Can LLMs Inform Us About Predictive Processing During Natural Listening in MEG?

Sahel Azizpour[1,2], **Britta U. Westner**[2,3], **Jakub Szewczyk**[4], **Umut Güçlü**[1,2], **Linda Geerligs**[1,2]
[1]Donders Institute for Brain, Cognition, and Behaviour
[2]Radboud University
[3]Radboud University Medical Center
[4]Jagiellonian University

Abstract

The brain uses contextual information and prior knowledge to anticipate upcoming content during language comprehension. Recent research has shown that predictive signals can be revealed in pre-onset electrocorticography (ECoG) activity during naturalistic narrative listening, by building encoding models based on word embeddings from large language models (LLMs). Similarly, evidence for long-range predictive encoding has been observed in functional magnetic resonance imaging (fMRI) data, where incorporating embeddings for multiple upcoming words in a narrative improves alignment with brain activity. This study examines whether similar predictive information can be detected in MEG, a technique with higher temporal resolution than fMRI but a lower signal-to-noise ratio than ECoG. Our findings indicate that MEG captures pre-onset representations up to 1 second before word onset, consistent with ECoG results. However, unlike fMRI findings, incorporating future word embeddings did not enhance encoding in MEG, not even for one word into the future, which suggests that the pre-onset encoding may not reflect predictive processing. This work demonstrates that MEG combined with LLMs is a valuable approach for studying language processing in naturalistic narratives and highlights the need to study further what constitutes evidence for prediction during natural listening.

Keywords: MEG; Language comprehension; Large Language Models; Predictive processing

Introduction

Predictive processing theories propose that the brain continuously anticipates upcoming input based on context and prior experience [Clark (2013); De Lange et al. (2018)]. Language processing is one domain where predictions are thought to play a crucial role, particularly in facilitating comprehension by pre-activating linguistic representations [Federmeier (2007); Kutas et al. (2011)].

Despite strong evidence that prediction shapes language processing, current methods fall short of capturing predicted representations themselves. For example, the N400, a well-known ERP marker of prediction error [Kutas & Federmeier (2011); Terporten et al. (2019)], reflects post-onset processing and does not directly reveal the contents of predictions. RSA-based EEG work provides more direct evidence by decoding features like animacy before word onset [Wang et al. (2020)]. However, many studies on prediction use artificial pacing or tightly controlled stimuli, reducing the richness of linguistic context and limiting ecological validity [Willems et al. (2020)].

Large-scale continuous speech datasets in combination with large language models (LLMs) now enable the study of predictive processing in naturalistic settings. LLMs provide context-sensitive embeddings that align with brain activity across different imaging modalities [Baroni (2022); Caucheteux & King (2022); Goldstein et al. (2022)].

In this study, we use LLMs to investigate predictive processing in MEG, drawing on two key prior works. [Goldstein et al. (2022)] found that ECoG signals encode upcoming words up to two seconds before onset during natural speech. We test (1) whether similar pre-onset encoding can be detected in MEG, which is non-invasive but has lower signal-to-noise. We also build on Caucheteux et al. (2023), which showed that adding future word embeddings improved LLM-to-fMRI mapping, suggesting long-range prediction. However, fMRI's low temporal resolution obscures word-level timing. Here, we leverage MEG's higher temporal resolution to ask: 2) Does incorporating future embeddings improve MEG–LLM alignment, as seen in fMRI?

