Investigating the Timescales of Language Processing with EEG and Language Models

Davide Turco (davide.turco@bristol.ac.uk)
Conor Houghton (conor.houghton@bristol.ac.uk)
Faculty of Engineering, University of Bristol, Bristol, BS8 1UB, UK

Abstract

This study explores the temporal dynamics of language processing by examining the alignment between word representations from a pre-trained transformer-based language model, and EEG data. Using a Temporal Response Function (TRF) model, we investigate how neural activity corresponds to model representations across different layers, revealing insights into the interaction between artificial language models and brain responses during language comprehension. Our analysis reveals patterns in TRFs from distinct layers, highlighting varying contributions to lexical and compositional processing. Additionally, we used linear discriminant analysis (LDA) to isolate part-of-speech (POS) representations, offering insights into their influence on neural responses and the underlying mechanisms of syntactic processing. These findings underscore EEG's utility for probing language processing dynamics with high temporal resolution. By bridging artificial language models and neural activity, this study advances our understanding of their interaction at fine timescales.

Keywords: EEG; neurolinguistics; language models; word representations; natural language processing.

Introduction

The representations of modern language models have been shown to linearly map to brain responses to the same linguistic stimulus (Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016; Caucheteux & King, 2022), as measured by fMRI or MEG. This may suggest that the two systems share similar mechanisms when processing language.

EEG, with its high temporal resolution, is an powerful and practical source of data for exploring the timescale of language processing in the brain. However, previous work has focused on simpler non-neural (Broderick, Anderson, Di Liberto, Crosse, & Lalor, 2018) or recurrent (Hale, Dyer, Kuncoro, & Brennan, 2018) language models, often using a link function, such as surprisal.

Here, we map the word representation of a pre-trained causal transformer to EEG data recorded from subjects listening to the same stimulus, via a linear convolution model. The objective of this preliminary work is to investigate the relationship between artificial models and the brain at small timescales, considering different layers of the language model and language aspects, such as syntax.

Methodology

We use publicly available EEG data (Bhattasali, Brennan, Luh, Franzluebbers, & Hale, 2020), recorded from 52 subjects listening to the first chapter of *Alice's Adventure in Wonderland*. We limited the analysis to 33 participants, excluding subjects with excessively noisy recordings or with poor scores in the post-experiment text comprehension test. For computational purposes, the signal has been segmented into 2 s windows, with a 10% overlap.

The same linguistic stimulus was given word-by-word to a pre-trained language model, GPT-2 (Radford et al., 2019), and word representations were extracted from the embedding layer and from a deep layer. We selected layer 9, as this has been previously shown to better predict brain responses (Caucheteux, Gramfort, & King, 2021). The activations were transformed to a vector with the same sampling frequency as the original data.

For aligning EEG activities $r_{t,e}$ and word representations $s_{t-\tau,i}$, we use a time-lagged linear regression model (Crosse, Di Liberto, Bednar, & Lalor, 2016):

$$r_{t,e} = \sum_i \sum_\tau w_{\tau,e,i}\, s_{t-\tau,i} \tag{1}$$

where e indicates electrode and i is the GPT-2 representation dimension ($i \in [0, 768]$). w is the linear filter kernel of length τ that, when applied to the stimulus s, it transforms it into the brain response r. This filter is known as the Temporal Response Function (TRF).

We implemented the model in PyTorch, using a kernel ranging from -100 ms to 1 s relative to the word onset. To prevent overfitting, a L2 regularisation was applied to the model weights: the value of the regularisation coefficient was chosen among ten log-spaced from 10^{-3} to 10^{5} using 5-fold cross-validation. For each participant, a model trained with the best parameter is then tested on held-out data from that specific subject, and the Pearson correlation score between the original and reconstructed EEG signal is computed.

To isolate syntactic factors in the language model representations, we used linear discriminant analysis (LDA), motivated by previous work showing that language models encode linguistic features in a linear manner (Linzen & Baroni, 2021). Specifically, we reduced the original 768-dimensional word representations to the same number of dimensions as the linguistic features of interest, in this case part-of-speech (POS).

