Collateral facilitation in humans and language models

James A. Michaelov
Department of Cognitive Science
University of California, San Diego
j1michae@ucsd.edu

Benjamin K. Bergen
Department of Cognitive Science
University of California, San Diego
bkbergen@ucsd.edu

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.

1 Introduction

Humans process words more easily when they more contextually predictable, whether predictability is determined by humans (Fischler and Bloom, 1979; Brothers and Kuperberg, 2021) or language models (McDonald and Shillcock, 2003; 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; De-Long 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 psycholinguistic 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.

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, 2019; 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, 2019; 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, 2019; 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 seman-

tically 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.

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

4   Experiment 1: Ito et al. (2016)

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 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
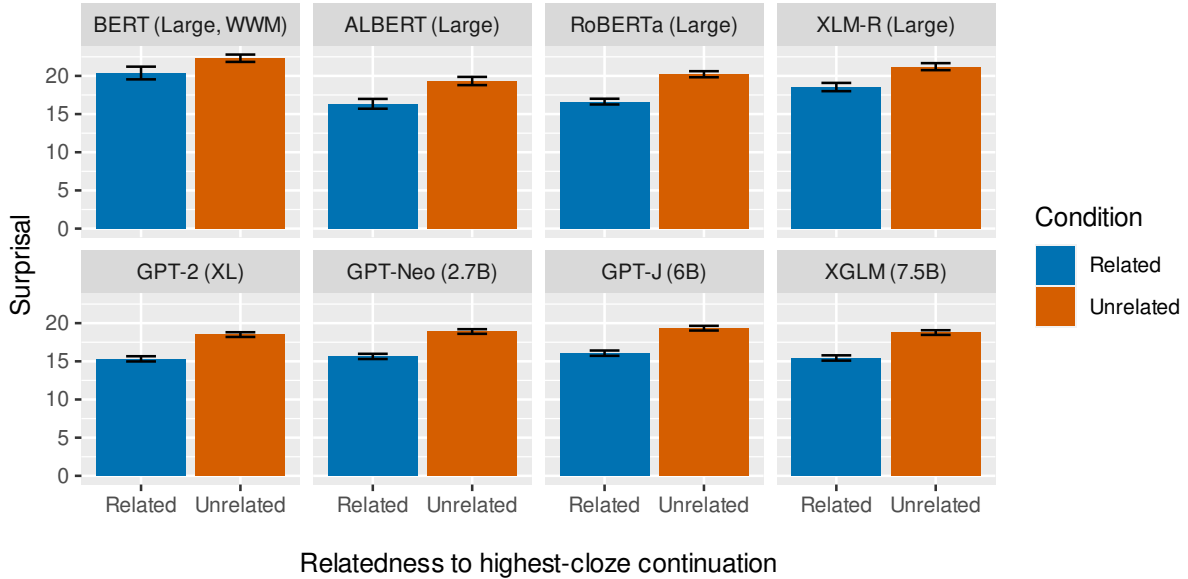
Figure 1: Mean surprisal elicited by each language model for the Ito et al. (2016) stimuli related and unrelated to the most probable (highest-cloze) continuation of each sentence. Error bars indicate standard error.

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

Table 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.

than the surprisal elicited by implausible stimuli unrelated to the highest-cloze continuation.

