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

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

***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 amplitude 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.**

***Index Terms*—N400, language, prediction, psycholinguistics, language comprehension, natural language processing, deep learning, neural language models, electrophysiology, electroencephalography (EEG), event-related brain potential (ERP).**

I. 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 [1], [2], [3], [4]). One prominent position is that the language processor does not waste resources on predictive processing [5]. 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 [6]. 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 [7], [8], some form of automatic spreading activation in the lexicon [9], [10], 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 [11], [12]. 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 [2], [13], [14]. 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 [15], [16], [17], [18], [14], [19], [20], [21].

What are these predictions based on? Since the early days of N400 research, cloze probability [22] has served as the chief metric of contextual word predictability [23], [2], [24]. 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 [22], 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 [23], [12], [14], as well as being read faster and recognized faster [24]—in fact, some work has shown that cloze probability and N400 amplitude are inversely correlated at a level of over 90% [25]. 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 [26]. 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 [27], [24].

Understanding the relationship between N400 amplitude and the statistics of language is vital to understanding the N400 [28]. Given the evidence that N400 amplitude is affected by linguistic input over the lifespan [12], 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

This work was partially supported by a 2020-2021 Center for Academic Research and Training in Anthropogeny Fellowship awarded to J.A. Michaelov.

J.A. Michaelov, S. Coulson, and B.K. Bergen are with the Department of Cognitive Science, University of California San Diego, La Jolla, CA 92093 USA (email: j1michae@ucsd.edu).

Published paper DOI: 10.1109/TCDS.2022.3176783

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 [29], [30], [31], [32] and the significant differences in N400 amplitude elicited by a range of experimental manipulations [28], but at least for some stimuli may be better at this than cloze probability [28], [32]. However, the two studies in which LM predictions outperform cloze have either looked at the effects without direct comparison to the N400 data [28] or targeted data from an experiment intended to show the N400 responds to factors other than cloze [32].

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 [33], 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 [33], and compared its performance to that of several pretrained neural network LMs [34], [35], [36], [37], [38], [39], [40], [41]. 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.

II. Background

A. Cloze probability

Cloze probability has long been used to asses a word's predictability in context [2], [42], [3], [24]. In addition to its use in understanding the N400 [23], [12], it has been shown to predict behavioural correlates of processing difficulty, such as word reading time [24]. In fact, when directly compared, cloze probability has previously be found to be better at predicting such behavioural metrics than LMs [24].

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 prediction. 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 [43], [44] 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 [13], [14], 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.

B. 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 [12], [14], 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 [45]. 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 [46].

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 [12], 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.

C. 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 (1).

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

In addition to theoretical claims behind surprisal theory as an explanation of predictability effects in language comprehension [47], [48], [49], there is also an array of evidence showing that LM surprisal correlates with behavioural metrics of processing difficulty such as reading time [50], [51], [52], [53], [27], [54], [55]. 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 [29], [30], [31], [32]. 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 [28], [32].

In previous work, operationalizing predictability as cloze probability generally appears to yield better results for human behavioural data than LM surprisal [24]; 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 [32]. 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 [23], [56], [57], [58], [59]. 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 [27], [60], 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 [28], 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.

D. 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) [33]. 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 [28], [32]). 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 [61], [59], and some of these studies have been successfully modeled using LM predictions [28], 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 [27], [60], [62].

III. Method

A. Original study and data

We use EEG data from a large-scale experiment by Nieuwland and colleagues [33]. 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 [33], [23], [12].

To calculate the cloze probability of items in the original study, each stimulus sentence was truncated before the target word [33]. 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 [60].

B. 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.

1) RNNs: Until the development of transformer LMs [63], recurrent neural network (RNN) language models long dominated the field. With their memory bottleneck and their incremental processing of words [64], [31], RNNs have often been used as cognitive models of human language processing [65], including prior efforts to model the N400 [29], [30], [28], [31], [32]. In the present study, we use two RNN LMs referred to in the literature (see, e.g., [66]) as GRNN [34] and JRNN [35]. Previous research has found JRNN surprisal to more closely resemble N400 amplitude than does GRNN surprisal [28]. GRNN and JRNN surprisal were calculated using the code accompanying Michaelov and Bergen [28].

