextual Cosine Similarity.

The AIC of each regression, normalized by the AIC of the null regression (which includes the same random effects structure as the other regressions, and only word frequency and orthographic neighborhood size as fixed effects) is presented in Figure 8.4.

Of the continuous predictors, Figure 8.4 indicates that the best-fitting regression is that including GPT-3 surprisal as a main effect, suggesting GPT-3 surprisal is the best predictor of N400 amplitude. GPT-3 surprisal is followed by human plausibility judgements, which are followed by fastText CCS, which in turn is followed by cloze probability and GloVe CCS. It is generally accepted that a difference in AIC of 4 indicates a substantial

difference (Burnham and Anderson, 2004), and the difference between cloze probability and GloVe CCS is only 0.9; thus it is not clear from our analysis which is the better predictor.

Figure 8.4 also indicates that the regression including experimental condition (a categorical variable with four levels: Best Completion, Related, Unrelated, and Implausible) has a lower AIC than the GPT-3 surprisal regression. However, experimental condition should not be considered to reflect a single variable in the way that the other individual predictors do because it includes information about predictability, plausibility, and relatedness to the best completion. Additionally, the experimental condition regression has an AIC of only 3 less than the GPT-3 surprisal regression; thus it is not clear that experimental condition is in fact a better predictor than GPT-3 surprisal.

We also ran likelihood ratio tests on each of the predictors listed in Figure 8.4, comparing each regression to a null regression, i.e., one without the predictor of interest but all other fixed and random effects. All variables were significant predictors of N400 amplitude (GloVe CCS: $\chi^2(1) = 30.6, p < 0.001$; Cloze: $\chi^2(1) = 31.6, p < 0.001$; fastText CCS: $\chi^2(1) = 49.1, p < 0.001$; Plausibility: $\chi^2(1) = 114.5, p < 0.001$; GPT-3 Surprisal: $\chi^2(1) = 126.6, p < 0.001$; Condition: $\chi^2(3) = 133.6, p < 0.001$).

8.3.2 Combined Accounts

The GPT-3 surprisal metric was chosen to model a prediction-based account of the N400, and GloVe and fastText contextual cosine similarity (CCS) were chosen to model the contextual semantic similarity accounts. As noted above, some authors have suggested the N400 indexes neurocognitive systems sensitive both to the predictability of a word and to its similarity to the semantic context. To investigate the viability of such combined accounts, we compare the AICs of regressions including a single variable corresponding to either prediction or contextual semantic similarity, with the AICs of regressions also including one

of the other. Thus, we look at all combinations of prediction (viz., Cloze Probability and GPT-3 Surprisal) with CCS metrics. The results are presented in Figure 8.5. A comparison of the AICs suggests that cloze probability and the two CCS metrics explain variance in N400 amplitude not explained by the other. This is borne out by the likelihood ratio tests: after correcting for multiple comparisons the cloze probability regression is improved by adding either GloVe ($\chi^2(1) = 21.0, p < 0.001$) or fastText CCS ($\chi^2(1) = 31.6, p < 0.001$) as a predictor; and conversely, the GloVe ($\chi^2(1) = 22.0, p < 0.001$) and fastText ($\chi^2(1) = 14.0, p < 0.001$) regressions are each improved by adding cloze probability as a predictor. This suggests cloze probability and the CCS metrics explain non-overlapping portions of the variance in N400 amplitude. However, the same is not true of GPT-3 surprisal—while adding GPT-3 surprisal improves both the GloVe ($\chi^2(1) = 96.3, p < 0.001$) and fastText ($\chi^2(1) = 77.8, p < 0.001$) CCS regressions, the GPT-3 surprisal regression is not improved by adding either GloVe ($\chi^2(1) = 0.4, p = 1.000$) or fastText CCS ($\chi^2(1) = 0.4, p = 1.000$). Thus GPT-3 explains variance left unexplained by the CCS measures, while the information provided by CCS was largely redundant with that provided by GPT-3.

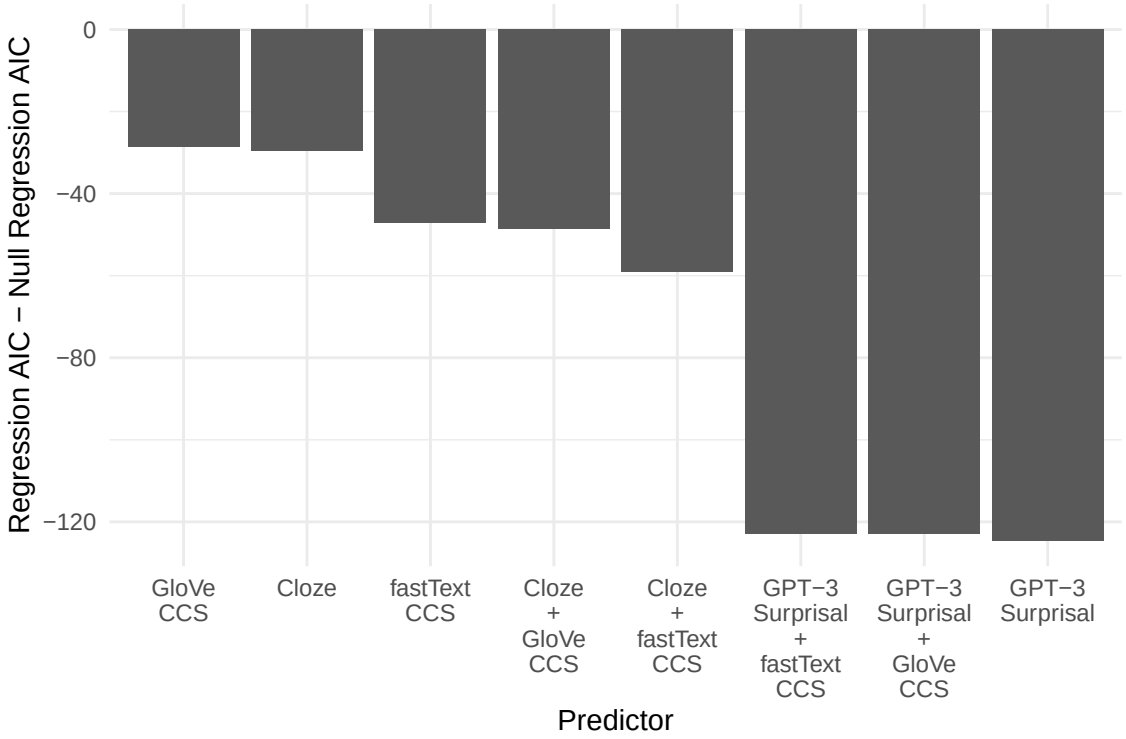

Figure 8.5: The AICs of the regressions resulting from the two-variable analyses corresponding to combined accounts. CCS refers to Contextual Cosine Similarity.

8.3.3 The plausibility effect

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

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

Shown in Figure 8.6, the results indicate that even when combined with cloze probability (and thus, when part of a combined model that takes into account predictability as well as contextual similarity), the AICs of the regressions including GloVe ($\chi^2(1) = 70.3, p < 0.001$) and fastText ($\chi^2(1) = 60.0, p < 0.001$) CCS are improved by the addition of plausibility as a predictor. By contrast, the GPT-3 surprisal regression is not improved by adding plausibility as a predictor ($\chi^2(1) = 1.9, p = 0.715$). Whereas neither CCS metric can model the N400 plausibility effect—even when combined with cloze—variance attributable to plausibility was captured by GPT-3 surprisal. Thus, predictability alone (operationalized by GPT-3 surprisal) can explain the apparent effect of plausibility on N400 amplitude.

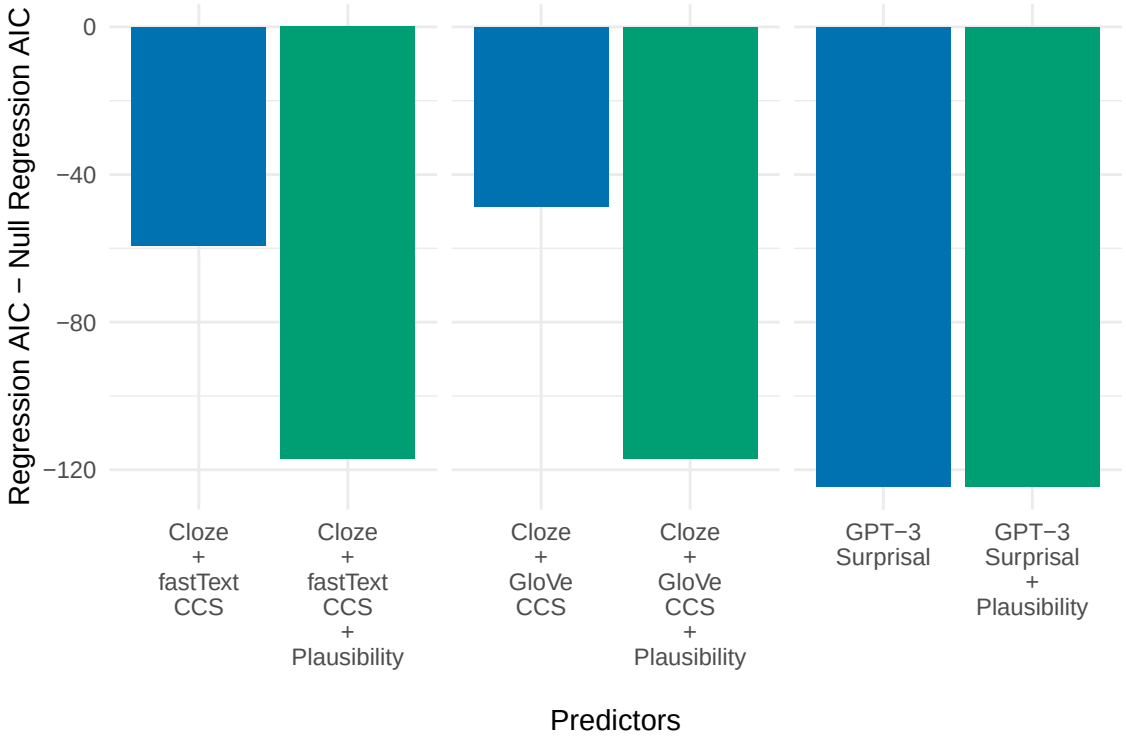

Figure 8.6: The AICs of the regressions resulting from the analyses investigating whether the single-factor and combined models account for the effect of plausiblity. CCS refers to Contextual Cosine Similarity.

8.3.4 The relatedness to the best completion effect

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

As with plausibility, we investigate whether our previous best regressions for each of our three statistical metrics—that is, GPT-3 surprisal, cloze + GloVe CCS, and cloze + fastText CCS—are improved by the addition of BCCS to the model. The results are shown in Figure 8.7. The addition of GloVe BCCS to either cloze + CCS regressions led to improvements in model performance (Cloze + GloVe CCS: $\chi^2(1) = 32.4, p < 0.001$; Cloze + fastText CCS: $\chi^2(1) = 24.8, p < 0.001$); likewise the addition of fastText BCCS to either cloze + CCS regression led to significant improvements (Cloze + GloVe CCS: $\chi^2(1) = 31.0, p < 0.001$; Cloze + fastText CCS: $\chi^2(1) = 23.6, p < 0.001$). These results show that even when combined with cloze, contextual similarity cannot explain the relatedness to best completion effect.

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

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

fastText BCCS ($\chi^2(1) = 5.7, p = 0.087$; uncorrected $p = 0.017$) significantly improved the regression already including GPT-3 surprisal.

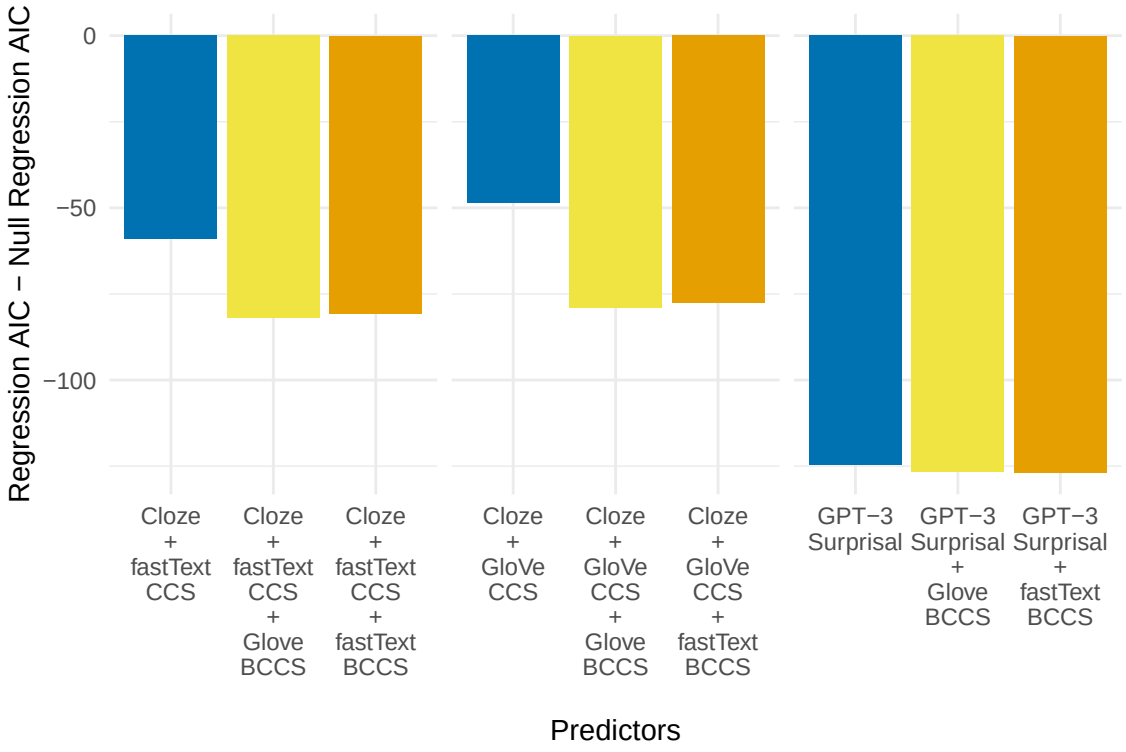

Figure 8.7: The AICs of the regressions resulting from the analyses investigating whether the single-factor and combined models account for the effect of the relatedness to the best completion. CCS refers to Contextual Cosine Similarity and BCCS refers to Best Completion Cosine Similarity.

8.4 Discussion

The aim of this paper was to use current state-of-the-art language models to compare the predictions of two accounts of the neural activation underlying the N400 response—predictive preactivation versus contextual semantic similarity. To do this, we investigated how well GPT-3 surprisal—our best approximation of the kinds of predictions

neurocognitive systems may make based on the statistics of language—predicts N400 amplitude. We compared this with the performance of GloVe and fastText contextual cosine similarity, our two best approximations of contextual semantic similarity based on the statistics of language. Finally, we compared this with the performance of combined models including both kinds of metrics. Based on this approach, we found that the predictive preactivation account explains more variance in N400 amplitude than the two models of contextual semantic similarity.

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

8.4.1 Expectancy Effects

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

(Yamada et al., 2020), (Uchida et al., 2021) show that high cloze sentence continuations from a number of ERP language studies are more similar to their contexts than their less predictable counterparts.

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

Of course, this same finding—that regressions with CCS measures are improved by the cloze probability factor—replicates work that supports the multiple systems account of the N400 (Lau et al., 2013; Federmeier, 2021). However, GPT-3 surprisal out-performed even these regressions (see Figure 8.5), suggesting that the predictive preactivation account of N400 is superior to both a pure contextual semantic similarity account and to a combined systems account.

8.4.2 Plausibility Effects

GPT-3 surprisal also accounts for more variance in N400 amplitude than our human-derived measure of cloze probability (in line with Michaelov et al., 2022), presumably due to its ability to capture subtle differences between highly unexpected items. Indeed, as (Nieuwland et al., 2020) note, plausibility effects on the N400 might result because less plausible stimuli are also less predictable. Because cloze probability measures

are limited in the extent to which they can adequately capture the predictability of highly improbable words, plausibility ratings may serve as a proxy for their predictability, allowing us to differentiate *very* low-probability completions from *extremely* low-probability ones. Of course, plausibility effects can also be accounted for in principle via contextual semantic similarity, since we would expect less plausible stimuli to be less related to their context.

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

8.4.3 Relatedness to Best Completion

As described in the Introduction, the RBC effect is not trivially explained by either predictability or contextual similarity; however, in principle it can be accommodated by either account, and there is some evidence for each. Under a predictability perspective, if semantic prediction is taking place, then we should expect words with a similar meaning to the best completion to be preactivated along with the best completion (DeLong et al.,

2019). Consistent with this account, the predictions of computational language models have been used to successfully model the RBC effect (Michaelov and Bergen, 2020). Specifically, (Michaelov and Bergen, 2020) report that two language models (Gulordava et al., 2018; Jozefowicz et al., 2016) find related words to be more predictable than unrelated overall when modeling the stimuli from an experiment by (Ito et al., 2016), and that one of these language models also shows the same pattern for stimuli from (Kutas, 1993).

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

The present study provides a conceptual replication of results reported both by (Michaelov and Bergen, 2020) and by (Ettinger et al., 2016). Using GPT-3 surprisal we find that our Related completions were more predictable than the Unrelated ones (in line with Michaelov and Bergen, 2020); using fastText CCS we find that Related completions were more similar to the preceding context than were the Unrelated ones (in line with Ettinger et al., 2016). However, results in Figure 8.2—like those in both (Michaelov and Bergen, 2020) and (Ettinger et al., 2016)—only demonstrate that overall, there is a significant difference in the predictability and in the contextual semantic similarity of Related and Unrelated completions as estimated by these computational language models; there is no direct comparison with human data.

The strength of the present study lies in our efforts to do just this. Direct comparison with the human N400 data suggests that the predictability metric from GPT-3 explains more variance in N400 amplitude than does either metric of semantic similarity to the context. Moreover, in our attempts to probe how well each metric captures the RBC effect, we utilized two computational measures of the semantic similarity between each best completion and the other three completions for the sentence frame: GloVe BCCS and fastText BCCS. As both the graphs in Figure 8.2 and the high correlation coefficient in Figure 8.3 suggest, the two BCCS measures were virtually identical with each other and both captured the human intuition that Related words were closer in meaning to the Best Completions than Unrelated words.

Regression models of N400 data indicate that the addition of either GloVe or fastText BCCS metrics to models already including cloze probability and GloVe or fastText CCS improves model fit (see Figure 8.7). This suggests that neither of our contextual semantic similarity metrics could fully account for the RBC effect—even when combined with cloze probability. On the other hand, the GPT-3 surprisal regression of N400 data was not substantially improved by the addition of either BCCS metric (see Figure 8.7), suggesting the variance associated with our measure of similarity to the best completion was largely redundant with that captured by GPT-3 surprisal. Moreover, the regression model including only GPT-3 surprisal out-performed all of the regression models with additive combinations of CCS, cloze probability, and BCCS. GPT-3 surprisal provides a better account of the RBC effect than does either a pure contextual semantic similarity account or a combination of prediction and contextual similarity.

While the superiority of GPT-3 over the contextual similarity measures is unambiguous, there is a bit of uncertainty regarding whether GPT-3 is improved by the addition of the BCCS metrics. In our statistical model comparisons, we do not consider regressions

with a difference in AIC of less than 4 to differ meaningfully in their fit (following Burnham and Anderson, 2004). However, it is the case that numerically, the regressions including both GPT-3 surprisal and either GloVe or fastText BCCS have a lower AIC than that *only* including surprisal. Unfortunately, the outcome of the relevant likelihood ratio tests was also somewhat equivocal on this matter. After correcting for multiple comparisons, neither GloVe nor fastText BCCS explain a significant amount of the variance in N400 amplitude above and beyond what is explained by GPT-3 surprisal. Before correction, however, those comparisons were both significant at the 0.05 level. It is thus important to consider what might explain this (marginally) better fit to the data.

One straightforward explanation can be arrived at by further inspection of Figure 8.2. As can be seen, GPT-3 surprisal provides a good account of the difference in the expectancy between Best Completions and the Unrelated condition, and a good account of the difference between the Unrelated and the Implausible condition—impressions borne out by the analyses comparing surprisal to human-derived metrics of cloze probability and plausibility. The disconnect between GPT-3 surprisal and N400 data lies mainly in failing to fully capture the similarity in N400 amplitude between the Best Completions and the Related condition, as the latter elicit more positive N400 in humans than the GPT-3 regression model fits suggest. Thus, the addition of another variable that captures the difference between Related and Unrelated completions—variance not present in cloze probability or plausibility, and unreliable in the CCS metrics—may explain the improved fit with the addition of BCCS metrics. This may also explain the slightly lower AIC of the regression including the categorical variable of experimental condition in Figure 8.4.

Crucially, however, even if GPT-3 does not fully account for the RBC effect, the RBC effect observed here supports predictive preactivation as at least a partial account of the brain activity underlying the N400. If words semantically related to the best com-

pletion are facilitated in virtue of being related to the best completion, this presupposes the preactivation of information related to the best completion (DeLong and Kutas, 2020 see also Kuperberg et al., 2020). For example, it may be the case that the reason for the greater facilitation for related than unrelated words is that predictive processing involves the preactivation of conceptual semantic features rather than lexical items (Thornhill and Van Petten, 2012). Alternatively, it may be that there is a separate associative mechanism that activates words related to the best completion. In the first case, the preactivation of the related word occurs as part of a single predictive process; in the second, as a consequence. Both possibilities require the preactivation of the best completion—either the lexical item itself or its semantic features. Regardless, the present study clearly shows that, as operationalized here, predictive preactivation provides a better account of the RBC N400 effect than does contextual semantic similarity (see Figure 8.7).

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

8.4.4 Implications for Neural Mechanisms

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

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

As (Kuperberg et al., 2020) note, this account does not fit neatly into either re-

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

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

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

The results of using a language model to model the study carried out by (Kim and Osterhout, 2005) may also be valuable in better understanding the content of the preactivation underlying the N400 response. For example, a number of accounts argue that the

preactivation underlying the N400 response is at the level of the semantic features of words (Federmeier, 2021; Kuperberg et al., 2020). While there is evidence that N400 amplitude is sensitive to phonological and grammatical features (DeLong et al., 2005; Van Berkum et al., 2005; Otten et al., 2007; Nicenboim et al., 2020; Urbach et al., 2020; Fleur et al., 2020), it may be that the shared semantic features between, for example, *devouring* and *devoured*, are sufficient to preactivate both words equally. Thus a semantically-augmented language model may be able to better model the effect.

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

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

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

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

8.5 Data and Code Availability Statements

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

8.6 Acknowledgements

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

Chapter 8, in full, is a reprint of the material as it appears in Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S., "Strong Prediction: Language model surprisal explains multiple N400 effects", *Neurobiology of Language*, 2024. The dissertation author was the primary investigator and author of this paper. Minor edits have been made to bring the formatting in line with the dissertation template.

Part III

The mathematical relationship between contextual probability and the N400

Chapter 9

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

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. This approach for the first time allows for the computation of metrics that mathematically model a variety of alternate theories of the preactivation of the stimulus word itself as well as all other words. 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.

9.1 Introduction

Perhaps the best studied neural signal of language comprehension, the N400 is a negative component of the event-related brain potential peaking roughly 400ms after the presentation of a stimulus (Kutas and Hillyard, 1980, 1984; Kutas and Federmeier, 2011). 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 (Kutas and Hillyard, 1984; Kutas et al., 2011; Kutas and Federmeier, 2011; Van Petten and Luka, 2012; Federmeier, 2021; Kuperberg et al., 2020). 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 (Van Petten and Kutas, 1990, 1991; Van Petten, 1993; Dambacher et al., 2006; Van Petten and Luka, 2012; Payne et al., 2015; Federmeier, 2021). The predictability effect has been replicated numerous times when predictability is operationalized as cloze probability (Kutas and Hillyard, 1984; Kutas and Federmeier, 2011), the propor-

tion 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* (Frank et al., 2015; Aurnhammer and Frank, 2019b; Yan and Jaeger, 2020; Merkx and Frank, 2021; Szewczyk and Federmeier, 2022; Michaelov et al., 2022, 2023), computational systems designed to predict the probability of a word in context based on the statistics of language (Jurafsky and Martin, 2021).

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 (Federmeier et al., 2007; Kutas et al., 2011; Van Petten and Luka, 2012; Thornhill and Van Petten, 2012; DeLong et al., 2014b; DeLong and Kutas, 2020; Kuperberg et al., 2020; Federmeier, 2021). 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 key 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., 2007, 2002; Otten and Berkum, 2008; Van Petten et al., 1999; Wlotko and Federmeier, 2007; Vissers et al., 2006; Federmeier, 2007; for review see Van Petten and Luka,

2012; Kuperberg et al., 2020; Federmeier, 2021). 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 target word, 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 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 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 (Hale, 2001; Hoeks et al., 2004; Debruille, 2007; Fitz and Chang, 2019). 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. Metusalem et al., 2012; Ito et al., 2016; DeLong et al., 2019). 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 and Martin, 2021).

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 (Frank et al., 2015; Aurnhammer and Frank, 2019b; Yan and Jaeger, 2020; Merkx and Frank, 2021; Szewczyk and Federmeier, 2022; Michaelov et al., 2022, 2023). To the best of our knowledge, no study has thus far attempted to directly test whether N400 amplitude can be predicted by probability assigned to any word other than the stimulus itself by a language model, let alone 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 all words other than the stimulus itself were predicted has not been directly investigated. 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. Kuperberg et al., 2020; Federmeier, 2021), 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. Metusalem et al., 2012; Ito et al., 2016; DeLong et al., 2019). This may not just be a methodological issue; as discussed in subsection 9.2.1, 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 out-perform 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.

9.2 Past Approaches

9.2.1 Constraint

Since early work on the N400 (Kutas and 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 inhibition-inclusive 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. However, as discussed, Federmeier et al. (2007) did not find any effect of constraint, leading them, and many other researchers (Federmeier et al., 2002; Otten and Berkum, 2008; Van Petten et al., 1999; Wlotko and Federmeier, 2007; Vissers et al., 2006; Federmeier, 2007; Kutas et al., 2011; Van Petten and Luka, 2012; Thornhill and Van Petten, 2012; Kuperberg et al., 2020; DeLong and Kutas, 2020; Federmeier, 2021) 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 Staub et al., 2015; Brothers and Kuperberg, 2021). 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 Kutas and Hillyard, 1984; Federmeier and Kutas, 1999; Metusalem et al., 2012; Amsel et al., 2015; Ito et al., 2016; DeLong et al., 2019). 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 and 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 Smith and Levy, 2011; Michaelov et al., 2022).

9.2.2 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 and 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, as shown in Equation 9.1.

$$S = -\log(p(w_i) \tag{9.1}$$

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 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 (Parviz et al., 2011; Frank et al., 2015; Aurnhammer and Frank, 2019b; Michaelov and Bergen, 2020; Merkx and Frank, 2021; Szewczyk and 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.

9.2.3 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., MacDonald, 2013; Pickering and Garrod, 2013; Kuperberg et al., 2020; Fitz and Chang, 2019; Federmeier, 2021). 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 9.9.1. Specifically, L^1 distance is has relationship to $p(w_i)$ shown in Equation 9.2.

$$L^1 = 2 - 2p(w_i) \tag{9.2}$$

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.

9.2.4 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 is given in Equation 9.3, where $\hat{p}(w_i)$ is the predicted probability of w_i in context.

$$-\sum_i \hat{p}(w_i) \log \hat{p}(w_i) \tag{9.3}$$

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 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 (Huizeling et al., 2022; Brodbeck 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), and thus is in line with some theories of the N400 (e.g., Fitz and 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. Aurnhammer and Frank (2019b) test how well 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. The two Lookahead Information Gain metrics are also both based on this probability distribution for the following word, and thus are also not relevant to the present study. It should also be noted that based

on the results, Aurnhammer and Frank (2019b) argue that none of these three are good metrics for modeling word reading effort.

9.3 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 (Frank et al., 2015; Aurnhammer and Frank, 2019b; Michaelov and Bergen, 2020; Merkx and Frank, 2021; Michaelov et al., 2021; Szewczyk and Federmeier, 2022; Michaelov and Bergen, 2022a; Michaelov et al., 2023).

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 et al., 2013b), 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., Wang et al., 2019b,a; Nie et al., 2020; Srivastava et al., 2022), 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 (Rao and Ballard, 1999; de Lange et al., 2018; Sherman and 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. Saffran et al., 1996; de Marneffe et al., 2012; Sherman et al., 2020; Pickering and Garrod, 2007, 2013; MacDonald, 2013; Ambridge et al., 2014; Elman, 2009; Newport and Aslin, 2004; Romberg and Saffran, 2010; Seidenberg, 1997; Gómez and Gerken, 2000; Gerken, 2006, 2007). 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 (Kutas and Federmeier, 2011; Van Petten and Kutas, 1990; Van Petten, 1993; Dambacher et al., 2006; Rugg, 1990; Fischer-Baum et al., 2014) 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.