4.2 Results

The results of the experiment are shown in Figure 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 1. As can be seen, condition is a significant predictor of the surprisal from every language model, confirm-
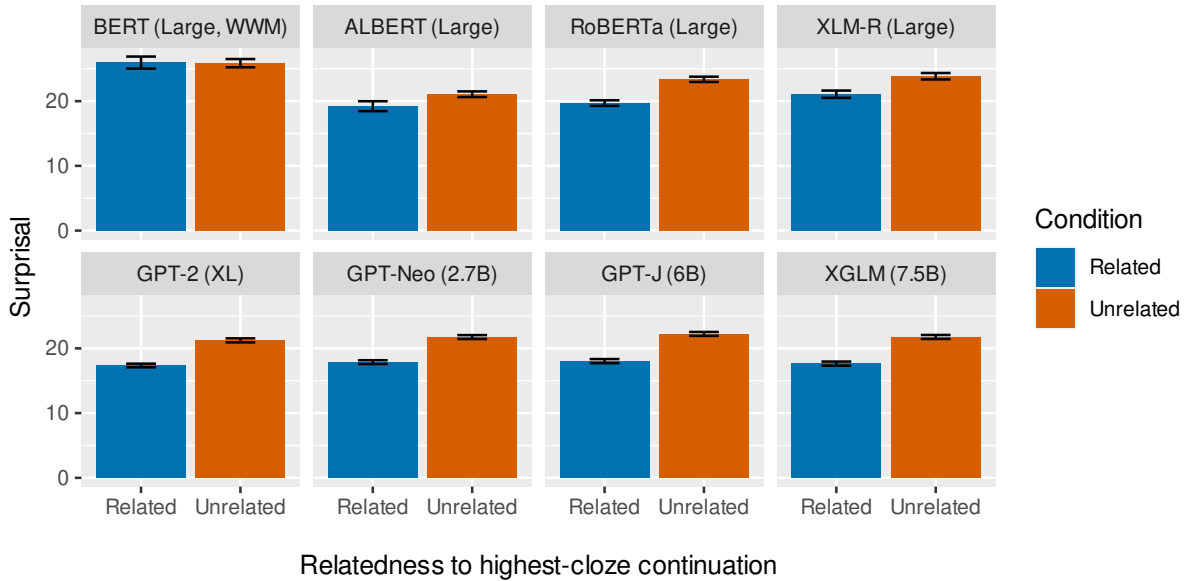
Figure 2: Mean surprisal elicited by each language model for the DeLong et al. (2019) stimuli related and unrelated to the most probable (highest-cloze) continuation of each sentence. Error bars indicate standard error.

ing 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.

5 Experiment 2: DeLong et al. (2019)

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.

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

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

Table 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.

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

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 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 experimental 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 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.

6 Experiment 3: Metusalem et al. (2012)

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.

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 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.
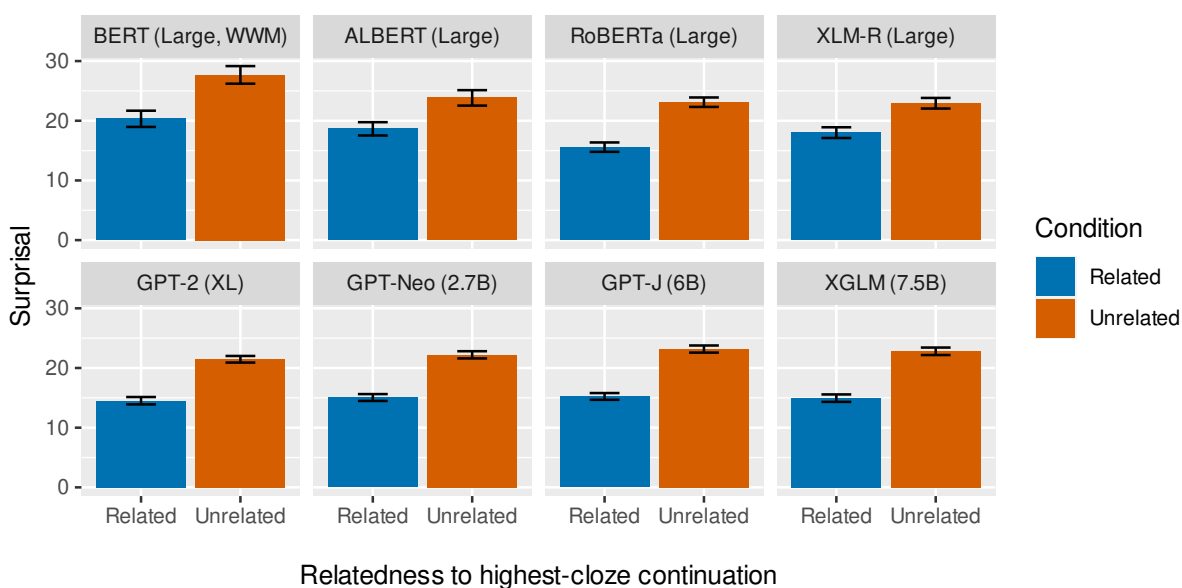

Figure 3: Mean surprisal elicited by each language model for the Metusalem et al. (2012) stimuli related and unrelated to the most probable (highest-cloze) continuation of each sentence. Error bars indicate standard error.

