Different kinds of cognitive plausibility: why are transformers better than RNNs at predicting N400 amplitude?

James A. Michaelov (j1michae@ucsd.edu)
Megan D. Bardolph (mbardolph@ucsd.edu)
Seana Coulson (scoulson@ucsd.edu)
Benjamin K. Bergen (bkbergen@ucsd.edu)
Department of Cognitive Science, University of California, San Diego
9500 Gilman Dr, La Jolla, CA 92093, USA

Abstract

Despite being designed for performance rather than cognitive plausibility, transformer language models have been found to be better at predicting metrics used to assess human language comprehension than language models with other architectures, such as recurrent neural networks. Based on how well they predict the N400, a neural signal associated with processing difficulty, we propose and provide evidence for one possible explanation—their predictions are affected by the preceding context in a way analogous to the effect of semantic facilitation in humans.

Keywords: Language Comprehension; Electroencephalography (EEG); Neural Networks; Machine Learning; Cognitive Architectures

Introduction

Neural language models (NLMs) are valuable tools in understanding human language comprehension because they learn to predict language based on the surface-level statistics of language alone. As such, they are inherently models both of what can be predicted and what can be learned about language based only on linguistic input. For this reason, they can be used to test hypotheses about how such knowledge may be used in the human language comprehension system that would be impossible to test experimentally.

Since the early days of their implementation, recurrent neural network language models (RNN-LMs) have been used to investigate human cognition (Elman, 1990). However, the recent development of the transformer network (Vaswani et al., 2017) has largely overshadowed them in the field of machine learning. Research has shown that language model performance (operationalized as perplexity) tends to correlate with how human-like RNN-LMs are in their processing of language (e.g., Aurnhammer & Frank, 2019). Thus, the fact that transformer language models (T-LMs) perform better than the previously state-of-the-art RNN-LMs despite their vastly different architectures suggests that they may be a viable alternative model of human language processing.

It is therefore unsurprising that in the last year, researchers interested in the cognition of language have compared RNN-LMs and T-LMs in terms of how well they capture human linguistic behavior (Merkx & Frank, 2020; Ettinger, 2020; Misra, Ettinger, & Rayz, 2020; Wilcox, Gauthier, Hu, Qian, & Levy, 2020; Eisape, Zaslavsky, & Levy, 2020). To the best of our knowledge, only one (currently unpublished) study has attempted to investigate whether RNN-LMs or T-LMs are better for predicting the N400, a neural response reflecting semantic retrieval demands. In their study, Merkx and Frank (2020) find that the surprisal of T-LMs fit the human N400 data better than the surprisal of RNN-LMs. This is somewhat unexpected because, as Merkx and Frank (2020) note, intuitively, RNN-LMs more closely match what we believe about the human language comprehension system—they process language word-by-word and have a limited 'working memory'.

Here we ask whether the success T-LMs evince in predicting N400 amplitude (Merkx & Frank, 2020) is connected to other findings; for example, they appear to show semantic priming effects and, like the N400 (Nieuwland & Kuperberg, 2008), they are rather insensitive to negation (Ettinger, 2020; Misra et al., 2020). To do so, we investigate whether the predictions of T-LMs show an analog of semantic priming phenomena that have been argued to impact N400 amplitude to a greater or lesser extent (e.g., Brouwer, Fitz, & Hoeks, 2012; Lau, Holcomb, & Kuperberg, 2013). If this is the case, it would suggest that T-LMs may be analogous to the human language comprehension system in a different way to RNN-LMs, and may offer an insight into the relationship between these phenomena. Such a result would also indicate that we need to update our conception of what makes a computational language model more or less cognitively plausible.