9.4 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 9.2). 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; Wlotko and Federmeier, 2012; Hubbard et al., 2019; Szewczyk et al., 2022) 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 (Yan and Jaeger, 2020; Szewczyk and Federmeier, 2022), especially for the best-performing models (Michaelov and Bergen, 2022b). However, Szewczyk and Federmeier (2022), analyzing the same dataset that we analyze, found that un-transformed 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 and Komatsuzaki, 2021). First, we test whether any of the 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.

9.5 Experiment 1

9.5.1 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 and 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 and 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 data on the same order of magnitude: Pythia 6.9B (Biderman et al., 2023b), 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.

9.5.2 Method

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; Wlotko and Federmeier, 2012; Hubbard et al., 2019; Szewczyk et al., 2022). 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 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 300-500ms time window. These N400 amplitudes are not baseline-corrected; instead, a baseline—the mean amplitude in the 100ms before the presentation of the stimulus—is included as a variable, and in the original analysis is included as a covariate (Szewczyk and 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 2x2 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 and Federmeier, 2007) were selected to be plausible and vary 'continuously through the full range of cloze probability' (Wlotko and 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 2x2 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 and 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).

Models

The details of the seven models tested are provided in Table 9.1. 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

Table 9.1: 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	0.24B	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	0.36B	33B
StableLM-Base-Alpha 7B	7.9B	800B

transformers (Wolf et al., 2020) *Python* (Van Rossum and Drake, 2009) package.

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; Wlotko and Federmeier, 2012; Hubbard et al., 2019; Szewczyk et al., 2022; Szewczyk and Federmeier, 2022). 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`.

9.5.3 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 Figure 9.1, 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 difference in AIC of 4 or more is generally considered to indicate that the lower-AIC regression

has 'considerably' more evidential support (Burnham and Anderson, 2004). Thus, the regression including GPT-J surprisal and probability is clearly the best-performing regression.

9.5.4 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 and 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.

9.6 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 and 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 and 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 and 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 (Dumais et al., 1988; Deerwester et al., 1990; Landauer et al., 1998; Chwilla and Kolk, 2005; Parviz et al., 2011; Mikolov et al., 2013b,a; Van Petten, 2014; Ettinger et al., 2016): 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.

9.6.1 Method

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.

Metrics

All metrics used in this analysis are defined in Table 9.2. The correlations between all metrics is shown in Figure 9.2.

Table 9.2: 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	$\sum_i(\frac{(p(w_i)-\hat{p}(w_i))^2}{\hat{p}(w_i)})$	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 (2021)
Entropy	$-\sum_i \hat{p}(w_i)\log\hat{p}(w_i)$	Jurafsky and Martin (2021)
Constraint (p)	$\hat{p}(w_{BC})$	-
Constraint (S)	$-\log(\hat{p}(w_{BC}))$	-

9.6.2 Results

First, we compared how well each of the metrics performs compared to surprisal, probability, and an overall 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 Figure 9.3, 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 Figure 9.4. 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 and 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 ($\chi^2(1) = 3.0969, p = 0.0784$), and neither does χ^2 distance ($\chi^2(1) = 1.8036, p = 0.1793$), entropy ($\chi^2(1) = 0.5557, p = 0.4560$), $L^{0.5}$ distance ($\chi^2(1) = 0.4025, p = 0.5258$), Hellinger distance ($\chi^2(1) = 0.1774, p = 0.6737$), either constraint metric (probability: $\chi^2(1) = 0.0113, p = 0.9153$; surprisal: $\chi^2(1) = 0.0145, p = 0.9042$), or L^2 distance ($\chi^2(1) = 0.0072, p = 0.9324$). Thus, no distribution-dependent metric explains any variance in N400 amplitude above and beyond that explained by predictability.

9.6.3 Discussion

Our results replicate and extend several findings. First, as in previous work (Frank et al., 2015; Aurnhammer and Frank, 2019b; Szewczyk and Federmeier, 2022; Michaelov and Bergen, 2022b), 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 and 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 et al., 2007, 2002; Otten and Berkum, 2008; Van Petten et al., 1999; Wlotko and Federmeier, 2007; Vissers et al., 2006; 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.

9.7 General Discussion

It has long been widely believed (with a few exceptions, e.g., Hoeks et al., 2004; Debruille, 2007; Fitz and Chang, 2019) 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. Kutas et al., 2011; Van Petten and Luka, 2012; Brouwer et al., 2012; Brouwer and Hoeks, 2013; DeLong et al., 2014b; Kuperberg and Jaeger, 2016; Delogu et al., 2019; Kuperberg et al., 2020; Federmeier, 2021). But, as discussed in section 9.1, this never been fully tested—previous work has looked at constraint (Federmeier et al., 2007, 2002; Otten and Berkum, 2008; Van Petten et al., 1999; Wlotko and Federmeier, 2007; Vissers et al., 2006; 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.

9.7.1 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 metric 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 (Kuperberg et al., 2003; Kim and Osterhout, 2005; Nieuwland and Van Berkum, 2005; Kuperberg, 2007; Paczynski and Kuperberg, 2011, 2012; Szewczyk and Schriefers, 2011, 2013; Nieuwland et al., 2013; Wang et al., 2020; Vega-Mendoza et al., 2021).

9.7.2 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 9.3, 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 (Lewis and Bastiaansen, 2015; Bornkessel-Schlesewsky and Schlesewsky, 2019; Kuperberg et al., 2020). 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 (Rao and Ballard, 1999; Huang and Rao, 2011; Clark, 2013). 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 and 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 and Schlesewsky, 2019; Kuperberg et al., 2020; Hodapp and Rabovsky, 2021).

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 and Martin, 2019). 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.

9.7.3 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 (Kutas and Federmeier, 2011; Kutas et al., 2011; Van Petten and Luka, 2012; DeLong et al., 2014b; Kuperberg and Jaeger, 2016; Kuperberg et al., 2020; DeLong and Kutas, 2020; Federmeier, 2021). So what happens to words that are preactivated but not encountered? One possibility is that the metabolic resources required for preactivation (see, e.g. Levy, 2008; Brothers and Kuperberg, 2021) 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 and Federmeier, 2018).

The other mechanistic account 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.

9.8 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.

9.9 Appendix

9.9.1 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 $w_{\neg i}$):

$$L^1 = E(w_i) + \sum E(w_{\neg i}) \tag{9.4}$$

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}:

$$E(w_i) = 1 - p(w_i) \tag{9.5}$$

For all other words, the absolute error is a negative prediction error, the predicted proba-

bility of the false word $p(w_{\neg i})$ minus the true probability, 0:

$$E(w_{\neg i}) = p(w_{\neg i}) - 0 \tag{9.6}$$

This simplifies to:

$$E(w_{\neg i}) = p(w_{\neg i}) \tag{9.7}$$

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

$$p(w_i) + \sum p(w_{\neg i}) = 1 \tag{9.8}$$

This means that the following is also the case:

$$\sum p(w_{\neg i}) = 1 - p(w_i) \tag{9.9}$$

We can substitute Equation 9.5 and Equation 9.9 into the equation for total Manhattan distance Equation 9.4, getting:

$$L^1 = (1 - p(w_i)) + (1 - p(w_i)) \tag{9.10}$$

Which can be simplified to:

$$L^1 = 2 - 2p(w_i) \tag{9.11}$$

9.10 Acknowledgements

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

Chapter 9, in full, is a reprint of the material as it appears in Michaelov, J. A. &

Bergen, B. K., "Ignoring the alternatives: The N400 is sensitive to stimulus preactivation alone", *Cortex*, 2023. The dissertation author was the primary investigator and author of this paper. Minor edits have been made to bring the formatting in line with the dissertation template.

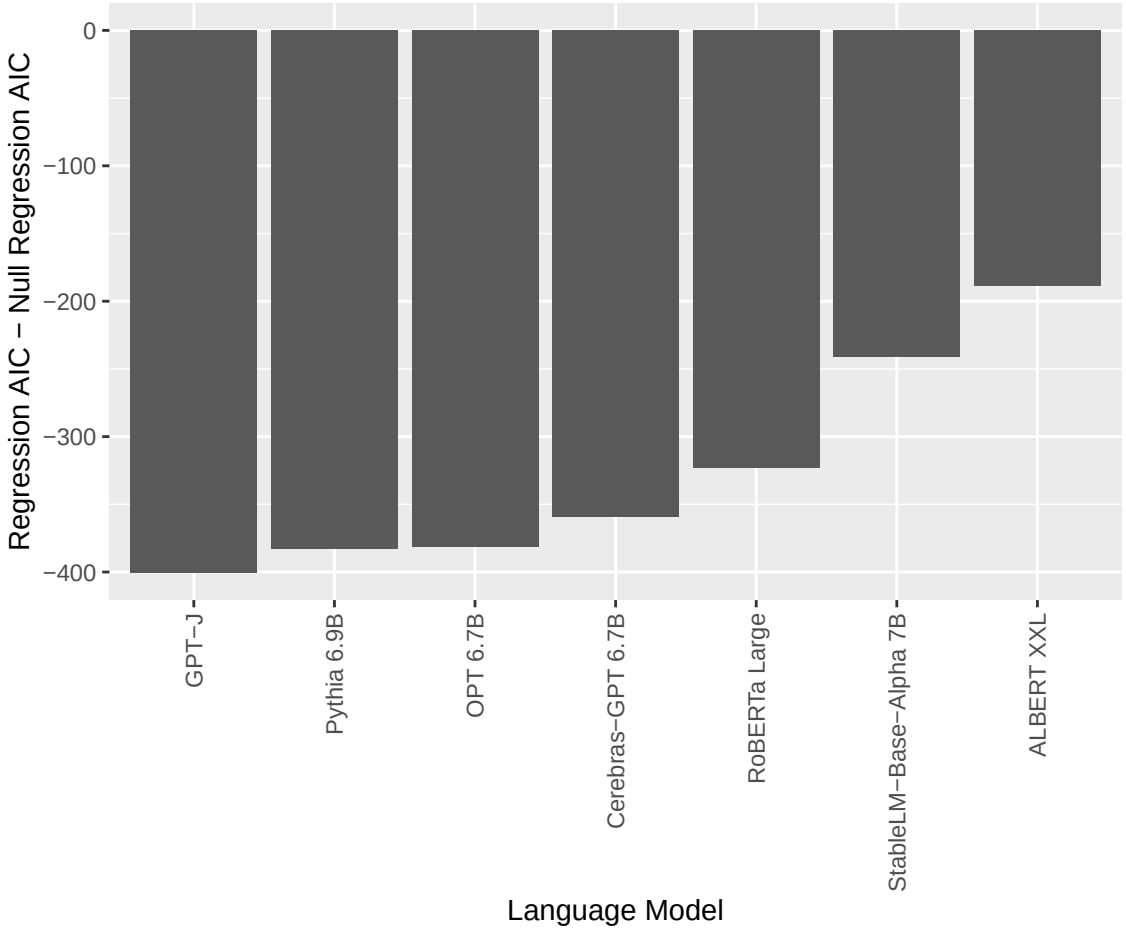

Figure 9.1: AICs of regressions including the probability and surprisal calculated from the indicated model as predictors. A lower AIC indicates a better fit.

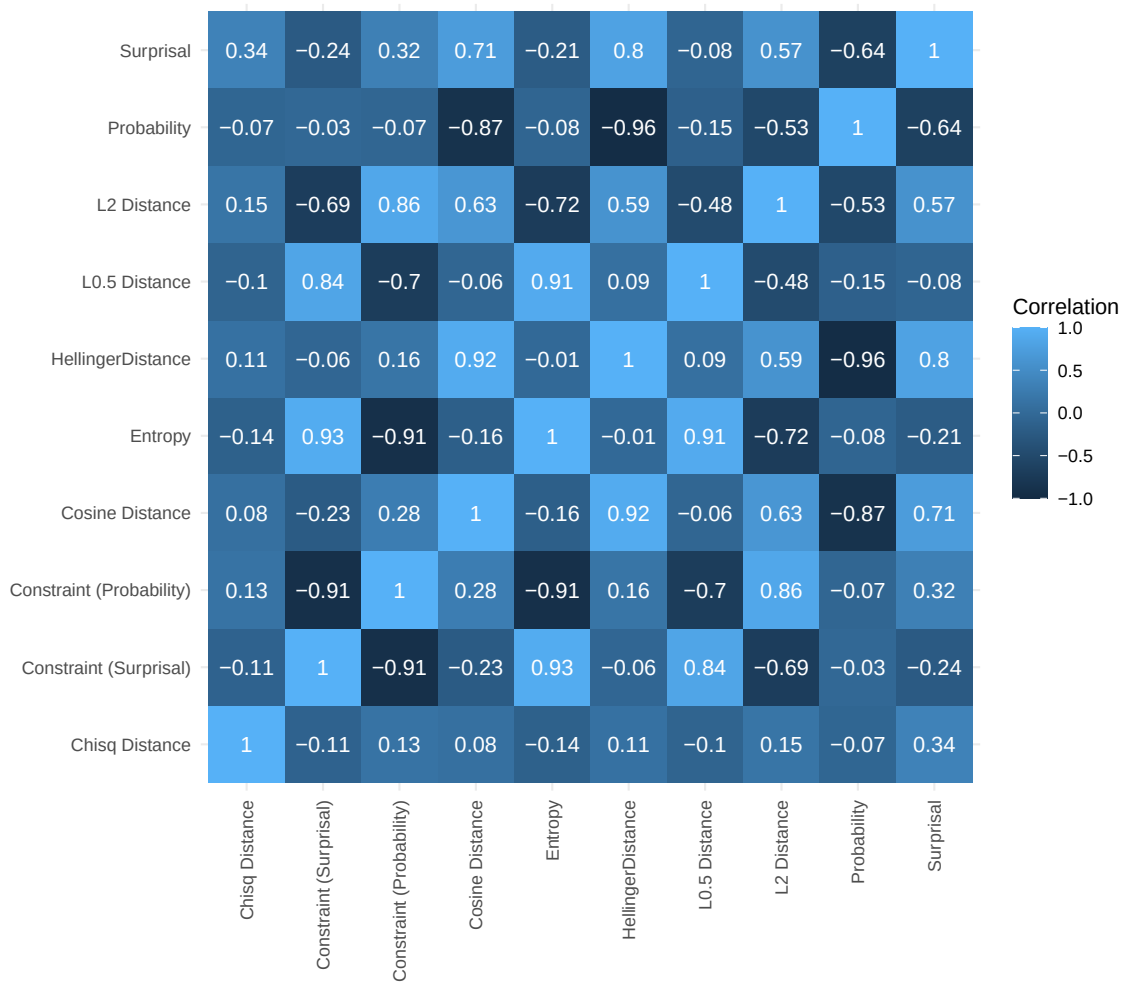

Figure 9.2: The Pearson Correlation r between all variables of interest in our study for all critical words that were single tokens for GPT-J.

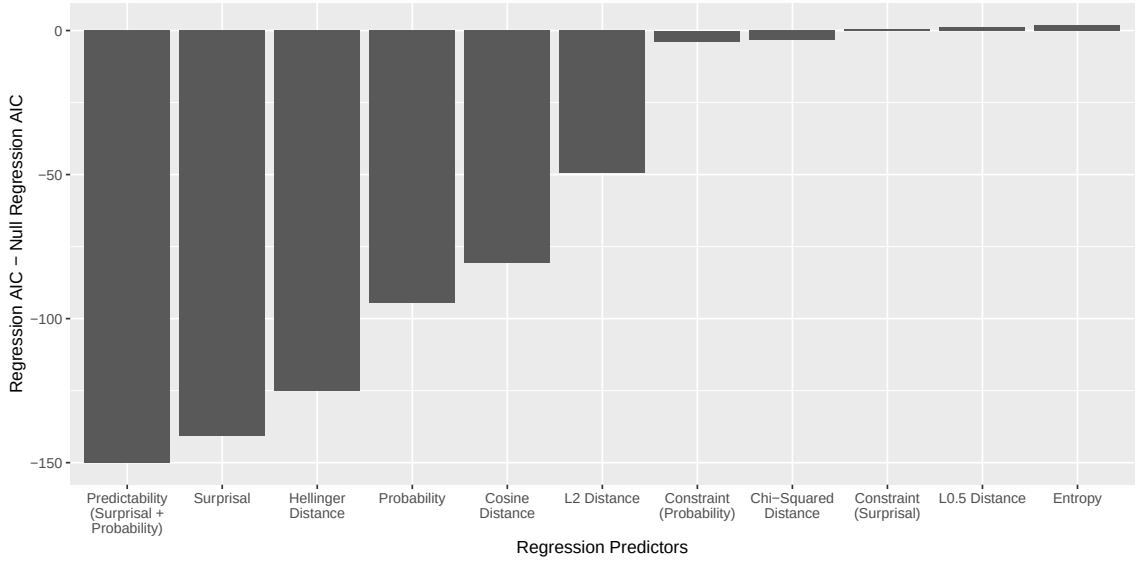

Figure 9.3: The AICs of all regressions including a single metric of interest as a predictor, as well as one including both predictability metrics (probability and surprisal).

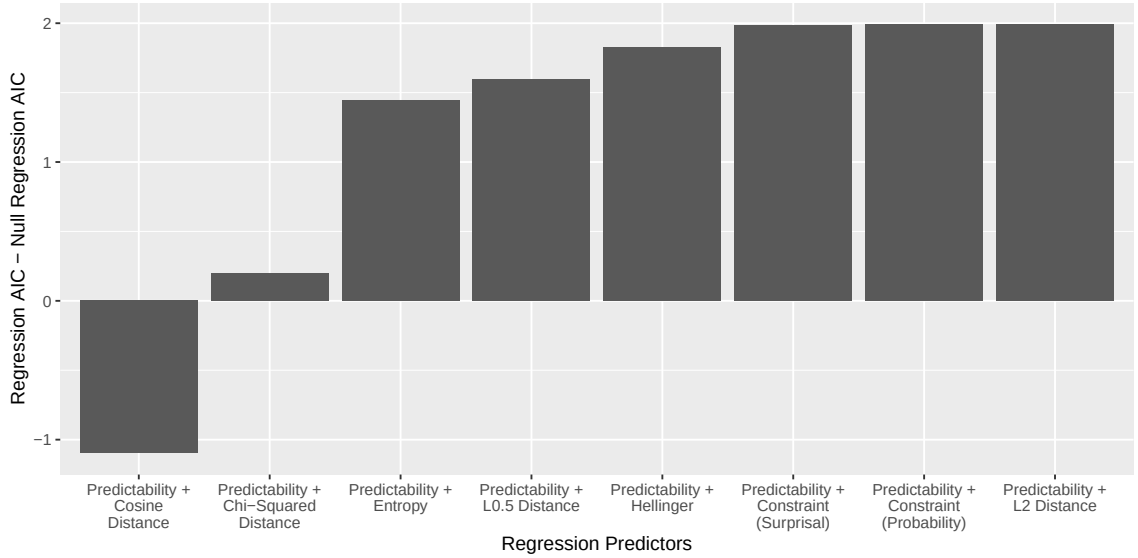

Figure 9.4: The AICs of all regressions including a single metric of interest as a predictor, as well as one including both predictability metrics (probability and surprisal).

Chapter 10

On the mathematical relationship between contextual probability and N400 amplitude

Abstract

Accounts of human language comprehension propose different mathematical relationships between the contextual probability of a word and how difficult it is to process, including linear, logarithmic, and super-logarithmic ones. However, the empirical evidence favoring one of these over another is mixed, appearing to vary depending on the index of processing difficulty used and the approach taken to calculate contextual probability. To help disentangle these results, we focus on the mathematical relationship between corpus-derived contextual probability and the N400, a neural index of processing difficulty. Specifically, we use 37 contemporary transformer language models to calculate the contextual probability of stimuli from 6 experimental studies of the N400, and test whether N400

amplitude is best predicted by a linear, logarithmic, super-logarithmic, or sub-logarithmic transformation of the probabilities calculated using these language models, as well as combinations of these transformed metrics. We replicate the finding that on some datasets, a combination of linearly and logarithmically-transformed probability can predict N400 amplitude better than either metric alone. In addition, we find that overall, the best single predictor of N400 amplitude is sub-logarithmically-transformed probability, which for almost all language models and datasets explains all the variance in N400 amplitude otherwise explained by the linear and logarithmic transformations. This is a novel finding that is not predicted by any current theoretical accounts, and thus one that we argue is likely to play an important role in increasing our understanding of how the statistical regularities of language impact language comprehension.

10.1 Introduction

The N400 (Kutas and Hillyard, 1980, 1984) is a negative component of the event-related brain potential that peaks around 400ms after the presentation of stimulus and is associated with lexical and semantic processing difficulty (Kutas et al., 2006; Thornhill and Van Petten, 2012; Aurnhammer and Frank, 2019b; Brouwer et al., 2021; Federmeier, 2021). Specifically, the amplitude of the N400 response to a stimulus has been found to be large by default, and is reduced (becomes less negative) when neural representations of the stimulus are preactivated—that is, activated by the preceding context (Van Petten and Luka, 2012; DeLong et al., 2014b; DeLong and Kutas, 2020; Kuperberg et al., 2020; Federmeier, 2021). In the language domain, it is by now well-established that the amplitude of the N400 response to a word is highly correlated with the word's contextual probability, whether this is operationalized based on human judgements (Kutas and Hillyard, 1984; for reviews see Kutas et al., 2011; Kutas and Federmeier, 2011; Van Petten and Luka,

2012; DeLong et al., 2014b; Kuperberg et al., 2020; Federmeier, 2021) or the statistics of language (Parviz et al., 2011; Frank et al., 2015; Aurnhammer and Frank, 2019b; Merkx and Frank, 2021; Szewczyk and Federmeier, 2022; Michaelov et al., 2022, 2024).

The canonical way to operationalize contextual probability is using the cloze task, where the cloze probability of a word in a given context is the proportion of participants in a norming study who fill in the gap in a sentence with that word (Taylor, 1953, 1957). Since the relationship between cloze probability and the N400 was first discovered (Kutas and Hillyard, 1984), the finding has been replicated numerous times (see, e.g., Kutas and Van Petten, 1994; Kutas and Federmeier, 2011; Van Petten and Luka, 2012; DeLong et al., 2014b; Kuperberg et al., 2020; Federmeier, 2021), with some studies finding as high a correlation between the two as $r = 0.9$ (Kutas and Van Petten, 1994; Kutas and Federmeier, 2011).

A more recent approach uses the contextual probabilities calculated by computational language models. Language models are systems designed to calculate the probability of a word given a context based on the statistics of language (Jurafsky and Martin, 2024b), and like cloze, these probabilities have also been found to be highly correlated with N400 amplitude (Parviz et al., 2011; Frank et al., 2015; Aurnhammer and Frank, 2019a,b; Yan and Jaeger, 2020; Michaelov et al., 2021; Merkx and Frank, 2021; Szewczyk and Federmeier, 2022; Michaelov et al., 2022, 2024). In addition to sometimes displaying a closer fit to the N400 data than cloze (Michaelov et al., 2022, 2024), these language model probabilities have a higher degree of explanatory power from an information-processing perspective—they allow researchers to specifically test to what extent the statistics of language may influence the preactivation underlying the N400 response.

The precise mathematical relationship between these language-model-derived word probabilities and the amplitude of the N400 responses elicited by the same words in humans

is of prime theoretical importance because it can in principle adjudicate among mechanistic accounts of language processing and of the N400 specifically. But the nature of that relationship is currently unknown, and we focus on this key question in the present study.

In the related literature on reading time, another type of measure which is also thought to reflect processing difficulty and to be impacted by contextual probability, the question has been studied extensively, and a range of possibilities have been proposed and tested. Most notably, the relationship between word probability and processing difficulty has been argued to be be linear (Brothers and Kuperberg, 2021), logarithmic (Smith and Levy, 2013; Shain et al., 2024; Wilcox et al., 2023b), or an exponential transformation of the logarithmically-transformed values (Levy and Jaeger, 2006; Meister et al., 2021; Hoover et al., 2023). The results from large-scale meta-analyses of reading time suggest a linear relationship between cloze and processing difficulty (Brothers and Kuperberg, 2021), but for corpus-derived probabilities (i.e., calculated using language models), there appears to be evidence for both a logarithmic (Smith and Levy, 2013; Shain et al., 2024) and a 'super-logarithmic' relationship (Meister et al., 2021; Hoover et al., 2023), the latter being a term used to describe a super-linear relationship between log-probability and processing difficulty (see Levy and Jaeger, 2006; Smith and Levy, 2013; Shain et al., 2024).

In contrast to reading time, there has been comparatively little work investigating the mathematical nature of the relationship between lexical probability and N400 amplitude, and the results are far from conclusive. Of the three studies that we are aware of that have looked at the relationship between cloze probability and the N400, two (Aurnhammer et al., 2021; Michaelov et al., 2022) found log-transformed cloze probability to correspond slightly more closely to N400 amplitude, while the other (Szewczyk and Federmeier, 2022) found the reverse. Thus far, only two studies have investigated the relationship between statistical (i.e., corpus-derived) lexical probability and N400 amplitude. Yan and

Jaeger (2020), using the probabilities derived from a hybrid model based on a mixture of a 5-gram model and 'skip bi-gram' (see Frank and Willems (2017)), find that surprisal (negative log-probability) better predicts N400 amplitude than un-transformed probability does. This finding is also replicated by Szewczyk and Federmeier (2022), who find that overall, surprisal derived from the GPT-2 language model (Radford et al., 2019) out-performs un-transformed GPT-2 probability as a predictor of N400 amplitude. However, Szewczyk and Federmeier (2022) also find that GPT-2 probability explains variance in N400 amplitude above and beyond that explained by GPT-2 surprisal, and may be a better predictor for more expected words (cloze > 0.05). Thus, the question of the mathematical relationship between the language-model-derived probability of a word and the amplitude of the N400 response to the word is still far from resolved.

In order to address this, we expand upon previous work in several ways. First, the question of the mathematical relationship between language-model-derived probability and N400 amplitude has only been tested for two language models; we analyze data from 37 contemporary transformer language models. Additionally, previous work has only compared the extent to which probability and surprisal predict N400 amplitude. In the current study, we also investigate a range of sub-logarithmic (surprisal to a power < 1) and super-logarithmic (surprisal to a power > 1) relationships in the same vein as some previous work on reading time (Meister et al., 2021; Shain et al., 2024). We use these approaches to re-analyze the five datasets used in the Szewczyk and Federmeier (2022) study (Federmeier et al., 2007; Wlotko and Federmeier, 2012; Hubbard et al., 2019; Szewczyk et al., 2022; Szewczyk and Federmeier, 2022), along with data from a large-scale study carried out by Nieuwland et al. (2018b).

10.2 Theoretical accounts and their mathematical formulations

Theoretical accounts of the N400 and processing difficulty in general differ in the mathematical relationships that they propose may hold between contextual probability and processing difficulty. In this section, we describe a range of such theoretical accounts and the mathematical relationships they propose.

10.2.1 Contextual Probability

There is a long history of studies using cloze probability as a predictor of N400 amplitude (since Kutas and Hillyard, 1984) or of behavioral measures such as reading time (since Fischler and Bloom, 1979). As Brothers and Kuperberg (2021) note, using cloze probability as a predictor of N400 amplitude implicitly assumes a linear relationship between contextual probability and processing difficulty. Brothers and Kuperberg (2021) use this previous work as a basis upon which to build a theoretical framework supporting such a linear relationship, which they term the *proportional preactivation* account.

Mechanistically, processing difficulty as described by the proportional preactivation account aligns with the majority of the contemporary accounts of the N400. As a basic principle, processing difficulty reflects the effort required to activate neural representations driven by the stimulus encountered. Difficulty is reduced by the extent to which these representations were preactivated—that is, already activated at the time that the stimulus was encountered (Kutas and Federmeier, 2011; Federmeier, 2021). Under the proportional preactivation account (as in DeLong et al., 2005; Kutas et al., 2011; Van Petten and Luka, 2012; DeLong et al., 2014b; DeLong and Kutas, 2020; Kuperberg et al., 2020), preactivation is largely driven by prediction based on the preceding context. And crucially, under

the account, words are preactivated in direct proportion to their contextual probabilities. And so, given that processing difficulty reflects the difference between the extent to which a word is preactivated and its full activation state, we should expect probability to be linearly related to processing difficulty (Brothers and Kuperberg, 2021).

10.2.2 Distribution update

Under the proportional preactivation and other contextual probability accounts, processing difficulty arises from the extent to which the stimulus itself is predictable based on its context. An alternative idea is that processing difficulty also reflects the probability of alternatives and the difficulty in disconfirming them. This idea forms the basis of accounts (e.g., Hale, 2001; Levy, 2008; Frank et al., 2013; Smith and Levy, 2013) which we term *distribution update* accounts, and which are often grouped under the category of *surprisal theory* because they posit a linear relationship between processing difficulty and *surprisal*, the negative log-probability of a word given its preceding context.

