Evaluating Computational Metrics for Predicting N400 Amplitude during Reading Comprehension

Anonymous ACL submission

Abstract

Given the interest recent research showed towards cognitive modeling of ERPs, we explored whether traditional word-level features such as position, word frequency, and number of strokes overlap with probability-based metrics such as surprisal, entropy, and entropy reduction. Analyzing and comparing different generalized linear models we found that the mathematical metrics do represent the same information as some of the "traditional" features overpowering them. A new cognitive-motivated computational feature is proposed.

1 Introduction

Advancements in eyetracking and EEG technologies have enabled the investigation of the psychological and cognitive dynamics of linguistic domains, such as reading (Just et al., 1982; Bizas et al., 1999) and listening (Friederici, 2002), as well as tasks like lexical decision and elicitation (Kuperman et al., 2013; Ganushchak et al., 2011). For example, syntactic anomalies often lead to increased fixations (and regressions) on the disambiguation area of a sentence (Meseguer et al., 2002). Similarly, semantic mismatches can induce a longer reading time for unexpected items (Rayner et al., 2004). Furthermore, brain activity modulates in response to specific words or sounds, reflecting as event-related potentials (ERPs) that arise in response to specific events at a time windows, such as 400 or 600 ms from the onset of the stimulus.

N400 is typically associated with cognitive overload during semantic integration, stemming from low semantic coherence between a word and its preceding context (Berkum et al., 1999) or the low frequency of the target term (Rugg, 1990). N400 modulation is thus dependent on both word-level features, such as word frequency, and the sentence-word relationship, namely the contextual likelihood of a word. This latter aspect has been modeled using metrics from the domain of information theory, such as surprisal, entropy, and entropy reduction. These metrics can be computed using probability distribution provided by language models (Hale, 2016). Recent studies have used these computational techniques to model reading times (Lowder et al., 2018; Salicchi et al., 2023) and ERP amplitudes (Michaelov et al., 2024; Hollenstein et al., 2023; Frank et al., 2013). However, the individual contributions of these features are still overlooked. Previous works have focused on only a single feature (Frank and Aumeistere, 2023) or, when using multiple metrics together in cognitive modeling, failed to provide psycholinguistic explanations for their results based on the interplay between the features (Van Schijndel and Linzen, 2021).

Thus, we aim to i) investigate the separate contribution of "traditional" features (i.e., word frequency, word complexity, and word position within a sentence), and probability-based metrics (i.e., surprisal, entropy, and entropy reduction), ii) examine overlapping functions between these two groups of features, and within the computational metrics, and iii) propose a new feature based on both probability and psycholinguistic observations. Furthermore, while most literature focuses on English or alphabetic languages, our experiments are focused on Chinese, using data from Jap et al. (2024)

2 Related Work

Van Schijndel and Linzen (2021) investigated the cognitive basis of reading behavior in garden-path sentences. They proposed a one-stage account of syntactic disambiguation, where the shift in the *probability distribution* of multiple, parallel parses explains the longer reading times in disambiguation regions of the garden-path sentences. They implemented and compared linear models to predict reading times using word frequency, word length, word position within the sentence, surprisal, entropy, and entropy reduction. Their results showed that probability-based computational features statistically significantly contribute to predicting the presence of syntactic ambiguity, but not its magnitude. However, they treated surprisal, entropy, and entropy reduction as equivalent metrics, limiting the discussion to a mere performance comparison. Other studies have successfully modeled reading times and ERP amplitudes, such as N400, using computational probability metrics, particularly surprisal. Frank and Aumeistere (2023) successfully modeled N400 amplitude recorded alongside with eyetracking data during naturalistic reading of Dutch sentences. They created linear mix-effects regression models using word frequency, word length, word position, and surprisal of the word given the previous context, finding a significant role of surprisal in predicting the amplitude of N400. This work inspired our study regarding the features, models, and metrics in computational modeling. However, we focused on a Sinitic language, Mandarin Chinese, extended Frank and Aumeistere's idea by employing entropy and entropy reduction, and attempted to provide a deeper explanation of each regression feature's contribution to computational prediction.

3 Method

3.1 Data

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

Given the different goals of the original study and the current one, customized event lists were created to compute ERPs for each word. The EEG data was re-referenced to the two mastoid electrodes, and the bad channels were interpolated. We then followed the typical ERP data filtering procedure by using a high-pass filter with a 0.1Hz cutoff frequency for data preprocessing. N400 was computed using the classical 300-500 ms window and, following Frank and Aumeistere (2023), included only signals from Cz, C3, C4, CP1, CP2, Pz, P3, and P4.

3.2 Model

We implemented and compared 127 generalized linear models, using N400 amplitude as the dependent variable and different combinations of 7 word-level features and computational metrics as independent variables (details below in 3.2.1).

