The language network occupies a privileged position among all brain voxels predicted by a language-based encoding model

Eyas Ayesh
eayesh3@gatech.edu
Georgia Tech

Shailee Jain
Shailee.Jain@ucsf.edu
UCSF

Josleen St Luce
josleen@mit.edu
MIT

Alexander Huth
huth@cs.utexas.edu
UT Austin

Anna A. Ivanova
a.ivanova@gatech.edu
Georgia Tech

Abstract

We report preliminary results from the project to systematically analyze the relationship between language voxels and voxels that are well-predicted by a GPT2-based encoding model (EM). Language voxels are defined as those that respond significantly more to sentences than non-words, as identified via an auditory language localizer task. We find that $\sim$half of the language voxels are well-predicted by the EM, although $>$90% of well-encoded voxels are not language voxels. Language voxels, on average, have significantly better EM performance than non-language voxels, both among all cortical voxels and among well-predicted voxels. Finally, we project the EM voxelwise weights into a 3-PC subspace and find the the language voxels tend to have a positive bias along each PC. Consequently, we find a separating plane in the 3-PC space that separates language and non-language voxels predicted by an EM with a $>$75% accuracy both within-subjects and across-subjects, suggesting that language models have a clearly identifiable EM signature.

Keywords: Language network, Encoding Models, fMRI, naturalistic paradigms

Introduction

Compelling neural evidence points to the existence of a specialized language network focused on language processing and associated tasks (Fedorenko et al., 2024). However, language-based encoding models (EM) can predict neural activity to linguistic stimuli not only in the language network (Schrimpf et al., 2021), but also in many other cortical regions, with many voxels exhibiting semantically selective responses (Huth et al., 2016). Here, we bridge these two lines of evidence by investigating the relationship between language-responsive voxels and voxels predicted by a language-based encoding model.

Methods

We examined the relationship between language-responsive voxels (hereafter, **language voxels**) and voxels that are significantly predicted by a GPT2-based EM (hereafter, **well-predicted voxels**) within two subjects, UTS03 and UTS08 (LeBel et al., 2023).

Language Voxels Language voxels are defined using the fMRI data collected on an auditory language localizer task (Scott et al., 2017), by contrasting voxel responses to short story excerpts and to their acoustically degraded versions. Any cortical voxel that passed a one-sided t-test with a significance threshold of $p <= 0.001$ (uncorrected) was classified as a language voxel.

Well-Predicted Voxels The EMs were built using fMRI data collected during a passive naturalistic story listening task on 27 stories, 26 of which were used to train and cross-validate the EMs while 1 held-out story was used to report the test-set prediction performance. To determine the (statistically significant) well-predicted voxels, we also fit 26 EMs by leaving one of the train stories out. Then, the EM performance for held-out stories across all 26 EMs was measured as the correlation (r) between the predicted and actual response time-series. Statistical significance of r was measured using a blockwise permutations test with correction for multiple comparisons ($q < 0.001$) (Jain and Huth, *in prep*; Vo et al., 2023).

Deriving the principal components of EM weights Following the approach in Huth et al. (2016), we characterize semantic selectivity across cortex by decomposing each subject's EM weights into principal components (PCs). We independently de-mean the weights of every significantly predicted voxel in the subject and then apply principal components analysis. The first three PCs explain a significant amount of variance across the EM weights (bootstrap test; $p < 0.05$) and are highly correlated between subjects.

To reduce the influence of low-level features on the PC space, we filter out voxels that are also significantly predicted by a low-level EM (features include: word rate, phoneme rate, and phonemic content) (Jain and Huth, *in prep*; Vo et al., 2023). The PC analysis is therefore conducted on the well-predicted voxels not also predicted by the low-level EM.

Classifying language vs. non-language voxels Finally, we trained a support-vector-machine (SVM) to classify well-encoded voxels as either language or non-language using their 3-PC space coordinates. The model was trained with a linear kernel and cost factor of 1. To compensate for the imbalance of language to non-language voxels, we up-sampled the language voxels before splitting the data into training and testing sets. 85% of the data was used for training.