2) Transformers: Transformer language models are a neural network LM architecture [63] that has been found to outperform RNNs at the standard language modeling task (predicting words from context, see [39] for review), as well as a range of other tasks [36], [38]. Transformer LMs have also been shown to do better than RNNs at predicting N400 amplitude [31], [32]. 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 [39], GPT-2 [38], and GPT-3 [41]. The three masked LMs that we use to calculate word probability are BERT [36], RoBERTa [37], and ALBERT [40]. For all transformer LMs except for GPT-3, we use the implementation of each model made available through the *transformers* [67] package to calculate surprisal. GPT-3 predictions were accessed via the OpenAI API [68].

C. Language model predictions

The aforementioned LMs were thus used to predict the probability of the target nouns from the original study [33]. 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 (1). We used a logarithm of base 2 so that surprisal can be measured in bits [66]. For fair comparison, only words appearing in all models' vocabularies were included in the analysis.

D. 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.

TABLE I
SUMMARY OF LANGUAGE MODELS USED

Model	Parameters[1]	Corpus size[2]	Ref.
GRNN	71.8M	90M	[34]
JRNN	1.04B	1B	[35]
Transformer-XL[3]	285M	103M	[39]
GPT-2 (XL)	1.56B	∼8B	[38]
GPT-3 (Davinci)	175B	∼300B	[41]
BERT (large, cased, WWM[4])	334M	3.3B	[36]
RoBERTa (large)	355M	∼33B	[37]
ALBERT (XXLarge v2)[5]	206M	3.3B	[40]

[1] The number of free parameters for the *transformers* [67] implementations of Transformer-XL, GPT-2, BERT, RoBERTa, and ALBERT were calculated using *pytorch* [69]. 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 [35], [41]. While we use the author-provided GRNN, no estimate of model parameters is given in the original paper [34], so we calculated this with *pytorch* [69].
[2] Number of words in training corpus is reported in the original papers [34], [35], [39], [36], or estimated (denoted by '∼'). ALBERT is trained on the same data as BERT [40]. 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 [41], [38], [70], the actual number of words is lower.
[3] We use the *transformers* [67] implementation of Transformer-XL; some models reported in the original paper [39] have a higher number of parameters.
[4] Whole-word masking, see [71].
[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 [40].

Statistical analysis and data manipulation were carried out in *R* [72] using *Rstudio* [73] and the *tidyverse* [74], *lme4* [75], and *ggh4x* [76] packages, and the code provided by Nicenboim et al. [19] for preparing the data [33]. To reduce the risk of Type I errors, all p-values in our analyses are corrected for multiple comparisons based on false discovery rate [77].

IV. RESULTS

A. 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 [33]. 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.

B. 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) [78] 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 [79]). 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 [80]. 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.

C. 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 II. 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 II
SIGNIFICANT PREDICTORS OF N400 AMPLITUDE

Predictor	$\chi^2(\text{df} = 1)$	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**

D. 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 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 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 [80]. 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.

