Different Reading Processing Stages or Different Brain Areas? A Computational Cognitive Investigation on N400, P600, and PNP

Lavinia Salicchi, Yu-Yin Hsu
The Hong Kong Polytechnic University
{la-vinia.salicchi, yu-yin.hsu}@polyu.edu.hk

The classical distinction between **N400** as an index of semantic processing and **P600** as a marker of syntactic processing has been challenged by studies reporting P600 effects in response to semantic violations. This has led to debates about the functional roles of these event-related potentials (ERPs), particularly the frontal P600 (**PNP**) and its relationship with N400 and posterior P600 [1]. Computational metrics like surprisal, entropy, and semantic similarity, mathematically representing cognitive dynamics, have been employed to model these ERPs [2, 3], to directly test which mechanisms take place at different reading stages. However, little computational research has been conducted on P600 and PNP, especially in non-alphabetic languages like Mandarin Chinese.

We analyzed EEG data from 38 participants reading 280 grammatical **Mandarin Chinese** sentences without semantic violations. This type of data allows us to extend previous research to Sinitic languages and create a general baseline for future investigations. We extracted N400, P600, and PNP, employing the channels selected in [4]. Using a Chinese GPT-2 model for conditional probabilities and word embeddings, we computed surprisal, entropy, entropy variation, and three semantic similarity metrics: *sentword*, a context-word similarity employed in [5], the semantic similarity between the upcoming word and the most expected word (*cosk1*) or a general concept based on the five most expected words (*cosk5*). We created 10 **linear mixed-effect models**: a baseline model, including word-level features only, 6 models employing the word-level regressors and one computational metric, and three general models, including all the features. As in [6], the baseline signal was as a covariate of no interest, and word ID and participant ID were random intercepts. To assess each metric's predictive power, we computed the target model - baseline model log likelihood difference (ΔLL).

Surprisal was the strongest predictor of N400 amplitude (ΔLL = 6.94, significantly different from zero - $p < 0.001$), suggesting that in early processing stages, readers are sensitive to the absence of expected lexical items. **Entropy variation** and **expectation-driven semantic similarity** (*cosk5*) predicted PNP (ΔLL = 4.43, $p = 0.001$ and ΔLL = 3.30, $p = 0.004$), suggesting that in later stages, readers perform a higher-level semantic evaluation and suppress previous expectations. The context-word **semantic similarity** predicted both P600 and PNP, indicating a semantic integration happening in later stages and involving a wide network. **Entropy** significantly modulated all ERPs.

Our findings support a multi-stage model: In the early stages of word processing, a centro-parietal network assesses whether w_n matches the predictions generated by the preceding context (C_{n-1}), with unexpected words requiring greater cognitive resources. Simultaneously, the number of possible continuations maintained in working memory increases cognitive effort. In later stages, if w_n introduces new sentence constraints, the resulting cognitive demand can be traced as frontal brain activity. Meanwhile, the reader evaluates the degree to which w_n fits C_{n-1}, with poorer matches inducing a higher cognitive load across frontal and posterior areas. Finally, a frontal network compares w_n's semantics to the predicted general concept, with conceptual mismatches being cognitively more expensive.

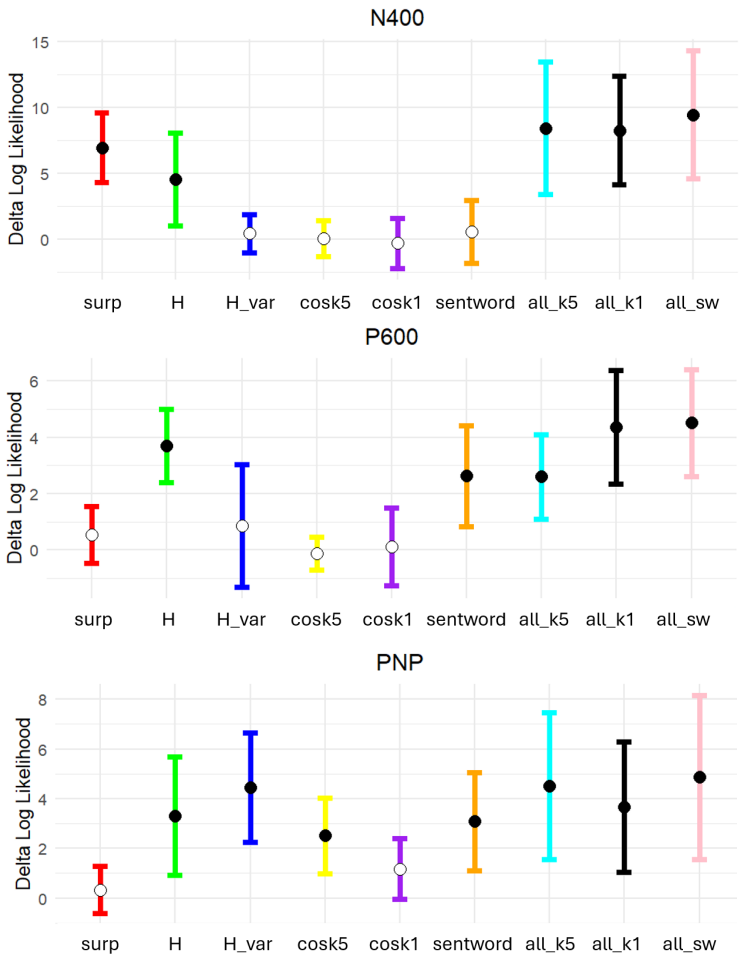

Figure 1: ΔLL values of models employing one or all computational metrics in predicting ERPs (from left to right: surprisal, entropy, entropy variation, cosine similarity between target word and top-five predictions, cosine similarity between target word and preferred prediction, cosine similarity between target word and context, and computational models employing all the first three metrics and one type of cosine). Error bars indicate 95% confidence intervals. Full dots indicate a ΔLL statistically different from zero.

References

[1] Aurnhammer, C. (2024). *Expectation-based retrieval and integration in language comprehension*.

[2] Li, J., & Futrell, R. (2023). A decomposition of surprisal tracks the N400 and P600 brain potentials. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

[3] Xu, H., Nakanishi, M., & Coulson, S. (2024). Revisiting Joke Comprehension with Surprisal and Contextual Similarity: Implication from N400 and P600 Components. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

[4] 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.

[5] Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2024). Strong Prediction: Language model surprisal explains multiple N400 effects. *Neurobiology of language*, *5*(1), 107–135.

[6] Frank, S. L., & Aumeistere, A. (2024). An eye-tracking-with-EEG coregistration corpus of narrative sentences. *Language Resources and Evaluation*, *58*(2), 641–657.