Results

Well-predicted language voxels constitute a minority of all well-predicted voxels We found partial overlap between the language voxels and the well-predicted voxels (Figure 1A). For UTS03 and UTS08, the percentage of language voxels that are well-predicted by the EM is $62\%$and 51% respectively. However, the percentage of well-predicted voxels that are also language-responsive is only 6.7% and 5.5% respectively, meaning that most well-predicted voxels are not identified with the language localizer contrast. As shown in Figure 1B, voxels that are both language and well-predicted are located in canonical language areas; language voxels that are not well-predicted are mostly located outside canonical language areas and might be driven by low-level artifacts; and well-predicted voxels that are not language mainly occupy large swaths of the associative cortex.

EM predictivity is higher for language voxels than for non-language voxels We then asked whether language voxels have better EM predictivity than non-language voxels, both among all cortical voxels (Figure 1C) and among well-predicted voxels specifically (Figure 1D). Permutation tests showed significantly higher predictivity for language voxels than for non-language voxels, both among all voxels (UTS03: language $r_{mean} = 0.24$, non-language $r_{mean} = 0.11$, $p =$

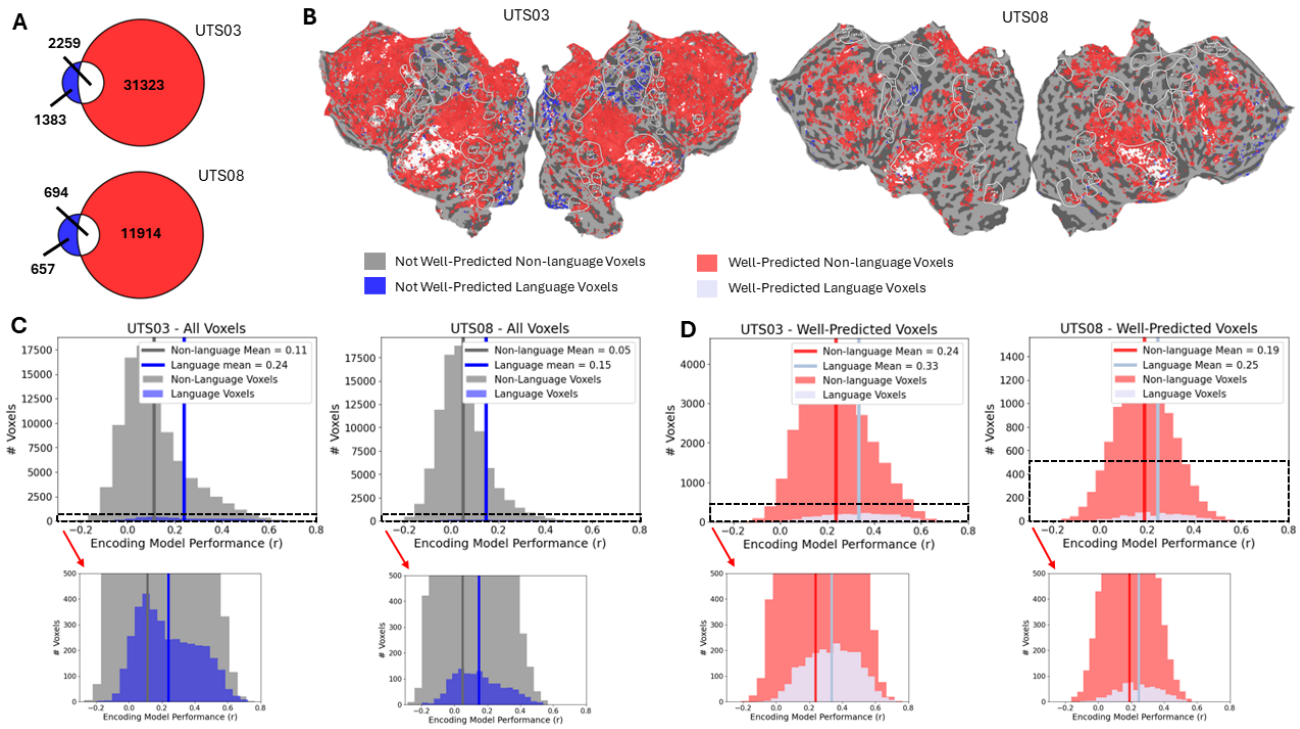

Figure 1: A) Venn diagrams demonstrating the relative numbers of language voxels and well-predicted voxels for each subject. B) Subject-level brain maps showing the locations of the language voxels (blue), the well-predicted voxels (red), and their overlap (white). C) EM predictivity histograms for language voxels vs non-language voxels for all voxels. D) EM predictivity histograms of language voxels vs non-language voxels for well-predicted voxels only. In all 4 histograms, language voxels have significantly better predictivity than non-language voxels.

