Surprisal is a good predictor of the N400 effect, but not for semantic relations

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

The N400, perhaps the best characterized neural index of semantic processing, has recently attracted the attention of computational modelers (Laszlo & Plaut, 2012; Laszlo & Armstrong, 2014; Rabovsky & McRae, 2014; Cheyette & Plaut, 2017; Brouwer *et al.*, 2017; Rabovsky *et al.*, 2018; Venhuizen *et al.*, 2018; Fitz & Chang, 2019). Such work has shown that the surprisal (negative log probability) of a word, as calculated using a recurrent neural network language model (RNN-LM), is a significant predictor of N400 amplitude (Frank *et al.*, 2015). Here we evaluate whether the RNN-LM is a candidate model for the semantic retrieval operations indexed by the N400 component of the scalp recorded ERP. If human sensitivity to surface level statistics in language arises from a cortical implementation of an RNN-LM, we might expect N400 amplitude to be well-predicted by surprisal values derived from an RNN-LM. While we expect surprisal values to provide an excellent model of cloze probability effects, we question whether surprisal and other measures derived from surface-level statistics of language can adequately account for more conceptual factors known to modulate the N400.

EEG was recorded from 29 scalp sites as 44 healthy adults read English sentences one word at a time. Materials (adapted from Thornhill & Van Petten, 2012) were 290 sentence frames completed with words in four conditions: *It's hard to admit when one is* ***wrong*** (BEST COMPLETION) / ***incorrect*** (RELATED to best completion) / ***lonely*** (UNRELATED to best completion) / ***screened*** (IMPLAUSIBLE). While RELATED and UNRELATED completions were matched for cloze, we hypothesized that the overlap between semantic features associated with BEST COMPLETIONS and RELATED words would lead to a less negative N400 amplitude in RELATED than UNRELATED words. We also hypothesized that conceptual factors would impact N400 amplitude, specifically, that IMPLAUSIBLE words would elicit larger N400 amplitudes than UNRELATED words. To calculate surprisal values for the target words, the stimuli were run through Jozefowicz *et al.*'s 2016 BIG LSTM+CNN INPUTS RNN-LM. Conditions differed significantly in mean surprisal ($p < .001$): BEST COMPLETION: 6.26, RELATED: 9.55 UNRELATED: 10.2, IMPLAUSIBLE: 19.8.

N400 amplitude was operationalized as mean amplitude 300-500ms post-word onset in each electrode on each trial. As expected, N400 amplitude was largest for IMPLAUSIBLE completions, followed by UNRELATED, RELATED, and BEST COMPLETION words. Linear mixed effects models were constructed to estimate the impact of surprisal ($M_{Surprisal}$), experimental condition ($M_{Condition}$), and their combination ($M_{Combined}$). With fixed effects of surprisal, electrode, and their interaction, $M_{Surprisal}$ showed that surprisal predicted single trial N400 amplitude well ($p < .001$; as did surprisal x electrode: $p < .001$), as in Frank *et al.* (2015). Employing fixed effects of condition, electrode, and their interaction, $M_{Condition}$ indicated experimental condition significantly predicts N400 amplitude ($p < .001$; condition x electrode: $p < .001$), replicating Thornhill & Van Petten (2012). However, conceptual factors indexed by experimental condition cannot be reduced to surprisal. While $M_{Condition}$ is significantly improved by adding surprisal ($p < .001$) and its interaction with electrode ($p < .001$), $M_{Surprisal}$ is significantly improved by including experimental condition ($p < .001$) and its interaction with electrode ($p < .001$).

To reveal unique variance explained by each factor we examined the models' predictions of N400 amplitudes for data held out from the statistical analysis (54,940 measurements; 15% of total data). Figure 1 shows that $M_{Surprisal}$ is far worse at distinguishing between UNRELATED and RELATED completions than $M_{Condition}$ (True mean difference: -0.606; $M_{Surprisal}$ predicted difference: -0.075; $M_{Condition}$ predicted difference: -0.499). To conclude, while neural activity underlying the N400 may be based in part on our knowledge of surface-level language statistics, other factors must explain the neural response to the RELATED condition. These may include predictions based on world knowledge or spreading activation from the most likely completion.