Such accounts vary in their specific details and formalizations, but at their core they share the idea that as humans comprehend linguistic input, we allocate our neurocognitive resources among different possible parses or interpretations of the current input (Hale, 2001; Levy, 2008), or among different possible next words in the utterance (Frank et al., 2013; Aurnhammer and Frank, 2019b). Specifically, resources are divided such that more likely candidates are allocated a larger amount than less likely candidates, in proportion to their probability. Processing difficulty, then, is the effort required to update the distribution over possible candidates after encountering a given stimulus. Notably, whether this is directly formalized as surprisal (Hale, 2001; Frank et al., 2013) or as the Kullback-Leibler divergence (Kullback and Leibler, 1951) between the probability distribution before and after a word is encountered (Levy, 2008; Aurnhammer and Frank, 2019b), it can be

shown formally that this effort can mathematically be described as surprisal.

In this study, we specifically focus on the formulation of the distribution update account presented (among others accounts) by Aurnhammer and Frank (2019b). Under this account, the state of the language comprehension system before encountering a lexical stimulus can be modeled as a probability distribution over possible next words given the preceding context, and the state after can be modeled as the true probability distribution—a distribution where the actual stimulus has a probability of 1 and all other words have a probability of 0. Processing difficulty under this account is precisely the effort required to 'collapse' the predicted probability distribution to the true probability distribution after encountering a word. This effort can be modeled as the Kullback-Leibler divergence between the two probability distributions, which, as Aurnhammer and Frank (2019b) note, is mathematically equivalent to surprisal.

10.2.3 Composite processing difficulty of sub-word features

The second, related family of logarithmic accounts posits that the relationship between contextual probability and processing difficulty arises not from a direct relationship between the contextual probability of a word and processing difficulty, but rather between sub-components of the word and processing difficulty (Smith and Levy, 2008, 2013). A key account of this kind is the 'highly incremental' account presented by Smith and Levy (2013). The account proposes that rather than occurring at the word level, the effect of probability on lexical processing difficulty might instead arise at the sub-word level, that is, in the processing of each consecutive sub-word fragment of a word. Crucially, the probability of each consecutive fragment of a word impacts word probability multiplicatively (i.e., the probability of a word w_i made up of chunks $c_1...c_k$ is given by $p(w_i) = p(c_1) \times ... \times p(c_k)$) but if each chunk is processed sequentially, its impact on reading time t is additive (i.e.,

$t(w_i) = t(c_1)+...+t(c_n)$). As Smith and Levy (2013) prove and demonstrate with examples, as k increases, any function f that relates $p(w_i)$ to $t(w_i)$ tends toward a linear function of $log(p(w_i))$. Thus, the account argues for a logarithmic relationship between contextual probability and processing difficulty.

While the account makes sense in the context of reading time, it is less clear whether it can account for the N400. Specifically, the highly incremental account relies on additional time taken for each sub-word chunk processed. When measuring processing difficulty using the N400, on the other hand, the focus is on amplitude over a given time period (generally 300-500ms after stimulus presentation), and it is not straightforward to imagine a mechanism whereby the difficulty in processing of incremental sub-word fragments would increase the amplitude in the same fixed-time period. Szewczyk and Federmeier (2022), however, propose an alternative account along the same lines that focuses on semantic features rather than sub-word chunks. Under this account, the probability of a word in a given context is the product of the probability of each of its semantic features, but the effect of the probability of each feature on N400 amplitude is linear. Following Smith and Levy (2013), therefore, if there are a sufficient number of semantic features associated with a given word—and it is difficult to imagine cases where words are not associated with many semantic features—we should expect N400 amplitude to be logarithmically related to contextual probability (Szewczyk and Federmeier, 2022).

While it is in principle possible to view the reading-time variants of this account (Smith and Levy, 2008, 2013) as identifying possible mechanisms by which distribution update accounts such as those provided by Hale (2001) or Levy (2008) could occur, this is not the case with N400-focused accounts. Crucially, under the formulation of the distribution update account provided by Aurnhammer and Frank (2019b), surprisal indexes lexical predictions, while Szewczyk and Federmeier (2022) argue instead that lexical predictions

lead to differences in N400 amplitude that are linearly related to contextual probability.

10.2.4 Uniform Information Density

In addition to linear and logarithmic relationships, it is also possible that processing difficulty is non-linearly related to log-probability. The main argument for this comes from considering how information is distributed throughout a sentence. Intuitively, one would expect that sentences where all the information is concentrated into a small number of words would be harder to comprehend than those where information is more evenly spread out (Levy, 2005), and it is possible that this may result in a pressure towards more uniform information density in sentence production (Levy and Jaeger, 2006; Smith and Levy, 2013; Meister et al., 2021; Shain et al., 2024). Specifically, Levy and Jaeger (2006)[1] demonstrate mathematically that if there is a super-logarithmic relationship between lexical probability and processing difficulty (i.e., difficulty= $log(p)^k$ where $k > 1$), then uniform information density minimizes the effort required to process a whole utterance.

While there is a substantial body of both theoretical and empirical work arguing in favor of uniform information density in general (see, e.g., Fenk and Fenk-Oczlon, 1980; Genzel and Charniak, 2002; Aylett and Turk, 2004; Maurits et al., 2010; Coupé et al., 2019; Clark et al., 2023), the question of whether it arises from comprehender-oriented principles (i.e., a form of audience design) is still an open question. Thus, a super-logarithmic relationship between contextual probability and processing difficulty could help explain why language users produce the utterances that they do.

[1] In the supplementary Appendix to the paper, available at `https://www.researchgate.net/publication/221618546`

10.2.5 Multiple sub-components

In addition to accounts proposing one specific mathematical relationship between contextual probability and processing difficulty, some have proposed that a combination of relationships holds between the two. Szewczyk and Federmeier (2022) propose, for example, that the N400 is related to contextual probability both logarithmically and linearly. The account provided by Szewczyk and Federmeier (2022) follows from accounts of the N400 under which the response is posited to generally reflect the overlap between the semantic features of the stimulus and its context, and to only under some conditions (for example, when more attention is paid to stimuli) reflect explicit lexical prediction (see, e.g., Lau et al., 2013; Federmeier, 2021). Szewczyk and Federmeier (2022) specifically argue that the logarithmic relationship reflects the effect of semantic feature overlap (proposing the aforementioned account of logarithmic composite processing difficulty based on sub-word semantic features), and that the linear relationship reflects the the effect of lexical prediction (following contextual probability accounts like the proportional preactivation account). While they do not provide direct evidence for the correspondence between these components and the proposed mechanisms indexed, Szewczyk and Federmeier (2022) do provide direct evidence of potentially separable linear and logarithmic effects. Specifically, Szewczyk and Federmeier (2022) find that the linear effect is only significant above and beyond the logarithmic for expected items (cloze $> 5\%$) when analyzing the whole N400 time window (300-500ms); and is only significant when predicting the N400 response to all tokens when analyzing the first half of the time window (300-400ms).

10.3 Analysis 1: Powers of Surprisal

10.3.1 Introduction

The aim of the present study is to investigate whether the relationship between contextual probability and N400 amplitude is linear, logarithmic, super-logarithmic, or sub-logarithmic. In Analyses 2 and 3 below we will compare how well each of these transformations of probability predict N400 amplitude. This first analysis sets the stage by identifying the super- or sub-logarithmic transformation of probability that best correlates with N400 amplitude (in order to subsequently compare this with probability and surprisal).

As previously discussed, the current evidence from reading time suggests either a logarithmic (Smith and Levy, 2013; Shain et al., 2024; Wilcox et al., 2023b) or super-logarithmic relationship (Meister et al., 2021) between contextual probability and processing difficulty. However, it is also in principle possible that there is a sub-logarithmic relationship, and as this type of analysis has never been carried out for the N400, we account for both possibilities. Specifically, we calculate all surprisal^k, where k covers all 0.1 increments between 0.1 and 2 (inclusive), as well as -1,-0.5, and -0.1, for comparison. We then test how well each of these predicts N400 amplitude.

10.3.2 Method

Language models

Recent research shows that among contemporary language model architectures, N400 amplitude is best predicted by transformers (Merkx and Frank, 2021; Michaelov et al., 2022). We therefore restricted our analysis to contemporary transformer language models made available through the *transformers* (Wolf et al., 2020) Python (Van Rossum

and Drake, 2009) package. We further restrict our analyses to only include autoregressive (unidirectional) transformer language models, as the best-performing model in Michaelov et al. (2022) was of this type, and because they produce well-defined probabilities for critical words made up of multiple tokens, allowing us to calculate surprisal for any word. These included 37 models of the GPT-2 (Radford et al., 2019), GPT-Neo (Black et al., 2021), GPT-J (Wang and Komatsuzaki, 2021), Pythia (Biderman et al., 2023b), OPT (Zhang et al., 2022), XGLM (Lin et al., 2021), and BLOOM (BigScience, 2022) architectures, as well as DistilGPT2, a 'distilled' form of GPT-2 (see Sanh et al., 2020).

Stimuli and N400 data

We use the stimuli and experimental data from 5 previously-published N400 studies (Federmeier et al., 2007; Wlotko and Federmeier, 2012; Nieuwland et al., 2018b; Hubbard et al., 2019; Szewczyk et al., 2022), and one unpublished dataset released as part of a recent meta-analysis (Szewczyk and Federmeier, 2022).

In the 5 datasets (Federmeier et al., 2007; Wlotko and Federmeier, 2012; Hubbard et al., 2019; Szewczyk et al., 2022; Szewczyk and Federmeier, 2022) preprocessed and released by Szewczyk and Federmeier (2022), N400 amplitude was operationalized as the mean voltage from four centro-parietal electrodes (MiCe, MiPa, LMCe, RMCe), and the mean was taken over the 300-500ms time window. In contrast to much of the work on the N400, in this dataset, N400 amplitudes are not corrected using a baseline amplitude; instead, the -100-0ms mean amplitude baseline is included as a covariate in analysis (see discussion in Szewczyk and Federmeier, 2022). The details of these datasets are presented in Table 10.1. The stimuli from the study by Federmeier et al. (2007) follow a 2×2 design, with sentences either having a high or low constraint, and the N400 being recorded from either an expected (highest-cloze) or unexpected (low-cloze) continuation. The stimuli

Table 10.1: Details of all datasets analyzed

Dataset	Stimuli	Participants	Total trials
Federmeier et al. (2007)	564	32	7856
Wlotko and Federmeier (2012)	300	16	4440
Hubbard et al. (2019)	192	32	5705
Szewczyk et al. (2022)	672	32	4939
Szewczyk and Federmeier (2022)	600	26	4822
Nieuwland et al. (2018b)	160	334	25978

from the other studies were generally selected from Federmeier et al. (2007), with Wlotko and Federmeier (2012) including additional sentences with critical words that varied more continuously in terms of their cloze probability, and Szewczyk et al. (2022) adding adjectives that either reduced or increased the cloze probability of critical items.

The remaining dataset is a large-scale study carried out by Nieuwland et al. (2018b). We take the subset of the data corresponding to N400 amplitudes elicited by nouns. We use the preprocessed data provided by Nieuwland et al. (2018b), who operationalize N400 amplitude as mean voltage between 200-500ms after stimulus presentation at 6 centro-parietal electrodes (Cz, C3, C4, Pz, P3, and P4), baseline-corrected by subtracting the mean amplitude in the -100-0ms time window. The details of this dataset are described in Table 10.1. In this study, stimuli had either expected (highest-cloze) or unexpected (low-cloze) critical words.

Calculating the metrics

To investigate how well the language models' predictions correlate with N400 amplitude, we ran each of the stimulus sentences from the six studies (Federmeier et al., 2007; Wlotko and Federmeier, 2012; Nieuwland et al., 2018b; Hubbard et al., 2019; Szewczyk et al., 2022; Szewczyk and Federmeier, 2022) up until the critical noun through each of the

37 language models, calculating surprisal for each of the critical words. As is commonly the case with transformer language models, not all critical nouns were in each model's vocabulary as individual tokens. Because all models are autoregressive, calculating the surprisal of multi-token words is straightforward—we calculate the surprisal of each sub-word token of the critical word given the preceding context (including preceding sub-word tokens) and take their sum, which is equivalent to taking the product of the probabilities. We then exponentiate these surprisal values, using these values in our analyses.

Procedure for Statistical Analysis

In order to test how well each exponentiated form of surprisal calculated using each language model predicts N400 amplitude, we construct linear mixed-effects regression models using these variables as predictors and N400 amplitude as the dependent variable, comparing the Akaike's Information Criterion (AIC; Akaike, 1973) of these regressions. AIC provides a measure of a regression's fit to the data, with a lower AIC indicating a better fit. We run further analyses on these AIC values to compare how well each metric performs across models.

When analyzing the data from the Nieuwland et al. (2018b) study, our statistical analysis approach aimed to match the original as much as possible. In these models, N400 amplitude was the dependent variable. The variable of interest for each language model (i.e., surprisal^k) was included as a fixed effect. The original study was carried out at multiple laboratories, with previous work showing that depending on the subset of the data used, laboratory can be a significant predictor of N400 (Nieuwland et al., 2018b; Michaelov et al., 2022), so we also included this as a fixed effect. In order to be able to compare regression fit across language models and metrics, the random effects structure needs to be consistent across regressions, and the maximal random effects structure that fulfils this

requirement in addition to converging and not resulting in any singular fits includes a random intercept for each subject. All numerical variables were z-scored.

The 5 other datasets analyzed (Federmeier et al., 2007; Wlotko and Federmeier, 2012; Nieuwland et al., 2018b; Hubbard et al., 2019; Szewczyk et al., 2022; Szewczyk and Federmeier, 2022) were all preprocessed in the same way by Szewczyk and Federmeier (2022). We kept our statistical analysis as close to those in Szewczyk and Federmeier (2022) as possible. As in Szewczyk and Federmeier (2022), un-baselined N400 amplitude was the dependent variable, with the baseline amplitude included as a fixed effect. The other covariates included in the original analyses and provided by Szewczyk and Federmeier (2022) were concreteness (Brysbaert et al., 2014), frequency (logarithmically transformed; Brysbaert and New, 2009), orthographic neighborhood (OLD20; Yarkoni et al., 2008), and sentence position, which we also included as fixed effects. We included the maximal random effects structure that would allow model convergence, result in no singular fits, and be consistent across regressions. The resulting random effects structure included random slopes for the baseline voltage for each subject and item, as well as a random intercept for each item. All numerical values were z-scored.

All graphs were created and statistical analyses carried out in *R* (R Core Team, 2023) using *Rstudio* (RStudio Team, 2020) and the *tidyverse* (Wickham et al., 2019), *lme4* (Bates et al., 2015), *lmerTest* (Kuznetsova et al., 2017), *mgcv* (Wood, 2017), *ggh4x* (van den Brand, 2021), *tidytext* (Silge and Robinson, 2016), *ggtext* (Wilke and Wiernik, 2022), *RColorBrewer* (Neuwirth, 2022), and *osfr* (Wolen et al., 2020) packages. All figures except Figure 10.1 use colorblind-friendly palettes (Chang, 2022). All reported p-values are corrected for multiple comparisons based on false discovery rate (Benjamini and Yekutieli, 2001) across all statistical tests carried out. We provide all data, code, and statistical analysis scripts at `https://osf.io/w5hez`.

10.3.3 Results

The fit of all regressions including surprisal^k are shown in Figure 10.1. For four of the six datasets (Federmeier et al., 2007; Hubbard et al., 2019; Szewczyk et al., 2022; Szewczyk and Federmeier, 2022), the lowest AICs are achieved by regressions with sub-linear transformations of surprisal as the predictor of N400 amplitude, indicating that these best fit the human data. For the remaining two datasets (Nieuwland et al., 2018b; Wlotko and Federmeier, 2012), the results are less clear—for the majority of language models, the best transformation of surprisal appears to be at $k = 1$ (i.e., no transformation) or slightly above.

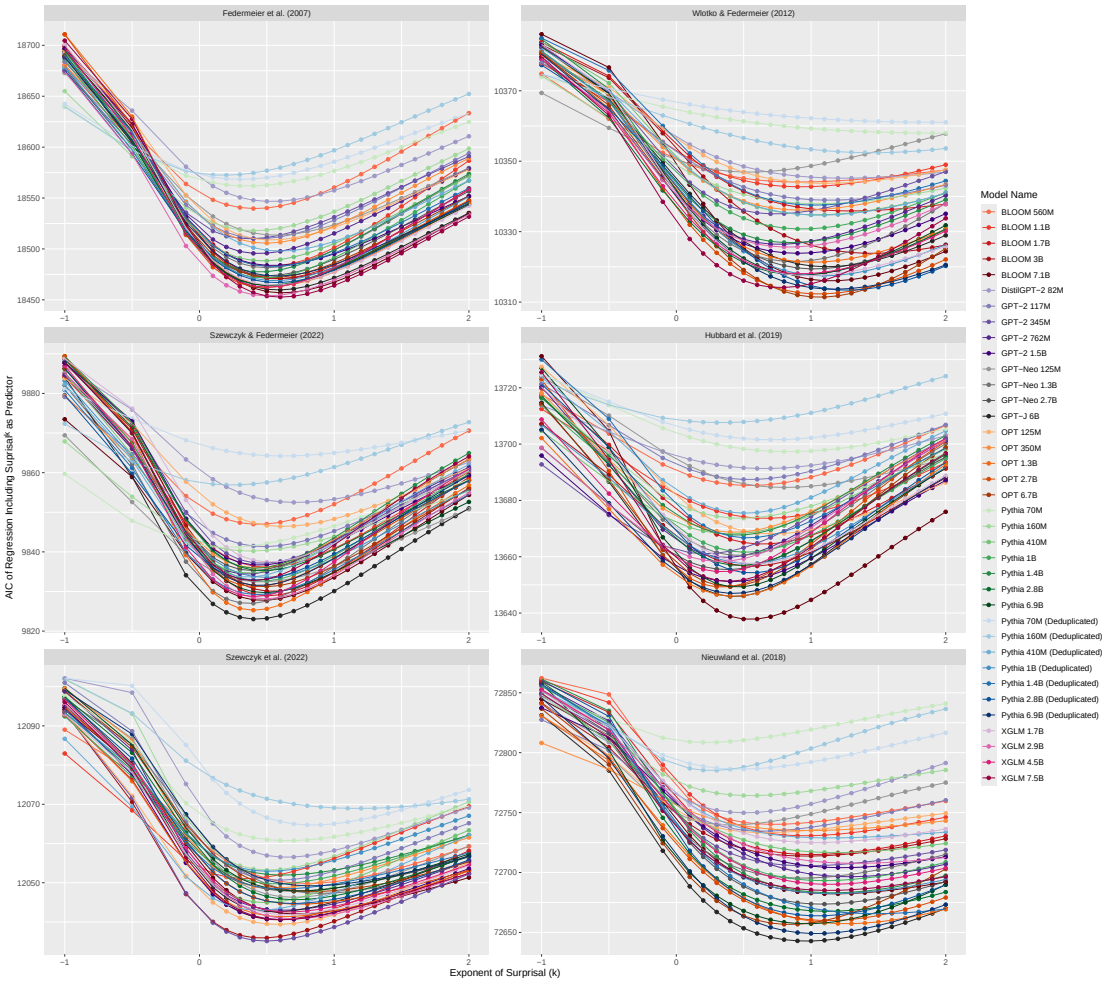

Figure 10.1: AIC of regressions predicting N400 amplitude with the exponentiated values of the surprisal calculated using 37 autoregressive transformer language models.

To investigate these results further, we fit a general additive model to these AIC values for each dataset, predicting them using the default thin plate regression splines provided by the *mgcv* (Wood, 2017) package to fit k as a predictor and including random effects terms for each language model. All fitted GAMs had adjusted R^2 values of greater than 0.9 (Federmeier et al., 2007: adjusted $R^2 = 0.94$; Wlotko and Federmeier, 2012: adjusted $R^2 = 0.92$; Szewczyk and Federmeier, 2022: adjusted $R^2 = 0.93$; Hubbard et al.,

2019: adjusted $R^2 = 0.90$; Szewczyk et al., 2022: adjusted $R^2 = 0.96$; Nieuwland et al., 2018b: adjusted $R^2 = 0.92$).

These general additive models were used to estimate the power of surprisal that produces the lowest AIC across language models, after accounting for differences between models. To do this, we generate a dummy dataset with values of k between 0 and 2 at 0.1 increments and an arbitrary (non-existent) language model. The general additive models were then used to estimate the regression AICs for an arbitrary language model based on k alone. The results replicate the numerical descriptions above—the general additive models estimate the lowest AIC for four of the datasets to occur when $0 < k < 1$ (Federmeier et al., 2007: $k = 0.5$; Hubbard et al., 2019: $k = 0.5$; Szewczyk et al., 2022: $k = 0.6$; Szewczyk and Federmeier, 2022: $k = 0.4$). Also matching the graphs, the lowest AIC for the Nieuwland et al. (2018b) and (Wlotko and Federmeier, 2012) datasets is estimated to occur when $k = 1$.

We also see a similar pattern if we look at the actual best-fitting predictability metric for each dataset overall: for the Federmeier et al. (2007) dataset this is surprisal$^{0.6}$ as calculated by XGLM-7.5B, for Hubbard et al. (2019) this is BLOOM 7.1B surprisal$^{0.5}$, for Szewczyk et al. (2022) this is GPT-2 345M surprisal$^{0.5}$, for Szewczyk and Federmeier (2022) this is GPT-J 6B surprisal$^{0.4}$, for Nieuwland et al. (2018b) this is GPT-J 6B surprisal^{1} (i.e., surprisal), and for Wlotko and Federmeier (2012) this is OPT 6.7B surprisal$^{1.1}$.

10.3.4 Discussion

The results of the analysis are clear: for four of the six datasets (Federmeier et al., 2007; Hubbard et al., 2019; Szewczyk et al., 2022; Szewczyk and Federmeier, 2022), the value of k that leads to the surprisal^k that best fits N400 amplitude is well below 1—in fact, they are closer to 0.5. Thus, for these datasets, the results suggest that there is a

sub-linear relationship between surprisal and N400 amplitude; that is, a sub-logarithmic relationship between probability and the N400. For the remaining 2 datasets, (Nieuwland et al., 2018b; Wlotko and Federmeier, 2012), the best values of k are close to 1, suggesting a linear relationship between surprisal and the N400, and so a logarithmic relationship between probability and the N400.

These results depart from previous work on reading time, where, depending on the dataset and method of analysis, the best values of k tend to fall on both sides of 1 (Shain et al., 2024), or even tend to be greater than 1 (Meister et al., 2021; Hoover et al., 2023), supporting the logarithmic or super-logarithmic accounts respectively. By contrast, in the present study, the evidence leans in the opposite direction—while two datasets appear to support a logarithmic relationship, the remaining four support a sub-logarithmic relationship. As far as we are aware, this is the first study with evidence most strongly supporting a sub-logarithmic relationship between contextual probability and processing difficulty.

10.4 Analysis 2: A comparison of metrics and language models

10.4.1 Introduction

In Analysis 1, we sought to quantify which exponential transformation of surprisal best predicts the N400, testing whether a logarithmic, sub-logarithmic, or super-logarithmic relationship between probability and the N400 best explains the data. We found that for four of the six datasets (Federmeier et al., 2007; Hubbard et al., 2019; Szewczyk et al., 2022; Szewczyk and Federmeier, 2022), a sub-logarithmic transformation best predicts N400 amplitude, while the results for the other two datasets suggest a log-

arithmic relationship (Nieuwland et al., 2018b; Wlotko and Federmeier, 2012). However, the approach in Analysis 1 raises two questions, which we address in this section.

First, how meaningful are the patterns observed in Analysis 1? Because the approach only estimates the overall best exponent from all language models, it is not clear how much better the best exponential transformation is compared to the alternatives, or how consistent any patterns are across language models. In this second analysis, we address both concerns, comparing surprisal with the best sub-logarithmic metric overall, which we find to be surprisal$^{0.6}$.

The second question is whether the previous finding that language model surprisal is a better predictor of N400 amplitude than language model probability (Yan and Jaeger, 2020; Szewczyk and Federmeier, 2022) holds for the larger number of language models and datasets that we test. As previously discussed, two studies thus far have directly compared how probability and surprisal predict N400 amplitude, finding surprisal to be a better predictor (Yan and Jaeger, 2020; Szewczyk and Federmeier, 2022). However, as noted, these studies only use one language model each. In this analysis, in addition to comparing the fit of surprisal and surprisal$^{0.6}$ to N400 amplitude, we also compare the fit of probability to both of these. Thus, we expand upon previous work by comparing how well probability and surprisal predict N400 amplitude by analyzing the predictions of 37 contemporary transformer language models, as well as carrying out the first analysis comparing the performance of either of these metrics to sub-logarithmically transformed probability (i.e., surprisal$^{0.6}$).

10.4.2 Method

Our analyses used the same datasets and language models as in Analysis 1. To calculate the best sub-logarithmic metric overall, we return to general additive models as in

Analysis 1, but instead fitting a single model to predict regression AIC across all datasets (and including dataset as a random effect). When we construct a dummy dataset as in Analysis 1, we find that the additive model predicts the best AIC overall for surprisal$^{0.6}$, so we use this as our best sub-logarithmic metric. We then calculate probability, surprisal, and surprisal$^{0.6}$ for each stimulus in each dataset using each language model.

10.4.3 Results

The AICs of the regressions including probability, surprisal, or surprisal$^{0.6}$ are presented in Figure 10.2. As can be seen visually, on the whole, the regressions including surprisal as a predictor perform better than those including probability. In addition, we see that regressions including surprisal$^{0.6}$ perform better than those including probability. As in Analysis 1, we see that for four out of six datasets (Federmeier et al., 2007; Hubbard et al., 2019; Szewczyk et al., 2022; Szewczyk and Federmeier, 2022), surprisal$^{0.6}$ is a better predictor of N400 amplitude than surprisal, while the reverse is true for the remaining two datasets (Nieuwland et al., 2018b; Wlotko and Federmeier, 2012).

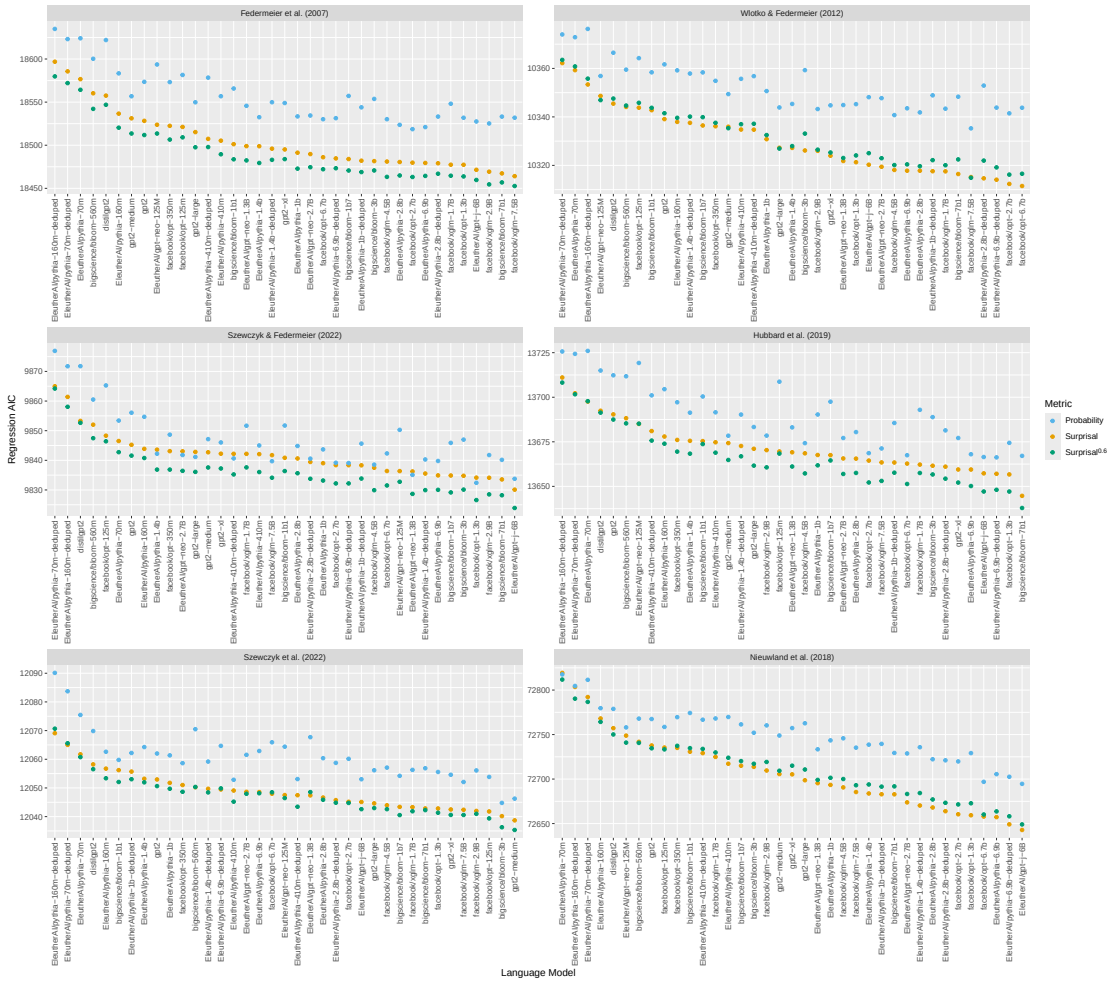

Figure 10.2: AIC of regressions predicting N400 amplitude using the probability, surprisal, or surprisal$^{0.6}$ calculated by 37 autoregressive transformer language models.