Background

The N400 (Kutas & Hillyard, 1980) is a negative deflection in the event-related brain potential (ERP), peaking roughly 400ms after the presentation of a stimulus. It is thought to index processing difficulty—if the preceding context activates semantic content associated with an upcoming word, the word is easier to process, and thus elicits a reduced amplitude N400. Recent accounts have hypothesized that the N400 specifically indexes the extent to which the upcoming word was not expected; a prediction error not affected by the strength of failed predictions (Van Petten & Luka, 2012; Luke & Christianson, 2016; DeLong & Kutas, 2020; Kuperberg, Brothers, & Wlotko, 2020).

The surprisal of an NLM towards a word is a clear conceptual analog of the N400—surprisal is the negative logarithm of the probability of an upcoming word given its context. NLM surprisal significantly predicts N400 amplitude, beating other metrics derived from NLMs (Frank, Ot-

ten, Galli, & Vigliocco, 2015; Aurnhammer & Frank, 2019; Merkx & Frank, 2020). Additionally, on more fine-grained analysis, surprisal appears to behave analogously to N400 amplitude—in many cases, experimental manipulations that impact N400 amplitude affect surprisal values in an analogous fashion (Michaelov & Bergen, 2020).

However, while surprisal is a good model of the extent to which a word is predicted in the context of a sentence, N400 amplitude is also modulated by other factors (Kutas & Federmeier, 2011; Kuperberg et al., 2020). One key finding is that the N400 to a target word is less negative in amplitude when it follows a semantically related word than an unrelated one (Kutas & Van Petten, 1988; Kutas & Federmeier, 2011). Additionally, the N400 response to a word is reduced if it is semantically related to the most predictable upcoming word—for example, the word *monopoly* in *"Checkmate," Rosaline announced with glee. She was getting to be really good at monopoly* elicits a less negative N400 than the word *football* by virtue of being more semantically related to the best completion, *chess* (Federmeier & Kutas, 1999). The N400 response to a word is also reduced if it is semantically related to the previous words in the utterance—for example, there is no difference in N400 response between the word *eat* in *for breakfast the boys would only eat...* and *for breakfast the eggs would only eat...*, despite its semantic implausibility in the second clause (Kuperberg, Sitnikova, Caplan, & Holcomb, 2003).

These effects have led some researchers to argue that the N400 can be explained by spreading activation, where words in the preceding context partially activate semantically-related upcoming words, regardless of whether they are appropriate completions to the sentence (Brouwer et al., 2012). This 'bag-of-words' approach to semantic pre-activation (Kuperberg, 2016) has also been reflected in computational modeling of the N400. Several researchers have used the cosine distance between the word embeddings of the target word and those for the preceding words to reflect the kind of semantic similarity that may lead to facilitated processing (Parviz, Johnson, Johnson, & Brock, 2011; Ettinger, Feldman, Resnik, & Phillips, 2016; Frank & Willems, 2017). This approach is also used to control for confounding effects of semantic similarity when explicitly investigating prediction (Kuperberg et al., 2020).

The present study has two main aims. First, to directly compare and quantify how well semantic facilitation (as operationalized by cosine distance between embeddings) and prediction (as operationalized by surprisal) predict N400 amplitude. Second, to identify the extent to which the two are correlated and how this varies by language model architecture, with the hope that this will inform why T-LMs have been found to predict N400 amplitude better than RNN-LMs despite their apparent cognitive implausibility.

Experiment 1: RNN-LM vs. T-LM surprisal

Modeling approach and details

This experiment follows the same general approach as previous research investigating surprisal as a predictor of N400 amplitude (Frank et al., 2015; Aurnhammer & Frank, 2019; Merkx & Frank, 2020), namely, comparing recorded ERP data to NLM surprisal for the same set of stimuli. We use stimuli from an ERP study whose data have been previously presented (Bardolph, Van Petten, & Coulson, 2018). These stimuli were run through two NLMs, one RNN-LM and one T-LM, and the predicted probability of target words was collected. This predicted probability was then transformed into surprisal, where 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).

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

This surprisal was then used as a predictor in a linear mixed-effects model to predict by-trial, by-electrode amplitude from the original ERP study.

