Decoding of Hierarchical Inference in the Human Brain during Speech Processing with Large Language Models

Joséphine Raugel (josephine.raugel@ens.fr)
Laboratoire de Neurosciences Cognitives et Computationnelles,
Institut National de la Santé et de la Recherche Médicale,
Département d'Études Cognitives, École Normale Supérieure, Université PSL,
Paris 75005, France

Valentin Wyart* (valentin.wyart@ens.fr)
Laboratoire de Neurosciences Cognitives et Computationnelles,
Institut National de la Santé et de la Recherche Médicale,
Département d'Études Cognitives, École Normale Supérieure, Université PSL,
Paris 75005, France

Jean-Rémi King* (jeanremi.king@gmail.com)
Laboratoire des Systèmes Perceptifs, Centre National de la Recherche Scientifique,
Département d'Études Cognitives, École Normale Supérieure, Université PSL,
Paris 75005, France

*shared senior authorship

Abstract

Many theories of language in the brain rely on the notion of predictions. Yet, little is known about how linguistic predictions effectively change the representations of language in the brain. Here, we investigate how two levels of representations in the language hierarchy vary with predictability: words and phonemes. For this, we rely on Large Language Models (LLMs) trained to predict incoming words and phonemes, and estimate the posterior probability of these features as speech unfolds. We then evaluate whether predictability impacts the representations of words and phonemes decoded from the MEG responses of 27 participants listening to two hours of stories. Our results show that both words and phonemes are best decoded from the brain if they are unexpected from a given context. This finding constrains the computational architecture underlying natural speech comprehension.

Keywords: inference; magnetoencephalography; natural language processing, large language model.

Introduction

Language is central to human cognition. It structures social interactions, and is the primary vehicle to accumulate knowledge within and across individuals. Yet, the precise biological and computational bases of language remain unknown. Specifically, to what extent does the human brain continuously predict future phonemes, words and concepts? How do such predictions shape neural representations? Previous results suggest the existence of an inferential process in the human brain during language processing, for words and phonemes (Caucheteux et al., 2023; Donhauser & Baillet, 2020; Forseth et al., 2020; Garrido et al., 2009; Goldstein et al., 2022; Heilbron et al., 2019, 2022; Lopopolo et al., 2017; Millet et al., 2023; Mousavi et al., 2020; Shain et al., 2020; Wacongne et al., 2011; Willems et al., 2016). However, the computational bases underlying this inferential process are not yet fully understood, specifically the interaction between the inferential computations taking place at these two levels of the language hierarchy. Large Language Models (LLMs), which are typically optimized to predict the next token based on an embedded context, offer a powerful framework to study how the human brain may implement hierarchical inferences about language.

To explore the inferential framework at play during language processing in the human brain, we rely on LLMs outputs with the same linguistic stimuli. We relate the language representations decoded from the human MEG data to the specific conditional expectations of speech stimuli computed by the LLM.

Materials and Methods

Neural recordings. We analyze magnetoencephalographic (MEG) recordings of 27 healthy participants listening to short stories (Gwilliams, Flick, et al., 2022). Participants are recorded with a 208 axial-gradiometer MEG scanner built by the Kanazawa Institute of Technology, and sampled at 1,000 Hz, and online band-pass filtered between 0.01 and 200 Hz.

Language models. We use GPT-2 (Radford et al., 2019). to estimate the contextual probabilities of each word given its past context, using the same stimuli to those heard by the participants.

Additionally, to get contextual probabilities at the scope of phonemics, we construct a GPT-Phonemic model by fine-tuning GPT on Wikipedia, transcribed into phonemes, with an adapted tokenizer.

Decoding. Speech (phonemic and semantic) features are decoded from a linear combination of MEG sensors using a ridge regression at each time sample relative to word onset (α ranging from 10^{-4} to 10^{4}, 5-fold cross-validation with sklearn-KFold). The input X is of size 27 subjects x 208 channels x 9000 words or 60000 phonemes. The output Y represents either the phonemes, which can be described via 6 phonetic features, or the words, which can be described via the 10 principal components of the 768-word features, as provided by Spacy (Honnibal & Montani, 2017).

Analysis. We relate the representations of words and phonemes decoded from MEG with stimuli posteriors approximated by the LLM – the expectation of the future word or phoneme given the context (Fig. 1).

Statistics. We apply Bonferroni-corrected T-tests on the regression of quintile-specific decoding scores.