We next quantify the exact degree to which the AICs of regressions including probability, surprisal, and surprisal$^{0.6}$ differ. To do this, we constructed linear mixed effects models comparing the performance of each pair of metrics (probability and surprisal, probability and surprisal$^{0.6}$, and surprisal and surprisal$^{0.6}$) as predictors. These linear mixed-effects models all had regression AIC as the dependent variable, metric as the predictor, and language model as a random intercept. We show the linear mixed ef-

fects models' estimates of the difference in AIC between regressions with each metric as a predictor in Table 10.2. A difference in AIC of 4 or more is generally taken to indicate a 'substantial' difference in support of the regression with the lower AIC over that with the higher AIC (Burnham and Anderson, 2004). Thus, we see that for all datasets, regressions using surprisal or surprisal$^{0.6}$ as a predictor tend to fit the N400 data substantially better than those using probability as a predictor. When comparing surprisal and surprisal$^{0.6}$, however, the results are less clear—for three of the datasets (Federmeier et al., 2007; Hubbard et al., 2019; Szewczyk and Federmeier, 2022), surprisal$^{0.6}$ as a predictor tends to lead to regressions that better fit the N400 data than surprisal, while for the Nieuwland et al. (2018b) dataset, the reverse is true. The estimated differences in AICs between regressions including surprisal and surprisal$^{0.6}$ as predictors on the remaining two N400 datasets (Wlotko and Federmeier, 2012; Szewczyk et al., 2022) are less than 4, and thus it is not clear that there is a meaningful difference between the two metrics on these datasets.

Table 10.2: Estimated differences between the AICs of regressions using probability, surprisal, and surprisal$^{0.6}$ as predictors. **P-S** reflects the difference between the AICs of regressions with probability (P) and surprisal (S) as predictors where the value reflects the extent to which regressions with surprisal as a predictor have a lower AIC than regressions with probability as a predictor. In the same way, **P-S**$^{0.6}$ reflects the difference in AIC between probability (P) and surprisal$^{0.6}$ (S$^{0.6}$) and **S-S**$^{0.6}$ the difference in AIC between surprisal (S) and surprisal$^{0.6}$ (S$^{0.6}$),

Exp.	**P-S**	**P-S**$^{0.6}$	**S-S**$^{0.6}$
Nieuwland et al. (2018b)	41.91	37.53	-4.38
Federmeier et al. (2007)	51.85	66.15	14.30
Wlotko and Federmeier (2012)	22.44	20.05	-2.39
Szewczyk and Federmeier (2022)	5.72	10.86	5.14
Hubbard et al. (2019)	18.23	24.43	6.19
Szewczyk et al. (2022)	11.89	13.41	1.52

In order to test how robust these estimates are, we carried out pairwise supplementary analyses evaluating whether metric (i.e., probability vs. surprisal, probability vs. surprisal$^{0.6}$, or surprisal vs. surprisal$^{0.6}$) was a significant predictor of regression AIC by

running likelihood ratio tests comparing the aforementioned linear mixed-effects models to equivalent models not including metric as a predictor. The results are reported in subsection 10.10.1. In all cases, metric was a significant predictor, and thus the findings that using surprisal$^{0.6}$ to predict N400 amplitude leads to the regression with the best fit on 3 of the datasets (Federmeier et al., 2007; Hubbard et al., 2019; Szewczyk and Federmeier, 2022) and that using surprisal leads to the regressions with the best fit on the Nieuwland et al. (2018b) dataset are statistically significant.

10.4.4 Discussion

This analysis has three findings. First, we find that the results of Yan and Jaeger (2020) and Szewczyk and Federmeier (2022) generalize across a larger number of models—overall, surprisal is a better predictor of N400 amplitude than probability is. Second, we find that, across language models, surprisal$^{0.6}$ is also a better predictor of N400 amplitude than probability is. Finally, when we compare the performance of regressions predicting N400 amplitude using surprisal or surprisal$^{0.6}$, we provide additional support for the results of Analysis 1: for the majority of datasets (Federmeier et al., 2007; Hubbard et al., 2019; Szewczyk et al., 2022; Szewczyk and Federmeier, 2022), using surprisal$^{0.6}$ to predict N400 amplitude leads to a numerically lower AIC than using surprisal; while for the remaining datasets (Nieuwland et al., 2018b; Wlotko and Federmeier, 2012), the reverse is true.

Analyzing the differences in more detail and running statistical tests adds further nuance to these results. Specifically, on three of the datasets (Federmeier et al., 2007; Hubbard et al., 2019; Szewczyk et al., 2022; Szewczyk and Federmeier, 2022), using surprisal$^{0.6}$ to predict N400 amplitude tends to lead to to a substantially better fit, and this difference is statistically significant across language models. In addition, for one dataset (Nieuwland et al., 2018b), the substantially better fit to the data of surprisal (compared to surprisal$^{0.6}$)

is statistically significant. Finally, for the remaining two datasets, while there are significant differences, the effect size of under 4 AIC suggests that there is no clear difference in whether using surprisal or surprisal$^{0.6}$ to predict N400 amplitude leads to a better fit to the data.

10.5 Analysis 3: Variance Explained

10.5.1 Introduction

A striking result of the present study so far has been that in contrast to work on reading time, the evidence best supports a sub-logarithmic relationship between probability and N400 amplitude. However, thus far, we have only analyzed differences in overall fit between models; the key quantitative question is to what extent the variables discussed can explain variance in N400 amplitude. That is the question we address in this section.

To do this, we turn to our best-performing language models. If we hope to understand the mathematical relationship between contextual probability and the N400, we should use the models whose metrics (i.e., transformed probability values) most closely correlate with N400 amplitude to avoid confounds. For example, because surprisal magnifies differences at the low end of the scale (i.e., when probability is close to zero), surprisal may magnify small differences in predictions across language models such that probabilities from poorly performing language models might actually spuriously outperform surprisal in some cases.

We therefore select the best overall language model for each dataset, and test how well probability, surprisal, and surprisal$^{0.6}$ each explain the variance in N400 amplitude in that dataset.

10.5.2 Method

Based on the results of Analyses 1 and 2, we use surprisal$^{0.6}$ as our sub-logarithmic metric. We select the best language models by looking at those that produced the regression with the lowest AIC for each dataset in Analysis 1: these were XGLM 7.5B on the Federmeier et al., 2007) dataset, BLOOM 7.1B on Hubbard et al. (2019), GPT-2 345M on Szewczyk et al. (2022), OPT 6.7B on Wlotko and Federmeier (2012), and GPT-J 6B on Nieuwland et al. (2018b) and Szewczyk and Federmeier (2022).

10.5.3 Results

First, we visualize the relationship between each metric and the N400. Figure 10.3 shows the relationship between baselined N400 amplitude and GPT-J 6B probability, surprisal, and surprisal$^{0.6}$ for all datasets. In line with the results of Analysis 2, we see that for the majority of datasets (Federmeier et al., 2007; Hubbard et al., 2019; Szewczyk and Federmeier, 2022; Szewczyk et al., 2022), surprisal$^{0.6}$ does indeed appear to have the most (approximately) linear relationship to N400 amplitude. By contrast, surprisal appears to have the most linear relationship with N400 amplitude in the Wlotko and Federmeier (2012) dataset. Finally, the results for the Nieuwland et al. (2018b) are less clear—visually, it is hard to tell which metric has the most linear relationship to N400 amplitude. We provide the equivalent graphs for the other language models in subsection 10.10.2.

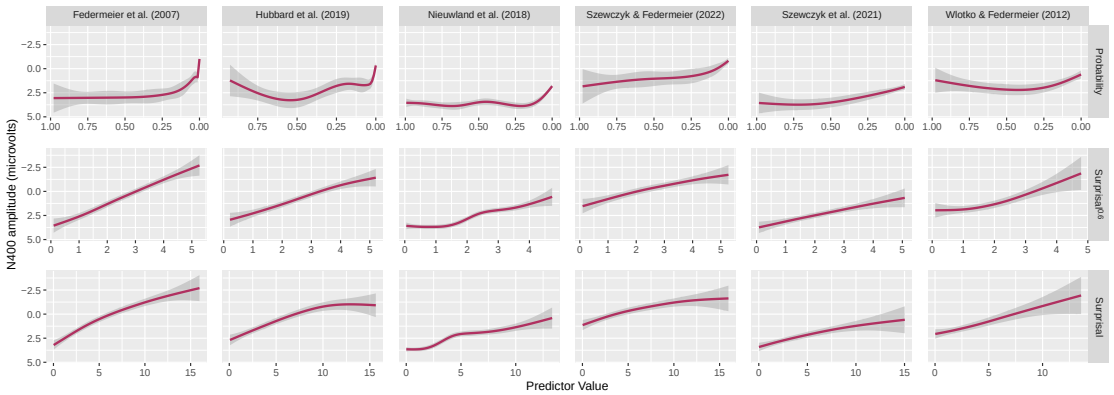

Figure 10.3: N400 amplitude as a function of GPT-J 6B probability, surprisal, and surprisal$^{0.6}$. The x-axis for probability is reversed for easier comparison with surprisal and surprisal$^{0.6}$.

We next investigate the extent to which each metric explains variance in N400 amplitude, and whether any variables explain additional variance once others are accounted for. Thus, in addition to looking at probability, surprisal, and surprisal$^{0.6}$ as in Analysis 2, we also look at the combination of surprisal and probability, following Szewczyk and Federmeier (2022). We show how the fit of regressions including these predictors compare in Figure 10.4. Given the drastically worse performance of the probability-only regressions on some of the datasets (in particular Federmeier et al., 2007 and Nieuwland et al., 2018b), we exclude these from the Figure 10.4; however, for completeness, we provide these in subsection 10.10.3.

Figure 10.4: The fit of regressions including probability, surprisal, surprisal$^{0.6}$, probability and surprisal, probability and surprisal$^{0.6}$, and surprisal and surprisal$^{0.6}$ as predictors of N400 amplitude. We look at the results for the 5 language models that best predict each of the 6 datasets.

We then test how well each of these predictors or sets of predictors calculated using each language model explains variance in N400 amplitude using likelihood ratio tests between linear mixed-effects models, testing the effect of adding probability, surprisal, surprisal$^{0.6}$, and both probability and surprisal to a regression already including one of these, thereby testing whether the added variable explains additional variance. Specifically, we test the effect of adding surprisal$^{0.6}$ to a linear mixed-effects model already including surprisal (S+$S^{0.6}$), adding probability to a model already including surprisal (S+p), adding surprisal to a model already including surprisal$^{0.6}$ ($S^{0.6}+S$), adding probability to a model already including surprisal$^{0.6}$ ($S^{0.6}+p$), adding surprisal to a model already including probability ($p+S$), adding surprisal$^{0.6}$ to a model already including probability ($p+S^{0.6}$), adding surprisal$^{0.6}$ to a model already including surprisal and probability ($(S+p)+S^{0.6}$), and adding both surprisal and probability to a model already including surprisal$^{0.6}$ ($S^{0.6}+(S+p)$).

Table 10.3: Results of the likelihood ratio tests testing the effect of adding GPT-J 6B probability (p), surprisal (S), surprisal$^{0.6}$ ($S^{0.6}$), or a combination of these to a linear mixed-effects model already including one or more other of these variables, thereby testing whether they explain any additional variance. As an example, $S + p$ refers to a likelihood ratio test of whether probability explains additional variance in N400 amplitude above and beyond that explained by surprisal. F07 refers to Federmeier et al. (2007), W12 to Wlotko and Federmeier (2012), SF22 to Szewczyk and Federmeier (2022), H19 to Hubbard et al. (2019), S22 to Szewczyk et al. (2022), and N18 to Nieuwland et al. (2018b).

	$S^{0.6} + S$		$S^{0.6} + p$		$S + S^{0.6}$		$S + p$	
Exp.	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**
F07	0.08	1.0000	0.03	1.0000	**11.87**	**0.0116**	**11.85**	**0.0116**
W12	5.25	0.2676	6.05	0.1827	0.49	1.0000	0.59	1.0000
SF22	1.50	1.0000	0.82	1.0000	7.69	0.0813	7.06	0.1104
H19	2.91	0.9155	0.59	1.0000	**13.1**	**0.0064**	**9.95**	**0.0254**
S22	0.05	1.0000	0.27	1.0000	2.55	1.0000	3.03	0.867
N18	6.26	0.1652	**9.31**	**0.0349**	0.01	1.0000	0.08	1.0000

	$p + S$		$p + S^{0.6}$		$(S + p) + S^{0.6}$		$S^{0.6} + (S + p)$	
Exp.	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**
F07	**67.88**	**<0.0001**	**67.85**	**<0.0001**	0.13	1.0000	0.20	1.0000
W12	**28.46**	**<0.0001**	**29.16**	**<0.0001**	0.14	1.0000	5.49	0.6927
SF22	**10.73**	**0.0182**	**10.67**	**0.0184**	0.71	1.0000	1.58	1.0000
H19	**19.11**	**0.0003**	**19.94**	**0.0002**	6.19	0.1697	5.95	0.5716
S22	**10.96**	**0.0164**	**10.71**	**0.0182**	0.47	1.0000	1.00	1.0000
N18	**51.87**	**<0.0001**	**54.85**	**<0.0001**	1.76	1.0000	8.09	0.2217

The results for the predictors calculated using GPT-J 6B are presented in Table 10.3. As previously noted, all p-values were corrected for multiple comparisons using the stringent false discovery rate method proposed by Benjamini and Yekutieli (2001), ensuring that we report only the most robust effects. The tables for the other 5 models are provided in subsection 10.10.4.

First, and perhaps least surprising, is the finding that in all the datasets both surprisal and surprisal$^{0.6}$ explain a significant amount of variance in N400 amplitude above and beyond that explained by probability. This is in line both with previous work and

the results presented in Figure 10.4, where we see that probability is a substantially worse predictor on its own than the other two metrics. As can be seen in Figure 10.4, this is also the case with the other 4 language models. However, there is one exception to this pattern—the difference for XGLM 7.5B on the Szewczyk and Federmeier (2022) dataset is not significant.

Next, we see that probability explains variance in N400 amplitude above and beyond that explained by surprisal$^{0.6}$ on the Nieuwland et al. (2018b) dataset. This is also the case with probabilities calculated using BLOOM 7.1B and XGLM 7.5B, but not OPT 6.7B or GPT-2 345M. It is also worth noting that on the same dataset, BLOOM 7.1B surprisal explains variance in N400 amplitude above and beyond that explained by surprisal$^{0.6}$. As can be seen in Figure 10.4, in each of these cases, the the combined surprisal$^{0.6}$ and probability regression has a lower AIC than the equivalent combined surprisal and probability regression, with the difference in AIC exceeding 4 in the case of OPT 6.7B.

We also see that both probability and surprisal$^{0.6}$ explain variance above and beyond that explained by surprisal on the Federmeier et al. (2007) and Hubbard et al. (2019) datasets. This is also true for other models with the exception of BLOOM 7.1B probabilities, with which only the Federmeier et al. (2007) dataset shows the effect. In addition, the effect is also found on the Szewczyk and Federmeier (2022) dataset for XGLM 7.5B probabilities.

Finally, while we see differences between how well they predict N400 amplitude numerically in Figure 10.4, after correction for multiple comparisons, we do not find surprisal$^{0.6}$ to explain any variance not already explained by both surprisal and probability, and neither do we find the reverse. This suggests that surprisal$^{0.6}$ and the combination of surprisal and probability explain very similar variance in N400 amplitude.

10.5.4 Discussion

Several clear results arise from these analyses. First, surprisal predicts N400 amplitude better than probability does. For five language models—each of which is the model that best predicts N400 amplitude on at least one dataset—surprisal explains a significant amount of the variance in N400 amplitude not explained by probability in every dataset, even after correction for multiple comparisons. Thus, these results replicate and expand on those of Szewczyk and Federmeier (2022). Specifically, Szewczyk and Federmeier (2022) find that GPT-2 1.5B surprisal explains variance in N400 amplitude not explained by probability on a combined dataset made up of four datasets (Wlotko and Federmeier, 2012; Szewczyk and Federmeier, 2022; Hubbard et al., 2019; Szewczyk et al., 2022), as well as for the unexpected (cloze $\leq 5\%$) completions in the Federmeier et al. (2007) dataset. We find the same pattern to hold when all the data from each of these datasets (as well as Federmeier et al., 2007 and Nieuwland et al., 2018b) are analyzed separately using the probabilities calculated from five additional language models (GPT-J 6B, BLOOM 7.1B, OPT 6.7B, XGLM 7.5B, and GPT-2 345M). Thus, our results suggest that this result is generalizable across language models and on at least one entirely new dataset (Nieuwland et al., 2018b).

We also find that for two of the datasets (Federmeier et al., 2007; Hubbard et al., 2019), probability conversely explains variance in N400 amplitude not explained by surprisal. Szewczyk and Federmeier (2022) report this for their aforementioned combined dataset, as well as the expected completions from Federmeier et al. (2007). Our results expand upon the latter finding. This effect is present with the probabilities calculated using five additional language models, even when considering all data from Federmeier et al. (2007)—i.e., not just expected completions—and even when correcting for multiple comparisons. Our results also pinpoint the Hubbard et al. (2019) dataset as a likely source

of the effect on the combined dataset—we see that the effect of probability significantly predicts N400 amplitude even when surprisal is accounted for with four of the five models.

These results raise the question of why we only see this pattern for two of the six datasets. One possibility is the experimental stimuli themselves—of the five datasets for which the stimulus selection process is reported, only Hubbard et al. (2019) directly use stimuli from Federmeier et al. (2007) without adding additional stimili; and thus there may be something about these stimuli that leads to the effect being detected. Whether this is because the effect is present but undetectable in the other stimuli or is caused by some as yet unidentified feature of the Federmeier et al. (2007) stimuli is a question for further research. However, it is worth noting that one piece of evidence in favor of the former is that one of the datasets that does not show this effect is the Nieuwland et al. (2018b), which has the smallest number of items and is predominantly made up of high-constraint sentences only, and the other is Wlotko and Federmeier (2012), which has the smallest number of experimental participants.

We also find two novel results based around surprisal$^{0.6}$. First, we find that like surprisal, surprisal$^{0.6}$ explains a significant amount of the variance in N400 amplitude not explained by probability across virtually all language models and datasets. In addition, we find that for the datasets where probability explains the variance in N400 amplitude above and beyond that explained by surprisal, surprisal$^{0.6}$ also explains variance not explained by surprisal; and in addition, probability does not explain variance not explained by surprisal$^{0.6}$. In fact, with the exception of the Nieuwland et al. (2018b) dataset, surprisal$^{0.6}$ explains all the statistically significant variance explained by both surprisal and probability across models and datasets. It is still important to note, however, that with the Nieuwland et al. (2018b) dataset, GPT-J 6B, BLOOM 7.1B, and XGLM 7.5B (but not OPT 6.7B or GPT-2 345M) probability explains variance in N400 amplitude not explained by

surprisal$^{0.6}$, as does BLOOM 7.1B surprisal. Given the aforementioned limitations of the Nieuwland et al. (2018b) dataset and the fact that this is only the case for three of the five language models on one of the six datasets tested, it is possible that this is simply an anomalous result, but this is something that would need to be tested by running similar analyses on a larger number of additional datasets. Taken together, our results for surprisal$^{0.6}$ may be taken to suggest that the individual effects of probability and surprisal reported by Szewczyk and Federmeier (2022) and replicated in our work could instead be empirically accounted for by a single sub-logarithmic relationship (i.e., $(-log(p))^{0.6}$ between language model probability and N400 amplitude. However, we are not aware of any previous theoretical work predicting such a relationship.

Overall, the results of this analysis showed three things. First, we replicate and expand Szewczyk and Federmeier's (2022) finding that language model surprisal explains variance in N400 amplitude above and beyond that explained by probability. We also replicate and extend the finding that probability explains variance not explained by probability on the Federmeier et al. (2007) and Hubbard et al. (2019) datasets. Finally, we found both of these effects are captured by a single variable—surprisal$^{0.6}$. In all cases, all the variance explained by surprisal that is not explained by probability is explained by surprisal$^{0.6}$, and all variance in Federmeier et al. (2007) and Hubbard et al. (2019) explained by probability that is not explained by surprisal is explained by surprisal$^{0.6}$. In fact, we see that on 27 of the 30 combinations of language models and datasets, surprisal$^{0.6}$ explains all the variance in N400 amplitude explained by either surprisal, probability, or their combination. Finally, in the 3 remaining cases, the combination of surprisal$^{0.6}$ and probability predicts N400 amplitude at least as well as the combination of surprisal and probability.

10.6 Interim General Discussion

The results of Analyses 1-3 show that the best single predictor of N400 amplitude is sub-logarithmically transformed probability, and that the same variance is explained by a combination of probability and surprisal. Thus, the empirical results can be considered to equally support a sub-logarithmic relationship or the combined relationship (in line with the multiple sub-component account of Szewczyk and Federmeier, 2022). Given the lack of any previously-proposed theory accounting for the sublogarithmic relationship, in this paper we focus on this new result and how it compares to Szewczyk and Federmeier's (2022) multiple sub-component account. Thus, we do not consider the additional possibilities of multiple sub-component accounts involving a sub-logarithmic relationship as well as either surprisal, probability, or both.

The present study shows the difficulty in distinguishing empirically between the sublogarithmic and multiple sub-component relationships—both predict N400 amplitude well. Both also explain seemingly disparate findings in previous work. For example, one well-established phenomenon is that N400 amplitude can differ greatly between words with matched cloze probabilities, especially between low (or zero) cloze items (Federmeier and Kutas, 1999; Metusalem et al., 2012; DeLong et al., 2019). More recent work, however, suggests that at least some of this variance in N400 amplitude can be captured by language model surprisal (Michaelov and Bergen, 2020, 2022a). Surprisal's success with low-probability words likely derives from the fact that it emphasizes differences in probability at the low end of the scale. But by the same token, it also reduces the differences at the high end of the probability scale. Empirical results suggests that this could be a problem for modeling the N400—indeed, part of the empirical motivation for Szewczyk and Federmeier's (2022) multiple sub-component account is that language model probability better predicts the amplitude of the N400 response elicited by high-probability items than

language model surprisal does. The question, then, is how to determine whether the theoretically unexplained but parsimonious sublogarithmic relationship or the theoretically-motivated but more complex multiple sub-component account is more strongly supported by the evidence. This is the aim of the remainder of this paper.

10.6.1 Towards a multiple sub-component account

First, we consider the evidence that exists or would be need to exist to support the multiple sub-component account as proposed by Szewczyk and Federmeier (2022). This account has three aspects: the idea that the N400 is sensitive to both contextual predictability and similarity (or more specifically, overlap of semantic features in long-term memory); the empirical result that the N400 can be predicted by contextual probability and surprisal; and the linking hypothesis that the effect of contextual predictability on N400 amplitude can be operationalized by a linear relationship between contextual probability and N400 amplitude and that the effect of contextual similarity on N400 amplitude can be operationalized by a logarithmic relationship between contextual probability and N400 amplitude (i.e., a linear relationship between surprisal and N400 amplitude).

There is some evidence for the first of these. Lau et al. (2013), for example, investigate this question through experimental manipulation. Their study used the well-established word-pair priming paradigm under which words preceded by related words elicit smaller N400 responses than words preceded by unrelated words (see, e.g., Bentin et al., 1985; Rugg, 1985; Holcomb, 1988; Kutas and Hillyard, 1989; Kutas, 1993). Lau et al. (2013) found that participants who had been presented with a larger proportion of trials where the words were related showed increased reductions in the N400 for associated targets, as well as a difference in onset latency and topographic distribution of the priming effect compared to participants who were presented with fewer related word pairs. Lau et al.

(2013) follow previous work (Neely, 1977; Becker, 1980) in arguing that the greater degree of predictive validity in the high relatedness proportion stimuli is indicative of increased predictive processing, and thus, that the differences in the N400 effects between relatedness proportions are due to an effect of a predictive processes. As noted, the differences based on this manipulation are dissociable in latency and topographic distribution from the baseline N400 effects in word-pair priming that are generally thought to arise from associative processes, and thus, the effects of prediction and association are argued to be distinct and hypothesized to possibly arise from 'qualitatively different' neurocognitive processes (Lau et al., 2013).

An alternative approach is to test how much variance is explained either by contextual predictability or semantic feature overlap. In studies of this type (Parviz et al., 2011; Frank and Willems, 2017), in addition to using the predictions of language models (or augmented language models) to operationalize predictability, researchers use the contextual similarity of word vectors to model semantic feature overlap to model the N400 (Frank and Willems, 2017). Specifically, Parviz et al. (2011) used latent semantic analysis (LSA; Dumais et al., 1988; Landauer et al., 1998) to calculate word vectors representing the semantics of all the words in the experimental stimuli, operationalizing semantic feature overlap as the cosine distance between the word vector representing the critical word and a context vector made up of the elementwise product of all word vectors in the context. Frank and Willems (2017), on the other hand, use their own implementation of *word2vec* (Mikolov et al., 2013a,b) to calculate word vectors, basing their metric of 'semantic distance' on the cosine similarity of the critical word and the sum of the vectors of the words in its context. In both cases, semantic feature overlap was found to predict variance in N400 amplitude above and beyond predictability, supporting the idea that the two may arise from distinct sub-processes. It is also worth noting that there are a number

of other studies that have been argued to directly or indirectly support this perspective (see Federmeier, 2021 for review).

The second aspect of the multiple sub-component account presented by Szewczyk and Federmeier (2022) is an empirical one—the finding that probability and surprisal can explain separate variance in N400 amplitude. Szewczyk and Federmeier (2022) provide evidence of this for GPT-2 1.5B (XL) probability, and we provide evidence for this for BLOOM 7.1B, GPT-2 345M (Medium), GPT-J 6B, OPT 6.7B, and XGLM 7.5B.

The third and final aspect of the account—the linking hypothesis—is less well-evidenced, however. Intuitively, given the fact that both language model predictions and aggregated cloze responses can be formulated as probabilities, it seems natural to stipulate that they could be linearly related. And while there does not appear to be a large difference in the extent to which un-transformed or log-transformed cloze probabilities predict N400 amplitude (Aurnhammer et al., 2021; Michaelov et al., 2022; Szewczyk and Federmeier, 2022), the evidence appears to support a linear relationship between cloze probability and reading time (Brothers and Kuperberg, 2021), leading Brothers and Kuperberg (2021) to argue that cloze probabilities closely reflect the extent to which words are predicted in the brain (see also Smith and Levy, 2011) and that processing difficulty is linearly related to this degree of proportional (predictive) preactivation. The account presented by Szewczyk and Federmeier (2022) follows this intuition in that it argues that the extent to which words are predicted is linearly related to N400 amplitude. However, while Szewczyk and Federmeier (2022) quantify the correlation between cloze probability and GPT-2 1.5B probability for expected (cloze probability > 0.05) words ($r = 0.72$), this is not compared to the degree of correlation between GPT-2 1.5B surprisal and cloze probability in the same range. Thus, in principle, even if we fully accept the proportional preactivation account of Brothers and Kuperberg (2021), it is not a given that language model probability is more related

to these 'subjective probabilities' (Smith and Levy, 2011) that have been argued to be linearly related to processing difficulty (Brothers and Kuperberg, 2021) than is language model surprisal, for example. This is an empirical question that can be directly tested, and we do, in Analysis 4.

There is also limited evidence for the second part of the linking hypothesis, namely, that language model surprisal can operationalize the degree of semantic featural overlap between a critical word and its context. The main issue is that the composite processing difficulty account provided by Szewczyk and Federmeier (2022) could in principle apply to any featural overlap between a word and its context. Indeed, the explanation is the same as that given in the original version of the account presented by Smith and Levy (2013) but with word fragments replaced by semantic features. Thus, while it is in principle a plausible account, there is no direct evidence that contextual similarity is well-correlated with language model surprisal, even if the latter does in fact in part reflect the former. One possible indirect piece of evidence for this component is found in Michaelov et al. (2024), where the semantic featural overlap between a word and its context was operationalized as the cosine similarity between the GloVe (Pennington et al., 2014) or fastText (Mikolov et al., 2018) word embedding for the critical word and the mean of the embeddings of the words in the context. Crucially, GPT-3 surprisal was found to fully account for the variance in N400 amplitude explained by either metric of semantic featural overlap, which is consistent with the idea that the two are strongly related, as proposed by Szewczyk and Federmeier (2022). However, it is worth noting that it is also the case that GPT-3 surprisal explains all the variance explained by cloze probability (Michaelov et al., 2024); and thus, the evidence for the linking hypthosis is far from conclusive. We thus also investigate this relationship in Analysis 4.