Two NLMs were used. The RNN-LM was the BIG LSTM+CNN INPUTS model (Jozefowicz, Vinyals, Schuster, Shazeer, & Wu, 2016), henceforth the JRNN. The T-LM used was GPT-2 (Radford et al., 2019). Both models are very large—the JRNN has roughly 1 billion parameters, while GPT-2 has roughly 1.5 billion. One area in which the two differ is that the JRNN has a vocabulary size of roughly 800,000, while GPT-2 has a vocabulary size of roughly 50,000. Additionally, while the JRNN was trained on approximately one billion words, GPT-2 was trained on a dataset an order of magnitude larger.

Original ERP Study

The original study (Bardolph et al., 2018) used stimuli adapted from previous work (Thornhill & Van Petten, 2012). There were 290 sentence frames with target words in one of four conditions, for a total of 1160 sentences.

The conditions were the following. The BEST COMPLETION was the completion with the highest cloze probability (Taylor, 1957; cloze = 0.458 ± 0.261). The cloze probability of a word is the proportion of participants in a norming study that filled in a gap in the sentence with that word. RELATED completions were low-cloze (0.043 ± 0.058) words that are semantically related to the best completion. UNRELATED completions are low-cloze (0.024 ± 0.037) words that are semantically unrelated to the best completion. IMPLAUSIBLE completions are completions that were semantically implausible with a cloze of zero. The conditions can be illustrated with the following example: *It's hard to admit when one is **wrong*** (BEST COMPLETION) / ***incorrect*** (RELATED to best completion) / ***lonely*** (UNRELATED to best completion) / ***screened*** (IMPLAUSIBLE).

As expected from previous research, Bardolph et al. (2018) found that the BEST COMPLETION elicited the lowest-
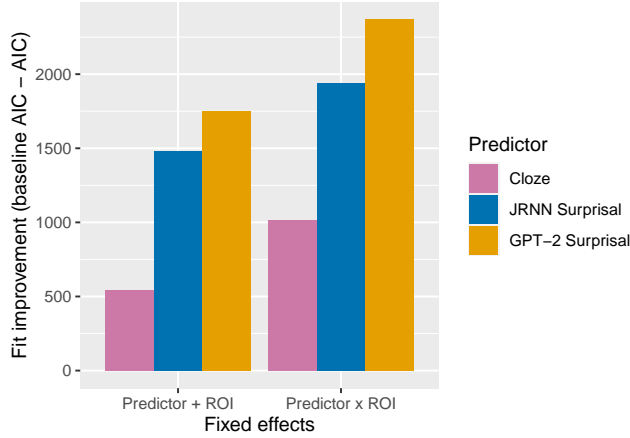
Figure 1: The improvement in AIC for each model compared to the baseline model with only ROI as a main effect. The 'Predictor + ROI' models have main effects of the predictor and ROI, while the 'Predictor x ROI' model also includes their interaction.

amplitude N400, followed by RELATED, UNRELATED, and IMPLAUSIBLE completions, in order of increasing amplitude.

In the study, 44 healthy adult experimental participants read sentences in English one word at a time. EEG was recorded from 29 scalp sites. In the present study, the mean amplitude at each site over the 300-500ms time period (the canonical N400 time-frame) was calculated for each electrode in each trial. These mean amplitude measurements thus served as the outcome measures in the regression models described below.

Results

Linear mixed-effects models were used to predict N400 amplitude. All models included region of interest (ROI; Prefrontal, Fronto-central, Central, Posterior, Left Temporal, Right Temporal) as a fixed effect and Subject, Sentence Frame, and Electrode as random intercepts (more complex random effects structures led to models that did not converge or had singular fits).

We evaluated the statistical significance of each predictor using likelihood ratio tests between nested models. All reported p-values are corrected for multiple comparisons based on the false discovery rate (Benjamini & Yekutieli, 2001; R Core Team, 2020). Adding the fixed effect of surprisal to a null model (a linear mixed-effects model with only ROI as a fixed effect and the aforementioned random intercepts) significantly improved model fit (JRNN: $\chi^2(1) = 1483, p < 0.0001$; GPT-2: $\chi^2(1) = 1752.5, p < 0.0001$). Further adding the interaction of surprisal and ROI also significantly improved the model fit (JRNN: $\chi^2(5) = 467.82, p < 0.0001$; GPT-2: $\chi^2(5) = 628.84, p < 0.0001$).

A comparison of AICs (Figure 1) shows that for equivalent models, GPT-2 surprisal fits N400 amplitude more closely. For comparison, we also include the AIC values for equiva-

lent linear mixed-effects models with cloze probability as the main predictor. While it may be unsurprising that cloze is a worse predictor of N400 amplitude than surprisal in this study due to the fact that the RELATED and UNRELATED conditions were matched for cloze, as far as we are aware, this is the first time that a corpus-derived metric has out-performed cloze as a predictor of human processing difficulty (see Brothers & Kuperberg, 2021, for discussion). To investigate the difference in AICs further, we compared the predictions of the best models (i.e. those including the Predictor x ROI interaction) to the real N400 in a held-out dataset of roughly 15% of the total data (46,280 measurements).

Figure 2 shows both the true and predicted amplitudes in the 300-500ms time window for electrodes in the Central and Posterior ROIs (the canonical N400 ROIs). As can be seen, the N400 amplitudes predicted by GPT-2 surprisal are closer to the true amplitudes than those predicted by JRNN surprisal in 3 out of 4 of the conditions: BEST COMPLETION (One-tailed t-test testing whether N400 amplitude predicted by the GPT-2 surprisal model < N400 amplitude predicted by the JRNN surprisal model: $t(10198) = 5.9324, p < 0.0001$), RELATED (GPT-2 < JRNN: $t(9105.3) = 3.3548, p = 0.0019$), and IMPLAUSIBLE (GPT-2 > JRNN: $t(9363.7) = -7.4523, p < 0.0001$). While there is a difference in means in the expected direction for UNRELATED completions, the difference is not significant (GPT-2 > JRNN: $t(9695.5) = -1.2441, p = 0.4584$).

We further tested whether the models successfully predicted the differences between conditions by running one-tailed t-tests between the predicted amplitudes for conditions closest in mean predicted amplitude (using the values for the Central and Posterior ROIs, as shown in Figure 2). GPT-2 surprisal successfully predicts that RELATED words elicit a higher-amplitude N400 than BEST COMPLETIONS ($t(9601.2) = 18.466, p < 0.0001$), that UNRELATED words elicit a higher-amplitude N400 than RELATED words ($t(9393.7) = 5.8936, p < 0.0001$), and IMPLAUSIBLE words elicit a higher-amplitude N400 than UNRELATED words ($t(9536.7) = 47.936, p < 0.0001$). On the other hand, while JRNN surprisal successfully predicts that RELATED words will elicit a higher-amplitude N400 than BEST COMPLETIONS ($t(9593.2) = 15.871, p < 0.0001$), and that IMPLAUSIBLE words elicit a higher-amplitude N400 than UNRELATED words ($t(9507.8) = 40.43, p < 0.0001$), it does not predict that UNRELATED completions elicit a higher-amplitude N400 than RELATED completions ($t(9392.6) = 1.2691, p = 0.4584$), which was observed in the ERP data and was successfully predicted by GPT-2 surprisal. In fact, while its predictions are closer, in terms of predicting significant differences between conditions, JRNN surprisal does no better than cloze (Predicted BEST COMPLETION N400 amplitude < predicted RELATED N400 amplitude: $t(9651.3) = 38.472, p < 0.0001$; RELATED < UNRELATED: $t(9384.9) = 0.5747, p = 1$; UNRELATED < IMPLAUSIBLE: $t(9528.4) = 3.7758, p = 0.0004$).
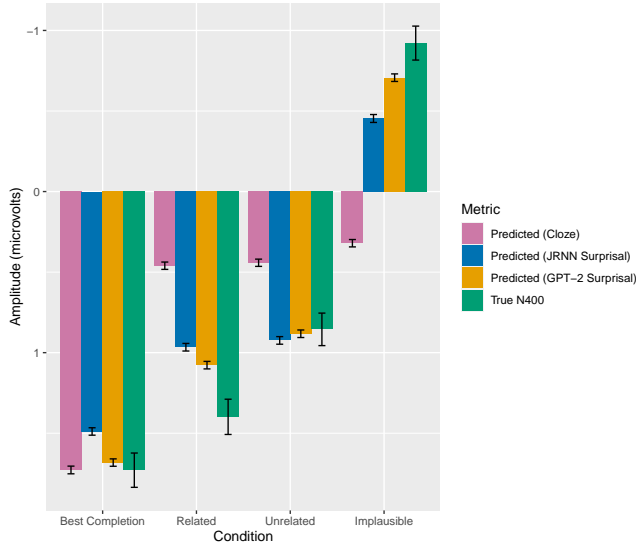
Figure 2: True and predicted amplitudes in the 300-500ms timeframe for electrodes in the Central and Posterior ROIs. Error bars indicate standard error. A more negative amplitude represents a higher-amplitude N400. It should be noted that the y-axis is reversed, following convention.

Discussion

There are several key results from this study. First, surprisal and its interaction with scalp ROI are significant predictors of N400 amplitude, replicating previous results (Frank et al., 2015; Aurnhammer & Frank, 2019; Merkx & Frank, 2020). Second, we replicate the finding that the surprisal of T-LMs better predicts N400 amplitude than that of RNN-LMs (Merkx & Frank, 2020). Finally, we find that while GPT-2 surprisal successfully predicts differences in N400 due to experimental manipulation—with BEST COMPLETION eliciting the most reduced N400, and RELATED, UNRELATED, and IMPLAUSIBLE completions each eliciting increasingly more negative (less positive) N400s—JRNN surprisal fails to distinguish between RELATED and UNRELATED completions. This difficulty for the JRNN to predict the difference between the two is consistent with the findings of Michaelov and Bergen (2020).

Why does GPT-2 surprisal fit N400 amplitude better in both models? It is likely that there are multiple reasons for this, some of which may involve the simple fact of larger size and larger training data. It is also important to note that it is possible that the better predictions of GPT-2 surprisal are due to different factors across the experimental conditions.

However, one architecture-related possibility is suggested by our replication of the finding that JRNN surprisal can struggle to predict the difference between the RELATED and UNELATED conditions for some sets of stimuli (Michaelov & Bergen, 2020) and our novel finding that this is not the case with GPT-2 surprisal. As can be seen in Figure 2, the inability of JRNN surprisal to predict the difference between the two

seems to be mostly driven by overestimating the N400 amplitude for the RELATED condition. This is something which is improved upon by using GPT-2 surprisal for prediction. It is important to note that as explained previously, the effect of semantic relatedness on N400 amplitude has been previously hypothesized to involve spreading activation or some other form of shallow semantic facilitation.

This is crucial because as Merkx and Frank (2020) note, one of the key differences between T-LMs and RNN-LMs is that T-LMs do not have the same memory bottleneck as RNN-LMs—they have direct access to all previous words in the sequence. RNN-LMs, on the other hand, only have one current state, which is adjusted with each new input. Thus, this increased memory capacity—the capacity to 'remember' exactly which words precede the current word—means that it is possible for the network to use specific previous words in predicting the next word, and that these could independently (or in a bag-of-words fashion) semantically facilitate predictions. This may also surface as susceptibility to priming—previous work has found that T-LMs are likely to repeat words that that they have already seen, and their predictions can be semantically primed by presenting them with an individual prime word (Misra et al., 2020).