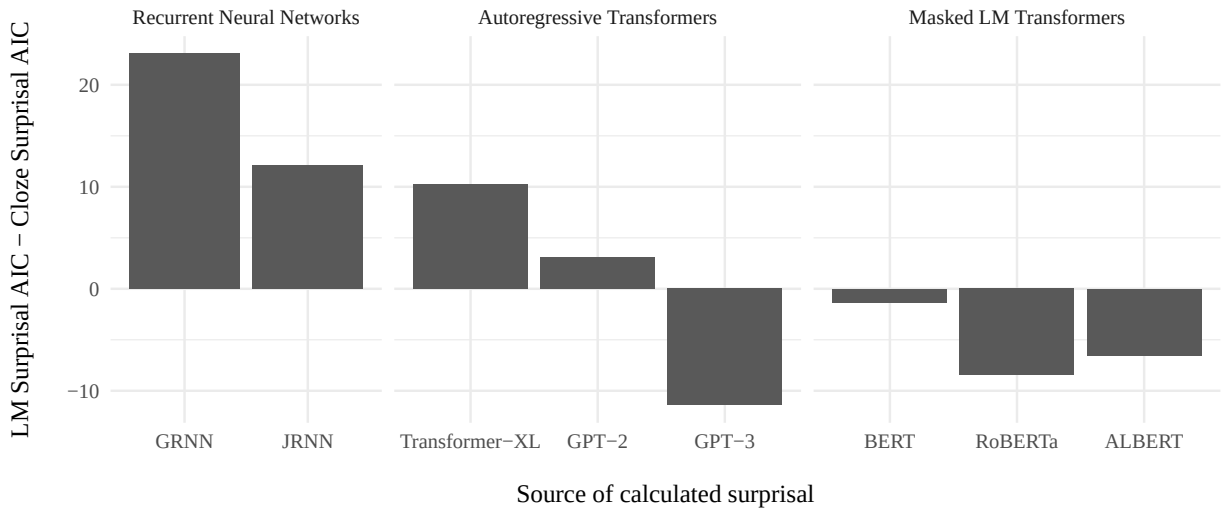

Fig. 1. AICs of all regressions including fixed effects of the denoted surprisal and laboratory, as well as random intercepts for each item and experimental participants. For easier comparison, AIC is scaled by subtracting the AIC of the regression including cloze surprisal, laboratory, and the aforementioned random intercepts. Lower AICs indicate better model fit [78].

E. 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 [24], 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 III.

As can be seen in Table III, 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.

F. 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 IV.

As can be seen in Table IV, 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.

V. 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

TABLE III
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	$\chi^2(\text{df}=1)$	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**

TABLE IV
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	$\chi^2(\text{df}=1)$	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

and N400 amplitude [29], [30], [28], [32], [31]. 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 [24], 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 [24] 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 [32], 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.

A. 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 [29], [81], [30], [31]. When compared by perplexity, a common evaluation metric for autoregressive transformer LMs, GPT-3 outperforms Transformer-XL and GPT-2 [39], [38], [41]. Similarly, ALBERT and RoBERTa each out-perform BERT at the GLUE benchmark [82], 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 [31], [32].

This finding may offer additional insight into why our results diverge from previous behavioral studies showing that cloze probability [24] and cloze surprisal [27] 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 [24]), with n-grams also being used [27], [24]. 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 [27], [24]. 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' [14], or the 'neural activation state at the time a physical stimulus is encountered' [13] 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. [59]:

> 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 [59], [58], [61]. Based on such experiments, it has been proposed that implausible continuations like *patient* are 'collaterally facilitated' by the preceding context [13], or, alternatively, that their preactivation is caused by a separate associative system [83].

However, recent work shows that the difference in N400 amplitude reported in Ito et al.'s [59] study can be successfully predicted based on GRNN and JRNN surprisal [28]. 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* [28] 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 [28], [32]. 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.

B. 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 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 [84], [85], [86], [87], 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 [88], [14], [89]. The present study suggests that surprisals derived from high-quality LMs should be used analogously in ERP investigations of language processing.

VI. Conclusion

Previous work has shown that LM predictions correlate with N400 amplitude when cloze does not [28], [32]. 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.

Acknowledgment

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

[1] M. Kutas, K. A. DeLong, and N. J. Smith, "A look around at what lies ahead: Prediction and predictability in language processing," in *Predictions in the Brain: Using Our Past to Generate a Future*, M. Bar, Ed. New York, NY, US: Oxford University Press, 2011, pp. 190–207.

[2] C. Van Petten and B. J. Luka, "Prediction during language comprehension: Benefits, costs, and ERP components," *Int. J. of Psychophysiol.*, vol. 83, no. 2, pp. 176–190, 2012.

[3] S. G. Luke and K. Christianson, "Limits on lexical prediction during reading," *Cogn. Psychol.*, vol. 88, pp. 22–60, 2016.

[4] G. R. Kuperberg and T. F. Jaeger, "What do we mean by prediction in language comprehension?" *Lang. Cogn. Neurosci.*, vol. 31, no. 1, pp. 32–59, 2016.

[5] K. I. Forster, "Priming and the effects of sentence and lexical contexts on naming time: Evidence for autonomous lexical processing," *Quart. J. Exp. Psychol. Sect. A*, vol. 33, no. 4, pp. 465–495, 1981.

[6] R. Jackendoff, *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, 2002.

[7] P. J. Schwanenflugel and E. J. Shoben, "The influence of sentence constraint on the scope of facilitation for upcoming words," *J. Mem. Lang.*, vol. 24, no. 2, pp. 232–252, 1985.