10.6.2 Towards a sublogarithmic account

As previously noted, no current theoretical account predicts a sublogarithmic relationship between contextual probability and any metric of processing difficulty, including the N400. Despite this, not all the accounts are inconsistent with such a relationship between statistical probability (as operationalized by language model probability) and the N400. This becomes clear if we consider the two accounts that explicitly posit that words are preactivated in the brain to an extent that correlates with contextual probability, namely, the contextual probability (Brothers and Kuperberg, 2021) and distribution update (Aurnhammer and Frank, 2019b) accounts. The linear (in the case of the contextual probability account) and logarithmic (in the case of the distribution update account) relationships described by these accounts are posited to be a result of these relative degrees of preactivation, which can be mathematically described as a probability distribution over candidate stimuli—the distribution of 'subjective probabilities' (Smith and Levy, 2011). Crucially, then, while the aforementioned three accounts make claims about the relationship between subjective probabilities and processing difficulty, they do not make any claim about the relationship between statistical probabilities and subjective probabilities.

Statistical probability does not directly correspond to cloze probability (for discussion, see Smith and Levy, 2011; Brothers and Kuperberg, 2021; Michaelov et al., 2022), and so it is possible that the subjective probability of a word in context is represented in the brain such that there is a nonlinear relationship between it and the word's statistical probability. In fact, such a relationship may be indirectly supported by previous work. Several researchers have argued that the same linguistic representations that underlie the predictions that occur during language comprehension are likely to be those drawn on when responding to the cloze task (Smith and Levy, 2011; Brothers and Kuperberg, 2021). It is therefore perhaps unsurprising that Brothers and Kuperberg (2021) find that the re-

lationship between cloze and processing difficulty is best described as linear. Meanwhile, however, recent empirical work has established the relationship between language probability and behavioral metrics of processing difficulty such as reading time to be decidedly non-linear (Smith and Levy, 2013; Shain et al., 2024; Wilcox et al., 2023b; Meister et al., 2021; Hoover et al., 2023).

Thus, it is perfectly possible that cloze probability and language model probability may have different relationships with N400 amplitude, and that the relationship between statistical probability (as approximated by language mode probability) and subjective word probability (as approximated by cloze probability) may be nonlinear. We test the viability of this possibility by investigating the relationship between language model probability and cloze probability.

The first thing to consider is that, as noted above, while behavioral evidence suggests a linear relationship between subjective probability and reading time (see Brothers and Kuperberg, 2021), the precise relationship between subjective probability and N400 amplitude is less clear. In Michaelov et al. (2022), cloze surprisal (i.e., log-transformed cloze probability) is found to predict N400 amplitude slightly better than un-transformed cloze probability in sentences where cloze > 0. Szewczyk and Federmeier (2022), on the other hand, report a slightly better performance from cloze probability for stimuli where cloze > 0.05.

For this reason, in the present study, we consider both possibilities—that the relationship between subjective probability and N400 amplitude is linear (as under the contextual probability account), and that the relationship is logarithmic (as under the distribution update account). Taken together, therefore, there are two possible ways in which the relationship between statistical probability and N400 amplitude is sublogarithmic.

The first of these is that the relationship between statistical probability and sub-

jective probability is sublogarithmic, and the relationship between subjective probability and N400 amplitude is linear, as under the contextual probability account. This would explain the evidence we find for a sublogarithmic relationship between language model probability and N400 amplitude. Given that this possibility involves a linear relationship between subjective probability and N400 amplitude and our finding that the best single characterization of the relationship between language model probability and N400 amplitude is linearly related to $(-\log p)^{0.6}$ (i.e., surprisal$^{0.6}$), we test how closely language model surprisal$^{0.6}$ correlates with cloze probability compared to un-transformed probability and surprisal.

Next we turn to the other possibility, namely, that in addition to a sublogarithmic relationship between statistical probability and N400 amplitude, there is also a logarithmic relationship between subjective probability and N400 amplitude, in line with the distribution update account. In this case, then, the relationship between language model probability and cloze probability would be linearly related to $e^{(-\log p)^{0.6}}$ (i.e., language model $e^{\text{surprisal}^{0.6}}$). We again compare how closely this correlates with cloze probability.

In summary, while our empirical results suggest that there may be a sublogarithmic relationship between statistical probability and N400 amplitude, this is not something that has been suggested in any previous work. However, most accounts theorize based on a relationship between subjective probability and N400 amplitude, which, as we note, is not the same as statistical probability. If cloze probability shows a linear relationship to language model surprisal$^{0.6}$ (in line with the contextual probability account) or language model $e^{\text{surprisal}^{0.6}}$ (in line with the distribution update account), this would provide independent empirical support for the sublogarithmic relationship between statistical probability and N400 amplitude. Future work would be needed to investigate why such a nonlinearity might characterize (or at least approximate) the relationship between statistical probability

and subjective probability; but to the best of our knowledge, only one study (Smith and Levy, 2011) has thus far studied the relationship between the two, and consequently, substantial work is needed in this area regardless. We hope that the investigations in Analysis 4 can help to further this much-needed line of research.

10.7 Analysis 4: Correlations of predictors

10.7.1 Introduction

In this analysis, we ask whether it is possible to determine whether the multiple sub-component account of Szewczyk and Federmeier (2022) or the sublogarithmic account proposed in the Interim Discussion is better supported by the evidence. As noted in the Interim Discussion, under the multiple sub-component account of Szewczyk and Federmeier (2022), we should expect that un-transformed statistical probability (operationalized by language model probability) is linearly related to subjective probability (operationalized as cloze probability) and language model surprisal is logarithmically related to semantic featural overlap with the context (as operationalized based on similarity between word vectors). Meanwhile, if the sublogarithmic account is true and the sublogarithmic relationship approximates a component of the relationship between statistical probability and subjective probability, then the relationship between language model probability and subjective probability should be either sublogarithmic ($(-\log p)^{0.6}$; i.e., $\text{surprisal}^{0.6}$) under the contextual probability account or an exponentation of this ($e^{(-\log p)^{0.6}}$; i.e., $e^{\text{surprisal}^{0.6}}$) under the distribution update account. This is what we test in this section.

10.7.2 Method

The N400 data and language models are the same as in Analysis 3. We also utilize the cloze probabilities provided by the authors (Nieuwland et al., 2018b; Szewczyk and Federmeier, 2022) for each item.

In addition, we calculate the featural overlap between critical words and their preceding context. To do this, we follow Ettinger et al. (2016) and Michaelov et al. (2024), calculating the cosine similarity between the mean of the word embeddings of the context and the word embedding of the critical word. Based on their performance at predicting N400 amplitude in previous work (Michaelov et al., 2024), we use fastText (Joulin et al., 2017) word embeddings. Departing from Michaelov et al. (2024), we use the 300-dimensional embeddings trained on a combination of English Common Crawl and Wikipedia data in Grave et al. (2018) because the authors provide the original model, allowing the word embedding of any critical or context word to be calculated.

10.7.3 Results

Correlation with cloze

First, we test how closely cloze probability correlates with the probability, surprisal, surprisal$^{0.6}$, and $e^{\text{surprisal}^{0.6}}$ calculated from each language model. To do this we calculate the absolute value of the correlation coefficient (Pearson's r) of cloze probability and each of the aforementioned variables. Following Szewczyk and Federmeier (2022), we only look at words with a cloze probability of greater than 0.05. The results are shown in Figure 10.5.

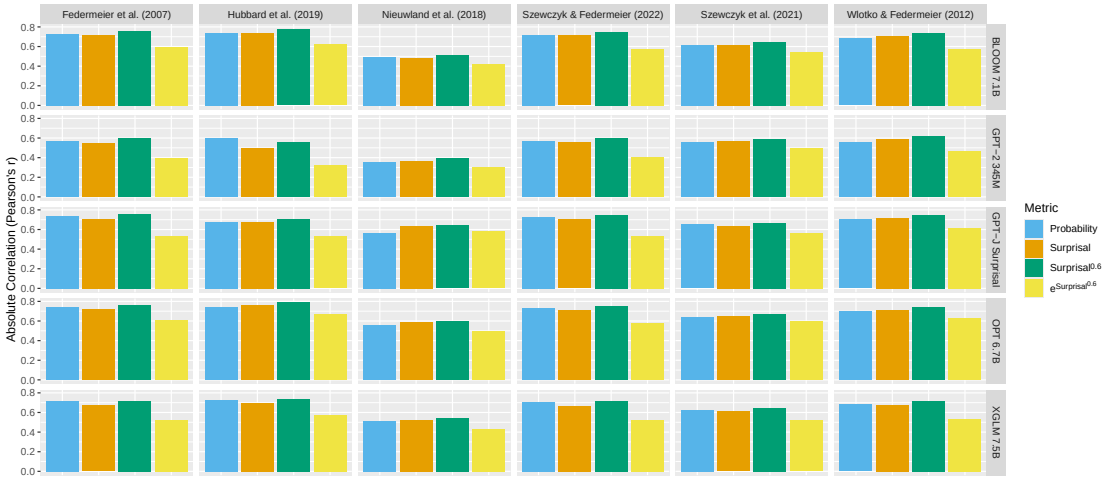

Figure 10.5: The absolute correlation coefficient between the probability, surprisal, surprisal$^{0.6}$, and $e^{surprisal^{0.6}}$ calculated from each language model and cloze probability. This analysis only includes data from stimuli with a cloze probability greater than 0.05.

These results align reasonably well with our formulation of Szewczyk and Federmeier's (2022) multiple sub-component account. Under this account, we would expect probability to correlate more closely with cloze probability than surprisal does. This is indeed the case for a narrow majority of the comparisons—19 of the total 30. Intriguingly, the exceptions seem to align with those found in the previous analyses of this paper—8 of the 11 cases where surprisal is more closely correlated with cloze probability are the Nieuwland et al. (2018b) and Wlotko and Federmeier (2012) datasets. Given that this analysis was carried out on the properties of the stimuli alone, this suggests that the fact that these datasets show different patterns in the N400 to the other datasets is likely due to the stimuli themselves.

We also tested how closely surprisal$^{0.6}$ and $e^{\text{surprisal}^{0.6}}$ correlate with cloze probability. The results of this comparison are clearer—with the exception of GPT-J 6B surprisal on the Nieuwland et al. (2018b) dataset, $e^{\text{surprisal}^{0.6}}$ is least closely correlated with cloze probability; and in all cases, surprisal$^{0.6}$ is mostly closely correlated with cloze probability.

Overall, then, the results of this analysis most strongly support the sub-logarithmic variant of the proportional preactivation account.

Correlation with contextual similarity

Figure 10.6: The absolute correlation coefficient between the probability, surprisal, surprisal$^{0.6}$, and $e^{surprisal^{0.6}}$ calculated from each language model and contextual similarity.

Second, we test how closely contextual similarity correlates with each of the same model-derived metrics. Again, we begin by considering how well these results align with our formulation of Szewczyk and Federmeier's (2022) multiple sub-component account. Under this account, we would expect surprisal to correlate more closely with contextual similarity than probability does. This is the case for 28 of the total 30 comparisons. Taken together with the previous results, this supports the idea that language model probability better captures lexical prediction (operationalized by cloze probability) than surprisal, and surprisal (to a greater extent) better captures contextual similarity.

However, it is also worth noting that surprisal$^{0.6}$ is always more closely correlated with contextual similarity than probability, and is more closely correlated than surprisal

in 20 of the 30 comparisons.

10.7.4 Discussion

The results of this analysis are mixed. The fact that in almost all (28 out of 30) comparisons, language model surprisal is more strongly correlated with contextual similarity than probability supports Szewczyk and Federmeier's (2022) claim that surprisal may operationalize the semantic featural overlap between a critical word and its context. The fact that language model probability is more strongly correlated with cloze probability in 19 out of 30 comparisons also supports the account, though less strongly.

We also see that $\text{surprisal}^{0.6}$ is generally the variable most strongly correlated with cloze probability, while $e^{\text{surprisal}^{0.6}}$ shows the weakest correlation. Thus, our results suggest if there is a sublogarithmic relationship between statistical probability and the N400, the sublogarithmic relationship lies between statistical and subjective probability. For this reason, they also provide indirect evidence for the contextual probability account in general, as the same mathematical relationship best characterizes both the relationship between language model probability and cloze probability and between language model probability and N400 amplitude.

Finally, we consider the extent to which these results shed light on the question of whether the multiple sub-component account or the sublogarithmic account better explains the data. On most of the models and datasets, we see the same patterns. First, we see that language model surprisal tends to be more correlated with contextual similarity than language model probability is. In addition, we see that language model probability tends to be more correlated with cloze probability than language model surprisal is. These findings are both in line with what we would expect to see under the multiple sub-component account. However, the evidential support for the account is somewhat undercut by the

fact that surprisal$^{0.6}$ is a better predictor of both. Thus, the results provide some evidence in favor of each possible account, without providing clear support for one over the other.

The fact that surprisal$^{0.6}$ is the best predictor of both our metric of semantic feature overlap (contetxual similarity) and lexical prediction (cloze probability) suggests that cloze probability may not clearly reflect lexical prediction as distinct from semantic feature overlap. This is further supported by previous work showing a correlation between contextual similarity and cloze probability (Michaelov et al., 2024) as well as between word association and cloze probability (Smith and Levy, 2011). Indeed, in the datasets we analyze, cloze probability and contextual similarity are generally weakly to moderately correlated (Federmeier et al. (2007): $r = 0.344$; Hubbard et al. (2019): $r = 0.294$ Nieuwland et al. (2018b): $r = 0.244$ Szewczyk and Federmeier (2022): $r = 0.343$ Szewczyk et al. (2022): $r = 0.179$ Wlotko and Federmeier (2012): $r = 0.301$). Whether this has implications for the extent to which semantic feature overlap and lexical prediction are dissociable in general is a question for future work.

10.8 General Discussion

The work described in this paper is novel in several ways. To the best of our knowledge, it is the first to compare linear, logarithmic, super-logarithmic, and sub-logarithmic relationships between probability and N400 amplitude in explaining experimental data. The key and surprising finding is that on the whole, sub-linearly-transformed surprisal is a better predictor of N400 amplitude than surprisal is—this is the case for four out of six datasets when comparing fit to the data across models. Moreover, we observed that when calculating the metrics using the model that best predicts N400 amplitude for each dataset, sub-linearly-transformed surprisal almost always explains all the variance explained by surprisal and probability. This is in contrast to previous work on reading time, where to date

there is no clear consensus on whether a linear, sub-linear, or super-linear relationship with surprisal best explains the data (Smith and Levy, 2013; Meister et al., 2021; Shain et al., 2024; Wilcox et al., 2023b).

Beyond this new finding, the current results also replicate and expand upon previous work. First, in line with Yan and Jaeger (2020) and Szewczyk and Federmeier (2022), surprisal out-performed un-transformed probability as a single predictor of N400 amplitude—across language models, using surprisal as a predictor results in significantly better-fitting linear mixed-effects models than probability. We show that this is not just the case for the Frank and Willems (2017) model used by Yan and Jaeger (2020) and for the GPT-2 (Radford et al., 2019) model used by Szewczyk and Federmeier (2022), but that it occurs generally across contemporary autoregressive transformer language models. While there are individual exceptions, overall across datasets, language model surprisal out-performs un-transformed probability.

The results reported here also further support the finding, reported by Szewczyk and Federmeier (2022), that language model probability can explain variance in N400 amplitude above and beyond that explained by surprisal. Szewczyk and Federmeier (2022) found this using GPT-2 1.5B; our results similarly show that for the Federmeier et al. (2007) and Hubbard et al. (2019) datasets, GPT-J 6B, GPT-2 345M, OPT 6.7B, and XGLM 7.5B probability explain variance in N400 not explained by the surprisal of the same language model; and the same is true for BLOOM 7.1B on the Federmeier et al. (2007) dataset. This suggests that the significant effect of probability (above and beyond surprisal) on N400 amplitude is not due to possible idiosyncrasies of GPT-2 1.5B, but may instead reflect a more general trend. As previously noted, however, the additional variance explained by probability can generally also be accounted for by a sub-linear transformation of surprisal.

It is worth briefly noting that previous work has generally shown that language models that are trained on more data and that are better at natural language processing tasks overall also tend to perform best at predicting N400 amplitude (Frank et al., 2015; Aurnhammer and Frank, 2019a,b; Michaelov and Bergen, 2020; Merkx and Frank, 2021; Michaelov et al., 2021, 2022). On the other hand, several recent studies have suggested that increasing model scale past a certain point leads to decreased performance in modeling reading time (Oh et al., 2022; Shain et al., 2024; de Varda and Marelli, 2023; Oh and Schuler, 2023b). The results of the present study are more in line with the former set of findings—the models that best predict N400 amplitude on five out of six datasets (GPT-J 6B, OPT 6.7B, BLOOM 7.1B, XGLM 7.5B) are among the largest we tested, both in terms of parameter count (6-7.5 billion parameters) and training data size (180 billion to 500 billion tokens).

10.8.1 Theoretical Implications

The results show that the single best predictor of N400 amplitude across all datasets is surprisal$^{0.6}$, that surprisal is generally a better predictor than un-transformed probability, and that, in line with the results of Szewczyk and Federmeier (2022), probability can explain variance in N400 amplitude not explained by surprisal. Overall, the empirical results are most consistent with either the multiple sub-component account of Szewczyk and Federmeier (2022) where there is both a linear and logarithmic relationship between statistical probability and N400 amplitude, or with a sub-logarithmic variant of the proportional preactivation account, where there is a sub-logarithmic relationship between statistical probability and the extent to which words are preactivated, with a linear relationship holding between this preactivation and the N400.

However, there is no clear single relationship suggested by the study. As noted, no

set of metrics derived from language model probability consistently predict N400 amplitude better than others in all cases. While surprisal$^{0.6}$ is the best single predictor of N400 amplitude, the probability and surprisal calculated by some language models can explain additional variance on one of the datasets (Nieuwland et al., 2018b).

We also see a similar lack of clarity when we compare how well different transformations of language model probability correlate with metrics thought to correlate with the factors that Szewczyk and Federmeier (2022) argue play important, dissociable roles in the neurocognitive processes underlying the N400—contextual similarity, which has been argued to model semantic feature overlap (Michaelov et al., 2024), and cloze probability, which can be used to model lexical prediction (Brothers and Kuperberg, 2021). Language model surprisal correlates more closely with the cosine similarity between the embeddings of critical words and their contexts than language model probability does, suggesting that it better reflects semantic feature overlap, in line with the account of Szewczyk and Federmeier (2022). However, we also find that in the majority of cases, surprisal$^{0.6}$ is more correlated with this metric of semantic feature overlap than is surprisal, which presents a problem for the account. Similarly, language model probability is more closely correlated than language model suprisal with cloze probability, a metric of lexical prediction, in the majority of cases, which is also in line with the account of Szewczyk and Federmeier (2022). Again, however, surprisal$^{0.6}$ is more closely correlated with cloze probability—in fact, this is so for all comparisons of language models and datasets carried out. The fact that surprisal$^{0.6}$ is more closely correlated with cloze probability and contextual similarity than the other metrics may be taken to suggest that the latter two metrics are correlated, and indeed we find that they are, but to less of an extent than surprisal$^{0.6}$ is to either.

Taken together, the results defy a single straightforward conclusion as to the mathematical relationship between language model probability and N400 amplitude. However,

we believe that the present study has brought us closer to characterizing this relationship by identifying several of its features. First, the relationship between language model probability and N400 amplitude does not appear to be a simple linear or logarithmic one. Second, the results suggest that the N400 is more sensitive to differences at the lower end of the scale than can captured by probability (and thus closer to those reflected in surprisal or surprisal$^{0.6}$), but also more sensitive to the differences at the higher end of the probability scale than can be captured by surprisal (and thus closer to those reflected in probability or surprisal$^{0.6}$). Finally, in our sample at least, we see that the metric that best predicts N400 amplitude alone is one that reflects both contextual predictability (as operationalized by cloze) and semantic feature overlap (as operationalized by contextual similarity; see also Michaelov et al., 2021).

This last finding is additionally worth highlighting because in virtually all previous work suggesting a nonlinear relationship between language model probability and processing difficulty, the nonlinearity is proposed to arise based on how subjective probabilities lead to differences in processing difficulty. Our results, however, suggest that the nonlinearity—which we find to sub-logarithmic—instead lies between statistical probabilities (as operationalized by language model probabilities) and these subjective probabilities.

10.9 Conclusions

In this study, we set out to compare how well linear, logarithmic, and exponentiated (i.e., super- and sub-) logarithmic transformations of contextual probability correlate with the N400, a neural index of processing difficulty. In line with previous work (Yan and Jaeger, 2020; Szewczyk and Federmeier, 2022), we find that that surprisal, a logarithmic transformation of probability, out-performs probability as a predictor of N400 amplitude. However, as has previously been reported (Szewczyk and Federmeier, 2022), we find that

probability can explain variance in N400 amplitude not explained by surprisal.

Our novel finding, and one that lies in contrast to previous work on reading time as an index of processing difficulty (Smith and Levy, 2013; Meister et al., 2021; Shain et al., 2024; Wilcox et al., 2023b), is that sub-logarithmically transformed probability is a better predictor of N400 amplitude than surprisal. Specifically, we find for almost all language models and datasets, surprisal$^{0.6}$ explains at least as much variance in N400 amplitude as both surprisal and probability; suggesting that the relationship between probability and N400 amplitude may in fact be sub-logarithmic.

The fact that this result is not accounted for by any previous work highlights the importance of not viewing language processing as a monolith—different metrics of contextual probability (cloze vs. corpus-derived) and different metrics of processing difficulty (reading time vs. the N400) show distinct patterns; and those different patterns may ultimately be key to understanding the processes of language comprehension.

10.10 Appendix

10.10.1 Statistical analysis of regression fit (Analysis 2)

In this appendix, we include the supplementary statistical analyses investigating whether linear mixed-effects regression models with probability, surprisal, or surprisal$^{0.6}$ as predictors are best able to predict N400 amplitude. We compare the AIC of regressions with probability or surprisal as predictors (Table 10.4), probability or surprisal$^{0.6}$ as predictors (Table 10.5), and surprisal or surprisal$^{0.6}$ as predictors (Table 10.6).

Table 10.4: Results of likelihood ratio tests and coefficients of linear mixed-effects models testing the difference in the AIC of regressions using probability or surprisal as a predictor of N400 amplitude on each dataset.

Likelihood ratio test			**Coefficient**		
Exp.	$\chi^2(1)$	$p_{corrected}$	**Estimate**	**SE**	**t**
Nieuwland et al. (2018b)	70.64	<0.0001	-41.91	2.87	-14.58
Federmeier et al. (2007)	109.23	<0.0001	-51.85	2.00	-25.91
Wlotko and Federmeier (2012)	95.82	<0.0001	-22.44	1.05	-21.35
Szewczyk and Federmeier (2022)	27.79	<0.0001	-5.72	0.89	-6.43
Hubbard et al. (2019)	62.70	<0.0001	-18.23	1.42	-12.82
Szewczyk et al. (2022)	76.75	<0.0001	-11.89	0.74	-16.05

Table 10.5: Results of likelihood ratio tests and coefficients of linear mixed-effects regression models testing the difference in the AIC of regressions using probability or surprisal$^{0.6}$ as a predictor of N400 amplitude on each dataset.

Likelihood ratio test			**Coefficient**		
Exp.	$\chi^2(1)$	$p_{corrected}$	**Estimate**	**SE**	**t**
Nieuwland et al. (2018b)	92.46	<0.0001	-37.53	1.85	-20.33
Federmeier et al. (2007)	137.87	<0.0001	-66.15	1.71	-38.69
Wlotko and Federmeier (2012)	110.11	<0.0001	-20.05	0.76	-26.24
Szewczyk and Federmeier (2022)	78.73	<0.0001	-10.86	0.66	-16.54
Hubbard et al. (2019)	106.17	<0.0001	-24.43	0.98	-25.47
Szewczyk et al. (2022)	108.02	<0.0001	-13.41	0.53	-25.47

Table 10.6: Results of likelihood ratio tests and coefficients of linear mixed-effects models testing the difference in the AIC of regressions using surprisal or surprisal$^{0.6}$ as a predictor of N400 amplitude on each dataset.

Likelihood ratio test			**Coefficient**		
Exp.	$\chi^2(1)$	$p_{corrected}$	**Estimate**	**SE**	**t**
Nieuwland et al. (2018b)	13.61	0.0049	4.38	1.08	4.06
Federmeier et al. (2007)	124.42	<0.0001	-14.30	0.45	-32.11
Wlotko and Federmeier (2012)	33.04	<0.0001	2.39	0.33	7.30
Szewczyk and Federmeier (2022)	85.82	<0.0001	-5.14	0.28	-18.42
Hubbard et al. (2019)	53.33	<0.0001	-6.19	0.57	-10.93
Szewczyk et al. (2022)	28.03	<0.0001	-1.52	0.23	-6.48

10.10.2 Language model probability and N400 amplitude plots (Analysis 3)

In this appendix, we plot the relationship between probability, surprisal, or surprisal$^{0.6}$ and N400 amplitude, with probabilities calculated using BLOOM 7.1B (Figure 10.7), OPT 6.7B (Figure 10.8), XGLM 7.5B (Figure 10.9), and GPT-2 345M (Figure 10.10). The plot for GPT-J 6B is included in the main body of the paper (Figure 10.3).

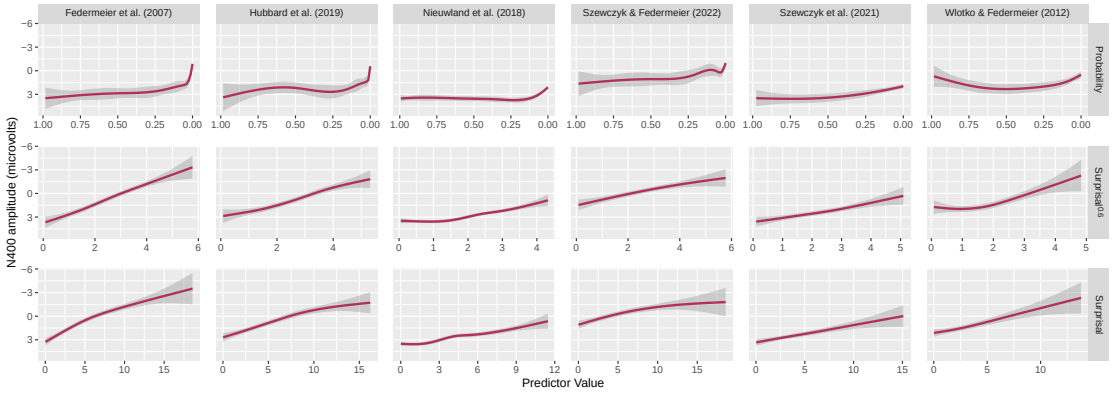

Figure 10.7: N400 amplitude as a function of BLOOM 7.1B probability, surprisal, and surprisal$^{0.6}$. The x-axis for probability is reversed for easier comparison with surprisal and surprisal$^{0.6}$.

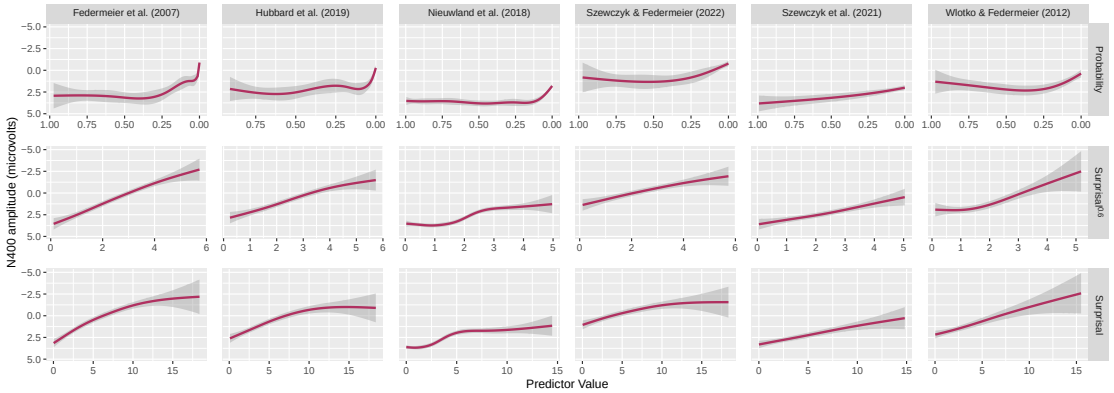

Figure 10.8: N400 amplitude as a function of OPT 6.7B probability, surprisal, and surprisal$^{0.6}$. The x-axis for probability is reversed for easier comparison with surprisal and surprisal$^{0.6}$.

