Model-based estimates of predictability reveal brain's robust sensitivity to variation in semantic fit even among unexpected words.

Jakub M. Szewczyk, Kara D. Federmeier (University of Illinois at Urbana-Champaign)

The brain's graded response to words that vary in their predictability in a sentence has been well-established: There is a monotonic relationship between the amplitude of the N400 ERP component and human production norms (cloze probability: CP), such that more predictable words elicit smaller N400s. However, this discriminability approaches a limit for words with CPs near zero because of limited variance in CP values and noisy estimation of CPs for weakly predictable items. Moreover, even comparisons between plausible but unexpected and wholly anomalous words yield only small N400 differences (e.g., Kuperberg et al., 2020). It is unclear whether this pattern is revealing of the mechanism underlying contextual facilitation – for example, that the language processor predicts only a small set of specific lexical items (e.g., Van Petten & Luka, 2012) – or simply because the CP metric is at floor and thus is unable to pull out variance that is actually in the signal. Two views provide opposing predictions: 1) the N400 is sensitive to differences in predictability of all words and it is related to the predictability on the log scale (the surprisal theory, Levy, 2008; Kuperberg & Jaeger, 2016); 2) the N400 is sensitive to predictability only in the range measurable with CP tests and the relationship is linear (Brothers & Kuperberg, 2021).

To adjudicate between these possibilities, in this study, in this study, we reanalyzed data from an ERP experiment in which 32 participants saw 282 simple English sentences that were completed by expected and unexpected (but plausible) words. We quantified the predictability of the sentence endings using GPT2-xl, a state-of-the-art machine learning model of language, which assigns a probability distribution across all possible sentence continuations. We first tested how model-derived predictability compares with predictability estimated by classic CP tests in explaining N400 amplitudes to expected endings, in which CP varied in the range 0.09-1.00 (mean CP=0.56). The mixed-effects regression revealed that both sources of predictability estimates are excellent predictors of the N400 amplitude to the sentence endings ($t=4.9$ for the GPT-2 model, $t=5.1$ for CPs), although, as revealed by model comparison, CP explained N400 variance over and beyond GPT2 ($\Delta\text{logLik} = 4$, $p < .01$) but not the other way around ($\Delta\text{logLik} = 1$, $p = .17$). Overall, both models explained N400 amplitude variance in the range of 5 μV (see Figure 1, left panel). Additional GAMM models showed that the relationship is linear (both with predictability estimated by CP tests and by the GPT2 model).

Next, we analyzed the response to unexpected endings. Here, CP could not explain N400 amplitude as all unexpected words had CP=0. However, a mixed-effects regression using the GPT2-based index of predictability revealed that the N400 was robustly sensitive to predictability even in this range ($t = 5.9$) and even though all the unexpected endings were fully plausible. Indeed, variance in N400 amplitude to unexpected endings was surprisingly large, exceeding variance observed to expected endings (see Figures 1 & 2). Additional GAMM models showed that the relationship between the N400 and predictability in this range is logarithmic. We replicated these findings using two other ERP experiments using similar items, involving 42 participants and 8602 data-points in total.

Because different functions related predictability and the N400 to expected and unexpected words, we made a final model using a single function that could jointly fit both types of words: $\beta_1 * p + \beta_2 * \log(p)$ (see Figure 2). We propose that the logarithmic component ($\log(p)$) reflects updating of conceptual representations, in line with the surprisal theory (Levy, 2008), while the linear component (p) corresponds to the degree to which lexical representation of the word was hierarchically preactivated by the representation of the context.

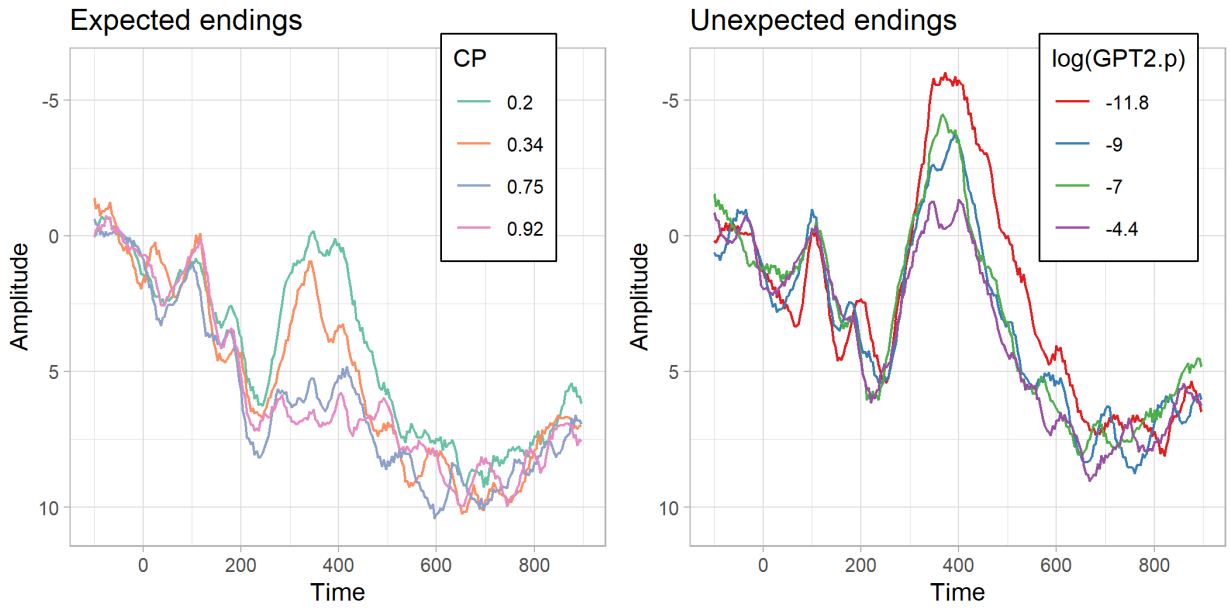

Figure 1. ERPs to expected (left panel) and unexpected (right panel) sentence endings, broken down by their predictability estimated by cloze probability tests (left panel, linear scale) or the GPT2 model (right panel; log scale). The bins were set to have an equal number of items. Values in the legend correspond to the mean predictability in each bin.

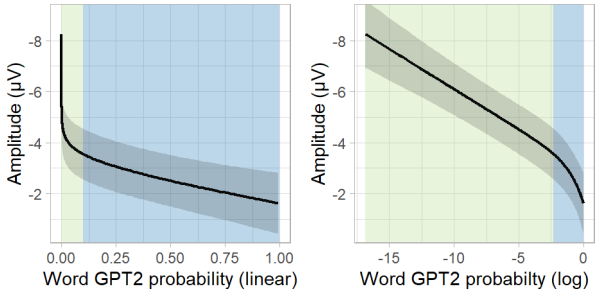

Figure 2. Predictions of the model of the N400 amplitude that includes predictors of word's probability both on the linear and logarithmic scale. Left panel: probability on the linear scale; right panel: probability on the logarithmic scale. The contrast between two colors of background corresponds to a threshold (arbitrarily set at p = .1) separating regions where the relationship between word probability and N400 is more linear (blue) and more logarithmic (green).

References:

Brothers & Kuperberg (2021). JML, v116, 104174
Levy, R. (2008). Cognition, v106, 1126–1177
Kuperberg, G. R., & Jaeger, T. F. (2016). LCN, v31, 32–59
Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). JoCN, v32, 1-35
Kutas, M., & Hillyard, S. A. (1984). Nature, v307, 161–163
Van Petten, C., & Luka, B. J. (2012). IJoP, v83, 176–190