Not Every Metric is Equal: Cognitive Models for Predicting N400 and P600 Components During Reading Comprehension

Lavinia Salicchi
The Hong Kong Polytechnic University
Hung Hom, Hong Kong
lavinia.salicchi@connect.polyu.hk

Yu-Yin Hsu
The Hong Kong Polytechnic University
Hung Hom, Hong Kong
yu-yin.hsu@polyu.edu.hk

Abstract

In recent years, numerous studies have sought to understand the cognitive dynamics underlying language processing by modeling reading times and ERP amplitudes using computational metrics like surprisal. In the present paper, we examine the predictive power of surprisal, entropy, and a novel metric based on semantic similarity for N400 and P600. Our experiments, conducted with Mandarin Chinese materials, revealed three key findings: 1) expectancy plays a primary role for N400; 2) P600 also reflects the cognitive effort required to evaluate linguistic input semantically; and 3) during the time window of interest, information uncertainty influences the language processing the most. Our findings show how computational metrics that capture distinct cognitive dimensions can effectively address psycholinguistic questions.

1 Introduction

Surprisal theory states that the cognitive effort needed to process a word is proportional to the probability of encountering such a word in a given context (Hale, 2001; Levy, 2008). This concept has been modeled in computational psycholinguistics and Natural Language Processing (NLP) as the *surprisal*, i.e., the negative logarithm of the conditional probability of a word given its context, which is typically computed using the probability distribution of words provided by language models (Hale, 2016). Several studies have found correlations between the surprisal of a word and its reading time (Smith and Levy, 2013; Frank, 2017; Salicchi et al., 2021), or event-related potentials (ERPs) amplitudes (Xu et al., 2024; Frank and Aumeistere, 2024). Most efforts in ERP modeling have focused on predicting the N400 effect, and only recently, some attention has been given to other components, such as P600 (de Varda et al., 2024; Krieger et al., 2024).

Aurnhammer et al. (2021) found that N400's sensitivity to lexical association reduces the expectancy effect when a priming term precedes unexpected items. P600, on the contrary, is less sensitive to lexical association and is elicited by unexpected words regardless of priming context. Aurnhammer et al. (2023) and Delogu et al. (2021) observed that when a strong semantic association exists between terms, implausible words do not elicit an N400 effect, instead triggering a P600 effect, with its amplitude modulated by the degree of plausibility. These observations suggest that P600 is more sensitive to sentence-level semantics than N400, which appears to be more influenced by local phenomena. Krieger et al. (2024) used materials from these three studies to investigate the effectiveness of surprisal in modeling N400 and P600 in different experimental settings. The results showed that surprisal from large language models (LLM) was useful in modeling most N400 effects, but consistently struggled to account for P600 amplitudes. These limitations of surprisal in modeling human cognition have led scholars to develop alternative metrics to predict ERPs; including approaches based on cosine similarity (Michaelov et al., 2024) and noise modeling (Li and Futrell, 2023).

In the present paper, we examine the predictive power of surprisal, entropy, and a new metric that combines predictability and semantics (*semantic similarity*), focusing on both N400 and P600 components. Our experiments use Mandarin Chinese materials, making this the first study, to our knowledge, to explore the interplay between these two ERPs and computational metrics in this language.

2 Related Work

In this study, we focused on N400 and P600. N400 is a negative deflection of EEG signal detected in the centro-parietal areas between 300 and 500 ms after stimulus onset. It is typically associated with semantic violations. P600 is a positive-going de-

Proceedings of the 31st International Conference on Computational Linguistics, pages 3648–3654
January 19–24, 2025.

flection traceable in the parietal regions between 600 and 1000 ms after stimulus onset. In the classical view, it is associated with syntactic violations. However, studies examining semantic violations, particularly those involving thematic role violations, have frequently reported a P600 effect instead of the expected N400. This has cast doubt on the exclusively syntactic nature of P600, and the late positive ERP was hypothesized to indicate reanalysis/repair (Van Herten et al., 2005, 2006) or semantic integration (Brouwer et al., 2012).