0.002; UTS08: language $r_{mean} = 0.15$, non-language $r_{mean} = 0.05$, $p = 0.002$) and among the set of well-predicted voxels (UTS03: language $r_{mean} = 0.33$, non-language $r_{mean} = 0.24$, $p = 0.002$; UTS08: language $r_{mean} = 0.15$, non-language $r_{mean} = 0.19$, $p = 0.002$). This indicates that, even though language voxels constitute only a small portion of well-encoded voxels, they stand out in terms of their EM predictivity.

Language voxels occupy a specific region in the PC space derived from EM weights Finally, we tested whether the language well-predicted voxels can be differentiated from non-language well-predicted voxels on the basis of EM weight patterns. To do so, for each subject, we projected all well-encoded voxels into the 3-dimensional PC space (see methods). Permutation tests showed that the language voxels have a significant positive bias along each of the 3 PCs.

An SVM trained to classify well-encoded voxels as language vs. non-language (Figure 2) achieved high performance (UTS03: accuracy 76.2%, F1 73.82%; UTS08: accuracy 80.4%, F1 79.7%). Moreover, each SVM could successfully generalize to the other subject (UTS03→UTS08: accuracy 80.0%, F1 79.3%; UTS08→UTS03: accuracy 76.0%, F1 73.02%). This result shows a high degree of consistency in the language voxels' location across subjects' PC spaces.

Conclusion and Future work

We show that GPT2-based EM predictivity is significantly higher for language voxels than for non-language voxels, even when only considering the voxels that are well-predicted by the EM. We also show that language voxels occupy a specific subspace of the EM weight PC-space, indicating that they have an identifiable EM signature. The goal of future work is to generalize these results to additional subjects and develop analysis tools to interpret the features that lead the EM to differentially represent language and non-language voxels.

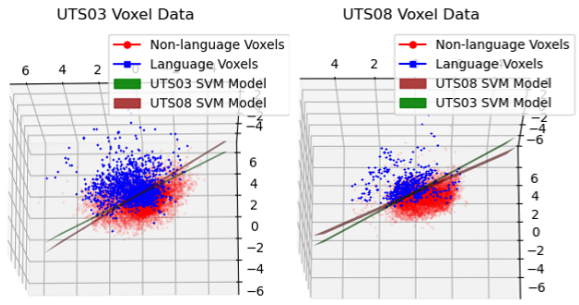

Figure 2: The scatter plots of well-predicted voxels in the 3PC space for each subject, along with the separating planes derived from subject-level SVM classifiers.

Acknowledgments

We thank Ev Fedorenko and Greta Tuckute for their comments on this project.

References

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*. https://doi.org/10.1038/s41583-024-00802-4

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Jain, S., & Huth, A. G. (*in prep*). The cortical organisation of language is jointly explained by semantics and integration timescales.

LeBel, A., Wagner, L., Jain, S., Adhikari-Desai, A., Gupta, B., Morgenthal, A., Tang, J., Xu, L., & Huth, A. G. (2023). A natural language fmri dataset for voxelwise encoding models. *Scientific Data*, *10*(1), 555.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45).

Scott, T. L., Gallée, J., & Fedorenko, E. (2017). A new fun and robust version of an fmri localizer for the frontotemporal language system. *Cognitive neuroscience*, *8*(3), 167–176.

Vo, V. A., Jain, S., Beckage, N., Chien, H.-Y. S., Obinwa, C., & Huth, A. G. (2023). A unifying computational account of temporal context effects in language across the human cortex [Pages: 2023.08.03.551886 Section: New Results]. https://doi.org/10.1101/2023.08.03.551886