A contextual encoding model for human ECoG responses to a spoken narrative

Kristijan Armeni (karmeni1@jhu.edu)
Johns Hopkins University

Tal Linzen (linzen@nyu.edu)
New York University

Christopher Honey (chris.honey@jhu.edu)
Johns Hopkins University

Abstract

Language understanding depends on context at multiple levels of linguistic granularity. How does this context-dependence differ across different cortical regions supporting language processing? Recently, it has been shown that electrophysiological responses to narrative stimuli can be predicted by contextualized word vector representations extracted from next-word prediction models. Here, we set out to apply and extend this approach within an electrocorticography (ECoG) dataset of 9 participants listening to a 7-minute narrative. For each word in the story, we predicted the neural response based on: (i) sensory features; (ii) non-contextualized word vectors; and contextualized word vectors with context scrambled at the word, sentence and paragraph levels. We show that contextualized embeddings, on average, are better predictors of broadband high-frequency (70+ Hz) power responses compared with non-contextualized embeddings. Moreover, the improved encoding performance of contextualized embeddings specifically depended on the preceding context being provided intact to the model. These initial results provide the basis for mapping the timescale of context-dependence (word, sentence, and paragraph level) for each intracranial site across the cortical surface.

Keywords: ECoG; language models; narratives; temporal receptive windows

Introduction

Human language understanding is contextual, as we interpret each moment in a narrative in terms of the words, sentences and paragraphs that came before. In the brain, it has been suggested that scaffolding of language context is reflected in a hierarchy of processing timescales: progressively increased sensitivity to longer-scale language context along the cortical processing hierarchy (Hasson, Yang, Vallines, Heeger, & Rubin, 2008; Hasson, Chen, & Honey, 2015). However, this prior timescale-mapping work depended on measuring neural responses to multiple narratives, each preserving different scales of context.

Recently, it has become possible to map neural timescales using next-word prediction algorithms known as *language models* (LMs). *Contextualized embeddings* – vector representations of words encoding their context-specific role in a sequence – have been a key ingredient of LMs (Elman, 1990; Peters et al., 2018). Goldstein et al. (2022) have reported that contextualized LM embeddings, relative to non-contextualized, are better predictors of ECoG responses in story-listening. Moreover, Caucheteux, Gramfort, and King (2021) developed a LM-based context-scrambling method to estimate TRWs in a story-listening fMRI dataset and largely replicated the experimental findings by Lerner, Honey, Silbert, and Hasson (2011). Here, we set out to combine the approaches of Goldstein et al. (2022) and Caucheteux et al. (2021), estimating the context-dependence timescales of neural responses in ECoG responses to a 7-minute narrative.

Methods

Preprocessing

The raw ECoG recordings of 9 subjects were first visually inspected and channels with excessive noise were excluded. Data were re-referenced to the average reference and high-pass filtered at 0.1 Hz. To estimate the **broadband high frequency** component of the signal, we applied a bank of Morlet wavelets (6 cycles) to story-epoched data with frequencies ranging from 70 to 200 Hz in steps of 5 Hz. After z-scoring and log-transforming, we averaged the power time-series across the frequencies. To focus on electrodes with high signal-to-noise ratio, we selected, per participant, the top 5 electrodes with highest *repeat reliability* (Pearson correlation of ECoG responses to two repeats of the same story, Fig. 2, A).

Encoding models

Predictors ($\mathbf{X}$) As a **sensory predictor**, we constructed a 3-dimensional vector representation for each word with dimensions: word length (no. of characters), word duration (msec), and the mean amplitude of the audio envelope. To obtain **non-contextualized word vectors**, we used 50-dimensional GloVe (Pennington, Socher, & Manning, 2014) word vectors. To obtain **contextualized word vectors**, we used the 8th layer contextualized embeddings (Caucheteux & King, 2022) of the 12-layer GPT-2 transformer (Radford et al., 2019). We reduced the GPT-2 embedding dimensionality from 768 to 50 by retaining the top 50 principal components. For all three predictors, **control predictors** (yellow in Fig. 1) were constructed by randomly permuting word vectors of the same sequence across the time-steps before refitting the encoding model. **Context scrambling predictors** were obtained by

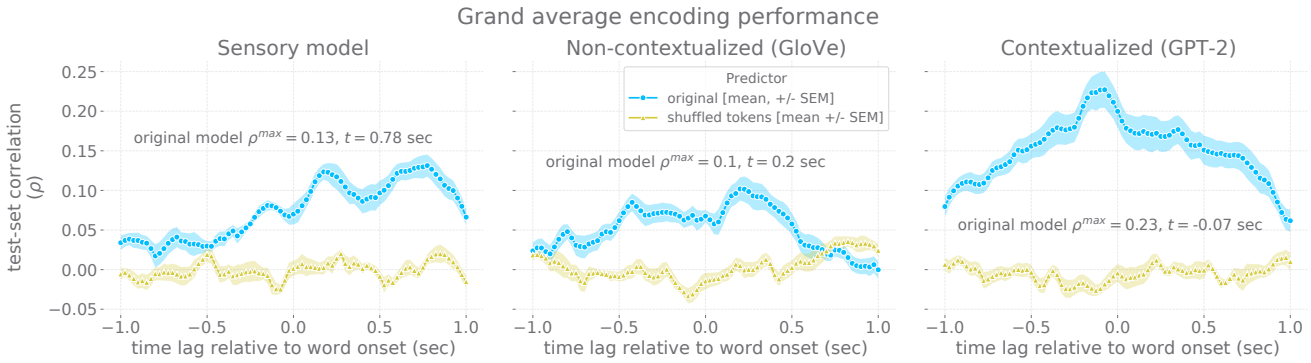

Figure 1: Encoding performance for sensory, non-contextualized, and contextualized word embeddings.