[8] M. J. Traxler and D. J. Foss, "Effects of sentence constraint on priming in natural language comprehension," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 26, no. 5, pp. 1266–1282, 2000.

[9] R. F. West and K. E. Stanovich, "Source of inhibition in experiments on the effect of sentence context on word recognition," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 8, no. 5, pp. 385–399, 1982.

[10] A. M. Collins and E. F. Loftus, "A spreading-activation theory of semantic processing," *Psychol. Rev.*, vol. 82, no. 6, pp. 407–428, 1975.

[11] M. Kutas and S. A. Hillyard, "Reading senseless sentences: Brain potentials reflect semantic incongruity," *Science*, vol. 207, no. 4427, pp. 203–205, 1980.

[12] M. Kutas and K. D. Federmeier, "Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP)," *Annu. Rev. Psychol.*, vol. 62, no. 1, pp. 621–647, 2011.

[13] K. A. DeLong and M. Kutas, "Comprehending surprising sentences: Sensitivity of post-N400 positivities to contextual congruity and semantic relatedness," *Lang. Cogn. Neurosci.*, vol. 35, no. 0, pp. 1044–1063, 2020.

[14] G. R. Kuperberg, T. Brothers, and E. W. Wlotko, "A Tale of Two Positivities and the N400: Distinct Neural Signatures Are Evoked by Confirmed and Violated Predictions at Different Levels of Representation," *J. Cogn. Neurosci.*, vol. 32, no. 1, pp. 12–35, 2020.

[15] K. A. DeLong, T. P. Urbach, and M. Kutas, "Probabilistic word preactivation during language comprehension inferred from electrical brain activity," *Nat. Neurosci.*, vol. 8, no. 8, pp. 1117–1121, 2005.

[16] J. J. A. Van Berkum, C. M. Brown, P. Zwitserlood, V. Kooijman, and P. Hagoort, "Anticipating Upcoming Words in Discourse: Evidence From ERPs and Reading Times." *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 31, no. 3, pp. 443–467, 2005.

[17] M. Otten, M. S. Nieuwland, and J. J. Van Berkum, "Great expectations: Specific lexical anticipation influences the processing of spoken language," *BMC Neurosci.*, vol. 8, no. 1, p. 89, 2007.

[18] N. Kwon, P. Sturt, and P. Liu, "Predicting semantic features in Chinese: Evidence from ERPs," *Cognition*, vol. 166, pp. 433–446, 2017.

[19] B. Nicenboim, S. Vasishth, and F. Rösler, "Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data," *Neuropsychologia*, vol. 142, p. 107427, 2020.

[20] T. P. Urbach, K. A. DeLong, W.-H. Chan, and M. Kutas, "An exploratory data analysis of word form prediction during word-by-word reading," *Proc. Nat. Acad. Sci.*, vol. 117, no. 34, pp. 20 483–20 494, 2020.

[21] D. S. Fleur, M. Flecken, J. Rommers, and M. S. Nieuwland, "Definitely saw it coming? The dual nature of the pre-nominal prediction effect," *Cognition*, vol. 204, p. 104335, 2020.

[22] W. L. Taylor, ""Cloze Procedure": A New Tool for Measuring Readability," *Journalism Quart.*, vol. 30, no. 4, pp. 415–433, 1953.

[23] M. Kutas and S. A. Hillyard, "Brain potentials during reading reflect word expectancy and semantic association," *Nature*, vol. 307, no. 5947, pp. 161–163, 1984.

[24] T. Brothers and G. R. Kuperberg, "Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension," *J. Mem. Lang.*, 2021.

[25] M. Kutas and C. Van Petten, "Psycholinguistics electrified: Event-related brain potential investigations," in *Handbook of Psycholinguistics*, 1st ed., M. A. Gernsbacher, Ed. San Diego: Academic Press, 1994, pp. 83–143.

[26] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. [Online Draft], 2021.

[27] N. J. Smith and R. Levy, "Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing," in *Proc. 33rd Annu. Meeting Cogn. Sci. Soc. (CogSci 2011)*, 2011, p. 7.