On the computational side, over the years, various approaches have been proposed to evaluate surprisal's (and other metrics') psychometric predictive power (PP) on reading times and ERPs. Many studies have used computational metrics in linear mixed-effects models to assess their efficacy in accounting for variations in psychometric variables. Van Schijndel and Linzen (2018) reported that surprisal and entropy were significant predictors of reading times when used together, indicating that they capture different cognitive phenomena of reading. Wilcox et al. (2023) found that while surprisal alone had a higher PP than entropy alone, their joint usage yielded the best performances. Similar results were reported by Haller et al. (2024) with German reading materials. Frank and Aumeistere (2024) reported that surprisal significantly influenced both eye-tracking and EEG metrics in Dutch.

Michaelov et al. (2024) examined the sensitivity of N400 to predictability and semantic similarity between the context and target word, using surprisal and cosine similarity, respectively. The latter was computed between the embedding of the target word and the vector representing its left context. The investigation showed that N400 was better predicted through surprisal, highlighting the primary role of predictability in this phase of language processing. The same approach was followed by Xu et al. (2024), who found that, although the best model in predicting N400 employed both surprisal and the context-target cosine similarity, only the latter was statistically significant. Moreover, the model relying on cosine similarity exclusively was the one accounting for P600 variations. Li and Futrell (2024)[1] successfully modeled N400 and P600, decomposing surprisal into *heuristic surprisal* and *discrepancy signal*. The first element, representing the amount of cognitive effort to create heuristic interpretations, was based on the canonical surprisal formula and successfully accounted for the N400 effect. The second component, representing the cognitive load required to process the veridical interpretation, included a semantic similarity metric computed as the cosine similarity between the heuristic interpretation of the sentence and the real sentence. A similar approach, but in the context of conversations and reading times, was proposed by Giulianelli et al. (2023) with the concept of *information value*.

All previous investigations focused on Indo-European languages, such as English, German, and Dutch. To our knowledge, no study focused on the PP of computational metrics representing different cognitive dynamics using Mandarin Chinese materials. Moreover, little computational research has been conducted on language processing underlying the P600 effect. For these reasons, adopting the method employed in previous cognitive modeling investigations, and introducing a new metric, our goal is to explore which computational metrics better predict N400 and P600 recorded during Mandarin sentence comprehension, and to identify the cognitive dynamics that take place in the 300-500 ms and 600-1000 ms time windows.

3 Method

Following previous studies, we implemented 5 linear mixed-effects models. 1) A **baseline** (BL) model, employing word-level features such as number of strokes, log-unigram frequency, and position of the word within the sentence, and, in accordance with the setting in Frank and Aumeistere (2024), we included the signal baseline, computed from the same channels used for calculating N400 and P600, in the 100ms preceding the word onset. 2) The **Surprisal** (surp) model, employing the baseline features and the surprisal computed using GPT-2. 3) The **Entropy** (H) model, employing the baseline features and the contextual entropy. 4) The **semantic similarity** (cos) model, employing the baseline features and the expectations-driven semantic similarity. 5) A model employing **all the features** (all): baseline features, surprisal, entropy, and semantic similarity. Details on these metrics are provided below.

We then split our dataset into training and test sets. To avoid overfitting, we performed a 10-fold cross-validation, computing the log-likelihood of the models on the held-out portion of data. The PP of the models was assessed in terms of average

[1] see also Li and Futrell (2023); Li and Ettinger (2023)

log-likelihood difference (ΔLL) between the target model and the baseline. The higher the ΔLL, the greater the metric's PP. The ΔLL values were then subject to a t-test to determine whether each difference was statistically different from zero at $\alpha = 0.05$.

Understanding the information represented by the computational metrics will allow us to interpret the performances of the models, informing us of the type of cognitive dynamics taking place in the two different time windows we focus on.

3.1 Computational metrics

For all the computational metrics, we employed a Chinese version of GPT2 Large[2]. The decision of employing GPT2 was twofold: on the one hand, the autoregressive nature of GPT ensures that only the left context influences a word's metric or embedding; on the other, recent studies investigating the predictive power of computational metrics found GPT2 to be a better choice than bigger and more recent language models (Kuribayashi et al., 2023; Haller et al., 2024).