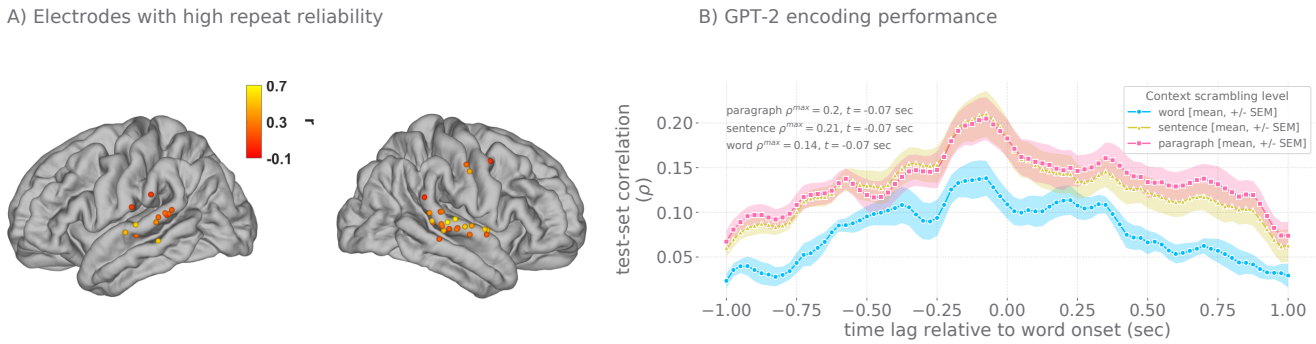

Figure 2: A) Selected electrodes with high repeat reliability (across 8 participants, 1 participant not shown due to missing information). B) Test-set encoding performance as a function of time lag for word, sentence and paragraph context scrambling.

randomly shuffling the context words/sentences/paragraphs *prior* to each sequence element in the original story, following Caucheteux et al. (2021).

Time-shifted neural response ($\mathbf{Y}$) For every token, we averaged the broadband power ECoG time-series in the window of 0-100 msec post word onset. To model the time-course of encoding performance, we time-shifted this response relative to word onset from -1 to +1 seconds in steps of 25 msec.

Regression model ($f(\mathbf{X})$) For every electrode, participant, time lag, and predictor, we fit a separate ridge regression model. We used 10-fold cross-validation to split the dataset ($N^{samples} = 748$) into training and test sets. Test-set performance (ρ, Pearson correlation between predicted and observed neural response) was averaged, per electrode and participant, across the ten test-folds and subsequently across the 5 electrodes, per participant. Finally, grand mean performance was computed across the 9 participants.

Results and Discussion

Contextualized embeddings were superior predictors of neural responses (Fig. 1). Contextualized predictors performed best overall with the mean test-set correlation peak at 0.23, while non-contextualized embeddings (middle) peak at 0.1, and the sensory model (left panel) at 0.13. The *time-courses* of encoding model performance, relative to the time of word onset, revealed that the contextualized embedding performance peaks before word onset (-0.07 sec). This observation is consistent with the predictive encoding processes reported by Goldstein et al. (2022), in which neural responses before word onset contain information about the upcoming word.

The context-dependence effects we observed were primarily driven by context *within* a sentence, rather than by context beyond a sentence (Fig. 2, B). Thus, the advantage of contextualized embeddings in our data likely derives from relatively local semantics and syntactic effects.

Conclusions

Consistent with prior fMRI and ECoG studies (Goldstein et al., 2022; Jain et al., 2020; Caucheteux et al., 2021), we find that neural responses captured by neural language models are predictive and context-dependent. We further show that the context advantage in the superior and middle temporal gyri are largely driven by within-sentence context effects. In order to map processing timescales across the cortical surface, we are now quantifying the contextual encoding performance within individual electrodes, and are extending these analyses to frontal cortical sites.

References

Caucheteux, C., Gramfort, A., & King, J.-R. (2021, November). Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3635–3644). Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved 2022-05-20, from https://aclanthology.org/2021.findings-emnlp.308 doi: 10.18653/v1/2021.findings-emnlp.308

Caucheteux, C., & King, J.-R. (2022, December). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134. Retrieved 2022-05-31, from https://www.nature.com/articles/s42003-022-03036-1 doi: 10.1038/s42003-022-03036-1

Elman, J. (1990, June). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. doi: 10.1016/0364-0213(90)90002-E

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... Hasson, U. (2022, March). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. Retrieved 2022-05-19, from https://www.nature.com/articles/s41593-022-01026-4 doi: 10.1038/s41593-022-01026-4

Hasson, U., Chen, J., & Honey, C. J. (2015, June). Hierarchical process memory: memory as an integral component of information processing. *Trends in Cognitive Sciences*, *19*(6), 304–313. Retrieved 2017-11-22, from http://www.sciencedirect.com/science/article/pii/S1364661315000923 doi: 10.1016/j.tics.2015.04.006

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008, March). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, *28*(10), 2539–2550. doi: 10.1523/JNEUROSCI.5487-07.2008

Jain, S., Mahto, S., Turek, J. S., Vo, V. A., LeBel, A., & Huth, A. G. (2020, October). *Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech* (preprint). Neuroscience. Retrieved 2021-06-14, from http://biorxiv.org/lookup/doi/10.1101/2020.10.02.324392 doi: 10.1101/2020.10.02.324392

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011, February). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*(8), 2906–2915. doi: 10.1523/JNEUROSCI.3684-10.2011

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. Retrieved 2020-08-08, from http://aclweb.org/anthology/D14-1162 doi: 10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018, June). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N18-1202 doi: 10.18653/v1/N18-1202

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners.* Retrieved from https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf