The more human-like the language model, the more surprisal is the best predictor of N400 amplitude

James A. Michaelov
Department of Cognitive Science
University of California San Diego
La Jolla, CA 92093
j1michae@ucsd.edu

Benjamin K. Bergen
Department of Cognitive Science
University of California San Diego
La Jolla, CA 92093
bkbergen@ucsd.edu

Abstract

Under information-theoretic accounts of language comprehension, the effort required to process a word is correlated with surprisal, the negative log-probability of that word given its context. This can (equivalently) be considered to reflect cognitive effort in proportion to the amount of information conveyed by a given word (Frank et al., 2015), or the amount of effort required to update the our incremental predictions about upcoming words (Levy, 2008; Aurnhammer and Frank, 2019). In contrast, others (e.g. Brothers and Kuperberg, 2021) have argued that processing difficulty is proportional to the contextual probability of a word, thus positing a linear (rather than logarithmic) relationship between word probability and processing difficulty. We investigate which of these two accounts best explain the N400, a neural response that provides some of the best evidence for prediction in language comprehension (Kutas et al., 2011; Van Petten and Luka, 2012; Kuperberg et al., 2020). To do this, we expand upon previous work by comparing how well the probability and surprisal calculated by 43 transformer language models predict N400 amplitude. We thus investigate both which models' predictions best predict the N400, and for each model, whether surprisal or probability is more closely correlated with N400 amplitude. We find that of the models tested, OPT-6.7B and GPT-J are reliably the best at predicting N400 amplitude, and that for these transformers, surprisal is the better predictor. In fact, we find that the more highly correlated the predictions of a language model are with N400 amplitude, the greater the extent to which surprisal is a better predictor than probability. Since language models that more closely mirror human statistical knowledge are more likely to be informative about the human predictive system, these results support the information-theoretic account of language comprehension.

1 Introduction

If the language comprehension system follows information-theoretic principles, we should expect that the information content conveyed by a word should play a crucial role in how that word is processed. This is a core feature of *surprisal theory*: the idea that the difficulty in processing a word in context is correlated with the surprisal of that word, that is, the negative logarithm of the probability of the word in context (Hale, 2001, 2003; Levy, 2008; Smith and Levy, 2011, 2013; Frank et al., 2015; Aurnhammer and Frank, 2019; Michaelov and Bergen, 2020). A key mathematical property of lexical surprisal is that it reflects the information content of a word (as formally defined information-theoretically), but also, as noted by Levy (2008), is equivalent to the Kullback-Leibler divergence (Kullback and Leibler, 1951) between the predicted probability distribution for the next word and the true probability distribution (1 for the actual word and 0 for all other possible words).

Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Thus, surprisal as a metric can be thought to account not only for the increased effort required to process a word with high information content (Hale, 2001; Frank et al., 2015), but also the effort required to update the entire predicted probability distribution based on encountering a word.

An alternative to surprisal theory is what Brothers and Kuperberg (2021) term the *proportional preactivation account*, which also posits that prediction plays a key role in language comprehension. However, instead of processing difficulty reflecting the surprisal of a word in context, it reflects the extent to which the word was not predicted—that is, instead of $-\log(p)$, scales with $1-p$. Thus, under surprisal theory, the mathematical relationship between lexical probability and processing difficulty is logarithmic, while under the proportional preactivation account, it is linear.

To measure real-time human language processing, we turn to the N400, a neural response that offers a snapshot into the neurocognitive system underlying language comprehension. It is widely agreed that the N400 reflects the brain activation driven by encountering a word and that it is modulated by the extent to which the word (or its semantic features) are preactivated by the preceding context (Kutas and Federmeier, 2011; Kutas et al., 2011; Van Petten and Luka, 2012; Kuperberg et al., 2020; Federmeier, 2021). Thus, the N400 is perhaps the most suitable processing difficulty metric for evaluating the predictions of surprisal theory (Michaelov and Bergen, 2020). Despite this fact, to the best of our knowledge, the question of whether surprisal or raw probability is a better predictor of N400 amplitude has only been tested for two language models thus far—Yan and Jaeger (2020) use a mixture of a 5-gram model and 'skip bi-gram' (both trained by Frank and Willems, 2017), and Szewczyk and Federmeier (2022) use GPT-2. Both find that surprisal better fits the N400 data, though on one dataset, Szewczyk and Federmeier (2022) find that when restricting their analysis to only expected (higher-probability) items, raw probability appears to be a better predictor; while when restricting their analysis to unexpected (lower-probability items), surprisal is the better predictor.

