Bayesian surprise during incremental anticipatory processing: a re-analysis of Nieuwland et al. (2017), based on DeLong et al. (2005)

Shaorong Yan[1], Gina R. Kuperberg[2,3], T. Florian Jaeger[1,4,5]
[1]Department of Brain and Cognitive Sciences, University of Rochester; [2]Department of Psychology, Tufts University; [3]Department of Psychiatry and the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School; [4]Department of Computer Science; [5]Department of Linguistics, University of Rochester

The extent to which language processing involves prediction of upcoming inputs remains a question of ongoing debate. One important data point comes from DeLong et al. (2005) who reported that an N400-like event-related potential (ERP) correlated with the cloze probabilities of articles whose form depended on an upcoming noun. This result is often cited as evidence for gradient probabilistic prediction of the semantics of the noun (mediated through prediction of its form), prior to its bottom-up input becoming available. However, a recent 9-lab study reports a failure to replicate this effect (Nieuwland et al., 2017).

We spell out the computational nature of predictive processes that one might expect to correlate with ERPs evoked by a functional element whose form is dependent on an upcoming predicted word. From this we derive, both conceptually and formally, a principled measure of anticipatory processing of the noun's semantics upon encountering the article: *Bayesian surprise,* i.e., the relative entropy or change in semantic expectations. We argue that this is a more appropriate measure of gradient semantic prediction than the cloze probability of the article.

We then formally show that this measure of Bayesian surprise closely approximates the article's *surprisal* (i.e., its log-transformed inverted probability). We present a large-scale corpus analysis investigating the relationship between the article's surprisal and the Bayesian surprise over noun semantics. Across several corpora of written and spoken American and British English, we find that the two measures are strongly correlated (*r*s range from .95 to .98). (This correlation is considerably higher than the correlation between the article's raw probability and Bayesian surprise, which ranged from -.65 to -.75 across the same corpora.)

Finally, we re-analyze the ERP data from Nieuwland and colleagues using the article's cloze surprisal, rather than its raw cloze probability, as an index of prediction. We find that surprisal is gradiently correlated with the amplitude of the N400 evoked by the article ($p < 0.05$).

Our results suggest that it is premature to conclude that Nieuwland et al.'s dataset provides no evidence for probabilistic anticipatory processing. Rather, in Nieuwland et al.'s data, we find that a measure that closely approximates a principled index of probabilistic semantic prediction—the Bayesian surprise over the noun semantics incurred while processing the article—*is* correlated with the neural index of semantic prediction: the N400. Our approach does, however, emphasize the need for future studies to further clarify the nature and degree of prediction at the level of both semantics and form, as well as its neural signatures, during language comprehension.

References:

1. DeLong et al. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*.

2. Nieuwland et al. (2017). Limits on prediction in language comprehension: A multi-lab failure to replicate evidence for probabilistic pre-activation of phonology. *bioRxiv*.