Language modeling using a neural network shows effects on N400 beyond just surprisal

Don Bell-Souder, Shannon McKnight, Vladimir Zhdanov, Sean Mullen, Akira Miyake, Phillip Gilley, and Albert Kim
(University of Colorado Boulder, Institute of Cognitive Science)

Electroencephalography (EEG) has provided evidence that the brain makes word-level predictions (Kuperberg & Jaeger, 2016; Van Berkum et al., 2005). However, such evidence comes from target words appearing in highly constraining sentence contexts, raising important questions about the generalizability of the effects. Here, we examined brain activity elicited by sentence-embedded words that varied substantially in their contextual support. We used a Long Short-Term Memory (LSTM) neural network to generate context-driven predictions about each word, modelling the predictions that human comprehenders might make (Sundermeyer et al., 2015). By comparing the model-generated predictions to the brain activity at each word, we evaluated the degree to which comprehenders were actually predicting words during sentence comprehension (inspired by earlier work by Frank et al., 2015; Frank & Hoeks, 2019).

The LSTM predictions were tested against EEG data collected from 190 young adults (ages 18-30) reading 400 experimental sentences (RSVP format), of which 240 were well-formed. For 5440 words, we quantified the amplitude of the N400 ERP component as mean voltage 300-500 ms post-stimulus-onset averaged across seven central-parietal channels. Substantial past research indicates that the N400 amplitude reflects the ease with which a word is accessed (Kutas and Federmeier, 2011). N400's for each word were averaged across participants.

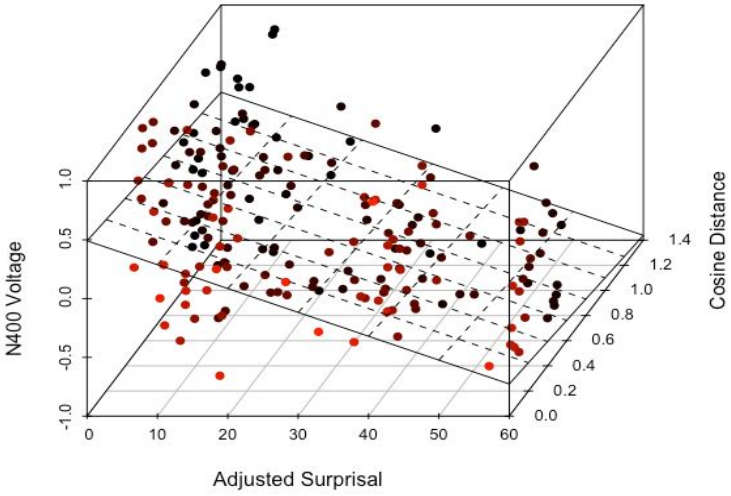
We initialised our LSTM on pretrained 300-dimensional Word2Vec embeddings, which were pretrained on the English Wikipedia corpus, and fine-tuned the LSTM using 1043 well-formed sentences from prior studies that were not included in the current study.

We used the LSTM to generate four predictors of brain activity. First, we used the LSTM to generate a distribution of conditional probabilities for words, given the previous context, and from this calculated the surprisal of each presented word. Second, we developed a novel measure of adjusted surprisal, by subtracting the surprisal of the word most predicted by the LSTM from that of the presented word. This quantifies the surprisal of a word after accounting for the level of surprisal that might have been anticipated given the context. Third, we calculated each perplexity at each word with it's left context. Finally, we calculated the cosine distance between Word2Vec embeddings for the presented word and the LSTM predicted word, reflecting the semantic distance between the most likely and actually presented words.

Regressing the N400 measures on our four candidate predictors, we found that cosine distance, surprisal, and adjusted surprisal were significant predictors. Greater cosine distance predicted more negative N400s ($F(1,5437) = 11.9$, $p < 0.001$) and increased adjusted surprisal also predicted more negative N400s ($F(1,5437) = 1159.4$, $p<0.001$) when controlling for the other (Figure 1). Surprisal and adjusted surprisal were highly correlated and therefore could not be evaluated in the same multiple regression model, but a model with cosine distance and adjusted surprisal outperformed one with cosine distance and surprisal.

These analyses show that the LSTM framework is a useful tool to examine EEG responses. More importantly, it shows that adjusted surprisal is a valuable way to quantify how the N400 is reflecting the difference in what the brain was already prepared to process and what it actually received. We plan to continue this analysis to examine if the same measures also explain other EEG components like the P600 or if they are differentially explained by different measures.

Figure 1



The regression plane of N400 voltage on adjusted surprisal and cosine distance. Also plotted are a random selection of 200 points representing individual words in the analysis.

Definitions of computed measures

Surprisal

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

Perplexity

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_1 \ldots w_{i-1})}}$$

Cosine Distance

$$CD(u,v) = 1 - \frac{u \cdot v}{\|u\|_2 \cdot \|v\|_2}$$

Table 1 - example sentences

Type	Example
Simple	My brother came into the room and looked around.
Complex	The quarterback that the fat bully ran from yelled for help.
Well-formed Control	The webs were spun by **a spider** this morning.
Semantic Anomaly	The webs were *spun* by *a* **clown** this morning.
Syntactic Anomaly	The webs were spun by **from** *spider* this morning.

References

Frank, S. L., & Hoeks, J. C. (2019). The interaction between structure and meaning in sentence comprehension. Recurrent neural networks and reading times.

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.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension?. *Language, cognition and neuroscience*, *31*(1), 32-59.

Sundermeyer, M., Ney, H., & Schlüter, R. (2015). From feedforward to recurrent LSTM neural networks for language modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(3), 517-529.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443.