3.2.1 Features

The first group of independent variables are word-level psycholinguistically motivated features: **Number of strokes** (n. strokes): Since all the words in our materials were two syllables, instead of using word length, we used the number of strokes of characters of each word to define the word complexity. We retrieved the number of strokes for each character from *hanziDB*[1]. **Word frequency** (word_freq): Computed using the Python library *wordfreq*.[2] **Position**: Computed as the number of words preceding the target one.

The second group of variables contains computational metrics extracted by using the Chinese version of BERT base (Devlin et al., 2018). Specifically, we fed each sentence to the model, substituting the target word with the special token [MASK]. We then passed the word of interest as the only candidate for the `targets` parameter to obtain its probability given the preceding context.

Surprisal represents the extent to which the reader expects a certain word given the previous context. It is computed as the negative logarithm of the probability of the word given the preceding tokens.

$$surprisal(w_n) = -log(P(w_n|w_0, w_1, ..., w_{n-1}))$$

Entropy: represents the general extent to which the reader expects a certain word. It is computed as the negative product of the probability distribution of the target word over the vocabulary and the logarithm of such a probability. In this case, no previous context was provided to BERT.

$$H(w) = -P(w) * logP(w)$$

Entropy reduction (ent. reduct.): represents the influence of context in modulating the expectations in encountering a certain word. It is computed as the difference between the target word's general entropy and the word's entropy given the previous context.

$$Ent.Reduct. = H(w) - H(w|w_0, ...w_{n-1})$$

[1]http://hanzidb.org/
[2]https://github.com/rspeer/wordfreq/

Cosine: represents the similarity between the expected word, given the context, and the word being read. We used the BERT masking mechanism to select the word most likely to appear in the target word's position, compute the vector representation of both the target word and the candidate[3], and then computed the cosine similarity between the two embeddings using the *cosine_similarity* function of *sklearn*[4].

4 Results and Discussion

Single-feature models. Firstly, we examine the main effects of each feature on predicting N400 amplitude. As shown in Table 1, the number of strokes, position, cosine, and entropy reduction are significant predictors of N400 amplitude. The model's intercept alone shows significance for surprise, word frequency, and entropy. At this stage, both traditional word-level features (number of strokes, word frequency, and position) and computational metrics seem good predictors for the target value.

We then compared the single feature models using the corrected Akaike information criterion (AICc). The model employing position only was the one with the lowest score, followed - with a substantial deviation of 103 - by the model relying on cosine, and entropy reduction (Table 1). These results suggest that position may have a prime role in modulating the N400 response, a finding consistent with psycholinguistic studies where word position within a sentence is a significant factor in reading processing.

Model	Interc.	Feat.	AICc	Δ
position	<2e-16	<0.01	6818.91	0.00
cosine	0.52	0.01	6922.09	103.18
ent.red.	<2e-16	0.03	6923.42	104.51
n.strokes	<2e-16	0.03	6923.54	104.63
word freq.	<2e-16	0.19	6926.51	107.60
surprisal	<2e-16	0.59	6927.92	109.01
entropy	<2e-16	0.67	6928.03	109.13

Table 1: Performance of the single models. P-values for the intercept of the models, p-values for the target features, AICc and differences in AICc values are reported.

Full model. Our second analysis included all 7 features in a full model. As shown in Table 2, only position, surprisal, and entropy were significant in predicting the ERP amplitude. If the significance of position is not surprising, given the single-model performance, the relevance of surprisal and entropy was not as obvious. These findings led us to speculate that surprisal and entropy not only do not (completely) overlap in the information they represent but also provide unique information beyond what is captured by the number of strokes and word frequency. Moreover, the close relationship between number of strokes and word frequency, and therefore their tendency to be both overridden by surprisal in the full model, is explainable by Zipf's law, stating that simpler words (in our case, characters composed of a lower number of strokes) are used more frequently.

Feature	p-value	Estimate	Std. Err.
(Intercept)	<0.01	-20.430	0.615
word_freq	0.539	-0.009	0.015
n. strokes	0.454	0.006	0.008
position	<2e-16	0.282	0.025
Surprisal	0.002	-0.042	0.014
cosine	0.571	-0.321	0.567
entropy	0.024	-22.960	10.214
ent_reduct	0.148	-0.749	0.518

Table 2: P-value, estimate, and standard error of the features within the full model.