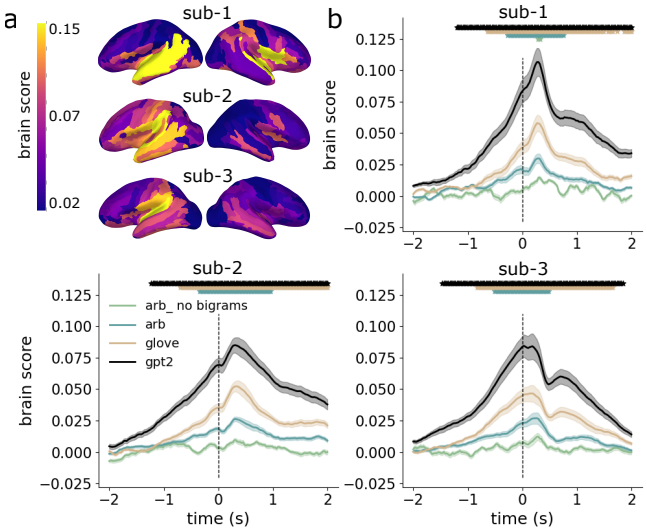

Figure 1: **Encoding of words relative to word onset. a.** Brain scores are significant across most regions across subjects, peaking in left temporal and inferior frontal areas linked to language processing. **b.** GPT-2 embeddings yield the highest brain scores, followed by GloVe, with arbitrary embeddings lowest. Pre-onset encoding persists with Glove and arbitrary embeddings but vanishes when repeated bigrams are removed, reducing both pre- and post-onset effects. Error bars denote standard error across MEG sources; stars indicate FDR-significant values.

Methods

Neural data and word embeddings

Here we used an openly available dataset containing MEG recordings collected while three native English-speaking participants (1 female; aged 35, 30, and 28 years) passively listened to 10 stories from the Adventures of Sherlock Holmes [Armeni et al. (2022)]. We extracted contextual word embeddings using the pre-trained GPT2-small model (12 layers), processing tokenized story chunks in 50-token windows and retrieving the final layer's hidden state for the last token. For non-contextual embeddings, we used 300-dimensional GloVe vectors. Embeddings were reduced to 50 dimensions via PCA to minimize computational load.

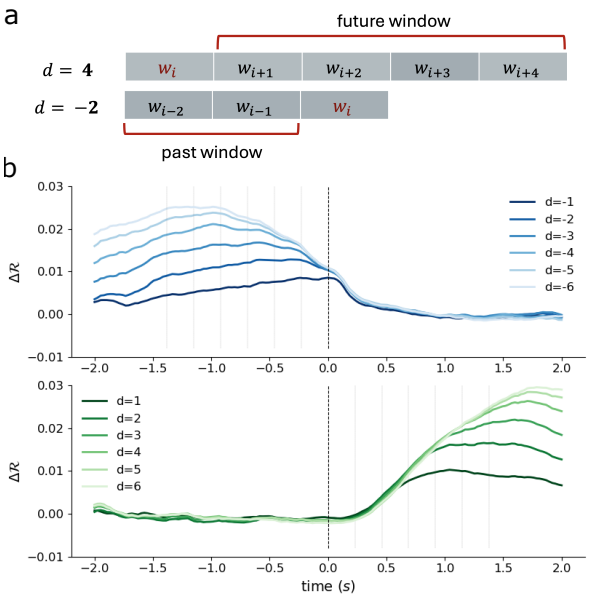

Figure 2: **Encoding of the future and past words. a.** The embedding vector is constructed by concatenating d future word embeddings ($d > 0$) or $|d|$ past word embeddings ($d < 0$) along with the embedding of the current word w_i. **b.** Encoding enhancement, $\Delta\mathcal{R}$, is shown for negative (upper) and positive (lower) values of d. Vertical gray lines mark the median inter-word interval values. Adding each successive future word embedding improves encoding only after that word is heard in the narrative, while including previous words consistently improves encoding beyond their offset.

Encoding model and brain score

The encoding model was built using linear ridge-regression with word embeddings as predictors and brain response as targets. A separate model was fit for each time point within the [-2 to +2 sec] window. We assessed model performance using 5-fold cross-validation, training on 80% of the story and testing on the remaining 20%. To evaluate model fit, we computed the Pearson correlation between the predicted and actual brain signals at each time point t around word onset across all words to compute the brain score $\mathcal{R}$. To compute the brain-wide brain score, we selected 30 MEG sources per participant. To prevent double dipping, we averaged scores from two participants and used the top 30 sources to evaluate the third.

Results

LLM embeddings align with MEG responses

Most MEG regions showed significant encoding with strongest effects in left temporal and inferior frontal areas (Fig. 1.b) associated with language processing. Brain scores peaked within 500 ms post-onset consistently across all 10 hours of the data.

To test whether pre-onset encoding reflects contextual prediction or arises from other properties of word embeddings, we conducted control analyses using static GloVe embeddings, which lack contextual information and arbitrary embeddings, which lack lexico-semantic information. Pre-onset encoding persisted with GloVe but was attenuated, suggesting that contextual information in gpt2 embeddings cannot explain all pre-onset encoding. Using arbitrary embeddings eliminated semantic and statistical dependencies while preserving word identity, yet significant pre-onset encoding persisted. To test whether this residual signal reflected learning of word co-occurrence patterns, we removed all repeated bigrams from the narrative. This abolished pre-onset encoding in the arbitrary condition. These controls indicate that statistical regularities, such as word co-occurrences, contribute to pre-onset encoding, cautioning against a straightforward interpretation of these signals as evidence of predictive processing.