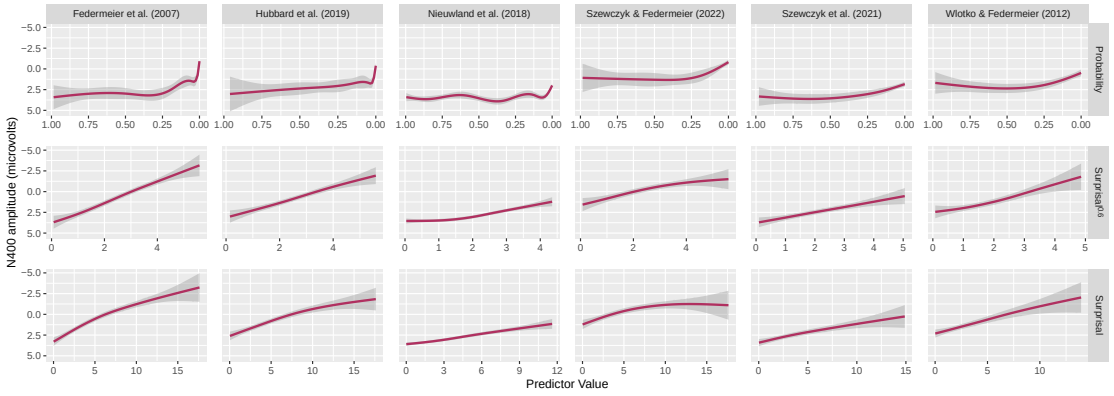

Figure 10.9: N400 amplitude as a function of XGLM 7.5B probability, surprisal, and surprisal$^{0.6}$. The x-axis for probability is reversed for easier comparison with surprisal and surprisal$^{0.6}$.

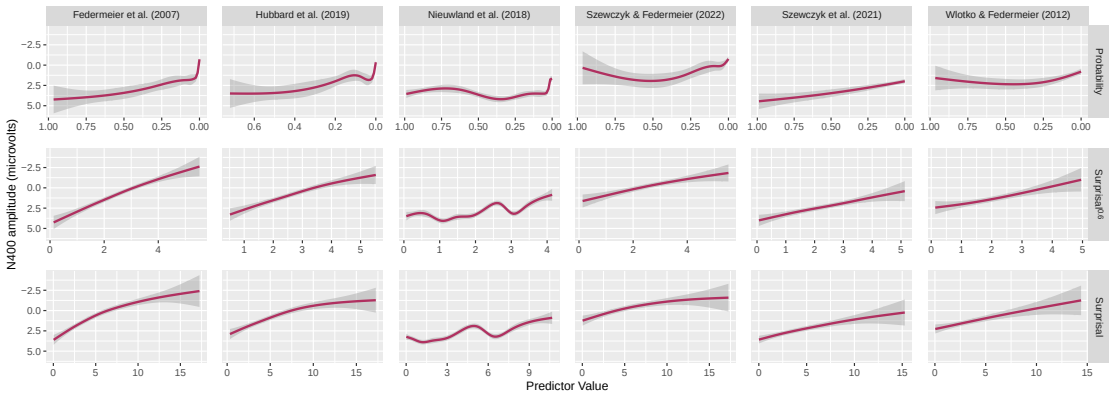

Figure 10.10: N400 amplitude as a function of GPT-2 345M probability, surprisal, and surprisal$^{0.6}$. The x-axis for probability is reversed for easier comparison with surprisal and surprisal$^{0.6}$.

10.10.3 Comparison of regression AICs including probability (Analysis 3)

In Analysis 3, we compare the fit of regressions of including probability, surprisal, surprisal$^{0.6}$, and combinations of these metrics as predictors of N400 amplitude. Because regressions with probability as a predictor (and not also either surprisal or surprisal$^{0.6}$) often perform far worse than the others, we do not include them in Figure 10.4 in the main body of the paper to increase the ease of comparing the performance of the other regressions. For completeness, we provide the full comparison in Figure 10.11.

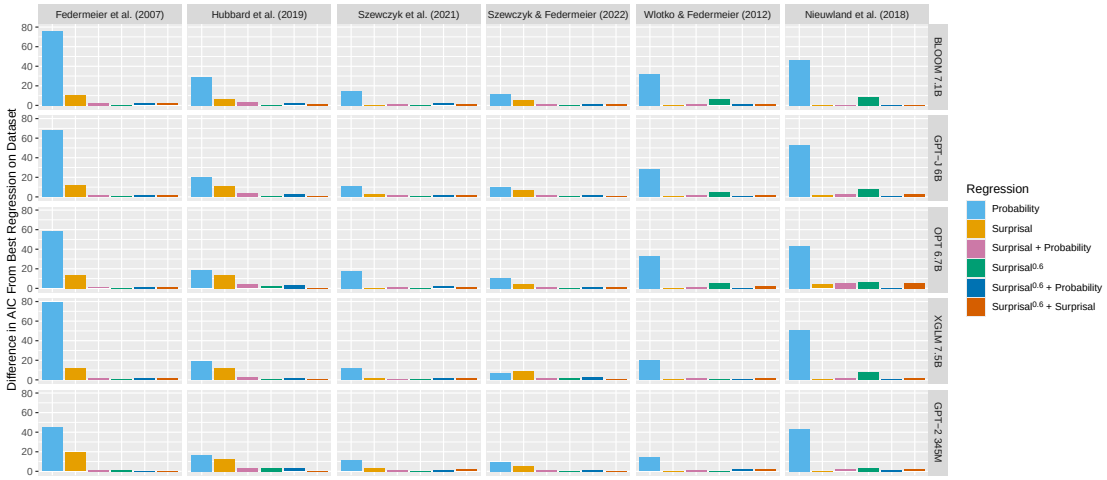

Figure 10.11: The fit of regressions including probability, surprisal, surprisal$^{0.6}$, and combinations of these as predictors of N400 amplitude. We look at the results for the 5 language models that best predict each of the 6 datasets.

10.10.4 Individual Language Model Statistical Analyses (Analysis 3)

In this appendix, we include the statistical analyses investigating the whether adding probability (p), surprisal (S), surprisal$^{0.6}$ ($S^{0.6}$), or a combination of these to a linear mixed-effects model already including one or more other of these variables as predictors significantly improves model fit. In this way, we test which of these metrics explain variance in N400 amplitude not explained by the others. We provide the results for the metrics calculated using BLOOM 7.1B (Table 10.7), OPT 6.7B (Table 10.8), XGLM 7.5B (Table 10.9), and GPT-2 345M (Table 10.10). The table for GPT-J 6B is included in the main body of the paper (Table 10.3).

Table 10.7: Results of the likelihood ratio tests testing the effect of adding BLOOM 7.1B probability (p), surprisal (S), surprisal$^{0.6}$ ($S^{0.6}$), or a combination of these to a linear mixed-effects model already including one or more other of these variables, thereby testing whether they explain any additional variance. As an example, $S + p$ refers to a likelihood ratio test of whether probability explains additional variance in N400 amplitude above and beyond that explained by surprisal. F07 refers to Federmeier et al. (2007), W12 to Wlotko and Federmeier (2012), SF22 to Szewczyk and Federmeier (2022), H19 to Hubbard et al. (2019), S22 to Szewczyk et al. (2022), and N18 to Nieuwland et al. (2018b).

	$S^{0.6} + S$		$S^{0.6} + p$		$S + S^{0.6}$		$S + p$	
Exp.	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**
F07	0.00	1.0000	0.01	1.0000	10.46	0.0203	10.15	0.0232
W12	6.96	0.1159	6.90	0.1186	0.94	1.0000	0.80	1.0000
SF22	0.91	1.0000	0.31	1.0000	6.25	0.1652	5.48	0.2419
H19	0.41	1.0000	0.00	1.0000	7.21	0.1024	5.69	0.2187
S22	0.26	1.0000	0.00	1.0000	0.84	1.0000	1.49	1.0000
N18	10.4	0.0205	10.74	0.0182	1.47	1.0000	1.20	1.0000

	$p + S$		$p + S^{0.6}$		$(S + p) + S^{0.6}$		$S^{0.6} + (S + p)$	
Exp.	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**
F07	76.01	<0.0001	76.32	<0.0001	0.32	1.0000	0.02	1.0000
W12	32.69	<0.0001	32.77	<0.0001	0.31	1.0000	7.13	0.3327
SF22	12.09	0.0104	12.25	0.0097	0.97	1.0000	1.11	1.0000
H19	28.11	<0.0001	29.23	<0.0001	2.77	0.9825	1.65	1.0000
S22	15.51	0.0021	14.6	0.0031	2.92	0.9155	3.83	1.0000
N18	47.67	<0.0001	48.28	<0.0001	0.68	1.0000	10.81	0.0672

Table 10.8: Results of the likelihood ratio tests testing the effect of adding OPT 6.7B probability (p), surprisal (S), surprisal$^{0.6}$ ($S^{0.6}$), or a combination of these to a linear mixed-effects model already including one or more other of these variables, thereby testing whether they explain any additional variance. As an example, $S + p$ refers to a likelihood ratio test of whether probability explains additional variance in N400 amplitude above and beyond that explained by surprisal. F07 refers to Federmeier et al. (2007), W12 to Wlotko and Federmeier (2012), SF22 to Szewczyk and Federmeier (2022), H19 to Hubbard et al. (2019), S22 to Szewczyk et al. (2022), and N18 to Nieuwland et al. (2018b).

	$S^{0.6} + S$		$S^{0.6} + p$		$S + S^{0.6}$		$S + p$	
Exp.	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**
F07	0.51	1.0000	0.6	1.0000	14.45	0.0033	14.93	0.0027
W12	5.46	0.2435	7.61	0.0837	0.41	1.0000	0.76	1.0000
SF22	0.82	1.0000	0.33	1.0000	5.69	0.2187	5.14	0.2817
H19	4.04	0.5019	0.92	1.0000	15.46	0.0021	11.4	0.0137
S22	0.75	1.0000	0.52	1.0000	0.23	1.0000	0.19	1.0000
N18	2.80	0.9702	8.30	0.0602	0.49	1.0000	0.10	1.0000

	$p + S$		$p + S^{0.6}$		$(S + p) + S^{0.6}$		$S^{0.6} + (S + p)$	
Exp.	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**
F07	59.01	<0.0001	58.62	<0.0001	0.02	1.0000	1.00	1.0000
W12	33.05	<0.0001	34.85	<0.0001	1.95	1.0000	7.76	0.2542
SF22	11.07	0.0157	11.13	0.0155	0.64	1.0000	0.91	1.0000
H19	16.11	0.0015	17.05	0.0010	7.81	0.0768	7.79	0.2522
S22	18.15	0.0006	17.95	0.0006	0.07	1.0000	0.78	1.0000
N18	39.10	<0.0001	44.99	<0.0001	4.88	0.3223	7.29	0.3126

Table 10.9: Results of the likelihood ratio tests testing the effect of adding XGLM 7.5B probability (p), surprisal (S), surprisal$^{0.6}$ ($S^{0.6}$), or a combination of these to a linear mixed-effects model already including one or more other of these variables, thereby testing whether they explain any additional variance. As an example, $S + p$ refers to a likelihood ratio test of whether probability explains additional variance in N400 amplitude above and beyond that explained by surprisal. F07 refers to Federmeier et al. (2007), W12 to Wlotko and Federmeier (2012), SF22 to Szewczyk and Federmeier (2022), H19 to Hubbard et al. (2019), S22 to Szewczyk et al. (2022), and N18 to Nieuwland et al. (2018b).

	$S^{0.6} + S$		$S^{0.6} + p$		$S + S^{0.6}$		$S + p$	
Exp.	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**
F07	0.01	1.0000	0.01	1.0000	11.44	0.0136	11.55	0.0131
W12	0.61	1.0000	1.18	1.0000	0.91	1.0000	0.6	1.0000
SF22	3.47	0.6804	1.18	1.0000	11.09	0.0156	9.42	0.0332
H19	2.94	0.912	0.99	1.0000	13.33	0.0057	10.56	0.0194
S22	0.01	1.0000	0.15	1.0000	1.82	1.0000	2.81	0.9699
N18	8.23	0.062	9.9	0.0258	0.61	1.0000	0.97	1.0000

	$p + S$		$p + S^{0.6}$		$(S + p) + S^{0.6}$		$S^{0.6} + (S + p)$	
Exp.	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**
F07	79.34	<0.0001	79.23	<0.0001	0.08	1.0000	0.20	1.0000
W12	20.69	0.0002	21.56	0.0001	1.53	1.0000	1.83	1.0000
SF22	7.38	0.0941	6.76	0.1270	2.36	1.0000	4.16	1.0000
H19	18.39	0.0005	19.21	0.0003	4.33	0.4323	4.50	1.0000
S22	12.5	0.0087	11.66	0.0125	3.19	0.7922	4.19	1.0000
N18	50.8	<0.0001	52.11	<0.0001	1.70	1.0000	10.29	0.0837

Table 10.10: Results of the likelihood ratio tests testing the effect of adding GPT-2 345M probability (p), surprisal (S), surprisal$^{0.6}$ ($S^{0.6}$), or a combination of these to a linear mixed-effects model already including one or more other of these variables, thereby testing whether they explain any additional variance. As an example, $S + p$ refers to a likelihood ratio test of whether probability explains additional variance in N400 amplitude above and beyond that explained by surprisal. F07 refers to Federmeier et al. (2007), W12 to Wlotko and Federmeier (2012), SF22 to Szewczyk and Federmeier (2022), H19 to Hubbard et al. (2019), S22 to Szewczyk et al. (2022), and N18 to Nieuwland et al. (2018b).

	$S^{0.6} + S$		$S^{0.6} + p$		$S + S^{0.6}$		$S + p$	
Exp.	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**
F07	3.63	0.6258	3.46	0.6804	21.26	0.0001	20.24	0.0002
W12	0.21	1.0000	0.27	1.0000	0.82	1.0000	0.91	1.0000
SF22	1.47	1.0000	0.49	1.0000	6.61	0.1375	6.0	0.1858
H19	5.20	0.2741	2.30	1.0000	14.67	0.003	11.21	0.015
S22	0.14	1.0000	0.60	1.0000	3.50	0.6757	4.10	0.4891
N18	3.74	0.5907	4.58	0.3779	0.01	1.000	0.00	1.0000

	$p + S$		$p + S^{0.6}$		$(S + p) + S^{0.6}$		$S^{0.6} + (S + p)$	
Exp.	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**	$\boldsymbol{\chi^2}$	**p**
F07	45.84	<0.0001	46.69	<0.0001	1.03	1.0000	3.64	1.0000
W12	14.37	0.0034	14.34	0.0034	0.04	1.0000	0.34	1.0000
SF22	10.44	0.0203	10.06	0.0241	0.64	1.0000	1.50	1.0000
H19	15.33	0.0022	15.9	0.0017	4.55	0.3812	6.29	0.4891
S22	11.67	0.0125	11.52	0.0131	0.58	1.0000	1.33	1.0000
N18	43.3	<0.0001	44.14	<0.0001	0.26	1.0000	4.00	1.0000

10.11 Acknowledgements

We would like to thank the members and affiliates of the Language and Cognition Lab at UCSD for their valuable discussion, and the editor and anonymous reviewers for their helpful feedback. All language models were run on hardware provided by the NVIDIA Corporation as part of an NVIDIA Academic Hardware Grant.

Chapter 10, in full, is a reprint of the material as it appears in Michaelov, J. A. & Bergen, B. K., "On the Mathematical Relationship Between Contextual Probability and

N400 Amplitude", *Open Mind*, 2024. The dissertation author was the primary investigator and author of this paper. Minor edits have been made to bring the formatting in line with the dissertation template.

Chapter 11

Conclusions

11.1 To what extent can the N400 be explained by the statistics of language?

In this dissertation, I have used language models to investigate the extent to which the statistics of language offer an explanation for a diverse array of N400 effects that have been reported in previous work.

The results of Chapter 2 show that language model predictions are highly correlated with N400 amplitude. Specifically, the surprisal calculated from four of the eight language models tested predicts N400 amplitude better than cloze does. While this high degree of correlation does not necessarily entail that the predictions indexed by the N400 in humans are primarily driven by language statistics, this result, combined with the previously-discussed evidence that language processing is sensitive to statistics, does lend support to the idea that language statistics may play a role in prediction. At the very least, the results suggest that the statistics of language reliably track the extent to which a word is contextually predictable. This idea is further strengthened by the finding that

larger models of each architecture (which generally perform better on natural language tasks) generate predictions with a higher degree of correlation to the N400 than smaller models—models that are better able to capture the statistics of language are better models of the prediction indexed by the N400.

The results of Chapter 3 push this result further. For models of all sizes, training a model on more data both makes it better at predicting the next word in a sequence and better at predicting N400 amplitude. Specifically, the results can be taken to show that models that are better at next-word prediction are better at modeling the N400. Thus, again, the results are consistent with the idea that the prediction that occurs during language processing could at the very least be partly based on input (i.e., language in this case) statistics, in line with predictive processing in other domains (Rao and Ballard, 1999; Huang and Rao, 2011; Clark, 2013; de Lange et al., 2018). The chapter additionally finds that performance at a range of natural language processing benchmarks requiring predictions that align with world knowledge (rather than simply the surface-level statistics of language) is also correlated with the extent to which language model predictions match those of the human predictive system in language. While this may be taken to be simply a consequence of the fact that models that are better at next-word prediction are better at such tasks, an alternative possibility is that it is precisely the acquisition of semantic knowledge by these language models that makes them so well-suited to modeling the N400. This is in line with the majority of accounts of the N400 that frame it as being primarily sensitive to word-level semantics (DeLong and Kutas, 2020; Kuperberg et al., 2020; Federmeier, 2021).

Chapter 4 investigates a range of N400 phenomena, finding that like the N400, surprisal is sensitive to semantic typicality (Urbach and Kutas, 2010), the extent to which a word is predictable based on the context (Kutas, 1993; Ito et al., 2016), whether a word is

semantically related to the most likely sentence continuation (Kutas, 1993; Ito et al., 2016), and whether a word is semantically anomalous (Osterhout and Mobley, 1995; Ainsworth-Darnell et al., 1998; Kim and Osterhout, 2005). And like the N400, surprisal is not sensitive to words such as *few* or *rarely* (Urbach and Kutas, 2010). There are several studies with less clear results, but two where language model surprisal clearly does not align with N400 amplitude. The first of these is pronoun mismatch—while humans do not show a difference between in N400 amplitude between *the aunt heard that* ***she*** and *the aunt heard that* ***he***, language models find the latter to be less likely. The other is that language models are more sensitive to the event structure violation in *the murder had been* ***witnessing*** (compared to *the murder had been* ***witnessed***)) than humans are.

Chapter 5 further expands upon these results. It finds that a range of contemporary transformers also show the related anomaly effect reported for recurrent neural networks in Chapter 4 (Ito et al., 2016), in that they show an increased prediction for words that are semantically related to the most likely continuation of a sentence than unrelated words, all else being equal (Ito et al., 2016; DeLong et al., 2019). Similarly, the results also show that language models replicate the variant of the related anomaly effect where words more related to the preceding context are more strongly predicted than unrelated words, all else being equal.

Chapter 6 also expands upon the results of Chapter 4. Specifically, the aim of the study was to test whether the lack of sensitivity to quantifiers such as *few* or *rarely* is generalizeable to contemporary models. This study finds that it is, more so for larger (and generally higher-quality) models than smaller models.

Chapter 7 expands upon the previous chapters by investigating a different effect—whether the context in which a word appears is able to override general world knowledge. We find that this is indeed the case—under the right conditions, a peanut can fall in love,

whether in the brain or in an LLM.

Next, Chapter 8 goes beyond Chapters 4–7 in testing whether surprisal can capture specific N400 effects, but specifically by testing whether it can account for all of the variance explained by the relevant experimental manipulation. In this study, we found that lexical surprisal calculated using GPT-3 can predict N400 amplitude better than cloze probability, plausibility, and the degree of association between a critical word and its context. However, GPT-3 surprisal does not appear to fully account for the effect explored in Chapter 5, namely, the fact that words semantically related to the most likely continuation are more strongly predicted than unrelated words.

Unlike the studies in Part 1 and 2 which mostly focus on the question of whether language statistics can explain N400 phenomena that we already know about, Chapter 9 focuses instead on whether we can use language models to gain new insights into the nature of the ERP component. Specifically, the study uses a new approach to investigate the question of whether the statistical probabilities of words other than the actual stimulus (and thus, the extent to which other words could be predicted based on the statistics of language) have an impact on the N400. Previous work on this question, which operationalizes contextual probability as cloze probability, suggests that the probabilities of non-stimulus words do not impact N400 amplitude (Van Petten et al., 1999; Federmeier et al., 2002; Vissers et al., 2006; Federmeier et al., 2007; Federmeier, 2007; Wlotko and Federmeier, 2007; Otten and Berkum, 2008). However, given that statistical probabilities are not the same as cloze probabilities (see, e.g., Smith and Levy, 2011 for discussion), the finding in earlier chapters that language model probability can show a closer fit to the N400 than cloze, and the fact that language models allow the probability of every alternative word to be calculated (i.e., not just those provided as responses to the cloze task), language models provide an opportunity to investigate this question in a new way. Evaluating the

fit of multiple metrics derived from the whole output probability distribution of language models (i.e., that takes into account the probabilities of all words), we found that no such metric individually displays a better fit to the N400 data, and that no such metric explains a significant amount of variance in N400 amplitude above and beyond a combination of surprisal and probability. Thus, the results of the study align with previous work showing that only the contextual probability that appears to impact N400 amplitude is that of the actual stimulus encountered.

The aim of Chapter 10 is to further investigate the mathematical relationship between statistical probability and the N400. The results of Chapter 9 show that surprisal is a better predictor of N400 amplitude than probability is, but also replicate Szewczyk and Federmeier's (2021) finding that on some datasets, a combination of the two predict N400 amplitude better than either alone. This chapter systematically investigates the relationship between contextual probability as calculated using language models and the N400, first by comparing probability, surprisal, and surprisal to the power of a set of numbers between zero and two (following Meister et al., 2021; Shain et al., 2024); and second, by comparing combinations of these to see how much variance is explained by each. The results show that the N400 is best predicted by a combination of surprisal and probability (as in Szewczyk and Federmeier, 2022), but also similarly well by surprisal$^{0.6}$, something that is not expected based on previous work. Overall, the results of the study suggest a single mathematical relationship that provides the best theoretical and empirical account of the N400 at the same time has yet to be discovered. What the results do show, however, is that the N400 is more sensitive to differences at the low end of the probability scale than would be predicted based on un-transformed probability, but also more sensitive to differences at the upper end of the scale than would be predicted based on surprisal.

All in all, the results show that with a few exceptions, the predictions of language

models are highly correlated with N400 amplitude, and they are both able show good quantitative fit to the N400 data—in the cases tested, better than either cloze probability or plausibility—and model specific qualitative effects. Furthermore, in virtually all cases, larger and higher-quality language models model the N400 better than smaller and lower-quality models. Taken together, these results suggest that at least in principle, a statistical account of the prediction indexed by the N400 is viable, and they provide support to the idea that human predictive processing of language is at least partly sensitive to the statistics of language.

11.2 What is missing in the statistics of language?

In the experiments reported in this dissertation, the predictions of language models show a remarkable degree of correlation with the N400. However, as previously noted, there are points of divergence. On the whole these appear to relate to word-level semantics. As noted in the previous section, one interpretation of the fact that better-trained and higher-quality language models are better predictors of N400 amplitude is that they are better able to capture the word-level semantics of critical words. Thus, the issue may be that the models are still limited in this respect.

For example, where the N400 does not show sensitivity to the event-structure violation in *the murder had been* ***witnessing*** but language models do, it is worth noting that other than the inflection of the verb (a linguistic feature), *witnessing* and *witnessed* are identical. Thus, in terms of word-level semantic representations, the two are virtually the same, and thus if it is indeed the case that N400 is primarily sensitive to word-level semantics, this would explain the lack of a significant difference in response between the two. The results of Chapter 8 could be explained similarly if the prediction of word-level semantic representations are the main driver of the N400 effects. Specifically, the word-

level semantics of the highest-probability continuation are typically shared with semantic associates, which would explain the results seen in Chapter 8 where related words elicit an N400 of virtually the same amplitude as the most likely continuations.

The question, then, is what is the cause of this divergence? There appear to be two main possibilities. The first is that the the word-level semantic representations of language models are limited compared to those of humans due to only deriving from the statistics of language. In this case, the differences in prediction would be due either to there being fewer (or less rich) semantic representations in the language models, or simply different representations due to the different sources of information on which these are based. On the other hand, it may be that the word-level semantic representations of humans and language models *are* similar enough, but simply that language models over-rely on the linguistic similarities since they are trained on text data only. Thus, one possible avenue would be to use multimodal training. Some research shows that visual grounding better enables models to learn the meanings of words (Zhuang et al., 2024), and given the role of sensorimotor information in human word learning (Barsalou, 1999), this may be precisely the kind of information that is missing in pure language models.

11.3 Conclusion

The aim of this dissertation was to investigate how well the predictions indexed by the N400 can be explained as deriving from the statistics of language. The specific approach taken was to explore the extent to which the predictions of language models correlate with N400 amplitude. The results are promising. First, language model surprisal is highly correlated with N400 amplitude—for some models better even than traditional metrics such as cloze and plausibility. In addition, larger and better language models are better for modeling the N400. Thus, we can expect that as language models continue to

advance, their predictions will likely even more closely correlate with N400 amplitude and thus give us an even better idea of the extent to which language statistics can account for the phenomena. Finally, this dissertation has highlighted one limitation in modeling the N400 using language model surprisal, namely, that the N400 seems more sensitive to the word-level semantic representations and less sensitive to linguistic representations than language model predictions.

Bibliography

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., and Søgaard, A. (2021). Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online. Association for Computational Linguistics.

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. In Van den Bussche, J. and Vianu, V., editors, *Database Theory — ICDT 2001*, Lecture Notes in Computer Science, pages 420–434, Berlin, Heidelberg. Springer.

Ainsworth-Darnell, K., Shulman, H. G., and Boland, J. E. (1998). Dissociating Brain Responses to Syntactic and Semantic Anomalies: Evidence from Event-Related Potentials. *Journal of Memory and Language*, 38(1):112–130.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In Petrov, B. N. and Csáki, F., editors, *Second International Symposium on Information Theory*, Springer Series in Statistics, pages 267–281, Budapest, Hungary. Akadémiai Kiadó.

Alabdulmohsin, I. M., Neyshabur, B., and Zhai, X. (2022). Revisiting Neural Scaling

Laws in Language and Vision. *Advances in Neural Information Processing Systems*, 35:22300–22312.

Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., and Muckli, L. (2010). Stimulus Predictability Reduces Responses in Primary Visual Cortex. *Journal of Neuroscience*, 30(8):2960–2966.

Allen, M. and Tsakiris, M. (2018). *The Body as First Prior: Interoceptive Predictive Processing and the Primacy of Self-Models*. Oxford University Press.

Ambridge, B., Pine, J. M., and Lieven, E. V. M. (2014). Child language acquisition: Why universal grammar doesn't help. *Language*, 90(3):e53–e90.

Amram, A., Ben David, A., and Tsarfaty, R. (2018). Representations and Architectures in Neural Sentiment Analysis for Morphologically Rich Languages: A Case Study from Modern Hebrew. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2242–2252, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Amsel, B. D., DeLong, K. A., and Kutas, M. (2015). Close, but no garlic: Perceptuomotor and event knowledge activation during language comprehension. *Journal of Memory and Language*, 82:118–132.

Amsel, B. D., Urbach, T. P., and Kutas, M. (2013). Alive and grasping: Stable and rapid semantic access to an object category but not object graspability. *NeuroImage*, 77:1–13.

Amsel, B. D., Urbach, T. P., and Kutas, M. (2014). Empirically grounding grounded cognition: The case of color. *NeuroImage*, 99:149–157.

An, A., Qian, P., Wilcox, E., and Levy, R. (2019). Representation of Constituents in Neural Language Models: Coordination Phrase as a Case Study. *arXiv:1909.04625 [cs]*.

Anderson, J. E. and Holcomb, P. J. (1995). Auditory and visual semantic priming using different stimulus onset asynchronies: An event-related brain potential study. *Psychophysiology*, 32(2):177–190.

Aravena, P., Courson, M., Frak, V., Cheylus, A., Paulignan, Y., Deprez, V., and Nazir, T. (2014). Action relevance in linguistic context drives word-induced motor activity. *Frontiers in Human Neuroscience*, 8.

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*.

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., and Crocker, M. W. (2021). Retrieval (N400) and integration (P600) in expectation-based comprehension. *PLOS ONE*, 16(9):e0257430.

Aurnhammer, C. and Frank, S. L. (2019a). Comparing Gated and Simple Recurrent Neural Network Architectures as Models of Human Sentence Processing. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci 2019)*.