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

Table 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.

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 3. As can be seen, experimental condition was a significant predictor of the surprisal of all models.

7 General Discussion

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 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.

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.

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 mono-

lingual 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, 2019; Michaelov and Bergen, 2020; Aurnhammer and Frank, 2019; 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).

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.

References

Adam Amram, Anat Ben David, and Reut Tsarfaty. 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.

Christoph Aurnhammer and Stefan L. Frank. 2019. Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, 134:107198.

Megan Bardolph, Cyma Van Petten, and Seana Coulson. 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.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Emily Bender. 2019. The #BenderRule: On naming the languages we study and why it matters. *The Gradient*.

Emily M. Bender. 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.

Emily M. Bender. 2011. On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology*, 6.

Yoav Benjamini and Yosef Hochberg. 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.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow. Zenodo.

Ina Bornkessel-Schlesewsky and Matthias Schlesewsky. 2019. Toward a Neurobiologically Plausible Model of Language-Related, Negative Event-Related Potentials. *Frontiers in Psychology*, 10.

Trevor Brothers and Gina R. Kuperberg. 2021. Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, 116:104174.

Harm Brouwer, Matthew W. Crocker, Noortje J. Venhuizen, and John C. J. Hoeks. 2017. A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive Science*, 41(S6):1318–1352.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Samuel J. Cheyette and David C. Plaut. 2017. Modeling the N400 ERP component as transient semantic overactivation within a neural network model of word comprehension. *Cognition*, 162:153–166.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 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.

Katherine A. DeLong, Wen-hsuan Chan, and Marta Kutas. 2019. Similar time courses for word form and meaning preactivation during sentence comprehension. *Psychophysiology*, 56(4):e13312.

Katherine A. DeLong and Marta Kutas. 2020. Comprehending surprising sentences: Sensitivity of post-N400 positivities to contextual congruity and semantic relatedness. *Language, Cognition and Neuroscience*, 35(0):1044–1063.

Katherine A DeLong, Thomas P Urbach, and Marta Kutas. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8(8):1117–1121.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

Tiwalayo Eisape, Noga Zaslavsky, and Roger Levy. 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.

Allyson Ettinger. 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.

Allyson Ettinger, Naomi Feldman, Philip Resnik, and Colin Phillips. 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.

Kara D. Federmeier. 2021. Connecting and considering: Electrophysiology provides insights into comprehension. *Psychophysiology*, n/a(n/a):e13940.

Kara D. Federmeier and Marta Kutas. 1999. A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, 41(4):469–495.

Fernanda Ferreira and Zoe Yang. 2019. The Problem of Comprehension in Psycholinguistics. *Discourse Processes*, 56(7):485–495.

Ira Fischler and Paul A. Bloom. 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.

Hartmut Fitz and Franklin Chang. 2019. Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, 111:15–52.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Stefan L. Frank and Roel M. Willems. 2017. Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, 32(9):1192–1203.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 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.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 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.

Aine Ito, Martin Corley, Martin J. Pickering, Andrea E. Martin, and Mante S. Nieuwland. 2016. Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, 86:157–171.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the Limits of Language Modeling. *arXiv:1602.02410 [cs]*.

Nora Kassner and Hinrich Schütze. 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.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-Aware Neural Language Models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Gina R. Kuperberg, Trevor Brothers, and Edward W. Wlotko. 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.

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

Marta Kutas, Katherine A. DeLong, and Nathaniel J. Smith. 2011. A look around at what lies ahead: Prediction and predictability in language processing. In Moshe Bar, editor, *Predictions in the Brain: Using Our Past to Generate a Future*, pages 190–207. Oxford University Press, New York, NY, US.