Interaction between traditional & computational metrics. To test the relationship between word frequency, number of strokes, and computational metrics, we created two sets of models with interacting features (number of strokes & each computational metric or word frequency & computational metrics). The number of strokes (Table 5 in Appendix) showed a significant interaction only with cosine and entropy reduction. The significance of number of strokes in the single-feature model and its lack of significance in the full model and interacting models suggests that the traditional feature is overpowered by surprisal and entropy. Similarly, word frequency interacts significantly with entropy and entropy reduction, indicating information sharing with surprisal and cosine similarity (Table 6 in Appendix). This suggests that the number of strokes is partially overridden by surprise and entropy, while word frequency is mostly influenced by surprise.

Interactions between computational metrics. We examined whether the computational metrics overlap with each other. Surprisal (Table 7 in Ap-

[3]The last layer of BERT was used to obtain the vectors.
[4]https://scikit-learn.org/

pendix) successfully interacts with cosine similarity only, thus having no fruitful interaction with the other two probability-based metrics, entropy and entropy reduction. These findings, together with the analysis of the full model (Table 2), where both surprisal and entropy were found to be significant predictors, suggest that both metrics are valuable in their main effect, but their similar calculation methods might limit their joint contribution in prediction models. Similarly, entropy reduction did not show significant interaction with other computational metrics. Theoretically, entropy reduction expresses how the context influences the expectations about a word's occurrence, and it thus should bring different information than surprisal or entropy. However, for the way it is mathematically expressed it may be seen as a hybrid, as a bridge between surprisal and entropy, since it includes both the probability of the word over the vocabulary (as entropy) and the probability of the item given the context (surprisal).

Best model(s). In the fifth step of our investigation, we considered all the possible models without feature interaction, from simple one-feature ones to the full one employing all 7 features. Relying on AICc, we explored which models best predicted N400 amplitude. Focusing on the top 5 models with the lowest AICc (Table 3), it is clear that surprisal, entropy, and position are constantly present in all the best models, followed by entropy reduction (3/5), cosine (2/5), and number of strokes (1/5). Overall, although with a very limited difference in terms of AICc, the best model was the one employing surprisal, position, entropy, and entropy reduction. These findings confirm some elements we noticed with the previous analyses: i) position is the only traditional feature that is not overridden by mathematical metrics, ii) therefore, probability-based metrics seem not only to overlap but to give more information than word-level features, iii) surprisal and entropy are independent in the information they bring.

	AICc	Delta_AICc
S+P+E+ER	6775.82	0.00
S+P+E	6775.89	0.06
S+P+C+E+ER	6777.52	1.70
S+NS+P+E+ER	6777.53	1.71
S+P+C+E	6777.55	1.72

Table 3: AICc of all models. Top 5 reported. S = surprisal, P = position, E = entropy, ER = entropy reduction, C = cosine, NS = number of strokes.

Checking the significance of the features within the best model we notice that entropy reduction is not statistically significant (Table 4). The best model outperforms the one having surprisal, position, and entropy by only 0.06 AICc points, confirming our speculation that the contribution of entropy reduction partially overlaps with the information brought by entropy and surprisal.

Feature	p-value	Estimate	Std. Err.
(Intercept)	<2e-16	-2.366	0.121
Surprisal	<0.01	0.290	0.023
position	<2e-16	-0.042	0.012
entropy	0.002	-26.647	8.382
ent. reduct.	0.151	-0.745	0.517

Table 4: P-value, estimate, and standard error of the features within the best model.

Cosine similarity. We finally focused on the performance and contribution of the proposed feature. From what we already noticed, cosine appears in two of the top 5 best performing models, revealing its potential. We then analyzed its interaction patterns (Table 10 in Appendix): cosine successfully interacts with number of strokes, position, surprisal, and entropy, while its joint contribution with word frequency and entropy reduction does not seem beneficial. These observations revealed how the new approach is both theoretically and technically valid: it takes into account the previous context, the semantics of both the expected word and the input one, and the difference between the two. This is achieved without mathematically explicitly relying on likelihood, making "cosine" suitable to be used in bigger models together with other metrics. Moreover, as shown in Table 1, the cosine similarity model outperforms the other computational metrics, proving how the proposed approach successfully models the cognitive dynamics beneath the elicitation of an N400 response.

5 Conclusion

The results of our analyses showed how traditional features such as number of strokes and word frequency overlap with - and are overpowered by - surprisal and entropy, and surprisal and cosine respectively in predicting N400 amplitude in Chinese sentences.

6 Limitations

To ensure a proper multilingual comparison between our findings and the ones presented in our study of reference, i.e., Frank and Aumeistere (2023), our next step will be i) the employment of a linear mix-effects regression model, instead of a generalized linear model, ii) repeat our analysis using English, Dutch, Mandarin Chinese, and Indonesian. Also, it would be interesting to investigate how the features we focused on in this paper interact in the prediction of a different ERP, namely P600, which is typically related to syntactic processing, instead of semantic one as N400, or with other neurocognitive data, like eye-tracking or fMRI.

