Is the relationship between word probability and processing difficulty linear or logarithmic?

James Michaelov, Megan Bardolph, Seana Coulson, Benjamin Bergen (UC San Diego)
j1michae@ucsd.edu

Research has found that the surprisal of a word—the negative logarithm of the probability of the word being the next word in an utterance—predicts multiple behavioral metrics of processing difficulty such as reading time [8, 10, 3, 6, 12, 17], as well as the N400, a neural index of semantic processing difficulty [7, 5, 1, 11]. However, a recent high-powered study by Brothers & Kuperberg [4] finds a linear relationship between word predictability and two behavioral metrics of processing difficulty (self-paced reading time and picture naming latency), and finds additional evidence of the same relationship in previous eye-tracking literature [13, 14, 15, 16].

One possible explanation is based on the different measures of predictability used between studies. Most of the surprisal studies use corpus-based metrics of word predictability such as trigram probability [17], or the probability estimated by recurrent neural network language models (RNN-LMs) trained on text corpora [7, 1, 11]. By contrast, Brothers & Kuperberg [4] use cloze probability, the probability that participants in a norming study will fill a specific gap in a sentence with a specific word. Thus, these two metrics represent different kinds of predictability—we should not necessarily expect the relationship between them and processing difficulty to be the same.

If part of the language system consists of a neurocognitive implementation of an RNN-LM-like system, that is, a system that predicts the next word in a sequence based on long-term knowledge of the statistics of language and the preceding context, there is no need in principle for its output to be directly proportional to the raw probability output of an RNN-LM. If this output is proportional to the negative log-transformed probability of an RNN-LM, as is supported by the aforementioned behavioural and neural evidence, then we should expect downstream tasks such as the cloze task to use these outputs. Thus, we would expect cloze probability to have a linear relationship with processing difficulty metrics, as was found by Brothers & Kuperberg [4]; and we would expect raw RNN probability to have a logarithmic relationship with both processing difficulty and cloze. In the present study, we investigate whether this is the case with N400 amplitude as our operationalization of processing difficulty.

To test this, we first investigate how well cloze and RNN-LM predicted probability and their log-transformed counterparts (i.e., their surprisals) predict N400 amplitude. To do this, we use a subset of the stimuli and EEG recordings from a previous ERP study [2]. We used the cloze probabilities for the sentence completions collected in the original study, and used a pretrained RNN-LM [9] to calculate corpus-based probability. Each of these was also log-transformed to get surprisal values. We used linear-mixed effects models to predict the by-trial, by-electrode mean amplitudes over the 300-500ms period after stimulus presentation (the canonical N400 time period). As can be seen in Figure 1A, we see that the models using raw cloze probabilities fit N400 amplitude better than those with the log-transformed probabilities, but that the reverse is true for the RNN-LM-derived probabilities. This shows that our initial hypothesis that the two probabilities may differ in this way is supported by the evidence.

To further investigate the hypothesis, we test whether RNN-LM probability or surprisal best predicts cloze probability by using linear mixed-effects models to predict cloze probability based on these two metrics. As can be seen in Figure 1B, we find that RNN-LM surprisal better fits cloze probability than raw RNN-LM probability. Thus, cloze probability more closely reflects corpus-derived probabilities that have been log-transformed than those that have not.

Therefore, if a neurocognitive system that predicts upcoming words based only on the surface-level statistics of previous linguistic input is either used in the cloze task or underlies the N400 response, we provide evidence that its output is closer to RNN-LM surprisal than raw probability.