[28] J. A. Michaelov and B. K. Bergen, "How well does surprisal explain N400 amplitude under different experimental conditions?" in *Proc. 24th Conf. Comp. Natural Lang. Learn. (CoNLL 2020)*. Online: Association for Computational Linguistics, 2020, pp. 652–663.

[29] S. L. Frank, L. J. Otten, G. Galli, and G. Vigliocco, "The ERP response to the amount of information conveyed by words in sentences," *Brain and Lang.*, vol. 140, pp. 1–11, 2015.

[30] C. Aurnhammer and S. L. Frank, "Evaluating information-theoretic measures of word prediction in naturalistic sentence reading," *Neuropsychologia*, vol. 134, p. 107198, 2019.

[31] D. Merkx and S. L. Frank, "Human Sentence Processing: Recurrence or Attention?" in *Proc. Workshop Cogn. Model. and Comp. Ling. (CMCL 2021)*. Online: Association for Computational Linguistics, 2021, pp. 12–22.

[32] J. A. Michaelov, M. D. Bardolph, S. Coulson, and B. K. Bergen, "Different kinds of cognitive plausibility: Why are transformers better than RNNs at predicting N400 amplitude?" in *Proc. 43rd Annu. Meeting Cogn. Sci. Soc. (CogSci 2021)*, University of Vienna, Vienna, Austria (Hybrid), 2021, pp. 300–306.

[33] M. S. Nieuwland, S. Politzer-Ahles, E. Heyselaar, K. Segaert, E. Darley, N. Kazanina, S. Von Grebmer Zu Wolfsthurn, F. Bartolozzi, V. Kogan, A. Ito, D. Mézière, D. J. Barr, G. A. Rousselet, H. J. Ferguson, S. Busch-Moreno, X. Fu, J. Tuomainen, E. Kulakova, E. M. Husband, D. I. Donaldson, Z. Kohút, S.-A. Rueschemeyer, and F. Huettig, "Large-scale replication study reveals a limit on probabilistic prediction in language comprehension," *eLife*, vol. 7, p. e33468, 2018.

[34] K. Gulordava, P. Bojanowski, E. Grave, T. Linzen, and M. Baroni, "Colorless Green Recurrent Networks Dream Hierarchically," in *Proc. 2018 Conf. North Amer. Chapter Assoc. Comp. Ling.: Human Lang. Technol. (NAACL-HLT 2018), Vol. 1*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 1195–1205.

[35] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the Limits of Language Modeling," *ArXiv160202410 Cs*, 2016.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comp. Ling.: Human Lang. Technol. (NAACL 2019), Vol. 1*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186.

[37] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *ArXiv190711692 Cs*, 2019.

[38] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," p. 24, 2019.

[39] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-XL: Attentive Language Models beyond a Fixed-Length Context," in *Proc. 57th Annu. Meeting Assoc. Comput. Ling. (ACL 2019)*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 2978–2988.

[40] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," in *Int. Conf. on Learn. Representations (ICLR 2020)*, 2020.

[41] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," in *Advances in Neural Inf. Process. Syst. (NeurIPS 2020)*, vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[42] K. A. DeLong, M. Troyer, and M. Kutas, "Pre-Processing in Sentence Comprehension: Sensitivity to Likely Upcoming Meaning and Structure," *Lang. Linguist. Compass*, vol. 8, no. 12, pp. 631–645, 2014.

[43] A. S. Meyer, F. Huettig, and W. J. Levelt, "Same, different, or closely related: What is the relationship between language production and comprehension?" *J. Mem. Lang.*, vol. 89, pp. 1–7, 2016.

[44] P. Hendriks, *Asymmetries between Language Production and Comprehension*, ser. Studies in Theoretical Psycholinguistics. Dordrecht: Springer Netherlands, 2014, vol. 42.

[45] A. Rogers, O. Kovaleva, and A. Rumshisky, "A Primer in BERTology: What We Know About How BERT Works," *Trans. Assoc. Comput. Ling. (TACL)*, vol. 8, pp. 842–866, 2021.

[46] G. S. Marmor, "Age at onset of blindness and the development of the semantics of color names," *J. Exp. Child Psychol.*, vol. 25, no. 2, pp. 267–278, 1978.