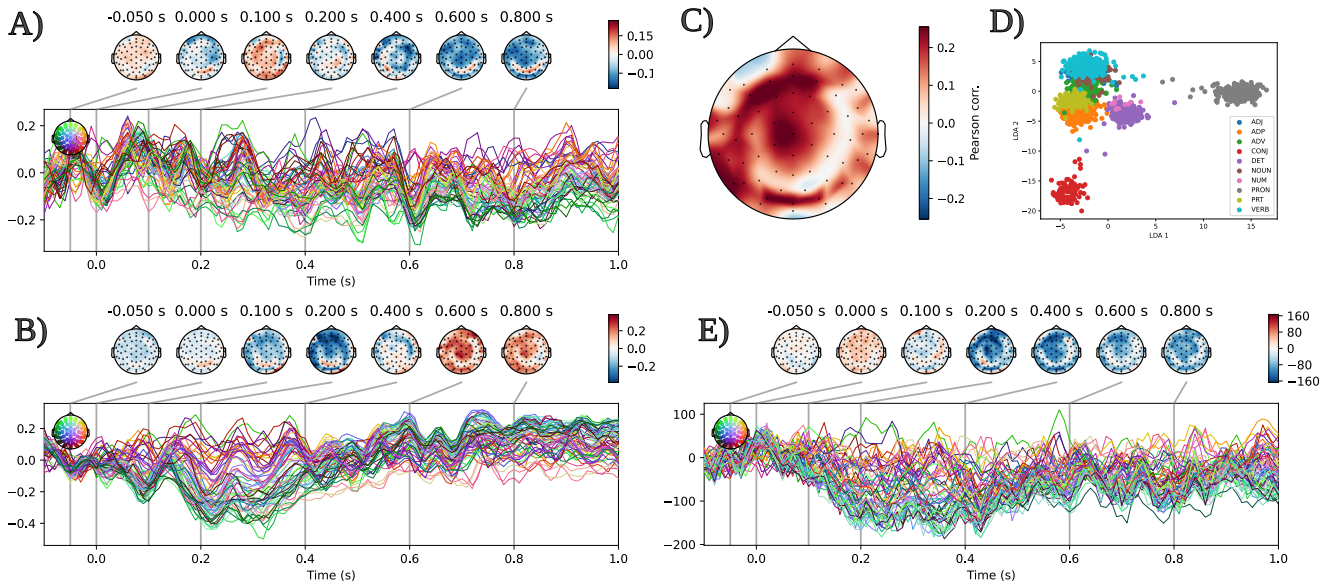

Figure 1: A) TRF obtained using the embedding layer (l=0). B) TRF obtained using a deep layer (l=9). C) Topographic map showing correlations between original and reconstructed EEG signal for a subject with high comprehension score. D) LDA-reduced representation space, with samples coloured by POS tag. E) TRF obtained using LDA-reduced representations.

Results

Our results show that mapping embedding (l=0) and deep (l=9) layer activations to EEG data lead to noticeably different TRFs. As shown in Fig. 1A), the TRF for the embedding layer mostly displays negativities, especially in late time scales. On the contrary, the TRF for the deep layer (Fig. 1B)) shows stronger negativity in the 200 ms region, and a positive effect in the post 600 ms region, compatible with a P600 event-related potential (Coulson, King, & Kutas, 1998). This shows a fundamental distinction between the embedding layer, that encodes only lexical information, and a deep layer, that encodes compositional information as well.

In Fig. 1C) we show the correlation between original and reconstructed EEG signal for a subject with high comprehension score, using activations from layer 9. The topographic plot shows higher scores in the central and left-temporal regions, normally associated with language processing. The correlation for all subjects is 0.03 ($p \ll 10^{-5}$); per-subject correlations and p-values have been aggregated using the Fisher's method.

To motivate the use of LDA to isolate syntactic representation, we plotted the reduced representation space in Fig. 1D). Samples corresponding to different POS tags are tightly clustered. Interestingly, samples corresponding to content words (e.g. `NOUN`, `VERB`, `ADV`) appear closer together than short function words such as conjunctions (`CONJ`) and pronouns (`PRON`).

We then fitted our model on this reduced space and the corresponding TRF is shown in Fig. 1E). The representation dimensions related to part of speech appear to negatively correlate with EEG activities in the post 200 ms segment.

Discussion

In this paper, we have introduced an approach for investigating the timescale of language processing by mapping the word representations of a transformer-based language model to high-temporal-resolution EEG data from human participants. We have shown that embedding and deep layers lead to different responses, both in their topographic distribution and in their timescale. We have also presented a simple technique for isolating POS representations in the language model activations, and shown that these are negatively correlated with EEG activity.

In future work, we plan to improve the mapping model by adding non-linear components to the architecture. We would also like to extend the analysis to other aspects of language, like semantics.

Acknowledgments

DT is funded by an UKRI Centre of Doctoral Training grant (EP/S022937/1). This work was carried out using the HPC facilities of the ACRC, University of Bristol. We are grateful to Dr Stewart whose philanthropy supported the purchase of GPU nodes.

References

Bhattasali, S., Brennan, J., Luh, W.-M., Franzluebbers, B., & Hale, J. (2020). The Alice datasets: fMRI & EEG observations of natural language comprehension. In *Proceedings of*

the twelfth international conference on language resources and evaluation (LREC 2020) (pp. 120–125).

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018, March). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803–809.e3. Retrieved from http://dx.doi.org/10.1016/j.cub.2018.01.080 doi: 10.1016/j.cub.2018.01.080

Caucheteux, C., Gramfort, A., & King, J.-R. (2021, 18–24 Jul). Disentangling syntax and semantics in the brain with deep networks. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning* (Vol. 139, pp. 1336–1348). PMLR.

Caucheteux, C., & King, J.-R. (2022, February). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1). Retrieved from http://dx.doi.org/10.1038/s42003-022-03036-1 doi: 10.1038/s42003-022-03036-1

Coulson, S., King, J. W., & Kutas, M. (1998, January). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes*, *13*(1), 21–58. Retrieved from http://dx.doi.org/10.1080/016909698386582 doi: 10.1080/016909698386582

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016, November). The multivariate temporal response function (mtrf) toolbox: A matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*. Retrieved from http://dx.doi.org/10.3389/fnhum.2016.00604 doi: 10.3389/fnhum.2016.00604

Hale, J., Dyer, C., Kuncoro, A., & Brennan, J. (2018, July). 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). Melbourne, Australia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P18-1254 doi: 10.18653/v1/P18-1254

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016, April). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. Retrieved from http://dx.doi.org/10.1038/nature17637 doi: 10.1038/nature17637

Linzen, T., & Baroni, M. (2021, January). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*(1), 195–212. Retrieved from http://dx.doi.org/10.1146/annurev-linguistics-032020-051035 doi: 10.1146/annurev-linguistics-032020-051035

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.