In the present study, we substantially expand upon this previous work. We run the stimuli from a previously-published N400 study (Nieuwland et al., 2018) through 43 neural language models, calculate the probability and surprisal for each item with each model, and use these to predict N400 amplitude. First, we ask which model is the best predictor of N400 amplitude, and whether surprisal or raw probability is a better predictor in all models. Additionally, because models that behave more similarly to humans overall are also more likely to be more informative about the human predictive system, we also compare surprisal and probability across models as a function of how well they predict the human N400 response overall.

2 Method

Following research showing that N400 amplitude is best predicted by transformers (Merkx and Frank, 2021; Michaelov et al., 2022), we restricted our analysis to contemporary transformer language models made available through the *transformers* (Wolf et al., 2020) Python package. These included 43 models of the GPT-2 (Radford et al., 2019), GPT-Neo (Black et al., 2021), GPT-J (Wang and Komatsuzaki, 2021), OPT (Zhang et al., 2022), (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), XGLM (Lin et al., 2021), BLOOM (BigScience, 2022), Multilingual BERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), and XLM-R (Conneau et al., 2020; Goyal et al., 2021) architectures. All models used are shown in Figure 1.

For our N400 data, we used a subset of the data from an experiment carried out by (Nieuwland et al., 2018), a large-scale study with 334 participants. Specifically, we look at the N400 amplitudes elicited by nouns that are either more or less contextually predictable. Nieuwland et al. (2018) operationalize N400 amplitude as mean voltage between 200-500ms after stimulus presentation at 6 electrodes in their region of interest (Cz, C3, C4, Pz, P3, and P4).

To investigate how well the language models' predictions correlate with N400 amplitude, we ran each of the 160 stimulus sentences from the Nieuwland et al. (2018) study up until the critical noun through each of the the 43 language models, calculating the probability and surprisal of each of the critical nouns. Not all critical nouns were in each model's vocabulary. The fairest approach is to compare only words appearing in the vocabulary of all models (see Michaelov and Bergen, 2020; Michaelov et al., 2022 for similar work following this approach). However, with stimuli in this study, this means that only data from 37 sentences could be analyzed. An alternative approach is to predict each sub-word token of the critical noun given the preceding context and the preceding sub-word tokens, and then either taking the product of probabilities or sum of the surprisals to get an overall

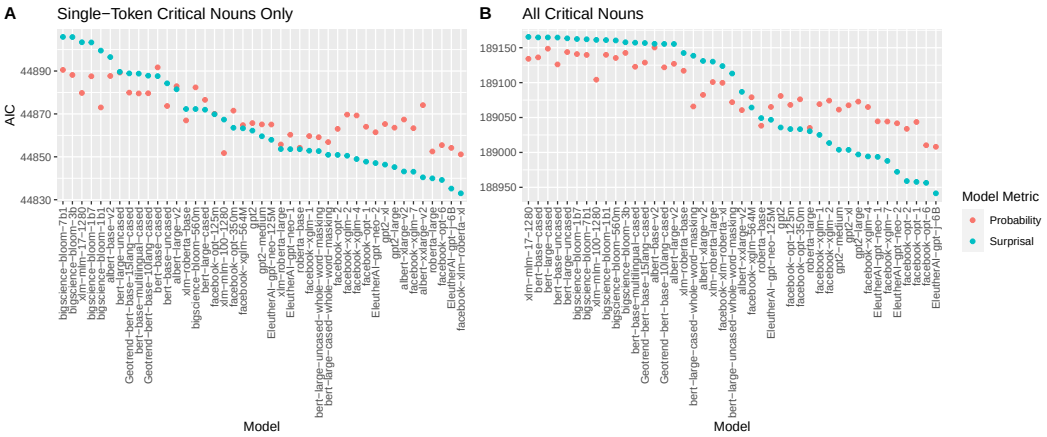

Figure 1: Comparison of the AICs of the regressions with the probability or surprisal calculated using each language model as a main effect.