Surprisal: representing the extent to which a word is unexpected, given the previous context.

$$Surprisal(w_n) = -log(P(w_n|C_{n-1})) \quad (1)$$

Where C_{n-1} is the context preceding the target word $(w_0, w_1, ..., w_{n-1})$ and $P(w_n|C_{n-1})$ is the conditional probability of w_n provided by the language model.

Contextual entropy: representing the level of uncertainty about the upcoming linguistic input at word w_n.

$$H(w_n) = -\sum_{w \in \overline{\Sigma}} P(w|C_{n-1}) log_2(P(w|C_{n-1})) \quad (2)$$

Where $\overline{\Sigma} = \Sigma \cup \{EOS\}$ is the Σ vocabulary of the language model enriched with the special token EOS indicating the end of the string.

Semantic similarity: representing the semantic closeness between the expected word given the context, and the actually upcoming one. The expected item is not however a unique, precise word, but a *general concept*, created upon the five most likely upcoming characters[3].

$$candidate = mean(\overrightarrow{\arg\max^5_{w \in \overline{\Sigma}} P(w|C_{n-1})}) \quad (3)$$

$$semantic\_sim.(w_n) = \cos(\overrightarrow{w_n}; candidate) \quad (4)$$

For each word in the vocabulary ($w \in \overline{\Sigma}$) we computed the conditional probability $P(w|C_{n-1})$. We selected the five words with the highest probability ($\arg\max^5$), inserted them in the context, and extracted their contextual word embeddings using the language model. We then computed the centroid ($mean$) between these five vectors (i.e., the *candidate*). Finally, the embedding of the word present in the stimuli is computed, and the cosine similarity (i.e., semantic similarity) between such an embedding and the vector representing the most salient expectations is computed[4] (Eq. 4). This metric represents a cognitive process at a higher level of the local phenomena surprisal may account for. It implies an evaluation of the semantics of the linguistic input, its integration within the context, and a comparison between the input's meaning and the expectations' semantics. If the semantic similarity is useful in the prediction of N400, it may suggest an early integration of local and sentence information; if, on the contrary, it better accounts for the P600 effect, it means that such complex semantic dynamics take place in later stages of language processing.

Prior to the regression model creation, we checked Pearson's correlations between computational metrics through a Monte Carlo estimation. We found a strong positive correlation ($r = 0.64$) between surprisal and entropy, a weak negative correlation between semantic similarity and surprisal ($r = -0.35$), and a very weak correlation between semantic similarity and entropy ($r = -0.08$). All the regressors were z-transformed before fitting the models.

3.2 Materials

We adopted the full set of items used in Jap et al. (2024), which contains 38 participants' ERP recordings while reading 280 sentences in Mandarin Chi-

[2]https://huggingface.co/uer/gpt2-large-chinese-cluecorpussmall

[3]Different values have been tested, from 1 to 10 top-prediction. The best performance was reached by computing the candidate as the centroid of the 5 top predictions, as reported here

[4]As in Li and Futrell (2024), we used GPT-2 to find the most likely tokens to appear in the given context. However, differently from their work, we computed the cosine similarity only between embeddings of single words, instead of vectors representing whole sentences.

nese. Each word was automatically and individually presented on the screen. Each sentence contains about 12 to 14 words, as shown in (1).

(1) 在学校组织的郊游途中，小婷被石头砸伤的状况让人着急。'In a trip organized by the school, Xiaoting's getting hurt by a rock made everyone worried.'

All the sentences were grammatical and without semantic violations.

EEG data was re-referenced to the two mastoid electrodes, and the bad channels were interpolated. We then used a high-pass filter with a 0.1Hz cutoff frequency for data preprocessing. N400 and P600 were computed using a 300-500 ms and 600-1000 ms window respectively. Following Frank and Aumeistere (2024), we included only signals from Cz, C3, C4, CP1, CP2, Pz, P3, and P4.

The final data included N400 and P600 amplitudes for each word and each participant.