[47] J. Hale, "A probabilistic earley parser as a psycholinguistic model," in *2nd Meeting North Amer. Chapter Assoc. Comp. Ling. Lang. Technol. (NAACL '01)*. Pittsburgh, Pennsylvania: Association for Computational Linguistics, 2001, pp. 1–8.

[48] R. Levy, "Expectation-based syntactic comprehension," *Cognition*, vol. 106, no. 3, pp. 1126–1177, 2008.

[49] N. J. Smith and R. Levy, "The effect of word predictability on reading time is logarithmic," *Cognition*, vol. 128, no. 3, pp. 302–319, 2013.

[50] M. F. Boston, J. Hale, R. Kliegl, U. Patil, and S. Vasishth, "Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus," *J. Eye Mov. Res.*, vol. 2, no. 1, 2008.

[51] V. Demberg and F. Keller, "Data from eye-tracking corpora as evidence for theories of syntactic processing complexity," *Cognition*, vol. 109, no. 2, pp. 193–210, 2008.

[52] B. Roark, A. Bachrach, C. Cardenas, and C. Pallier, "Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing," in *Proc. 2009 Conf. on Empirical Methods in Natural Lang. Process. (EMNLP 2009)*, vol. 1. Singapore: Association for Computational Linguistics, 2009, p. 324.

[53] J. Mitchell, M. Lapata, V. Demberg, and F. Keller, "Syntactic and semantic factors in processing difficulty: An integrated measure," in

Proc. 48th Annu, Meeting Assoc. Comput. Ling. (ACL 2010), 2010, pp. 196–206.

[54] I. F. Monsalve, S. L. Frank, and G. Vigliocco, "Lexical surprisal as a general predictor of reading time," in *Proc. 13th Conf. Eur. Chapter Assoc. Comput. Ling. (EACL 2012)*. Association for Computational Linguistics, 2012, pp. 398–408.

[55] R. M. Willems, S. L. Frank, A. D. Nijhof, P. Hagoort, and A. van den Bosch, "Prediction During Natural Language Comprehension," *Cereb. Cortex*, vol. 26, no. 6, pp. 2506–2516, 2016.

[56] M. Kutas, "In the company of other words: Electrophysiological evidence for single-word and sentence context effects," *Lang. Cogn. Process.*, vol. 8, no. 4, pp. 533–572, 1993.

[57] K. D. Federmeier and M. Kutas, "A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing," *J. Mem. Lang.*, 1999.

[58] D. E. Thornhill and C. Van Petten, "Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components," *Int. J. Psychophysiol.*, vol. 83, no. 3, pp. 382–392, 2012.

[59] A. Ito, M. Corley, M. J. Pickering, A. E. Martin, and M. S. Nieuwland, "Predicting form and meaning: Evidence from brain potentials," *J. Mem. Lang.*, vol. 86, pp. 157–171, 2016.

[60] M. W. Lowder, W. Choi, F. Ferreira, and J. M. Henderson, "Lexical Predictability During Natural Reading: Effects of Surprisal and Entropy Reduction," *Cogn. Sci.*, vol. 42, pp. 1166–1183, 2018.

[61] R. Metusalem, M. Kutas, T. P. Urbach, M. Hare, K. McRae, and J. L. Elman, "Generalized event knowledge activation during online sentence comprehension," *J. Mem. Lang.*, vol. 66, no. 4, pp. 545–567, 2012.

[62] N. Delaney-Busch, E. Morgan, E. Lau, and G. R. Kuperberg, "Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming," *Cognition*, vol. 187, pp. 10–20, 2019.

[63] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," *Adv. Neural Inf. Process. Syst. (NeurIPS 2017)*, vol. 30, pp. 5998–6008, 2017.

[64] F. Keller, "Cognitively Plausible Models of Human Language Processing," in *Proc. Assoc. Comput. Ling. 2010 (ACL 2010)*. Uppsala, Sweden: Association for Computational Linguistics, 2010, pp. 60–67.

[65] J. L. Elman, "Finding Structure in Time," *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, 1990.

[66] R. Futrell, E. Wilcox, T. Morita, P. Qian, M. Ballesteros, and R. Levy, "Neural language models as psycholinguistic subjects: Representations of syntactic state," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comp. Ling.: Human Lang. Technol. (NAACL-HLT 2019), Vol. 1*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 32–42.