Experiment 2: Quantifying semantic facilitation

In Experiment 2, we investigate whether the surprisal of GPT-2 incorporates something roughly analogous to spreading activation. That is, the finding that words are at least partly predicted because they have been activated by the semantics of previous words in the sequence. We do so by comparing the extent to which surprisal in each of the two models correlates with estimates of semantic similarity from each of the two models. A higher correlation would indicate the model is more biased towards predicting words that are semantically related to preceding words in the utterance, i.e., that it exhibits behavior akin to that commonly attributed to semantic spreading activation in humans.

Method

The high-level process for calculating semantic similarity in the NLMs was similar to that used for calculating surprisal. Stimuli from the same experiment (Bardolph et al., 2018) were run through the NLMs, and the activation states of the model were recorded. For this study, however, we used the context-free word embeddings for each of the two NLMs (JRNN and GPT-2). As in previous work, we calculated the mean embeddings of all words preceding the target word, and calculated the cosine distance between this and the target word embedding. We then compared the extent to which cosine distance and surprisal were correlated for each network.

Results and Discussion

As can be seen in Figure 3, cosine similarity and surprisal are substantially more correlated for GPT-2 ($r = -0.480$) than for the JRNN ($r = -0.200$), supporting the idea that semantic
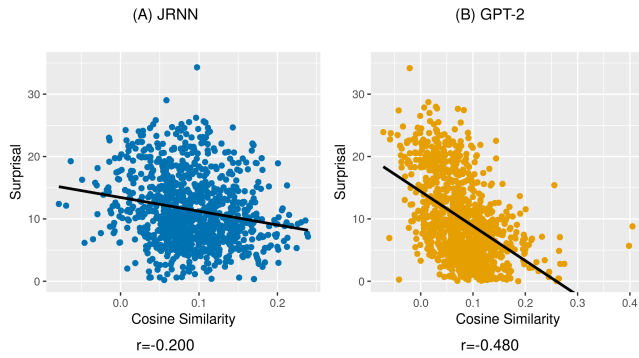
Figure 3: Cosine similarity and surprisal in the JRNN and GPT-2. *r* is Pearson's correlation coefficient.

similarity is more directly correlated with surprisal in GPT-2 than the JRNN.

To the best of our knowledge the approach here is novel. Moreover, it is valuable because it is self-contained. We utilize each model's own semantic representations to evaluate the degree of semantic spreading activation. Since our own semantic representations are the only ones we have access to during language comprehension, the approach can be seen as cognitively plausible. Further, it can easily be applied to any language model and any data set, without necessarily requiring carefully constructed stimuli.

Experiment 3: Testing the implicit semantic facilitation account

Experiment 1 showed that GPT-2 surprisal is a better predictor of N400 amplitude than JRNN surprisal is. Experiment 2 showed that GPT-2 surprisal is more highly correlated with semantic similarity than JRNN surprisal is. In Experiment 3, we directly test whether the latter finding explains the former, that is, whether the fact that GPT-2 surprisal is more correlated with semantic similarity (as operationalized by cosine similarity) leads to its better prediction of N400 amplitude.

Method

To do this, we investigated the extent to which each NLM's surprisal and cosine similarity metrics explain different proportions of the variance in N400 amplitude. If it is indeed the case that GPT-2 surprisal predicts N400 amplitude better than JRNN surprisal because its surprisal is more correlated with cosine similarity, then we should expect that adding cosine similarity as a predictor to a linear mixed-effects model with GPT-2 surprisal as a predictor should lead to less improvement than adding cosine similarity to the equivalent JRNN surprisal model.