Marta Kutas and Steven A. Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163.

Marta Kutas, Timothy E Lindamood, and Steven A Hillyard. 1984. Word expectancy and event-related brain potentials during sentence processing. In S. Kornblum and J. Requin, editors, *Preparatory States and Processes*, pages 217–237. Lawrence Erlbaum, Hillsdale, NJ.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82:1–26.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Sarah Laszlo and Blair C. Armstrong. 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.

Sarah Laszlo and David C. Plaut. 2012. A neurally plausible Parallel Distributed Processing model of Event-Related Potential word reading data. *Brain and Language*, 120(3):271–281.

Ellen F. Lau, Phillip J. Holcomb, and Gina R. Kuperberg. 2013. Dissociating N400 Effects of Prediction from Association in Single-word Contexts. *Journal of Cognitive Neuroscience*, 25(3):484–502.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot Learning with Multilingual Language Models. *arXiv:2112.10668 [cs]*.

Alma Lindborg and Milena Rabovsky. 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).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Scott A. McDonald and Richard C. Shillcock. 2003. Eye Movements Reveal the On-Line Computation of Lexical Probabilities During Reading. *Psychological Science*, 14(6):648–652.

Danny Merkx and Stefan L. Frank. 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.

Ross Metusalem, Marta Kutas, Thomas P. Urbach, Mary Hare, Ken McRae, and Jeffrey L. Elman. 2012. Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, 66(4):545–567.

James A. Michaelov, Megan D. Bardolph, Seana Coulson, and Benjamin K. Bergen. 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).

James A. Michaelov and Benjamin K. Bergen. 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.

James A. Michaelov, Seana Coulson, and Benjamin K. Bergen. 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*.

Sabrina J. Mielke. 2016. Language diversity in ACL 2004 - 2016.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 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.

Robert Munro. 2015. Languages at ACL this year.

Mehdi Parviz, Mark Johnson, Blake Johnson, and Jon Brock. 2011. Using Language Models and Latent Semantic Analysis to Characterise the N400m Neural Response. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 38–46, Canberra, Australia.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 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.

Milena Rabovsky, Steven S. Hansen, and James L. McClelland. 2018. Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, 2(9):693–705.

Milena Rabovsky and Ken McRae. 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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. page 24.

Joost Rommers, Antje S. Meyer, Peter Praamstra, and Falk Huettig. 2013. The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. *Neuropsychologia*, 51(3):437–447.

RStudio Team. 2020. *RStudio: Integrated Development Environment for r*. RStudio, PBC., Boston, MA.

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. 2020. The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *bioRxiv*, page 2020.06.26.174482.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Jakub M. Szewczyk and Kara D. Federmeier. 2022. Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language*, 123:104311.

Wilson L. Taylor. 1953. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433.

Wilson L. Taylor. 1957. "Cloze" readability scores as indices of individual differences in comprehension and aptitude. *Journal of Applied Psychology*, 41(1):19–26.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications.

Reut Tsarfaty, Djamé Seddah, Sandra Kübler, and Joakim Nivre. 2013. Parsing Morphologically Rich Languages: Introduction to the Special Issue. *Computational Linguistics*, 39(1):15–22.

Takahisa Uchida, Nicolas Lair, Hiroshi Ishiguro, and Peter Ford Dominey. 2021. A Model of Online Temporal-Spatial Integration for Immediacy and Overrule in Discourse Comprehension. *Neurobiology of Language*, 2(1):83–105.

Cyma Van Petten and Barbara J. Luka. 2012. Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, 83(2):176–190.

Noortje J. Venhuizen, Matthew W. Crocker, and Harm Brouwer. 2019. Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Processes*, 56(3):229–255.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pretrained Transformer Language Models.

A 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.

B Models used

The details of the models used in this study are provided in Table 4.

Model Name	Full Name on the Hugging Face Model Hub	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)

Table 4: Transformer langauge models used in the present study. All were accessed using the transformers (Wolf et al., 2020) package.