[67] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-Art Natural Language Processing," in *Proc. 2009 Conf. on Empirical Methods in Natural Lang. Process.: System Demonstrations*. Online: Association for Computational Linguistics, 2020, pp. 38–45.

[68] OpenAI, "OpenAI API," https://beta.openai.com, 2021.

[69] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Inf. Process. Syst. (NeurIPS 2019)*, vol. 32. Curran Associates, Inc., 2019.

[70] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in *Proc. 54th Annu. Meeting Assoc. Comput. Ling. (ACL 2016), Vol 1*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1715–1725.

[71] Google Research, "BERT," https://github.com/google-research/bert.

[72] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020.

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

[74] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani, "Welcome to the tidyverse," *J. Open Source Softw.*, vol. 4, no. 43, p. 1686, 2019.

[75] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015.

[76] T. van den Brand, *Ggh4x: Hacks for 'Ggplot2'*, 2021.

[77] Y. Benjamini and D. Yekutieli, "The Control of the False Discovery Rate in Multiple Testing under Dependency," *Ann. Stat.*, vol. 29, no. 4, pp. 1165–1188, 2001.

[78] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, ser. Springer Series in Statistics, B. N. Petrov and F. Csáki, Eds. Budapest, Hungary: Akadémiai Kiadó, 1973, pp. 267–281.

[79] E.-J. Wagenmakers and S. Farrell, "AIC model selection using Akaike weights," *Psychonomic Bull. & Rev.*, vol. 11, no. 1, pp. 192–196, 2004.

[80] K. P. Burnham and D. R. Anderson, "Multimodel Inference: Understanding AIC and BIC in Model Selection," *Sociol. Methods & Res.*, vol. 33, no. 2, pp. 261–304, 2004.

[81] A. Goodkind and K. Bicknell, "Predictive power of word surprisal for reading times is a linear function of language model quality," in *Proc. 8th Workshop Cogn. Model. Comput. Ling. (CMCL 2018)*. Salt Lake City, Utah: Association for Computational Linguistics, 2018, pp. 10–18.

[82] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," in *Int. Conf. Learn. Representations (ICLR 2019)*, 2019.

[83] S. L. Frank and R. M. Willems, "Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension," *Lang. Cogn. Neurosci.*, vol. 32, no. 9, pp. 1192–1203, 2017.

[84] D. J. Chwilla and H. H. J. Kolk, "Accessing world knowledge: Evidence from N400 and reaction time priming," *Cogn. Brain Res.*, vol. 25, no. 3, pp. 589–606, 2005.

[85] M. Parviz, M. Johnson, B. Johnson, and J. Brock, "Using Language Models and Latent Semantic Analysis to Characterise the N400m Neural Response," in *Proc. Australas. Lang. Technol. Assoc. Workshop 2011*, Canberra, Australia, 2011, pp. 38–46.

[86] C. Van Petten, "Examining the N400 semantic context effect item-by-item: Relationship to corpus-based measures of word co-occurrence," *Int. J. of Psychophysiol.*, vol. 94, no. 3, pp. 407–419, 2014.

[87] A. Ettinger, N. Feldman, P. Resnik, and C. Phillips, "Modeling N400 amplitude using vector space models of word representation." in *Proc. 38th Annu. Conf. Cogn. Sci. Soc. (CogSci 2016)*, Philadelphia, USA, 2016.

[88] D. J. Chwilla, H. H. J. Kolk, and C. T. W. M. Vissers, "Immediate integration of novel meanings: N400 support for an embodied view of language comprehension," *Brain Res.*, vol. 1183, pp. 109–123, 2007.

[89] M. S. Nieuwland, D. J. Barr, F. Bartolozzi, S. Busch-Moreno, E. Darley, D. I. Donaldson, H. J. Ferguson, X. Fu, E. Heyselaar, F. Huettig, E. Matthew Husband, A. Ito, N. Kazanina, V. Kogan, Z. Kohút, E. Kulakova, D. Mézière, S. Politzer-Ahles, G. Rousselet, S.-A. Rueschemeyer, K. Segaert, J. Tuomainen, and S. Von Grebmer Zu Wolfsthurn, "Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials," *Philos. Trans. Roy. Soc. B: Biol. Sci.*, vol. 375, no. 1791, p. 20180522, 2020.