Results

We tested this hypothesis by running likelihood ratio tests comparing the previous best models (including surprisal, ROI, and their interaction as fixed effects) with models that

also included cosine similarity as a predictor and larger models that included both cosine similarity and its interaction with ROI as predictors. We found that each of these significantly improved the JRNN model fit (cosine similarity: $\chi^2(1) = 92.782, p < 0.0001$; cosine similarity x ROI: $\chi^2(6) = 93.594, p < 0.0001$), but neither improved GPT-2 model fit (cosine similarity: $\chi^2(1) = 0.1524, p = 1$; cosine similarity x ROI: $\chi^2(6) = 5.7346, p = 1$).

Discussion

The results show that cosine similarity explains additional variance in N400 amplitude beyond what is explained by JRNN surprisal. However, this is not the case with GPT-2 surprisal. The results, therefore, provide evidence that it is not only the case that GPT-2 surprisal is a better predictor of N400 amplitude and more correlated with semantic similarity than JRNN surprisal, but that the two are related. Specifically, they provide evidence for the hypothesis that it is the fact that GPT-2 surprisal correlates better with semantic similarity that makes it better at predicting N400 amplitude.

General Discussion

Our results replicate and build upon previous work. As in previous work, we find that the surprisal of both RNN-LMs and T-LMs significantly predicts N400 amplitude, that T-LM surprisal is a better predictor than RNN-LM surprisal, and that JRNN surprisal struggles to predict the difference between words that are semantically related and unrelated to the highest-cloze completion (Frank et al., 2015; Aurnhammer & Frank, 2019; Michaelov & Bergen, 2020; Merkx & Frank, 2020).

Our first novel finding is that GPT-2 surprisal can better predict the amplitude of N400s elicited by words that are semantically related to the highest-cloze completions than can JRNN surprisal. This allows GPT-2 surprisal to successfully distinguish between low-cloze words that are semantically related and unrelated to the best completion where JRNN surprisal struggles.

Our second and more important finding is that GPT-2 surprisal is more (inversely) correlated with GPT-2 semantic similarity than JRNN surprisal is with JRNN semantic similarity. We hypothesize that this is due to the difference in architecture—with access to the previous words in a given input, T-LMs are able to predict the next word based on any of these words, while RNN-LMs are limited to predict based on their single recurrent state, which may be multi-layered, but nonetheless does not store individual previous words explicitly. This is in concord with recent findings showing that BERT, another T-LM, is susceptible to priming (Ettinger, 2020; Misra et al., 2020). This suggests that 'bag-of-words' semantic spreading activation, while not the whole story behind the neurocognitive system or systems underlying the N400 response, may still play a part in it, and thus, in the whole language comprehension system.

As a computational modeling study, the results of our experiments do not directly demonstrate a specific way in which

language comprehension is implemented in the brain. However, they do demonstrate that it is not necessary to posit separate systems to explain the fact that N400 amplitude is affected both by how predictable an upcoming word is and the prior occurrence of semantically or associatively related words in the context. While this is been previously shown with other (more elaborate) modeling approaches (e.g. Rabovsky, Hansen, & McClelland, 2018), we show directly that lexical prediction could in principle implicitly incorporate semantic relatedness or similarity.

It should be noted, however, that we do not provide evidence that the the same system *must* underlie both kinds of N400 response. This is still an open research question. The one fMRI-based study on the topic, for example, suggests that the two may occur in distinct areas (Frank & Willems, 2017). By contrast, the ERP-based work on the topic suggests that while there are differences between the effects of semantic facilitation and prediction that might be used to dissociate the two, the N400 response to each exhibit a very similar time course and topography (Kutas, 1993; Van Petten, 1993; Lau et al., 2013; Broderick, Anderson, Di Liberto, Crosse, & Lalor, 2018).