word probability or surprisal. This produces a probability and surprisal for each model; however, given the differences in tokenization (including tokenizer algorithm) and training tasks (for example, traditional vs. masked language modeling, and the possible addition of whole-word masking to the latter task), this may introduce possible confounds. Thus, in our analyses, we try both approaches. Probability, surprisal, and N400 amplitude were all z-scored.

3 Results

3.1 Single-token critical nouns

First, we run our analyses for only the critical nouns that are single tokens in all models. We constructed linear mixed-effects regressions following (Michaelov et al., 2022), with the relevant language model probability or surprisal as a main effect, with the laboratory that the experiment was carried out in as a covariate, and random intercepts of experimental subject and stimulus item. We then calculated each regression's Akaike Information Criterion (AIC; Akaike, 1973) as a metric of fit. We then compared regression fit across model and metric. The results can be seen in Figure 1A.

First, in line with previous work, for the models that can be used to best predict N400 amplitude—XLM-R_{XL}, GPT-J, OPT-6B, etc.—surprisal better predicts N400 amplitude than raw probability. However, in contrast to previous work, probability is a better predictor for some other models. To test for trends across all models, we ran a likelihood ratio test using linear mixed-effects regressions. AIC was the dependent variable, and we compared a null model with only language model as random intercept with one also including metric (either probability or surprisal) as a fixed effect, finding no significant difference ($\chi^2(1) = 3.62, p = 0.0572$). Thus, there is no overall trend as to whether surprisal or probability is a better predictor of N400 amplitude.

However, as can be seen in the graph, it appears that for models whose predictions are better predictors of N400 amplitude, surprisal tends to be a better predictor of N400 amplitude than probability. In other words, it appears that as model performance at N400 amplitude prediction improves, the extent to which surprisal is a better predictor than probability increases. To test this, we constructed a simple least-squares regression with the surprisal regression AICs as the independent variable and the difference between surprisal and probability regression AICs as the dependent variable. We found a significant correlation between the two ($F(1, 41) = 11.06, p = 0.0019$) in the previously-discussed direction—as surprisal regression AIC decreases, the extent to which surprisal regression AIC is lower than probability regression AIC increases. We also constructed a least-squares regression in the same vein but with probability regression AIC as the independent variable; which also displayed a significant correlation ($F(1, 41) = 90.15, p < 0.0001$). Thus, we find that as language model performance at N400 prediction (whether based on surprisal or raw probability) increases, so does the extent to which surprisal is a better predictor.

3.2 All critical words

We then ran the same analyses on all the stimuli, where the critical nouns were either one or more tokens in all models. As before, we constructed linear mixed-effects regressions in the same way, using AIC as a metric of fit. The results are shown inFigure 1B.

On this analysis, the ranking of which models best predict N400 amplitude differs. But the higher-level patterns do not: there is no significant difference in performance between surprisal and probability overall ($\chi^2(1) = 0.97, p = 0.3258$), and regression AIC ($F(1, 41) = 218, p < 0.0001$) and probability regression AIC ($F(1, 41) = 30.6, p < 0.0001$) are correlated with the difference between the two in the same direction—that is, as the extent to which the predictions of a language model predict N400 amplitude improves, so does the extent to which surprisal is a better predictor.

4 Discussion

Clear patterns are seen in both sets of results. Under both analyses, neither surprisal nor probability is better overall, but for the models that best predict N400 amplitude (when either surprisal or probability is considered), surprisal is the best predictor. This in fact is shown to be a statistically significant regularity: and as the performance of language model predictions at predicting N400 amplitude improves, the extent to which surprisal is a better predictor than probability increases.