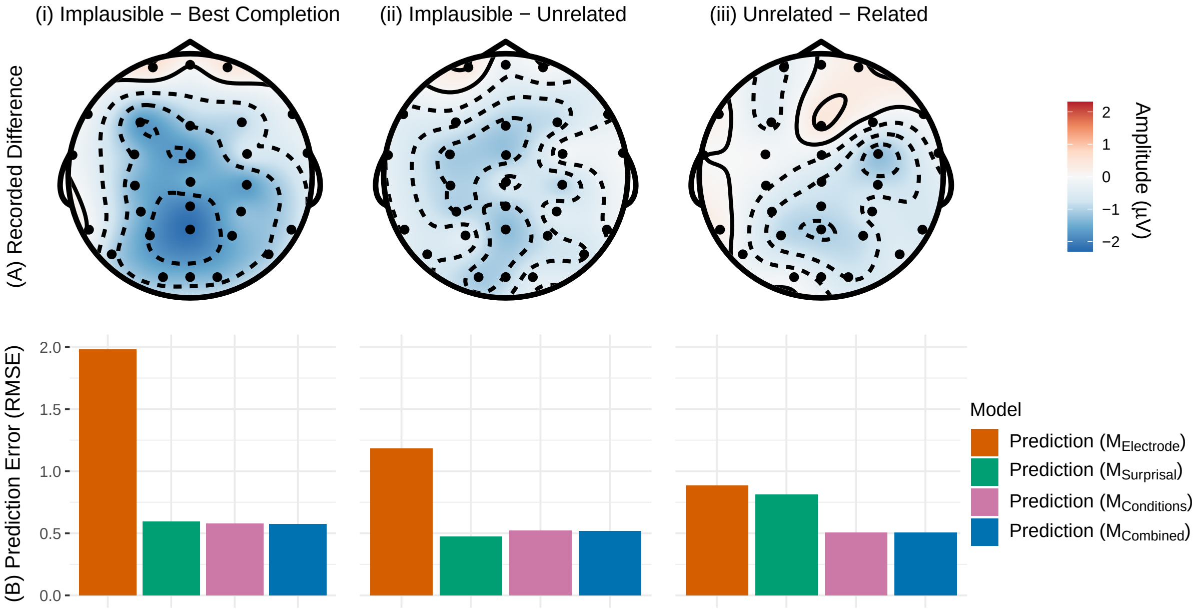

Figure 1: (A) Topographic plots representing the mean difference in potential between (i) the Implausible and Best Completion conditions, (ii) the Implausible and Unrelated conditions, and (iii) the Unrelated and Related conditions, at each electrode recording site. (B) The root mean-squared error of the predictions of each of the statistical models for the same mean differences at each electrode. $M_{Electrode}$ has Electrode as is its only fixed effect, $M_{Surprisal}$ adds Surprisal and its interaction with Electrode, $M_{Condition}$ instead adds Condition and its interaction with electrode, and $M_{Combined}$ includes all five aforementioned fixed effects.

References

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. (2017). *Cognitive science,* **41**, 1318–1352.

Cheyette, S. J. & Plaut, D. C. (2017). *Cognition,* **162**, 153–166.

Fitz, H. & Chang, F. (2019). *Cognitive Psychology,* **111**, 15 – 52.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). *Brain and Language,* **140**, 1–11.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). *arXiv:1602.02410 [cs],* .

Laszlo, S. & Armstrong, B. C. (2014). *Brain and language,* **132**, 22–27.

Laszlo, S. & Plaut, D. C. (2012). *Brain and language,* **120** (3), 271–281.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). *Nature Human Behaviour,* **2** (9), 693–705.

Rabovsky, M. & McRae, K. (2014). *Cognition,* **132** (1), 68–89.

Thornhill, D. E. & Van Petten, C. (2012). *International Journal of Psychophysiology,* **83** (3), 382–392.

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2018). *Discourse Processes,* **56** (3), 1–27.