Dissociable Neural Signatures for Surprisal and Entropy Reduction in Mandarin Speech Comprehension

Qifei Wang (qifei.wang@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, the Netherlands

Eva Berlot (eva.berlot@gmail.com)
Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, the Netherlands

Judit Fazekas (judit.fazekas@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, the Netherlands

Jakub Szewczyk (jakub.szewczyk@gmail.com)
Institute of Psychology, Jagiellonian University, Kraków, Poland

Floris de Lange (floris.delange@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, the Netherlands

Abstract

The neural response to words is modulated by the amount of information conveyed by the words. For example, it is well established that neural activity monotonically increases as a function of the surprisal of a word. Another important information-theoretic measure, entropy reduction (ER), quantifies how each incoming word constrains subsequent content interpretations. Here we examined whether and when surprisal and ER modulated neural activity during naturalistic speech comprehension. Through magnetoencephalographic recordings (MEG) of naturalistic Mandarin comprehension, we demonstrate that ER accounts for neural responses that cannot be explained by surprisal alone. Initial observations show that ER elicits a later (400-600 ms) cortical response compared to the N400 surprisal effect. These preliminary results suggest that ER may involve neural computations distinct from those underlying surprisal, with early resolution of surprisal followed by later ER. Our findings extend predictive processing frameworks to tonal languages and highlight entropy reduction as a key component of neural language models, operating beyond surprisal.

Keywords: Language comprehension; Large language model; Surprisal ; Entropy reduction

Introduction

Language comprehension requires the brain to continuously update its expectations as new words arrive. While surprisal (the extent to which a word came unexpected to reader or listener) robustly modulates neural responses (Kutas & Hillyard, 1980), the role of entropy reduction (the change in uncertainty about future words, ER) is less clear (Frank et al., 2015). These measures are mathematically related but computationally distinct: surprisal depends on a word's contextual probability, while entropy reduction quantifies how much that word constrains subsequent predictions (Hale, 2003, 2006).

Behavioral studies consistently show that ER affects reading times independently of surprisal (Frank, 2013; Lowder et al., 2018), yet neurophysiological evidence for this dissociability is mixed: neuroimaging results show ER's unique contribution to a widespread brain activation (Song et al., 2024), and electrophysiological studies in reading tasks report either grammar-based P600-like central effect (Hale et al., 2018) or null results (Frank et al., 2015). These discrepancies may reflect methodological differences in how ER is quantified. Grammar-based approaches capture structural uncertainty (Hale, 2006), whereas word-based methods reflect meaning-level expectations based on only small text samples (Frank, 2013).

Our study aims to resolve these challenges by examining naturalistic Mandarin comprehension with MEG. We quantified surprisal and ER using contextually-sensitive measures derived from Chinese GPT-2 while employing temporal response functions (TRFs) (Di Liberto et al., 2015) to isolate word-level neural responses. This approach allowed us to determine whether: (1) the well-established surprisal effects generalize to Mandarin's linguistic structure, (2) ER generates responses independently from surprisal and elicits distinct spatiotemporal dynamics.

Methods

Procedure. Thirty-four native Mandarin speakers with normal hearing were recruited for the study. Participants listened to a 50-minute Chinese audiobook while their brain activity was recorded using MEG.

Analysis. We employed a hierarchical approach to model how neural responses encode different linguistic features. Our baseline model incorporated control regressors for acoustic features (broadband amplitude envelope, spectrogram, acoustic edge, pitch) and lexical properties (phoneme, character and word onsets, lexical frequency). For each character in the narrative, we computed two key metrics using Chinese GPT-2: surprisal, calculated as -log p(character|context) to quantify prediction error, and entropy reduction (ER), computed over three-character lookahead sequences (Frank, 2013) (see Figure 1A). Surprisal and entropy reduction showed minimal correlation (r = -0.06), enabling independent examination of their neural contributions.

We employed time-resolved TRF analysis to model linguistic features as impulse functions aligned to character onsets. The performance of the TRF models with different feature combinations was evaluated through cross-validation, quantified by the Pearson correlation

between modeled and recorded MEG response (Figure 1B).

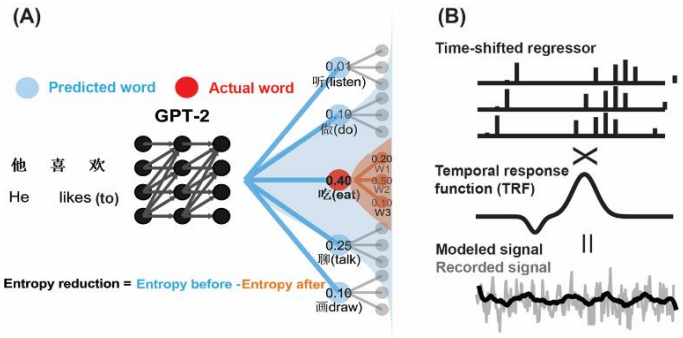

Figure 1. Schematic illustrations for entropy reduction and TRF. (A) Entropy reduction calculation using GPT-2. The model generates predicted characters (blue circles with probabilities) given preceding context. These predictions update the context to forecast subsequent characters (gray circles). The target character (red circle) is then observed, and the process repeats for subsequent characters (orange circles). Sequence probabilities are computed by multiplying constituent character probabilities (illustrated for 2-character sequences). Entropy reduction equals the difference between the entropy of future character sequences before (blue distribution) and after (orange distribution) observing the target character. **(B)** Linguistic features (one as example) were modeled as time-shifted impulses (top) to predict neural responses (bottom) via linear regression. Resulting TRF weights (middle) reflect the time course of neural response to each feature.