Evidence for postdiction and not prediction

Extending Caucheteux et al. (2023), we tested whether concatenating future word embeddings improves encoding. Unlike their fMRI results, MEG showed no improvement at word onset (Fig. 2.b); increases in brain score emerged only 300 ms later—after the current word onset. Given the average inter-word interval of 230 ms, this increase is likely due to the fact that the next word has already been heard by this point in the narrative. In contrast, including previous words improved encoding at word onset, strongest for the immediately prior word. This temporal asymmetry held for both GPT-2 and GloVe, and persisted across control analyses. These results further support the idea that past words remain active while future words are not pre-encoded in MEG signals.

Discussion

This study shows that word representations can be robustly encoded in MEG during naturalistic listening, even with just one hour of data. We observed pre-onset representations similar to those in ECoG, though we cannot rule out that these pre-onset signals are due to correlations between nearby embeddings and by word co-occurrences. Unlike fMRI findings, adding future word embeddings did not enhance encoding, even for a single word in the future, however we found robust evidence for postdiction. This suggests that many words are not immediately integrated into context, and that processing may not be strictly time-locked to word onset [Gwilliams et al. (2018); Hogendoorn (2022); Szewczyk et al. (2022)]. These results align with findings from Toneva et al. (2022), who found that MEG predominantly captured lexical (single-word) information of current and previous word. In contrast, fMRI signals showed robust encoding of supra-word representations (information arising from combinations of words independent of lexical information). Together these findings may suggest that predictive information may be encoded in a form not readily accessible to MEG—perhaps due to its supra-word, abstract, or spatially distributed nature, whereas postdictive information is likely re-accessed during the processing of the current word, making it more readily detectable with MEG.

Acknowledgments

Linda Geerligs was supported by a VIDI grant of the Netherlands Organization for Scientific Research (grant number VI.Vidi.201.150).

References

Armeni, K., Güçlü, U., van Gerven, M., & Schoffelen, J.-M. (2022). A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. *Scientific Data*, *9*(1), 278.

Baroni, M. (2022). On the proper role of linguistically oriented deep net analysis in linguistic theorising. In *Algebraic structures in natural language.* CRC Press.

Caucheteux, C., Gramfort, A., & King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, *7*(3), 430–441.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications biology*, *5*(1), 134.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181–204.

De Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in cognitive sciences*, *22*(9), 764–779.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... others (2022). Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, *25*(3), 369–380.

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In spoken word recognition, the future predicts the past. *Journal of Neuroscience*, *38*(35), 7585–7599.

Hogendoorn, H. (2022). Perception in real-time: predicting the present, reconstructing the past. *Trends in Cognitive Sciences*, *26*(2), 128–141.

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. *Predictions in the brain: Using our past to generate a future*, *190207*(10.1093).

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, *62*(1), 621–647.

Szewczyk, J. M., Mech, E. N., & Federmeier, K. D. (2022). The power of "good": Can adjectives rapidly decrease as well as increase the availability of the upcoming noun? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(6), 856.

Terporten, R., Schoffelen, J.-M., Dai, B., Hagoort, P., & Kösem, A. (2019). The relation between alpha/beta oscillations and the encoding of sentence induced contextual information. *Scientific Reports*, *9*(1), 20255.

Toneva, M., Mitchell, T. M., & Wehbe, L. (2022). Combining computational controls with natural text reveals aspects of meaning composition. *Nature computational science*, *2*(11), 745–757.

Wang, L., Wlotko, E., Alexander, E., Schoot, L., Kim, M., Warnke, L., & Kuperberg, G. R. (2020). Neural evidence for the prediction of animacy features during language comprehension: evidence from meg and eeg representational similarity analysis. *Journal of Neuroscience*, *40*(16), 3278–3291.

Willems, R. M., Nastase, S. A., & Milivojevic, B. (2020). Narratives for neuroscience. *Trends in neurosciences*, *43*(5), 271–273.