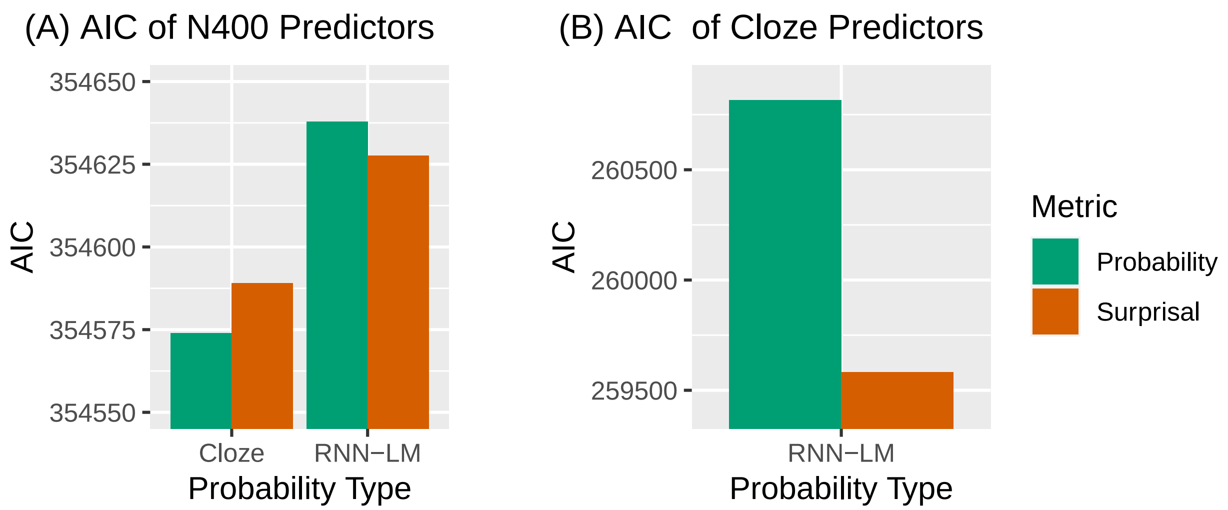

Figure 1: (A) AIC of linear mixed-effects models predicting N400 amplitude by predictor. (B) AIC of linear mixed-effects models predicting Cloze probability by predictor.

References

[1] Aurnhammer C., Frank S. L., 2019, Neuropsychologia, 134, 107198

[2] Bardolph M., Van Petten C., Coulson S., 2018, in Twelfth Annual Meeting of the Society for the Neurobiology of Language. Quebec City, Canada

[3] Boston M. F., Hale J., Kliegl R., Patil U., Vasishth S., 2008, Journal of Eye Movement Research, 2

[4] Brothers T., Kuperberg G. R., 2021, Journal of Memory and Language, 116, 104174

[5] Delaney-Busch N., Morgan E., Lau E., Kuperberg G., 2017, in Proceedings of the 39th Annual Conference of the Cognitive Science Society. London, UK, p. 6

[6] Demberg V., Keller F., 2008, Cognition, 109, 193

[7] Frank S. L., Otten L. J., Galli G., Vigliocco G., 2015, Brain and Language, 140, 1

[8] Hale J., 2001, in Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies 2001 - NAACL '01. Association for Computational Linguistics, Pittsburgh, Pennsylvania, pp 1–8, doi:10.3115/1073336.1073357

[9] Jozefowicz R., Vinyals O., Schuster M., Shazeer N., Wu Y., 2016, arXiv:1602.02410 [cs]

[10] Levy R., 2008, Cognition, 106, 1126

[11] Michaelov J. A., Bergen B. K., 2020, in Proceedings of the 24th Conference on Computational Natural Language Learning. Association for Computational Linguistics

[12] Monsalve I. F., Frank S. L., Vigliocco G., 2012, in Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp 398–408

[13] Rayner K., Well A. D., 1996, Psychonomic Bulletin & Review, 3, 504

[14] Rayner K., Li X., Juhasz B. J., Yan G., 2005, Psychonomic Bulletin & Review, 12, 1089

[15] Rayner K., Reichle E. D., Stroud M. J., Williams C. C., Pollatsek A., 2006, Psychology and Aging, 21, 448

[16] Sereno S. C., Hand C. J., Shahid A., Yao B., O'Donnell P. J., 2018, Quarterly Journal of Experimental Psychology, 71, 302

[17] Smith N. J., Levy R., 2013, Cognition, 128, 302