Another interesting question raised by our results is how to determine the cognitive plausibility of an NLM. As noted by Merkx and Frank (2020), intuitively, the RNN-LM architecture appears more cognitively plausible as a model of language comprehension than the T-LM. This is due to what Keller (2010) refers to as the distance-based memory cost of plausible language models—they have limited 'working memory', and like humans, struggle with long-distance linguistic phenomena (e.g. long-range dependencies). This is something inherent in the architecture of RNN-LMs, even those with features such as long short-term memory (LSTM) that help them remember and forget necessary input. Transformers, on the other hand, have perfect memory for their entire context window (1024 tokens in GPT-2).

However, as discussed, some aspects of language comprehension may involve facilitation based on semantic similarity, and as we demonstrate in Experiment 2, this is a feature that appears to be more present in GPT-2, and combined with the findings of Ettinger and colleagues (Ettinger, 2020; Misra et al., 2020), the present study suggests that this may be a widespread feature of T-LMs in general. Therefore, it appears that some aspects of human language comprehension, specifically, those associated with language processing with limited working memory, may be better modeled with RNN-LM surprisal; while those that involve more shallow semantic facilitation may be better modeled with T-LMs surprisal. This may help to explain why there are conflicting results regarding which is better for modeling reading time (Merkx & Frank, 2020; Wilcox et al., 2020; Eisape et al., 2020).

Understanding the differences between the two model architectures and how these relate to different aspects of human language comprehension may thus not only help us improve our language models, but also offer insight into the neurocog-

nitive systems involved in language.

Acknowledgments

References

Aurnhammer, C., & Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, *134*, 107198.

Bardolph, M., Van Petten, C., & Coulson, S. (2018). Single Trial EEG Data Reveals Sensitivity to Conceptual Expectations (N400) and Integrative Demands (LPC). In *Twelfth Annual Meeting of the Society for the Neurobiology of Language.* Quebec City, Canada.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165–1188.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Current Biology*, *28*(5), 803-809.e3.

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

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143.

DeLong, K. A., & Kutas, M. (2020). Comprehending surprising sentences: Sensitivity of post-N400 positivities to contextual congruity and semantic relatedness. *Language, Cognition and Neuroscience*, *0*(0), 1–20.

Eisape, T., Zaslavsky, N., & Levy, R. (2020). Cloze Distillation Improves Psychometric Predictive Power. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 609–619). Online: Association for Computational Linguistics.

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, *14*(2), 179–211.

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

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

Federmeier, K. D., & Kutas, M. (1999). A Rose by Any Other Name: Long-Term Memory Structure and Sentence

Processing. *Journal of Memory and Language*, *41*(4), 469–495.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

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

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the Limits of Language Modeling. *arXiv:1602.02410 [cs]*.

Keller, F. (2010, July). Cognitively Plausible Models of Human Language Processing. In *Proceedings of the ACL 2010 Conference Short Papers* (pp. 60–67). Uppsala, Sweden: Association for Computational Linguistics.

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, *31*(5), 602–616.

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

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

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

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

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Kutas, M., & Van Petten, C. (1988). Event-related brain potential studies of language. *Advances in psychophysiology*, *3*, 139–187.

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

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60.

Merkx, D., & Frank, S. L. (2020). Comparing Transformers and RNNs on predicting human sentence processing data. *arXiv:2005.09471 [cs]*.

Michaelov, J. A., & Bergen, B. K. (2020, November). How

well does surprisal explain N400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL 2020)*. Online: Association for Computational Linguistics.

Misra, K., Ettinger, A., & Rayz, J. (2020). Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4625–4635). Online: Association for Computational Linguistics.

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, *19*(12), 1213–1218.

Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2011, December). Using Language Models and Latent Semantic Analysis to Characterise the N400m Neural Response. In *Proceedings of the Australasian Language Technology Association Workshop 2011* (pp. 38–46). Canberra, Australia.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018, September). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693–705. doi: 10.1038/s41562-018-0406-4

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. , 24.

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

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

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

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 5998–6008). Curran Associates, Inc.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. P. (2020). On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci 2020)* (p. 7).