The question of which model best predicts N400 amplitude is more complicated. Masked language models appear to perform worse at modeling N400 amplitude when all tokens are considered compared to the when only single-token words are considered. This is striking given that under the single-token-only analysis, the best-performing model (XLM-R_{XL}) is a masked language model. This pattern is likely at least partly explained by training task. Specifically, predicting later sub-word tokens based only on their preceding context and preceding sub-word tokens occurs with every single multi-token word for autoregressive language models, but for masked language models, this is not necessarily the case. On the one hand, this suggests that if attempting to test which model best predicts a human comprehension metric like the N400, it may be best to use single-token-only analyses. On the other hand, however, from a cognitive modeling perspective, it is valuable to be able to analyze all stimuli, especially if the surprisal of a sequence of two or more words is needed. Our results suggest that for such experiments, autoregressive models such as GPT-J and OPT-6B (the best-performing in our experiment) should be used.

We now return to the question of how well an information-theoretic account of language processing, surprisal theory, accounts for the experimental neurocognitive data, compared to the proportional preactivation account. The results show that for models with more human-like predictions, surprisal better fits the data. Because these models are more likely to also reveal the mathematical relationship between lexical statistics and human predictive processing, this result supports the information-theoretic account.

Finally, we may want to ask why it is the case that for less human-like models, probability better models N400 amplitude. One possible explanation relates to Szewczyk and Federmeier's (2022) finding that language model probability is more closely correlated with N400 amplitude for higher-probability items, while surprisal is more closely correlated for lower-probability items. Because surprisal is a logarithmic metric, small differences at the extremely low-probability end of the scale are magnified. Thus, while there may only be small differences in probability for such items, for models whose predictions are less human-like, any divergence from human-like predictions are magnified at this end of the scale; while this would not be the case with raw probability. Further research is needed to investigate whether this or some other explanation explains these findings.

5 Conclusion

In this study, we investigated the mathematical relationship between the lexical predictions of language models and amplitude of the human N400 response to the same word. Specifically, we compared whether N400 amplitude is more closely correlated with word probability or surprisal, the information-theoretic metric of information content. For the models whose predictions more closely match those of humans, surprisal was the better metric, supporting an information-theoretic account of the neurocognitive system underlying the N400.

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In Petrov, B. N. and Csáki, F., editors, *Second International Symposium on Information Theory*, Springer Series in Statistics, pages 267–281, Budapest, Hungary. Akadémiai Kiadó.

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

BigScience (2022). BigScience Language Open-science Open-access Multilingual (BLOOM) Language Model. International, May 2021-May 2022.

Black, S., Gao, L., Wang, P., Leahy, C., and Biderman, S. (2021). GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow. Zenodo.

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

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Frank, S. L. and 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.

Goyal, N., Du, J., Ott, M., Anantharaman, G., and Conneau, A. (2021). Larger-Scale Transformers for Multilingual Masked Language Modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online. Association for Computational Linguistics.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies 2001 - NAACL '01*, pages 1–8, Pittsburgh, Pennsylvania. Association for Computational Linguistics.

Hale, J. (2003). The Information Conveyed by Words in Sentences. *Journal of Psycholinguistic Research*, 32(2):101–123.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

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

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

Kutas, M. and Federmeier, K. D. (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.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

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

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., Pasunuru, R., Shleifer, S., Koura, P. S., Chaudhary, V., O'Horo, B., Wang, J., Zettlemoyer, L., Kozareva, Z., Diab, M., Stoyanov, V., and Li, X. (2021). Few-shot Learning with Multilingual Language Models. *arXiv:2112.10668 [cs]*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Merkx, D. and Frank, S. L. (2021). Human Sentence Processing: Recurrence or Attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22, Online. Association for Computational Linguistics.

Michaelov, J. A. and Bergen, B. K. (2020). How well does surprisal explain N400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 652–663, Online. Association for Computational Linguistics.

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

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

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

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

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

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

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

Wang, B. and Komatsuzaki, A. (2021). GPT-J-6B: A 6 billion parameter autoregressive language model.

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

Yan, S. and Jaeger, T. F. (2020). (Early) context effects on event-related potentials over natural inputs. *Language, Cognition and Neuroscience*, 35(5):658–679.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. (2022). OPT: Open Pre-trained Transformer Language Models.