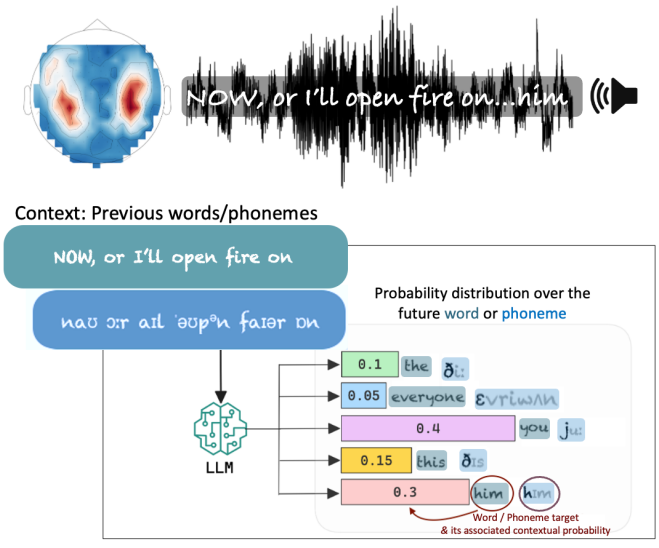

Figure 1: Experimental design. *Top: word and phonemic decoding. Bottom: approximation of stimuli's contextual expectancy through the LLM's posteriors.*

Results

Decoding. Phonetic features and word embeddings can be decoded between -20 and 1000ms (words) and 50ms and 350ms (phonemes) relative to their onset. In addition, basal as well as conditional expectancies of phonemes and words can also be decoded in similar time windows (Fig. 2). Overall, these results confirm that MEG can track the predictability as well as the content of phonemes and words during natural language listening.

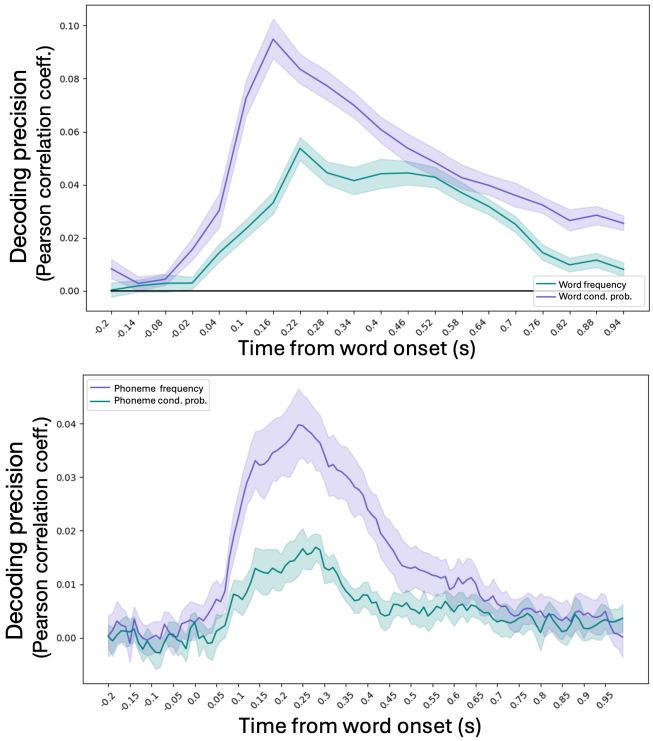

Figure 2: Temporal decoding of frequencies (from the lexicon) and conditional expectancies (from GPT2) using MEG activity patterns. *Top: for words. Bottom: for phonemes.*

Impact of predictability. We now test whether these decoded representations vary with predictability, as estimated with GPT2-Word and GPT2-Phoneme. The results show that both the words and the phonemes are better decoded when they are less expected. We observe this negative correlation between conditional expectations and the decoding precision of words and phonemes across five quintiles of decreasing expectations – from 1st level (the most expected words, resp. phonemes) to 5th level (the least expected words, resp. phonemes). Bonferroni-corrected T-tests on the regression of quintiles' decoding scores against frequencies and conditional expectancies confirm the significance of this relation for words and phonemes (P <0.01) (Fig. 3).

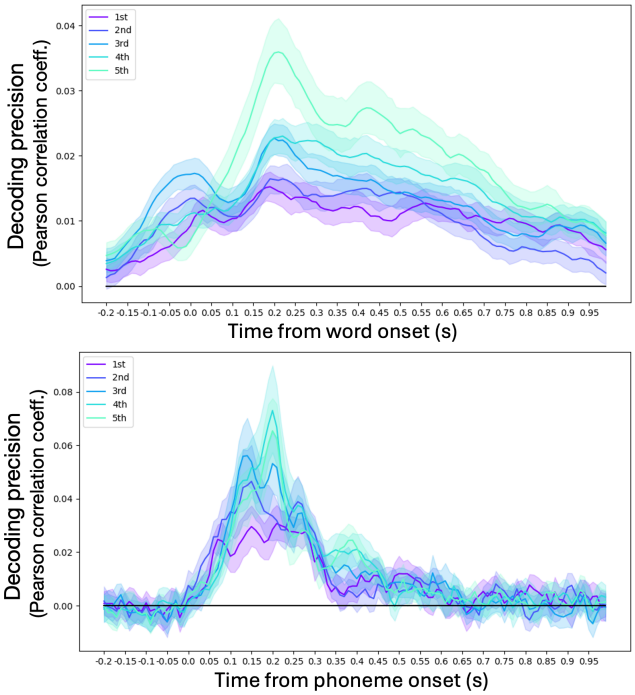

Figure 3: Temporal decoding of word and phonemic features as a function of conditional expectancies from GPT2 (split into quintiles). *Top: word features. Bottom: phonemic feature (vowel/consonant).*

Discussion

By combining neural data with a modern language model's output to the same linguistic stimuli, we show that two levels of representations in the language hierarchy get sharper when there are more surprising. These results complement previous work. For instance, (Gwilliams, King, et al., 2022; Heilbron et al., 2022) showed that the brain continuously encodes the three most recently heard speech sounds in parallel, and that high-level linguistic predictions can inform low-level ones. Here, we further show that contextual expectations can be better decoded than basal expectations, for both words and phonemes. Additionally, we find that the level of expectation with which humans predict a future word or phoneme impacts the capacity to decode this word or phoneme from their neural recordings. These findings point toward new inquiries regarding the link between expectations and the sharpening of neural representations, which we are currently exploring through additional analyses.

Conclusion

Overall, these findings provide empirical evidence to constrain the computational modeling of the human brain processing natural language. They further show how LLMs can be used as a powerful computational framework for studying the neural bases of human cognition.

Acknowledgements

This work was supported by a starting grant from the European Research Council awarded to V.W. (ERC-StG759341), and by an institutional grant from the Agence Nationale de la Recherche awarded to the Département d'Etudes Cognitives (ANR-17-EURE-0017, EUR FrontCog). The authors express their gratitude to (Gwilliams, Flick, et al., 2022) for sharing their MEG dataset.

References

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. https://doi.org/10.1038/s41562-022-01516-2

Donhauser, P. W., & Baillet, S. (2020). Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron*, *105*(2), 385-393.e9. https://doi.org/10.1016/j.neuron.2019.10.019

Forseth, K. J., Hickok, G., Rollo, P. S., & Tandon, N. (2020). Language prediction mechanisms in human auditory cortex. *Nature Communications*, *11*(1), 5240. https://doi.org/10.1038/s41467-020-19010-6

Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: A review of underlying mechanisms. *Clinical Neurophysiology*, *120*(3), 453–463. https://doi.org/10.1016/j.clinph.2008.11.029

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Gwilliams, L., Flick, G., Marantz, A., Pylkkanen, L., Poeppel, D., & King, J.-R. (2022). *MEG-MASC: A high-quality magneto-encephalography dataset for evaluating natural speech processing* (arXiv:2208.11488). arXiv. http://arxiv.org/abs/2208.11488

Gwilliams, L., King, J.-R., Marantz, A., & Poeppel, D. (2022). Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature Communications*, *13*(1), 6606. https://doi.org/10.1038/s41467-022-34326-1

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119. https://doi.org/10.1073/pnas.2201968119

Heilbron, M., Ehinger, B., Hagoort, P., & de Lange, F. P. (2019). Tracking Naturalistic Linguistic Predictions with Deep Neural Language Models. *2019 Conference on Cognitive Computational Neuroscience*. https://doi.org/10.32470/CCN.2019.1096-0

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Lopopolo, A., Frank, S. L., Van Den Bosch, A., & Willems, R. M. (2017). Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. *PLOS ONE*, *12*(5), e0177794. https://doi.org/10.1371/journal.pone.0177794

Millet, J., Caucheteux, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., & King, J.-R. (2023). *Toward a realistic model of speech processing in the brain with self-supervised learning* (arXiv:2206.01685). arXiv. http://arxiv.org/abs/2206.01685

Mousavi, Z., Kiani, M. M., & Aghajan, H. (2020). *Brain signatures of surprise in EEG and MEG data*. https://doi.org/10.1101/2020.01.06.895664

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI*.

Shain, C., Blank, I. A., Van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307

Wacongne, C., Labyt, E., Van Wassenhove, V., Bekinschtein, T., Naccache, L., & Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences*, *108*(51), 20754–20759. https://doi.org/10.1073/pnas.1117807108

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van Den Bosch, A. (2016). Prediction During Natural Language Comprehension. *Cerebral Cortex*, *26*(6), 2506–2516. https://doi.org/10.1093/cercor/bhv075