Results

Dissociable neural signatures for surprisal and entropy reduction. The baseline model explained significant variance in neural responses ($p<0.001$ comparing to zero), establishing a robust foundation for testing predictive mechanisms. Adding lexical surprisal alone significantly improved model fit ($p<0.001$. Cohen's $d = 2.04$), with TRF weights peaking between 200-400 ms post-character (Figure 2B). This deflection, maximal over temporal lobe sensors, corresponds to the canonical N400 prediction error effect observed across languages (Kutas & Hillyard, 1980; Wilcox et al., 2023).

The inclusion of ER on top of the surprisal further enhanced the model fit to the neural response ($p = 0.001$, $d = 0.63$, Figure 2A). Initial observations revealed that although ER-related activity shared a large overlap with that of the surprisal, it emerged later (400-600 ms) compared to the earlier N400 surprisal effect (Figure 2B-C). Note that although these temporal differences align with theoretical accounts of distinct predictive mechanisms, they require formal statistical verification.

Our study temporally dissociates the neural correlates of surprisal and entropy reduction, two important information-theoretic measures in language processing. The two temporal signatures suggest that the brain engages in sequential operations: first resolving immediate prediction conflicts, then updating future expectations.

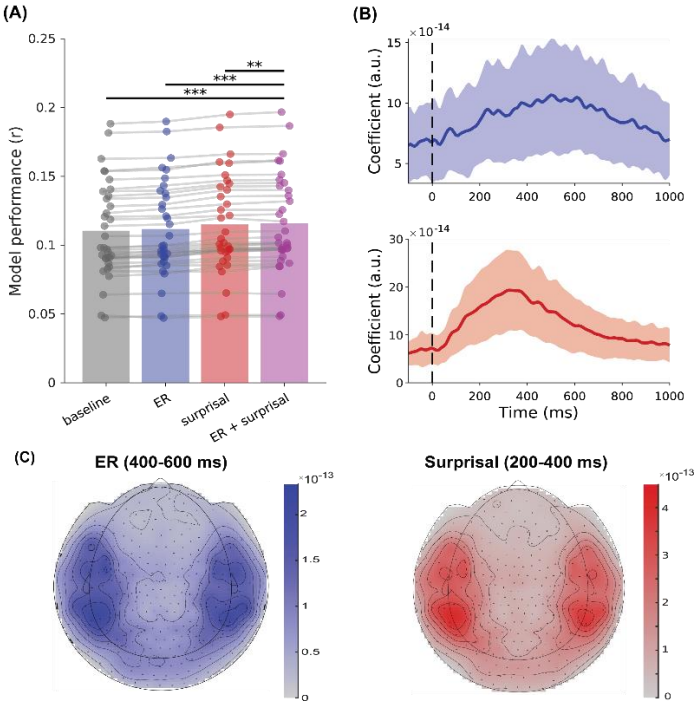

Figure 2. Distinct neural responses to surprisal and entropy reduction (A) Model performance of different models, model with ER and surprisal has the highest model performance, and adding ER, surprisal on top of each other also significantly improved model performance. (B) TRFs show earlier peak for surprisal (200-400 ms, red) vs later ER responses (400-600 ms, blue). (C) Topography of ER and surprisal.

References

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Current Biology*, *25*(19), 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

Frank, S. L. (2013). Uncertainty Reduction as a Measure of Cognitive Load in Sentence Comprehension. *Topics in Cognitive Science*, *5*(3), 475–494. https://doi.org/10.1111/tops.12025

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. https://doi.org/10.1016/j.bandl.2014.10.006

Hale, J. (2003). The Information Conveyed by Words in Sentences. *Journal of Psycholinguistic Research*, *32*(2), 101–123. https://doi.org/10.1023/A:1022492123056

Hale, J. (2006). Uncertainty About the Rest of the Sentence. *Cognitive Science*, *30*(4), 643–672. https://doi.org/10.1207/s15516709cog0000_64

Hale, J., Dyer, C., Kuncoro, A., & Brennan, J. (2018). Finding syntax in human encephalography with beam search. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2727–2736). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1254

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205. https://doi.org/10.1126/science.7350657

Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical Predictability During Natural Reading: Effects of Surprisal and Entropy Reduction. *Cognitive Science*, *42*(S4), 1166–1183. https://doi.org/10.1111/cogs.12597

Song, M., Wang, J., & Cai, Q. (2024). The unique contribution of uncertainty reduction during naturalistic language comprehension. *Cortex*, *181*, 12–25. https://doi.org/10.1016/j.cortex.2024.09.007

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470. https://doi.org/10.1162/tacl_a_00612