References

Jos JA van Berkum, Peter Hagoort, and Colin M Brown. 1999. Semantic integration in sentences and discourse: Evidence from the n400. *Journal of cognitive neuroscience*, 11(6):657–671.

E Bizas, PG Simos, CJ Stam, S Arvanitis, D Terzakis, and S Micheloyannis. 1999. Eeg correlates of cerebral engagement in reading tasks. *Brain Topography*, 12:99–105.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Stefan L Frank and Anna Aumeistere. 2023. An eye-tracking-with-eeg coregistration corpus of narrative sentences. *Language Resources and Evaluation*, pages 1–17.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2013. Word surprisal predicts n400 amplitude during reading.

Angela D Friederici. 2002. Towards a neural basis of auditory sentence processing. *Trends in cognitive sciences*, 6(2):78–84.

Lesya Y Ganushchak, Ingrid K Christoffels, and Niels O Schiller. 2011. The use of electroencephalography in language production research: a review. *Frontiers in psychology*, 2:208.

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

Nora Hollenstein, Marius Tröndle, Martyna Plomecka, Samuel Kiegeland, Yilmazcan Özyurt, Lena A Jäger, and Nicolas Langer. 2023. The zuco benchmark on cross-subject reading task classification with eeg and eye-tracking data. *Frontiers in Psychology*, 13:1028824.

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.

Marcel A Just, Patricia A Carpenter, and Jacqueline D Woolley. 1982. Paradigms and processes in reading comprehension. *Journal of experimental psychology: General*, 111(2):228.

Victor Kuperman, Denis Drieghe, Emmanuel Keuleers, and Marc Brysbaert. 2013. How strongly do word reading times and lexical decision times correlate? combining data from eye movement corpora and megastudies. *Quarterly Journal of Experimental Psychology*, 66(3):563–580.

Matthew W Lowder, Wonil Choi, Fernanda Ferreira, and John M Henderson. 2018. Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive science*, 42:1166–1183.

Enrique Meseguer, Manuel Carreiras, and Charles Clifton. 2002. Overt reanalysis strategies and eye movements during the reading of mild garden path sentences. *Memory & cognition*, 30(4):551–561.

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*, pages 1–29.

Keith Rayner, Tessa Warren, Barbara J Juhasz, and Simon P Liversedge. 2004. The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(6):1290.

Michael D Rugg. 1990. Event-related brain potentials dissociate repetition effects of high-and low-frequency words. *Memory & cognition*, 18(4):367–379.

Lavinia Salicchi, Emmanuele Chersoni, and Alessandro Lenci. 2023. A study on surprisal and semantic relatedness for eye-tracking data prediction. *Frontiers in Psychology*, 14:1112365.

Marten Van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive science*, 45(6):e12988.

A Appendix

Features	p-value
n_strokes:word_freq	0.19102
n_strokes:position	8.57e-13 ***
n_strokes:surprisal	0.856
n_strokes:cosine	0.0249 *
n_strokes:entropy	0.630
n_strokes:ent_reduct	0.0898 .

Table 5: Number of strokes interacting with other features in n_strokes*feat models.

Features	p-value
word_freq:n_stokes	0.19102
word_freq:surprisal	0.820
word_freq:position	4.96e-16 ***
word_freq:cosine	0.469
word_freq:entropy	3.39e-05 ***
word_freq:ent_reduct	0.00601 **

Table 6: Word frequency interacting with other features in word_freq*feat models.

Features	p-value
surprisal:n_strokes	0.856
surprisal:word_freq	0.820
surprisal:position	1.48e-05 ***
surprisal:cosine	1.46e-07 ***
surprisal:entropy	0.298
surprisal:ent_reduct	0.16512

Table 7: Surprisal interacting with other features in surprisal*feat models.

Features	p-value
entropy:n_strokes	0.630
entropy:word_freq	3.39e-05 ***
entropy:position	1.65e-08 ***
entropy:surprisal	0.298
entropy:cosine	0.004107 **
entropy:ent_reduct	0.5377

Table 8: Entropy interacting with other features in entropy*feat models.

Features	p-value
ent_reduct:n_strokes	0.0898 .
ent_reduct:word_freq	0.00601 **
ent_reduct:position	0.927
ent_reduct:surprisal	0.16512
ent_reduct:cosine	0.101320
ent_reduct:entropy	0.5377

Table 9: Entropy reduction interacting with other features in ent_red*feat models.

Feature	p-value
cosine:n_stokes	0.0249 *
cosine:word_freq	0.469
cosine:position	1.46e-06 ***
cosine:surprisal	1.46e-07 ***
cosine:entropy	0.004107 **
cosine:ent_reduct	0.101320

Table 10: Cosine similarity interacting with other features in cosine*feat models.