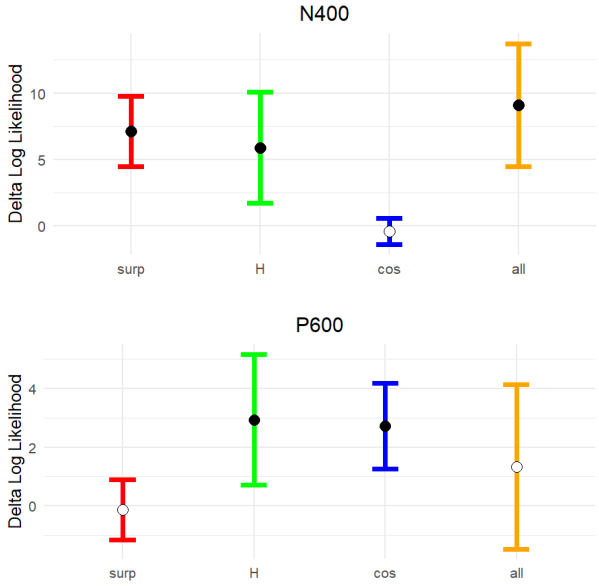

Figure 1: ΔLL values of models employing surprisal, entropy, semantic similarity, and all the features in predicting N400 and P600. Error bars indicate 95% confidence intervals. Full dots indicate a ΔLL statistically different from zero.

4 Results

As shown in Figure 1, surprisal benefits the model only in predicting N400, while for P600, the model registered a negative ΔLL. The usage of entropy, however, improved the model in predicting both the neurocognitive indexes. The probability-driven semantic similarity shows an opposite trend in comparison with surprisal: while it did not outperformed the baseline on the N400 prediction, it is as useful as entropy in predicting the P600 effect.

In both N400 and P600, the model employing all the features shows a positive ΔLL, meaning that the joint usage of surprisal, entropy, and semantic similarity helps predict the ERPs' amplitudes. However, while the all-features model had the highest predictive power for the N400 with a statistically significant ΔLL, it was outperformed by the *H* and *cos* models for the P600, and the difference between its log-likelihood and the log-likelihood of the baseline did not reach significance.

Focusing on the features within the *all* model, as shown in Figure 2, entropy accounts for the majority of the ERP signal variation in both N400 and P600, followed by surprisal for the first psychometric. For N400, both surprisal and entropy had a $p-value < 0.05$ within the *all* model, while in the prediction of P600, only entropy gave a statistically significant contribution. Although the semantic similarity is more impactful in predicting P600, it does not reach significance in such a general model.

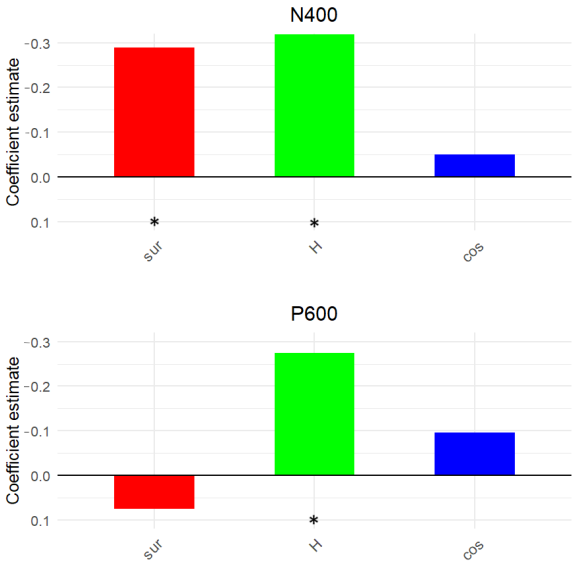

Figure 2: Coefficients of the regression factors within the *all* model, in the prediction of N400 and P600. Asterisks indicate that the factors showed a p-value < 0.05.

5 Discussion

Our results show the suitability of surprisal in predicting N400, but not of P600. This finding is in line with what was reported in Xu et al. (2024) and reveals how expectancy plays a major role in this

phase of language processing, in comparison with later stages.