Aurnhammer, C. and Frank, S. L. (2019b). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, 134:107198.

Aylett, M. and Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence, and Duration in Spontaneous Speech. *Language and Speech*, 47(1):31–56.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S.,

Olah, C., Mann, B., and Kaplan, J. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback.

Bardolph, M., Van Petten, C., and Coulson, S. (2018). Single Trial EEG Data Reveals Sensitivity to Conceptual Expectations (N400) and Integrative Demands (LPC). In *Twelfth Annual Meeting of the Society for the Neurobiology of Language*, Quebec City, Canada.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–660.

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2018). Parsimonious Mixed Models. *arXiv:1506.04967 [stat]*.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Becker, C. A. (1980). Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory & Cognition*, 8(6):493–512.

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

Bender, E. (2019). The #BenderRule: On naming the languages we study and why it matters. *The Gradient*.

Bender, E. M. (2009). Linguistically Naïve != Language Independent: Why NLP Needs Linguistic Typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Bender, E. M. (2011). On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology*, 6.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Bendixen, A., SanMiguel, I., and Schröger, E. (2012). Early electrophysiological indicators for predictive processing in audition: A review. *International Journal of Psychophysiology*, 83(2):120–131.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, 29(4):1165–1188.

Bentin, S., McCarthy, G., and Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology*, 60(4):343–355.

Berger, J. and Packard, G. (2022). Using natural language processing to understand people and culture. *American Psychologist*, 77:525–537.

Bicknell, K., Elman, J. L., Hare, M., McRae, K., and Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, 63(4):489–505.

Biderman, S., Prashanth, U., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., and Raff, E. (2023a). Emergent and Predictable Memorization in Large Language Models. *Advances in Neural Information Processing Systems*, 36:28072–28090.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and Wal, O. V. D. (2023b). Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In *Proceedings of the 40th International Conference on Machine Learning*, pages 2397–2430. PMLR.

BigScience (2022). BigScience Language Open-science Open-access Multilingual (BLOOM) Language Model. International, May 2021-May 2022.

Bisk, Y., Zellers, R., Bras, R. L., Gao, J., and Choi, Y. (2020). PIQA: Reasoning about Physical Commonsense in Natural Language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

Black, S., Gao, L., Wang, P., Leahy, C., and Biderman, S. (2021). GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow. Zenodo.

Bloom, P. A. and Fischler, I. (1980). Completion norms for 329 sentence contexts. *Memory & Cognition*, 8(6):631–642.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K. A., Davis, J., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L. E., Goel, K., Goodman, N. D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T. F., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M. S., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S. P., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J. F., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y. H., Ruiz, C., Ryan, J., R'e, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K. P., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M. A., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. (2021). On the opportunities and risks of foundation models. *ArXiv*.

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., and Liang, P. (2023a). The Foundation Model Transparency Index.

Bommasani, R., Klyman, K., Longpre, S., Xiong, B., Kapoor, S., Maslej, N., Narayanan, A., and Liang, P. (2024). Foundation Model Transparency Reports.

Bommasani, R., Liang, P., and Lee, T. (2023b). Holistic Evaluation of Language Models. *Annals of the New York Academy of Sciences*, 1525(1):140–146.

Bornkessel-Schlesewsky, I. and Schlesewsky, M. (2019). Toward a Neurobiologically Plausible Model of Language-Related, Negative Event-Related Potentials. *Frontiers in Psychology*, 10.

Boston, M. F., Hale, J., Kliegl, R., Patil, U., and Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2(1).

Bostrom, K. and Durrett, G. (2020). Byte Pair Encoding is Suboptimal for Language Model Pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Bransford, J. D., Barclay, J. R., and Franks, J. J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive Psychology*, 3(2):193–209.

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

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., and Lalor, E. C. (2018). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Current Biology*, 28(5):803–809.e3.

Brothers, T. and Kuperberg, G. R. (2021). Word predictability effects are linear, not

logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, 116:104174.

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

Brouwer, H., Delogu, F., Venhuizen, N. J., and Crocker, M. W. (2021). Neurobehavioral Correlates of Surprisal in Language Comprehension: A Neurocomputational Model. *Frontiers in Psychology*, 12.

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

Brouwer, H., Fitz, H., and Hoeks, J. (2012). Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, 1446:127–143.

Brouwer, H. and 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.

Brown, C. and Hagoort, P. (1993). The Processing Nature of the N400: Evidence from Masked Priming. *Journal of Cognitive Neuroscience*, 5(1):34–44.

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., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S.,

Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Bruner, J. S. (1951). Personality dynamics and the process of perceiving. In *Perception: An Approach to Personality*, pages 121–147. Ronald Press Company, New York, NY, US.

Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Brysbaert, M., Stevens, M., Mandera, P., and 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.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

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

Butcher, K. R. and Kintsch, W. (2012). Text Comprehension and Discourse Processing. In *Handbook of Psychology, Second Edition*, chapter 21. American Cancer Society.

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

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. (2022). Quan-

tifying Memorization Across Neural Language Models. In *The Eleventh International Conference on Learning Representations*.

Chang, W. (2022). Colors (ggplot2).

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. (2014). One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. *arXiv:1312.3005 [cs]*.

Cheyette, S. J. and Plaut, D. C. (2017). Modeling the N400 ERP component as transient semantic over-activation within a neural network model of word comprehension. *Cognition*, 162:153–166.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2022). PaLM: Scaling Language Modeling with Pathways.

Chwilla, D. J. and Kolk, H. H. J. (2005). Accessing world knowledge: Evidence from N400 and reaction time priming. *Cognitive Brain Research*, 25(3):589–606.

Chwilla, D. J., Kolk, H. H. J., and Vissers, C. T. W. M. (2007). Immediate integration of

novel meanings: N400 support for an embodied view of language comprehension. *Brain Research*, 1183:109–123.

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

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N. A. (2021). All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Clark, J. H., Garrette, D., Turc, I., and Wieting, J. (2022). Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Clark, T. H., Meister, C., Pimentel, T., Hahn, M., Cotterell, R., Futrell, R., and Levy, R. (2023). A Cross-Linguistic Pressure for Uniform Information Density in Word Order. *Transactions of the Association for Computational Linguistics*, 11:1048–1065.

Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428.

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

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Costa, J. K. D. and Chaves, R. P. (2020). Assessing the ability of Transformer-based Neural Models to represent structurally unbounded dependencies. In *Proceedings of the Society for Computation in Linguistics (SCiL)*, volume 3, page 10.

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

Coulson, S. and Lovett, C. (2004). Handedness, hemispheric asymmetries, and joke comprehension. *Cognitive Brain Research*, 19(3):275–288.

Coupé, C., Oh, Y. M., Dediu, D., and Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9):eaaw2594.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. (2019). Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Dambacher, M., Kliegl, R., Hofmann, M., and Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, 1084(1):89–103.

de Groot, A. M. B. (2011). *Language and Cognition in Bilinguals and Multilinguals an Introduction*. Psychology Press, New York [u.a.

de Lange, F. P., Heilbron, M., and Kok, P. (2018). How Do Expectations Shape Perception? *Trends in Cognitive Sciences*, 22(9):764–779.

de Marneffe, M.-C., Grimm, S., Arnon, I., Kirby, S., and Bresnan, J. (2012). A statistical model of the grammatical choices in child production of dative sentences. *Language and Cognitive Processes*, 27(1):25–61.

de Varda, A. and Marelli, M. (2023). Scaling in Cognitive Modelling: A Multilingual Approach to Human Reading Times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 139–149, Toronto, Canada. Association for Computational Linguistics.

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., and Nissim, M. (2019). BERTje: A Dutch BERT Model. *arXiv:1912.09582 [cs]*.

Debruille, J. B. (2007). The N400 potential could index a semantic inhibition. *Brain Research Reviews*, 56(2):472–477.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Delaney-Busch, N., Morgan, E., Lau, E., and Kuperberg, G. (2017). Comprehenders Rationally Adapt Semantic Predictions to the Statistics of the Local Environment: A Bayesian Model of Trial-by-Trial N400 Amplitudes. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, page 6, London, UK.

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

Delobelle, P., Winters, T., and Berendt, B. (2020). RobBERT: A Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Delogu, F., Brouwer, H., and Crocker, M. W. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain and Cognition*, 135:103569.

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

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

DeLong, K. A., Quante, L., and Kutas, M. (2014a). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, 61:150–162.

DeLong, K. A., Troyer, M., and Kutas, M. (2014b). Pre-Processing in Sentence Comprehension: Sensitivity to Likely Upcoming Meaning and Structure. *Language and Linguistics Compass*, 8(12):631–645.

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

DeLong, K. A., Urbach, T. P., and Kutas, M. (2017). Is there a replication crisis? Perhaps. Is this an example? No: A commentary on Ito, Martin, and Nieuwland (2016). *Language, Cognition and Neuroscience*, 32(8):966–973.

Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dey, N., Gosal, G., Chen, Z. C., Khachane, H., Marshall, W., Pathria, R., Tom, M., and Hestness, J. (2023). Cerebras-GPT: Open Compute-Optimal Language Models Trained on the Cerebras Wafer-Scale Cluster.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and 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*, pages 281–285, Washington, D.C., United States. ACM Press.

Egner, T., Monti, J. M., and Summerfield, C. (2010). Expectation and Surprise Determine Neural Population Responses in the Ventral Visual Stream. *Journal of Neuroscience*, 30(49):16601–16608.

Eisape, T., Zaslavsky, N., and Levy, R. (2020). Cloze Distillation: Improving Neural Language Models with Human Next-Word Prediction. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 609–619, Online. Association for Computational Linguistics.

Ekman, M., Kok, P., and de Lange, F. P. (2017). Time-compressed preplay of anticipated events in human primary visual cortex. *Nature Communications*, 8(1):15276.

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, 14(2):179–211.

Elman, J. L. (2009). On the Meaning of Words and Dinosaur Bones: Lexical Knowledge Without a Lexicon. *Cognitive Science*, 33(4):547–582.

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

Ettinger, A., Feldman, N., Resnik, P., and 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.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, 44(4):491–505.

Federmeier, K. D. (2021). Connecting and considering: Electrophysiology provides insights into comprehension. *Psychophysiology*, n/a(n/a):e13940.

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

Federmeier, K. D. and Kutas, M. (2001). Meaning and modality: Influences of context, semantic memory organization, and perceptual predictability on picture processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1):202–224.

Federmeier, K. D., Kutas, M., and Dickson, D. S. (2016). A Common Neural Progression

to Meaning in About a Third of a Second. In Hickok, G. and Small, S. L., editors, *Neurobiology of Language*, pages 557–567. Academic Press, San Diego.

Federmeier, K. D., McLENNAN, D. B., Ochoa, E. D., and 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(2):133–146.

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

Fedorenko, E. and Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in Cognitive Sciences*, 18(3):120–126.

Fenk, A. and Fenk-Oczlon, G. (1980). Konstanz im Kurzzeitgedächtnis - Konstanz im sprachlichen Informationsfluß? *Zeitschrift für experimentelle und angewandte Psychologie*, 27:400–414.

Ferguson, H. J. and Sanford, A. J. (2008). Anomalies in real and counterfactual worlds: An eye-movement investigation. *Journal of Memory and Language*, 58(3):609–626.

Ferguson, H. J., Sanford, A. J., and Leuthold, H. (2008). Eye-movements and ERPs reveal the time course of processing negation and remitting counterfactual worlds. *Brain Research*, 1236:113–125.

Ferreira, F. and Yang, Z. (2019). The Problem of Comprehension in Psycholinguistics. *Discourse Processes*, 56(7):485–495.

Filik, R. and Leuthold, H. (2008). Processing local pragmatic anomalies in fictional contexts: Evidence from the N400. *Psychophysiology*, 45(4):554–558.

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

Fischler, I. and Bloom, P. A. (1979). Automatic and attentional processes in the effects of sentence contexts on word recognition. *Journal of Verbal Learning and Verbal Behavior*, 18(1):1–20.

Fischler, I., Bloom, P. A., Childers, D. G., Arroyo, A. A., and Perry, N. W. (1984). Brain potentials during sentence verification: Late negativity and long-term memory strength. *Neuropsychologia*, 22(5):559–568.

Fitz, H. and Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, 111:15–52.

Fleur, D. S., Flecken, M., Rommers, J., and Nieuwland, M. S. (2020). Definitely saw it coming? The dual nature of the pre-nominal prediction effect. *Cognition*, 204:104335.

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

Forster, K. I. (1981). Priming and the effects of sentence and lexical contexts on naming time: Evidence for autonomous lexical processing. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):465–495.

Fossum, V. and Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 61–69.

Foucart, A., Martin, C. D., Moreno, E. M., and Costa, A. (2014). Can bilinguals see it coming? Word anticipation in L2 sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(5):1461–1469.

Frank, S. L. (2014). Modelling reading times in bilingual sentence comprehension. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, Austin,TX.

Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. (2013). Word surprisal predicts N400 amplitude during reading. In Schuetze, H., Fung, P., and Poesio, M., editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 878–883, Sofia, Bulgaria. Association for Computational Linguistics.

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

Frank, S. L. and Thompson, R. (2012). Early effects of word surprisal on pupil size during reading. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

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

Frisch, S. and Schlesewsky, M. (2005). The resolution of case conflicts from a neurophysiological perspective. *Cognitive Brain Research*, 25(2):484–498.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836.

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

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., and Fedorenko, E. (2021). The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55(1):63–77.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., and Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Ganis, G., Kutas, M., and Sereno, M. I. (1996). The Search for "Common Sense": An Electrophysiological Study of the Comprehension of Words and Pictures in Reading. *Journal of Cognitive Neuroscience*, 8(2):89–106.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. (2020). The Pile: An 800GB Dataset of Diverse Text for Language Modeling.

Gao, L., Tow, J., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., McDonell, K., Muennighoff, N., Phang, J., Reynolds, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. (2021). A framework for few-shot language model evaluation. Zenodo.

Garnham, A. (1981). Mental models as representations of text. *Memory & Cognition*, 9(6):560–565.

Gauthier, J., Hu, J., Wilcox, E., Qian, P., and Levy, R. (2020). SyntaxGym: An Online Platform for Targeted Evaluation of Language Models. In Celikyilmaz, A. and Wen, T.-H., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Genzel, D. and Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 199–206, USA. Association for Computational Linguistics.

Gerken, L. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition*, 98(3):B67–B74.

Gerken, L. (2007). Acquiring Linguistic Structure. In *Blackwell Handbook of Language Development*, chapter 9, pages 173–190. John Wiley & Sons, Ltd.

Gibbs, A. L. and Su, F. E. (2002). On Choosing and Bounding Probability Metrics. *International Statistical Review*, 70(3):419–435.

Girshick, A. R., Landy, M. S., and Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, 14(7):926–932.

Gómez, R. L. and Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, 4(5):178–186.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Goodkind, A. and Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Google Research (2019). BERT.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). Colorless Green Recurrent Networks Dream Hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Hagoort, P., Baggio, G., and Willems, R. M. (2009). Semantic unification. In Gazzaniga, M. S., editor, *The Cognitive Neurosciences*, pages 819–836. MIT Press, Cambridge, MA, 4 edition.

Hagoort, P., Hald, L., Bastiaansen, M., and Petersson, K. M. (2004). Integration of Word Meaning and World Knowledge in Language Comprehension. *Science*, 304(5669):438–441.

Hagoort, P. and van Berkum, J. (2007). Beyond the sentence given. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):801–811.

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*, pages 1–8, Pittsburgh, Pennsylvania. Association for Computational Linguistics.

Halgren, E., Dhond, R. P., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J. D., and Dale, A. M. (2002). N400-like Magnetoencephalography Responses Modulated by Semantic Context, Word Frequency, and Lexical Class in Sentences. *NeuroImage*, 17(3):1101–1116.

Hanna, J. and Pulvermüller, F. (2014). Neurophysiological evidence for whole form retrieval of complex derived words: A mismatch negativity study. *Frontiers in Human Neuroscience*, 8.

Hao, Y., Mendelsohn, S., Sterneck, R., Martinez, R., and Frank, R. (2020). Probabilistic Predictions of People Perusing: Evaluating Metrics of Language Model Performance for Psycholinguistic Modeling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–86, Online. Association for Computational Linguistics.

Havinga, Y. (2021). GPT2-Medium pre-trained on cleaned Dutch mC4.

Havinga, Y. (2022a). GPT-Neo 125M pre-trained on cleaned Dutch mC4.

Havinga, Y. (2022b). GPT Neo 1.3B pre-trained on cleaned Dutch mC4.

Havinga, Y. (2022c). GPT2-Large pre-trained on cleaned Dutch mC4.

Heilbron, F. C., Escorcia, V., Ghanem, B., and Niebles, J. C. (2015). ActivityNet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., and de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119.

Hendriks, P. (2014). *Asymmetries between Language Production and Comprehension*, volume 42 of *Studies in Theoretical Psycholinguistics*. Springer Netherlands, Dordrecht.

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., Hallacy, C., Mann, B., Radford, A., Ramesh, A., Ryder, N., Ziegler, D. M., Schulman, J., Amodei, D., and McCandlish, S. (2020). Scaling Laws for Autoregressive Generative Modeling.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Hodapp, A. and Rabovsky, M. (2021). The N400 ERP component reflects an error-based implicit learning signal during language comprehension. *European Journal of Neuroscience*, 54(9):7125–7140.

Hoeben Mannaert, L. and Dijkstra, K. (2021). Situation model updating in young and older adults. *International Journal of Behavioral Development*, 45(5):389–396.

Hoeks, J. C. J., Stowe, L. A., and Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, 19(1):59–73.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen,

E., Vinyals, O., Rae, J. W., and Sifre, L. (2022). Training Compute-Optimal Large Language Models. In *Advances in Neural Information Processing Systems*.

Hofmann, V., Pierrehumbert, J. B., and Schütze, H. (2021). Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words. *arXiv:2101.00403 [cs]*.

Holcomb, P. J. (1988). Automatic and attentional processing: An event-related brain potential analysis of semantic priming. *Brain and Language*, 35(1):66–85.

Holcomb, P. J. (1993). Semantic priming and stimulus degradation: Implications for the role of the N400 in language processing. *Psychophysiology*, 30(1):47–61.

Hoover, J. L., Sonderegger, M., Piantadosi, S. T., and O'Donnell, T. J. (2023). The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing. *Open Mind*, 7:350–391.

Huang, Y. and Rao, R. P. N. (2011). Predictive coding. *WIREs Cognitive Science*, 2(5):580–593.

Hubbard, R. J., Rommers, J., Jacobs, C. L., and Federmeier, K. D. (2019). Downstream Behavioral and Electrophysiological Consequences of Word Prediction on Recognition Memory. *Frontiers in Human Neuroscience*, 13.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, 1626:118–135.

Huizeling, E., Arana, S., Hagoort, P., and Schoffelen, J.-M. (2022). Lexical Frequency and Sentence Context Influence the Brain's Response to Single Words. *Neurobiology of Language*, 3(1):149–179.

Ito, A., Corley, M., Pickering, M. J., Martin, A. E., and Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, 86:157–171.

Ito, A., Martin, A. E., and Nieuwland, M. S. (2017a). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience*, 32(8):954–965.

Ito, A., Martin, A. E., and Nieuwland, M. S. (2017b). Why the A/AN prediction effect may be hard to replicate: A rebuttal to Delong, Urbach, and Kutas (2017). *Language, Cognition and Neuroscience*, 32(8):974–983.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press.

Jackson, R. (2016). 15-level colorblind-friendly palette – Jackson Lab.

Jain, A. K. (1976). On an Estimate of the Bhattacharyya Distance. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(11):763–766.

Jang, J., Ye, S., and Seo, M. (2023). Can Large Language Models Truly Understand Prompts? A Case Study with Negated Prompts. In *Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop*, pages 52–62. PMLR.

Jannai, D., Meron, A., Lenz, B., Levine, Y., and Shoham, Y. (2023). Human or Not? A Gamified Approach to the Turing Test.

Jelinek, F., Mercer, R. L., Bahl, L. R., and Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63.

JGraph (2024). Draw.io. JGraph.

Jiménez-Zafra, S. M., Morante, R., Teresa Martín-Valdivia, M., and Ureña-López, L. A. (2020). Corpora Annotated with Negation: An Overview. *Computational Linguistics*, 46(1):1–52.

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science*, 4(1):71–115.

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Number 6. Harvard University Press.

Jones, C. R. and Bergen, B. K. (2024a). Does GPT-4 pass the Turing test?

Jones, C. R. and Bergen, B. K. (2024b). People cannot distinguish GPT-4 from a human in a Turing test.

Jones, C. R., Chang, T. A., Coulson, S., Michaelov, J. A., Trott, S., and Bergen, B. K. (2022). Distrubutional Semantics Still Can't Account for Affordances. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44(44).

Jordan, M. I. (1986). Serial Order: A Parallel Distributed Processing Approach. Technical Report 8604, Institute for Cognitive Science, University of California, San Diego, La Jolla, California, USA.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. (2016). Exploring the Limits of Language Modeling. *arXiv:1602.02410 [cs]*.

Jurafsky, D. and Martin, J. H. (2019). *Speech and Language Processing*. [Online Draft].

Jurafsky, D. and Martin, J. H. (2021). *Speech and Language Processing*. [Online Draft], 3 edition.

Jurafsky, D. and Martin, J. H. (2024a). *Speech and Language Processing*. 3 edition.

Jurafsky, D. and Martin, J. H. (2024b). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd (draft) edition.

Kalouli, A.-L., Sevastjanova, R., Beck, C., and Romero, M. (2022). Negation, Coordination, and Quantifiers in Contextualized Language Models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3074–3085, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling Laws for Neural Language Models.

Kassner, N. and Schütze, H. (2020). Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Keller, F. (2010). Cognitively Plausible Models of Human Language Processing. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 60–67, Uppsala, Sweden. Association for Computational Linguistics.

Kendall, M. (1948). *Rank Correlation Methods*. Rank Correlation Methods. Charles Griffin, Oxford, England.

Kennedy, A., Hill, R., and Pynte, J. (2003). The Dundee Corpus. In *The 12th European Conference on Eye Movements*, Dundee, UK.

Kim, A. and Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, 52(2):205–225.

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

Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M. (2016). Character-Aware Neural Language Models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2):163.

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Comprehension: A Paradigm for Cognition. Cambridge University Press, New York, NY, US.

Kintsch, W. (2005). An Overview of Top-Down and Bottom-Up Effects in Comprehension: The CI Perspective. *Discourse Processes*, 39(2-3):125–128.

Kintsch, W. (2018). Revisiting the Construction—Integration Model of Text Comprehension and its Implications for Instruction. In *Theoretical Models and Processes of Literacy*. Routledge, 7 edition.

Kintsch, W. and Mangalath, P. (2011). The Construction of Meaning. *Topics in Cognitive Science*, 3(2):346–370.

Kintsch, W. and van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85(5):363–394.

Klein, S. and Tsarfaty, R. (2020). Getting the ##life out of living: How Adequate Are Word-Pieces for Modelling Complex Morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–209, Online. Association for Computational Linguistics.

Köbis, N. and Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, 114:106553.

Kochari, A. R. and Flecken, M. (2019). Lexical prediction in language comprehension: A replication study of grammatical gender effects in Dutch. *Language, Cognition and Neuroscience*, 34(2):239–253.

Kok, P., Brouwer, G. J., van Gerven, M. A. J., and de Lange, F. P. (2013). Prior Expectations Bias Sensory Representations in Visual Cortex. *Journal of Neuroscience*, 33(41):16275–16284.

Kok, P., Jehee, J. F. M., and de Lange, F. P. (2012). Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron*, 75(2):265–270.

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

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., and Niebles, J. C. (2017). Dense-Captioning Events in Videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 706–715.

Kuhn, M., Jackson, S., and Cimentada, J. (2022). *Corrr: Correlations in R.*

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, 1146:23–49.

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, 31(5):602–616.

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

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

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

Kuperberg, G. R., Sitnikova, T., Caplan, D., and Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, 17(1):117–129.

Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., and Inui, K. (2021). Lower Perplexity is Not Always Human-Like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217, Online. Association for Computational Linguistics.

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

Kutas, M., DeLong, K. A., and Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In Bar, M., editor, *Predictions in the Brain: Using Our Past to Generate a Future*, pages 190–207. Oxford University Press, New York, NY, US.

Kutas, M. and Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, 4(12):463–470.

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

Kutas, M. and Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

Kutas, M. and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163.

Kutas, M. and Hillyard, S. A. (1989). An Electrophysiological Probe of Incidental Semantic Association. *Journal of Cognitive Neuroscience*, 1(1):38–49.

Kutas, M., Lindamood, T. E., and Hillyard, S. A. (1984). Word expectancy and event-related brain potentials during sentence processing. In Kornblum, S. and Requin, J., editors, *Preparatory States and Processes*, pages 217–237. Lawrence Erlbaum, Hillsdale, NJ.

Kutas, M. and Van Petten, C. (1994). Psycholinguistics electrified: Event-related brain potential investigations. In Gernsbacher, M. A., editor, *Handbook of Psycholinguistics*, pages 83–143. Academic Press, San Diego, 1 edition.

Kutas, M., Van Petten, C. K., and Kluender, R. (2006). Psycholinguistics Electrified II (1994–2005). In Gernsbacher, M. and Traxler, M., editors, *Handbook of Psycholinguistics*, pages 659–724. Elsevier, New York, 2 edition.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82:1–26.

Kwon, N., Sturt, P., and Liu, P. (2017). Predicting semantic features in Chinese: Evidence from ERPs. *Cognition*, 166:433–446.

Lago, S., Namyst, A., Jäger, L. A., and Lau, E. (2019). Antecedent access mechanisms in pronoun processing: Evidence from the N400. *Language, Cognition and Neuroscience*, 34(5):641–661.

Lambert, N., Gyges, SE., Biderman, S., and Skowron, A. (2023). How the foundation model transparency index distorts transparency. .

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.

Laszlo, S. and Armstrong, B. C. (2014). PSPs and ERPs: Applying the dynamics of post-synaptic potentials to individual units in simulation of temporally extended Event-Related Potential reading data. *Brain and Language*, 132:22–27.

Laszlo, S. and Plaut, D. C. (2012). A neurally plausible Parallel Distributed Processing model of Event-Related Potential word reading data. *Brain and Language*, 120(3):271–281.

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

Lau, E. F., Phillips, C., and Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, 9(12):920–933.

Le Scao, T., Wang, T., Hesslow, D., Bekman, S., Bari, M. S., Biderman, S., Elsahar, H., Muennighoff, N., Phang, J., Press, O., Raffel, C., Sanh, V., Shen, S., Sutawika, L., Tae, J., Yong, Z. X., Launay, J., and Beltagy, I. (2022). What Language Model to Train if You Have One Million GPU Hours? In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 765–782, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lee, T. S. and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448.

Levesque, H., Davis, E., and Morgenstern, L. (2012). The Winograd Schema Challenge.

In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Levy, R. P. (2005). *Probabilistic Models of Word Order and Syntactic Discontinuity*. PhD thesis, Stanford University, Stanford, CA, USA.

Levy, R. P. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Levy, R. P. and Jaeger, T. F. (2006). Speakers optimize information density through syntactic reduction. In Schölkopf, B., Platt, J., and Hofmann, T., editors, *Advances in Neural Information Processing Systems*, volume 19. The MIT Press.

Lewis, A. G. and Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex*, 68:155–168.

Li, B. Z., Nye, M., and Andreas, J. (2021). Implicit Representations of Meaning in Neural Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics.

Lin, S., Hilton, J., and Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., Pasunuru, R., Shleifer, S., Koura, P. S., Chaudhary, V., O'Horo,

B., Wang, J., Zettlemoyer, L., Kozareva, Z., Diab, M., Stoyanov, V., and Li, X. (2021). Few-shot Learning with Multilingual Language Models. *arXiv:2112.10668 [cs]*.

Lindborg, A. and Rabovsky, M. (2021). Meaning in brains and machines: Internal activation update in large-scale language model partially reflects the N400 brain potential. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43(43).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Lowder, M. W., Choi, W., Ferreira, F., and Henderson, J. M. (2018). Lexical Predictability During Natural Reading: Effects of Surprisal and Entropy Reduction. *Cognitive Science*, 42:1166–1183.

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique*. A Bradford Book, Cambridge, MA, USA, 2 edition.

Luka, B. J. and Van Petten, C. (2014). Prospective and retrospective semantic processing: Prediction, time, and relationship strength in event-related potentials. *Brain and Language*, 135:115–129.

Luke, S. G. and Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, 88:22–60.

Luke, S. G. and Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

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

Mann, H. B. (1945). Nonparametric Tests Against Trend. *Econometrica*, 13(3):245.

Manning, C. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Mantegna, F., Hintz, F., Ostarek, M., Alday, P. M., and Huettig, F. (2019). Distinguishing integration and prediction accounts of ERP N400 modulations in language processing through experimental design. *Neuropsychologia*, 134:107199.

Marmor, G. S. (1978). Age at onset of blindness and the development of the semantics of color names. *Journal of Experimental Child Psychology*, 25(2):267–278.

Martin, C. D., Branzi, F. M., and Bar, M. (2018). Prediction is Production: The missing link between language production and comprehension. *Scientific Reports*, 8(1):1079.

Martin, C. D., Thierry, G., Kuipers, J.-R., Boutonnet, B., Foucart, A., and Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language*, 69(4):574–588.

Marvin, R. and Linzen, T. (2018). Targeted Syntactic Evaluation of Language Models. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Maurits, L., Navarro, D., and Perfors, A. (2010). Why are some word orders more common than others? A uniform information density account. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

McDonald, S. A. and Shillcock, R. C. (2003a). Eye Movements Reveal the On-Line Computation of Lexical Probabilities During Reading. *Psychological Science*, 14(6):648–652.

McDonald, S. A. and Shillcock, R. C. (2003b). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, 43(16):1735–1751.

McKenzie, I., Lyzhov, A., Parrish, A., Prabhu, A., Mueller, A., Kim, N., Bowman, S., and Perez, E. (2022a). The inverse scaling prize.

McKenzie, I., Lyzhov, A., Parrish, A., Prabhu, A., Mueller, A., Kim, N., Bowman, S., and Perez, E. (2022b). Inverse scaling prize: First round winners.

McKenzie, I. R., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A., Prabhu, A., McLean, E., Kirtland, A., Ross, A., Liu, A., Gritsevskiy, A., Wurgaft, D., Kauffman, D., Recchia, G., Liu, J., Cavanagh, J., Weiss, M., Huang, S., Droid, T. F., Tseng, T., Korbak, T., Shen, X., Zhang, Y., Zhou, Z., Kim, N., Bowman, S. R., and Perez, E. (2023). Inverse Scaling: When Bigger Isn't Better.

McRae, K., Brown, K. S., and Elman, J. L. (2019). Prediction-Based Learning and Processing of Event Knowledge. *Topics in Cognitive Science*, pages 1–18.

Medler, D. and Binder, J. (2005). MCWord: An On-Line Orthographic Database of the English Language.

Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., and Levy, R. (2021). Revisiting the Uniform Information Density Hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Menenti, L., Petersson, K. M., Scheeringa, R., and Hagoort, P. (2009). When Elephants Fly: Differential Sensitivity of Right and Left Inferior Frontal Gyri to Discourse and World Knowledge. *Journal of Cognitive Neuroscience*, 21(12):2358–2368.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2017). Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*.

Merkx, D. and Frank, S. L. (2021). Human Sentence Processing: Recurrence or Attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22, Online. Association for Computational Linguistics.

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

Meyer, A. S., Huettig, F., and Levelt, W. J. (2016). Same, different, or closely related: What is the relationship between language production and comprehension? *Journal of Memory and Language*, 89:1–7.

Michaelov, J. A., Bardolph, M. D., Coulson, S., and 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*, pages 300–306, University of Vienna, Vienna, Austria (Hybrid).

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., and Coulson, S. (2023). Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, pages 1–29.

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., and Coulson, S. (2024). Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, 5(1):107–135.

Michaelov, J. A. and Bergen, B. K. (2020). How well does surprisal explain N400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference*

on Computational Natural Language Learning, pages 652–663, Online. Association for Computational Linguistics.

Michaelov, J. A. and Bergen, B. K. (2022a). Collateral facilitation in humans and language models. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 13–26, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Michaelov, J. A. and 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*.

Michaelov, J. A. and Bergen, B. K. (2023). Ignoring the alternatives: The N400 is sensitive to stimulus preactivation alone. *Cortex*, 168:82–101.

Michaelov, J. A., Coulson, S., and 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*.

Mielke, S. J. (2016). Language diversity in ACL 2004 - 2016.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in Pre-Training Distributed Word Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In Burges, C. J. C.,

Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Miller, G. A. and Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, 2(3):217–228.

Misra, K., Ettinger, A., and Rayz, J. (2020). Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635, Online. Association for Computational Linguistics.

Mitchell, J., Lapata, M., Demberg, V., and Keller, F. (2010). Syntactic and semantic factors in processing difficulty: An integrated measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 196–206.

Mohebbi, H., Modarressi, A., and Pilehvar, M. T. (2021). Exploring the Role of BERT Token Representations to Explain Sentence Probing Results. *arXiv:2104.01477 [cs]*.

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

Muennighoff, N., Rush, A., Barak, B., Le Scao, T., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., and Raffel, C. A. (2023). Scaling Data-Constrained Language Models. In *Advances in Neural Information Processing Systems*, volume 36, pages 50358–50376.

Munro, R. (2015). Languages at ACL this year.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., But-

ton, K., Knight, M., Chess, B., and Schulman, J. (2022). WebGPT: Browser-assisted question-answering with human feedback.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, 106(3):226–254.

Neuwirth, E. (2022). *RColorBrewer: ColorBrewer Palettes*.

Newport, E. L. and Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48(2):127–162.

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

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and 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*, pages 4885–4901, Online. Association for Computational Linguistics.

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

Nieuwland, M. S., Martin, A. E., and Carreiras, M. (2013). Event-related brain potential evidence for animacy processing asymmetries during sentence comprehension. *Brain and Language*, 126(2):151–158.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., Donaldson, D. I., Kohút, Z., Rueschemeyer, S.-A., and Huettig, F. (2018a). Additional discussion of Yan, Kuperberg & Jaeger (2017). *Open Science Framework*.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., Donaldson, D. I., Kohút, Z., Rueschemeyer, S.-A., and Huettig, F. (2018b). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, 7:e33468.

Nieuwland, M. S. and Van Berkum, J. J. A. (2005). Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, 24(3):691–701.

Nieuwland, M. S. and van Berkum, J. J. A. (2006). When Peanuts Fall in Love: N400 Evidence for the Power of Discourse. *Journal of Cognitive Neuroscience*, 18(7):1098–1111.

Oh, B.-D., Clark, C., and Schuler, W. (2022). Comparison of Structural Parsers and Neural Language Models as Surprisal Estimators. *Frontiers in Artificial Intelligence*, 5.

Oh, B.-D. and Schuler, W. (2023a). Transformer-Based Language Model Surprisal Predicts Human Reading Times Best with About Two Billion Training Tokens. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921, Singapore. Association for Computational Linguistics.

Oh, B.-D. and Schuler, W. (2023b). Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Oh, B.-D., Yue, S., and Schuler, W. (2024). Frequency Explains the Inverse Correlation of Large Language Models' Size, Training Data Amount, and Surprisal's Fit to Reading Times. In Graham, Y. and Purver, M., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian's, Malta. Association for Computational Linguistics.

Oostdijk, N., Reynaert, M., Hoste, V., and Schuurman, I. (2013). The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch. In Spyns, P. and Odijk, J., editors, *Essential Speech and Language Technology for Dutch: Results by the STEVIN Programme*, Theory and Applications of Natural Language Processing, pages 219–247. Springer, Berlin, Heidelberg.

OpenAI (2021). OpenAI API.

OpenAI (2023). Model index for researchers.

Osterhout, L. and Mobley, L. A. (1995). Event-Related Brain Potentials Elicited by Failure to Agree. *Journal of Memory and Language*, 34(6):739–773.

Otten, M. and Berkum, J. J. A. V. (2008). Discourse-Based Word Anticipation During Language Processing: Prediction or Priming? *Discourse Processes*, 45(6):464–496.

Otten, M., Nieuwland, M. S., and Van Berkum, J. J. (2007). Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience*, 8(1):89.

Paczynski, M. and 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(9):1402–1456.

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

Paperno, D., Kruszewski, G., Lazaridou, A., Pham, N. Q., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernández, R. (2016). The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Parras, G. G., Nieto-Diego, J., Carbajal, G. V., Valdés-Baizabal, C., Escera, C., and Malmierca, M. S. (2017). Neurons along the auditory pathway exhibit a hierarchical organization of prediction error. *Nature Communications*, 8(1):2148.

Parviz, M., Johnson, M., Johnson, B., and Brock, J. (2011). Using Language Models and Latent Semantic Analysis to Characterise the N400m Neural Response. In *Proceed-*

ings of the Australasian Language Technology Association Workshop 2011, pages 38–46, Canberra, Australia.

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., and Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Perez, E., McKenzie, I., and Bowman, S. (2022). Announcing the Inverse Scaling Prize ($250k Prize Pool).

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Pezzelle, S., Steinert-Threlkeld, S., Bernardi, R., and Szymanik, J. (2018). Some of Them Can be Guessed! Exploring the Effect of Linguistic Context in Predicting Quantifiers.

In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 114–119, Melbourne, Australia. Association for Computational Linguistics.

Piantadosi, S. and Hill, F. (2022). Meaning without reference in large language models. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*.

Pickering, M. J. and Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, 11(3):105–110.

Pickering, M. J. and Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4):329–347.

Postman, L. (1951). Toward a general theory of cognition. In *Social Psychology at the Crossroads; the University of Oklahoma Lectures in Social Psychology*, pages 242–272. Harper, Oxford, England.

Prasad, G., van Schijndel, M., and Linzen, T. (2019). Using Priming to Uncover the Organization of Syntactic Representations in Neural Language Models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabovsky, M. (2020). Change in a probabilistic representation of meaning can account for N400 effects on articles: A neural network model. *Neuropsychologia*, 143:107466.

Rabovsky, M., Hansen, S. S., and McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, 2(9):693–705.

Rabovsky, M. and McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition*, 132(1):68–89.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. page 12.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. page 24.

Radvansky, G. A., Zwaan, R. A., Federico, T., and Franklin, N. (1998). Retrieval from temporally organized situation models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(5):1224–1237.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., van den Driessche, G., Hendricks, L. A., Rauh, M., Huang, P.-S., Glaese, A., Welbl, J., Dathathri, S., Huang, S., Uesato, J., Mellor, J., Higgins, I., Creswell, A., McAleese, N., Wu, A., Elsen, E., Jayakumar, S., Buchatskaya, E., Budden, D., Sutherland, E., Simonyan, K., Paganini, M., Sifre, L., Martens, L., Li, X. L., Kuncoro, A., Nematzadeh, A., Gribovskaya, E., Donato, D., Lazaridou, A., Mensch, A., Lespiau, J.-B., Tsimpoukelli, M., Grigorev, N., Fritz, D., Sottiaux, T., Pajarskas, M., Pohlen, T.,

Gong, Z., Toyama, D., d'Autume, C. d. M., Li, Y., Terzi, T., Mikulik, V., Babuschkin, I., Clark, A., Casas, D. d. L., Guy, A., Jones, C., Bradbury, J., Johnson, M., Hechtman, B., Weidinger, L., Gabriel, I., Isaac, W., Lockhart, E., Osindero, S., Rimell, L., Dyer, C., Vinyals, O., Ayoub, K., Stanway, J., Bennett, L., Hassabis, D., Kavukcuoglu, K., and Irving, G. (2022). Scaling Language Models: Methods, Analysis & Insights from Training Gopher.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Raji, D., Denton, E., Bender, E. M., Hanna, A., and Paullada, A. (2021). AI and the Everything in the Whole Wide World Benchmark. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.

Raji, I. D., Kumar, I. E., Horowitz, A., and Selbst, A. (2022). The Fallacy of AI Functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 959–972, New York, NY, USA. Association for Computing Machinery.

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

Roark, B., Bachrach, A., Cardenas, C., and Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 1, page 324, Singapore. Association for Computational Linguistics.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2021). A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Romberg, A. R. and Saffran, J. R. (2010). Statistical learning and language acquisition. *WIREs Cognitive Science*, 1(6):906–914.

Rommers, J. and Federmeier, K. D. (2018). Lingering expectations: A pseudo-repetition effect for words previously expected but not presented. *NeuroImage*, 183:263–272.

Rommers, J., Meyer, A. S., Praamstra, P., and Huettig, F. (2013). The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. *Neuropsychologia*, 51(3):437–447.

Rosenbach, A. (2008). Animacy and grammatical variation—Findings from English genitive variation. *Lingua*, 118(2):151–171.

RStudio Team (2020). *RStudio: Integrated Development Environment for r*. RStudio, PBC., Boston, MA.

Ruan, Y., Maddison, C. J., and Hashimoto, T. (2024). Observational scaling laws and the predictability of language model performance.

Rubin, J., Ulanovsky, N., Nelken, I., and Tishby, N. (2016). The Representation of Prediction Error in Auditory Cortex. *PLOS Computational Biology*, 12(8):e1005058.

Rugg, M. D. (1985). The Effects of Semantic Priming and Word Repetition on Event-Related Potentials. *Psychophysiology*, 22(6):642–647.

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

Ryskin, R., Stearns, L., Bergen, L., Eddy, M., Fedorenko, E., and Gibson, E. (2021). An ERP index of real-time error correction within a noisy-channel framework of human communication. *Neuropsychologia*, page 107855.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, 274(5294):1926–1928.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. (2019). WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *arXiv:1907.10641 [cs]*.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. (2020). WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740.

Sanders, A. F. (1966). Expectancy: Application and measurement. *Acta Psychologica*, 25:293–313.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter.

Sardana, N. and Frankle, J. (2023). Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws. In *The 3rd NeurIPS Workshop on Efficient Natural Language and Speech Processing (ENLSP-III)*, New Orleans, LA, USA.

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

Schäfer, R. and Bildhauer, F. (2012). Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, Turkey. European Language Resources Association (ELRA).

Schoelkopf, H. (2024). Comment on Issue #1269 of EleutherAI/lm-evaluation-harness.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J., and Fedorenko, E. (2020). The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *bioRxiv*, page 2020.06.26.174482.

Schwanenflugel, P. J. and Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language*, 24(2):232–252.

Seidenberg, M. S. (1997). Language Acquisition and Use: Learning and Applying Probabilistic Constraints. *Science*, 275(5306):1599–1603.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Seriès, P. and Seitz, A. (2013). Learning what to expect (in visual perception). *Frontiers in Human Neuroscience*, 7.

Shain, C. (2024). Word Frequency and Predictability Dissociate in Naturalistic Reading. *Open Mind*, 8:177–201.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., and Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Sherman, B. E., Graves, K. N., and 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.

Sherman, B. E. and Turk-Browne, N. B. (2020). Statistical prediction of the future impairs episodic encoding of the present. *Proceedings of the National Academy of Sciences*, 117(37):22760–22770.

Shipp, S., Adams, R. A., and Friston, K. J. (2013). Reflections on agranular architecture: Predictive coding in the motor cortex. *Trends in Neurosciences*, 36(12):706–716.

Shlegeris, B., Roger, F., Chan, L., and McLean, E. (2022). Language models are better than humans at next-token prediction.

Silge, J. and Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in R. *JOSS*, 1(3).

Sinclair, A., Jumelet, J., Zuidema, W., and Fernández, R. (2022). Structural Persistence in Language Models: Priming as a Window into Abstract Language Representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

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

Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Smith, N. J. and Levy, R. P. (2008). Optimal Processing Times in Reading: A Formal Model and Empirical Investigation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

Speer, R., Chin, J., Lin, A., Jewett, S., and Nathan, L. (2018). LuminosoInsight/wordfreq: V2.2. Zenodo.

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., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A., La, A., Lampinen, A., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ramírez, C. F., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C. E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodola, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G., Jaimovitch-López, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta,

H., Bogar, H., Shevlin, H., Schütze, H., Yakura, H., Zhang, H., Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Kocoń, J., Thompson, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Jones, J., Tenenbaum, J. B., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K. D., Gimpel, K., Omondi, K., Mathewson, K., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Colón, L. O., Metz, L., Şenel, L. K., Bosma, M., Sap, M., ter Hoeve, M., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Quintana, M. J. R., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M. V., Peng, N., Chi, N., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N. S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P., Eckersley, P., Htut, P. M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Delgado, R. R., Millière, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymaekers, R., Frank, R.,

Sikand, R., Novak, R., Sitelew, R., LeBras, R., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R., Lee, R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Shyamolima, Debnath, Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S. T., Shieber, S. M., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Telleen-Lawton, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V., Prabhu, V. U., Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, Y., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z., and Wu, Z. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models.

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., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A. J., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A. M., La, A., Lampinen, A., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A.,

Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ferri, C., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C. E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, C. D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Schrader, D., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodolà, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G. X., Jaimovitch-Lopez, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H. F. A., Schuetze, H., Yakura, H., Zhang, H., Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Kocon, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J. B., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K.,

Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K., Gimpel, K., Omondi, K., Mathewson, K. W., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Oliveros-Colón, L., Metz, L., Senel, L. K., Bosma, M., Sap, M., Hoeve, M. T., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Ramirez-Quintana, M. J., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Walker, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M. V., Peng, N., Chi, N. A., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Martinez, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N. S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P. W., Eckersley, P., Htut, P. M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Risco, R., Millière, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., Bras, R. L., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R. A., Lee, S. R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Debnath, S. S., Shakeri, S., Thormeyer, S., Melzi, S., Reddy,

S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S., Shieber, S., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V. V., Prabhu, V. U., Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, Y., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z., and Wu, Z. (2023). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Stability AI (2023). StableLM-Base-Alpha 7B.

Staub, A., Grant, M., Astheimer, L., and Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, 82:1–17.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. (2020). Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Stone, K., Vasishth, S., and von der Malsburg, T. (2021). Does entropy modulate the prediction of German long-distance verb particles? Data and code.

Stone, K., Vasishth, S., and von der Malsburg, T. (2022). Does entropy modulate the prediction of German long-distance verb particles? *PLOS ONE*, 17(8):e0267813.

Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M.-M., and Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience*, 11(9):1004–1006.

Szewczyk, J. and Federmeier, K. D. (2021). Context-Based Facilitation of Semantic Access Follows Both Logarithmic and Linear Functions of Stimulus Probability.

Szewczyk, J. M. and Federmeier, K. D. (2022). Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language*, 123:104311.

Szewczyk, J. M., Mech, E. N., and 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.

Szewczyk, J. M. and Schriefers, H. (2011). Is animacy special?: ERP correlates of semantic violations and animacy violations in sentence processing. *Brain Research*, 1368:208–221.

Szewczyk, J. M. and Schriefers, H. (2013). Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. *Journal of Memory and Language*, 68(4):297–314.

Talmor, A., Elazar, Y., Goldberg, Y., and Berant, J. (2020). oLMpics-On What Language Model Pre-training Captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Tannenbaum, P. H., Williams, F., and Hillier, C. S. (1965). Word predictability in the environments of hesitations. *Journal of Verbal Learning and Verbal Behavior*, 4(2):134–140.

Tay, Y., Dehghani, M., Abnar, S., Chung, H., Fedus, W., Rao, J., Narang, S., Tran, V., Yogatama, D., and Metzler, D. (2023). Scaling Laws vs Model Architectures: How does Inductive Bias Influence Scaling? In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12342–12364, Singapore. Association for Computational Linguistics.

Taylor, W. L. (1953). "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433.

Taylor, W. L. (1957). "Cloze" readability scores as indices of individual differences in comprehension and aptitude. *Journal of Applied Psychology*, 41(1):19–26.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhao, V., Zhou, Y., Chang, C.-C., Krivokon, I., Rusch, W., Pickett, M., Srinivasan, P., Man, L., Meier-Hellstern, K., Morris, M. R., Doshi, T., Santos, R. D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera-Arcas, B., Cui, C., Croak, M., Chi, E., and Le, Q. (2022). LaMDA: Language Models for Dialog Applications.

Thornhill, D. E. and Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology*, 83(3):382–392.

Tirumala, K., Markosyan, A., Zettlemoyer, L., and Aghajanyan, A. (2022). Memoriza-

tion Without Overfitting: Analyzing the Training Dynamics of Large Language Models. *Advances in Neural Information Processing Systems*, 35:38274–38290.

Todorovic, A. and de Lange, F. P. (2012). Repetition Suppression and Expectation Suppression Are Dissociable in Time in Early Auditory Evoked Fields. *Journal of Neuroscience*, 32(39):13389–13395.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models.

Traxler, M. J. and Foss, D. J. (2000). Effects of sentence constraint on priming in natural language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(5):1266–1282.

Tsarfaty, R., Seddah, D., Kübler, S., and Nivre, J. (2013). Parsing Morphologically Rich Languages: Introduction to the Special Issue. *Computational Linguistics*, 39(1):15–22.

Tulkens, S., Emmery, C., and Daelemans, W. (2016). Evaluating Unsupervised Dutch Word Embeddings as a Linguistic Resource. In *Proceedings of the Tenth International Con-*

ference on Language Resources and Evaluation (LREC'16), pages 4130–4136, Portorož, Slovenia. European Language Resources Association (ELRA).

Tulving, E. and Gold, C. (1963). Stimulus information and contextual information as determinants of tachistoscopic recognition of words. *Journal of Experimental Psychology*, 66(4):319–327.

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

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

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

Van Berkum, J. J. A. (2009). The Neuropragmatics of 'Simple' Utterance Comprehension: An ERP Review. In Breheny, R., Sauerland, U., and Yatsushiro, K., editors, *Semantics and Pragmatics: From Experiment to Theory*, pages 276–316. Palgrave Macmillan.

Van Berkum, J. J. A. (2010). The brain is a prediction machine that cares about good and bad - Any implications for neuropragmatics? *Italian Journal of Linguistics*, 22:181–208.

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

Van Berkum, J. J. A., Koornneef, A. W., Otten, M., and Nieuwland, M. S. (2007). Establishing reference in language comprehension: An electrophysiological perspective. *Brain Research*, 1146:158–171.

van den Brand, T. (2021). *Ggh4x: Hacks for 'Ggplot2'*.

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

van Dijk, T. A. and Kintsch, W. (1983). *Strategies of Discourse Comprehension*. Academic Press, New York.

van Erven, T. and Harremos, P. (2014). Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

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

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

Van Petten, C., Coulson, S., Rubin, S., Plante, E., and Parks, M. (1999). Time Course of Word Identification and Semantic Integration in Spoken Language. *Journal of Experimental Psychology: Learning Memory and Cognition*, 25(2):394–417.

Van Petten, C. and Kutas, M. (1988). Tracking the time course of meaning activation. In Small, S., Cottrell, G., and Tanenhaus, M., editors, *Lexical Ambiguity Resolution in the Comprehension of Human Language*, pages 431–475. Morgan Kaufmann Publishers, San Mateo, California.

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

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

Van Petten, C. and Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, 83(2):176–190.

Van Petten, C., Weckerly, J., McIsaac, H. K., and Kutas, M. (1997). Working Memory Capacity Dissociates Lexical and Sentential Context Effects. *Psychological Science*, 8(3):238–242.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

Vega-Mendoza, M., Pickering, M. J., and Nieuwland, M. S. (2021). Concurrent use of animacy and event-knowledge during comprehension: Evidence from event-related potentials. *Neuropsychologia*, 152:107724.

Venhuizen, N. J., Crocker, M. W., and Brouwer, H. (2019). Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Processes*, 56(3):229–255.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M.,

Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261–272.

Vissers, C. T. W. M., Chwilla, D. J., and Kolk, H. H. J. (2006). Monitoring in language perception: The effect of misspellings of words in highly constrained sentences. *Brain Research*, 1106(1):150–163.

von Helmholtz, H. (1867). *Handbuch der Physiologischen Optik [Handbook of Physiological Optics]*, volume 3 of *Allgemeinen Encyclopädie der Physik*. Leopold Voss, Leipzig.

Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T., Naccache, L., and Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences*, 108(51):20754–20759.

Wagenmakers, E.-J. and Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11(1):192–196.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019a). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 3266–3280. Curran Associates, Inc.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019b). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*.

Wang, B. and Komatsuzaki, A. (2021). GPT-J-6B: A 6 billion parameter autoregressive language model.

Wang, L., Wlotko, E., Alexander, E., Schoot, L., Kim, M., Warnke, L., and 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(16):3278–3291.

Warren, T., McConnell, K., and Rayner, K. (2008). Effects of context on eye movements when reading about possible and impossible events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(4):1001–1010.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., and Bowman, S. R. (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022a). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.

Wei, J., Tay, Y., and Le, Q. V. (2022b). Inverse scaling can become U-shaped.

West, R. F. and Stanovich, K. E. (1982). Source of inhibition in experiments on the effect of sentence context on word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(5):385–399.

White, M., Haddad, I., Osborne, C., Yanglet, X.-Y. L., Abdelmonsef, A., and Varghese, S. (2024). The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency, and Usability in Artificial Intelligence.

Wicha, N. Y., Bates, E. A., Moreno, E. M., and Kutas, M. (2003a). Potato not Pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters*, 346(3):165–168.

Wicha, N. Y., Moreno, E. M., and Kutas, M. (2003b). Expecting Gender: An Event Related Brain Potential Study on the Role of Grammatical Gender in Comprehending a Line Drawing Within a Written Sentence in Spanish. *Cortex*, 39(3):483–508.

Wicha, N. Y. Y., Moreno, E. M., and Kutas, M. (2004). Anticipating Words and Their Gender: An Event-related Brain Potential Study of Semantic Integration, Gender Expectancy, and Gender Agreement in Spanish Sentence Reading. *Journal of Cognitive Neuroscience*, 16(7):1272–1288.

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., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Wilcox, E., Levy, R., Morita, T., and Futrell, R. (2018). What do RNN Language Models Learn about Filler–Gap Dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Wilcox, E., Meister, C., Cotterell, R., and Pimentel, T. (2023a). Language Model Quality

Correlates with Psychometric Predictive Power in Multiple Languages. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511, Singapore. Association for Computational Linguistics.

Wilcox, E., Qian, P., Futrell, R., Ballesteros, M., and Levy, R. (2019). Structural Supervision Improves Learning of Non-Local Grammatical Dependencies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3302–3312, Minneapolis, Minnesota. Association for Computational Linguistics.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., and Levy, R. P. (2020). On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci 2020)*, page 7.

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., and Levy, R. P. (2023b). Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Wilke, C. O. (2020). *Cowplot: Streamlined Plot Theme and Plot Annotations for 'Ggplot2'*.

Wilke, C. O. and Wiernik, B. M. (2022). *Ggtext: Improved Text Rendering Support for 'Ggplot2'*.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., and van den Bosch, A. (2016). Prediction During Natural Language Comprehension. *Cerebral Cortex*, 26(6):2506–2516.

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Wlotko, E. W. and Federmeier, K. D. (2007). Finding the right word: Hemispheric asymmetries in the use of sentence context information. *Neuropsychologia*, 45(13):3001–3014.

Wlotko, E. W. and Federmeier, K. D. (2012). So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *NeuroImage*, 62(1):356–366.

Wolen, A. R., Hartgerink, C. H., Hafen, R., Richards, B. G., Soderberg, C. K., and York, T. P. (2020). osfr: An R interface to the open science framework. *Journal of Open Source Software*, 5(46):2071.

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., Drame, M., Lhoest, Q., and 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*, pages 38–45, Online. Association for Computational Linguistics.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman and Hall/CRC, Boca Raton, 2 edition.

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

Wu, Y. C. and Coulson, S. (2011). Are depictive gestures like pictures? Commonalities and differences in semantic processing. *Brain and Language*, 119(3):184–195.

Wyart, V., Nobre, A. C., and Summerfield, C. (2012). Dissociable prior influences of signal

probability and relevance on visual contrast sensitivity. *Proceedings of the National Academy of Sciences*, 109(9):3593–3598.

Xiang, M. and Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*, 30(6):648–672.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., and Matsumoto, Y. (2020). Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30, Online. Association for Computational Linguistics.

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

Yarkoni, T., Balota, D., and Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15(5):971–979.

Yehezkel, S. and Pinter, Y. (2023). Incorporating Context into Subword Vocabularies.

Zacks, J. M. and Ferstl, E. C. (2016). Discourse Comprehension. In Hickok, G. and Small, S. L., editors, *Neurobiology of Language*, pages 661–673. Academic Press, San Diego.

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

Zeileis, A., Hornik, K., and Murrell, P. (2009). Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, 53(9):3259–3270.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. (2019). HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. (2022). OPT: Open Pre-trained Transformer Language Models.

Zhuang, C., Fedorenko, E., and Andreas, J. (2024). Visual Grounding Helps Learn Word Meanings in Low-Data Regimes. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1311–1329, Mexico City, Mexico. Association for Computational Linguistics.

Zwaan, R. A. (2014). Embodiment and language comprehension: Reframing the discussion. *Trends in Cognitive Sciences*, 18(5):229–234.

Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review*, 23(4):1028–1034.

Zwaan, R. A., Langston, M. C., and Graesser, A. C. (1995a). The Construction of Situation Models in Narrative Comprehension: An Event-Indexing Model. *Psychological Science*, 6(5):292–297.

Zwaan, R. A. and Madden, C. J. (2004). Updating situation models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(1):283–288.

Zwaan, R. A., Magliano, J. P., and Graesser, A. C. (1995b). Dimensions of situation model construction in narrative comprehension. *Journal of experimental psychology: Learning, memory, and cognition*, 21(2):386.

Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162–185.