Semantic similarity, on the contrary, was not influential in modeling the N400 effect, but it was found to significantly improve the prediction of P600 in comparison with the baseline. Given that our semantic similarity metric aims to represent the closeness between the meaning of the expected concept and the meaning of the processed word, this finding suggests that a *semantic evaluation* of the linguistic input based on expectations occurs only in later stages of processing. This finding supports Li and Futrell (2024) previous finding but from a word-level perspective. Although both surprisal and our semantic similarity are expectations-based, they represent different ways of handling unexpected items: surprisal only tell us to what extent the word being read was expected based on the context, regardless the predictions previously made, while the semantic similarity provides information on how closely the expectations align with the upcoming words. This suggests two distinguished, although interdependent phases of semantic processing.

The good performances of entropy as a single model (H), together with the fact that in the joint models for both N400 and P600 such a metric is the factor with the largest coefficient, suggest that the level of uncertainty plays a role through the largest span of language processing stages. This is not completely surprising: as Stone et al. (2022) pointed out, a high entropy indicates a high number of possible, equally preferred continuations of the sentence, increasing the processing cost. On the contrary, low entropy indicates a few, strong constraints, making the processing easier. Therefore, the level of uncertainty, meaning the strength of previous expectations, impacts the amount of cognitive resources required to decide to what extent the encountered word is expected (N400) and the processing cost of comparing the strong (or weak) predictions with the linguistic input (P600).

To summarize, from a general perspective of a model of language processing, our results suggest that:

1) Between 300 and 500 ms from the onset of the stimulus, local phenomena, based on the conditional probability of words within context, are processed;

2) At later stages, an evaluation of the difference in terms of semantics between expectations and linguistic data is performed;

3) Through the whole 300-1000 ms window, the level of uncertainty about the upcoming linguistic information modulates the cognitive effort required to elaborate the input.

6 Conclusions

In the present paper, for the first time in the context of Mandarin Chinese, we analyzed the predictive power of surprisal, entropy, and a probability-driven semantic similarity on ERPs. Our results confirmed previous findings and shed further light on the cognitive dynamics characterizing different stages of language processing. In particular, our novel approach showed how between 600 and 1000 ms from the stimulus onset a high-level evaluation of the input semantics is performed.

Limitations

Our work presents some limitations. First, given the nature of the materials we conducted our tests on (i.e., grammatical sentences without any kind of semantic violation or manipulation), our findings can account for general trends only; in future work, we will apply the same approach and metrics on experimental materials, to test if our conclusions stand in front of lexical manipulations, prime effects, etc.

Furthermore, our study is a monolingual investigation. In the future, we will repeat the same study on other languages to test the generalizability of our findings.

Acknowledgments

This study was supported by the Departmental General Research Fund (4-ZZV0) funded by the Department of Chinese and Bilingual Studies and the Research Large Equipment Fund (1-BC7N) at the Hong Kong Polytechnic University. We would like to thank the anonymous reviewers for their feedback and suggestions.

References

Christoph Aurnhammer, Francesca Delogu, Harm Brouwer, and Matthew W Crocker. 2023. The P600 as a continuous index of integration effort. *Psychophysiology*, 60(9):e14302.

Christoph Aurnhammer, Francesca Delogu, Miriam Schulz, Harm Brouwer, and Matthew W Crocker. 2021. Retrieval (N400) and integration (P600) in expectation-based comprehension. *Plos one*, 16(9):e0257430.

Harm Brouwer, Hartmut Fitz, and John Hoeks. 2012. Getting real about semantic illusions: rethinking the functional role of the P600 in language comprehension. *Brain research*, 1446:127–143.

Andrea Gregor de Varda, Marco Marelli, and Simona Amenta. 2024. Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data. *Behavior Research Methods*, 56(5):5190–5213.

Francesca Delogu, Harm Brouwer, and Matthew W Crocker. 2021. When components collide: Spatiotemporal overlap of the N400 and P600 in language comprehension. *Brain Research*, 1766:147514.

Stefan L Frank. 2017. Word embedding distance does not predict word reading time. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Stefan L Frank and Anna Aumeistere. 2024. An eye-tracking-with-EEG coregistration corpus of narrative sentences. *Language Resources and Evaluation*, 58(2):641–657.

Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. 2023. Information value: Measuring utterance predictability as distance from plausible alternatives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5633–5653, Singapore. Association for Computational Linguistics.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

Patrick Haller, Lena S Bolliger, and Lena A Jäger. 2024. Language models emulate certain cognitive profiles: An investigation of how predictability measures interact with individual differences. *arXiv preprint arXiv:2406.04988*.

Bernard A. J. Jap, Yu-Yin Hsu, Lavinia Salicchi, and Yu Xi Li. 2024. What's in a name? electrophysiological differences in processing proper nouns in Mandarin Chinese. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024*, pages 79–85, Torino, Italia. ELRA and ICCL.

Benedict Krieger, Harm Brouwer, Christoph Aurnhammer, and Matthew W Crocker. 2024. On the limits of LLM Surprisal as functional explanation of ERPs. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. 2023. Psychometric predictive power of large language models. *arXiv preprint arXiv:2311.07484*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Jiaxuan Li and Allyson Ettinger. 2023. Heuristic interpretation as rational inference: A computational model of the N400 and P600 in language processing. *Cognition*, 233:105359.

Jiaxuan Li and Richard Futrell. 2023. A decomposition of surprisal tracks the N400 and P600 brain potentials. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Jiaxuan Li and Richard Futrell. 2024. Decomposition of surprisal: Unified computational model of ERP components in language processing. *Preprint*, arXiv:2409.06803.

James A Michaelov, Megan D Bardolph, Cyma K Van Petten, Benjamin K Bergen, and Seana Coulson. 2024. Strong Prediction: Language model surprisal explains multiple N400 effects. *Neurobiology of language*, 5(1):107–135.

Lavinia Salicchi, Alessandro Lenci, Emmanuele Chersoni, et al. 2021. Looking for a role for word embeddings in eye-tracking features prediction: does semantic similarity help? In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*. Association for Computational Linguistics.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Kate Stone, Shravan Vasishth, and Titus von der Malsburg. 2022. Does entropy modulate the prediction of German long-distance verb particles? *Plos one*, 17(8):e0267813.

Marieke Van Herten, Dorothee J Chwilla, and Herman HJ Kolk. 2006. When heuristics clash with parsing routines: ERP evidence for conflict monitoring in sentence perception. *Journal of cognitive neuroscience*, 18(7):1181–1197.

Marieke Van Herten, Herman HJ Kolk, and Dorothee J Chwilla. 2005. An ERP study of P600 effects elicited by semantic anomalies. *Cognitive brain research*, 22(2):241–255.

Marten Van Schijndel and Tal Linzen. 2018. Can entropy explain successor surprisal effects in reading? *arXiv preprint arXiv:1810.11481*.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Haoyin Xu, Masaki Nakanishi, and Seana Coulson. 2024. Revisiting Joke Comprehension with Surprisal and Contextual Similarity: Implication from N400 and P600 Components. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

A Appendix

A.1 Linear mixed-effects models

In this section, we will provide more details regarding the linear mixed-effects models implemented and compared.
We used Julia to create and train/validate the models.

```
BL = fit(MixedModel, @formula(target ~
    1+baseline+word_len+word_freq+position+
    (1+word_len+word_freq+position|ptpID) + (1|word_ID)), train_data)

surp = fit(MixedModel, @formula(target ~
    1+baseline+word_len+word_freq+position+sur+
    (1+word_len+word_freq+position+sur|ptpID) + (1|word_ID)), train_data)

H = fit(MixedModel, @formula(target ~
    1+baseline+word_len+word_freq+position+H+
    (1+word_len+word_freq+position+H|ptpID) + (1|word_ID)), train_data)

cos = fit(MixedModel, @formula(target ~
    1+baseline+word_len+word_freq+position+semsim+
    (1+word_len+word_freq+position+semsim|ptpID) + (1|word_ID)), train_data)

all = fit(MixedModel, @formula(target ~
    1+baseline+word_len+word_freq+position+sur+H+semsim+
    (1+word_len+word_freq+position+sur+H+semsim|ptpID) + (1|word_ID)), train_data)
```

Where *target* is either N400 or P600.
As it is possible to notice from the snippets, following Frank and Aumeistere (2024), the baseline signal was a covariate of no interest. Word ID and participant ID were used as random intercepts.