Multilingual Computational Models Reveal Shared Brain Responses to 21 Languages

Andrea Gregor de Varda (devar_ag@mit.edu), Saima Malik-Moraleda (saimamm@mit.edu)
Greta Tuckute (gretatu@mit.edu), Evelina Fedorenko (evelina9@mit.edu)
Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences
43 Vassar Street, Cambridge, MA 02139 USA

Abstract

How does the human brain process the rich variety of languages? Multilingual neural network language models (MNNLMs) offer a promising avenue to answer this question by providing a theory-agnostic way of representing linguistic content across languages. We combined existing and newly collected fMRI data from speakers of 21 languages to test whether MNNLM-based encoding models can predict brain activity in the language network. Across 20 models and 8 architectures, encoding models successfully predicted responses in the various languages, replicating and extending previous findings. Critically, models trained on a subset of languages generalized zero-shot to held-out ones, even across language families. This cross-linguistic generalization points to a shared component in how the brain processes language, plausibly related to a shared meaning space.

Keywords: brain encoding models; multilingual language models; fMRI

Introduction

Human languages vary remarkably in their surface features, yet all are supported by the same biological system: the language network (Fedorenko, Ivanova, & Regev, 2024). Prior research has shown that this network is functionally and anatomically consistent across languages (Malik-Moraleda, Ayyash et al., 2022), but it remains unclear whether linguistic representations themselves are also similar. Recent advances in multilingual neural network language models (MNNLMs) offer a unified, theory-agnostic way to represent linguistic content across languages. These models, trained on massive multilingual corpora with or without explicit cross-lingual supervision, project linguistic inputs into shared representational spaces, making it possible to quantify and compare linguistic content across languages.

To assess whether language representations are similar across languages in the human brain, we evaluated how well MNNLMs predict brain activity from fMRI during naturalistic language comprehension in 21 languages across 7 families. We first tested whether models reliably predicted brain responses within each individual language, in line with prior work (Antonello & Huth, 2024; Aw & Toneva, 2023; Caucheteux & King, 2022; Goldstein et al., 2022; Schrimpf et al., 2021, *inter alia*), and if model features known to improve encoding performance in English (e.g., autoregressive training, next-word prediction accuracy, model size) generalized across languages. Next and critically, we asked whether encoding models trained on a subset of languages could predict brain responses in held-out languages (Study I) even across different modalities and types of stimuli (Study II), probing the existence of a shared, language-general component in linguistic representations.

Results

In Study I, we evaluated whether MNNLMs can predict brain responses to 12 typologically diverse languages. We used pre-existing data (Malik-Moraleda, Ayyash et al., 2022) where participants listened to a short passage in their native language during fMRI scanning, and we extracted contextualized word embeddings from 20 MNNLMs spanning a range of training objectives, architectures, and sizes. Linear encoding models

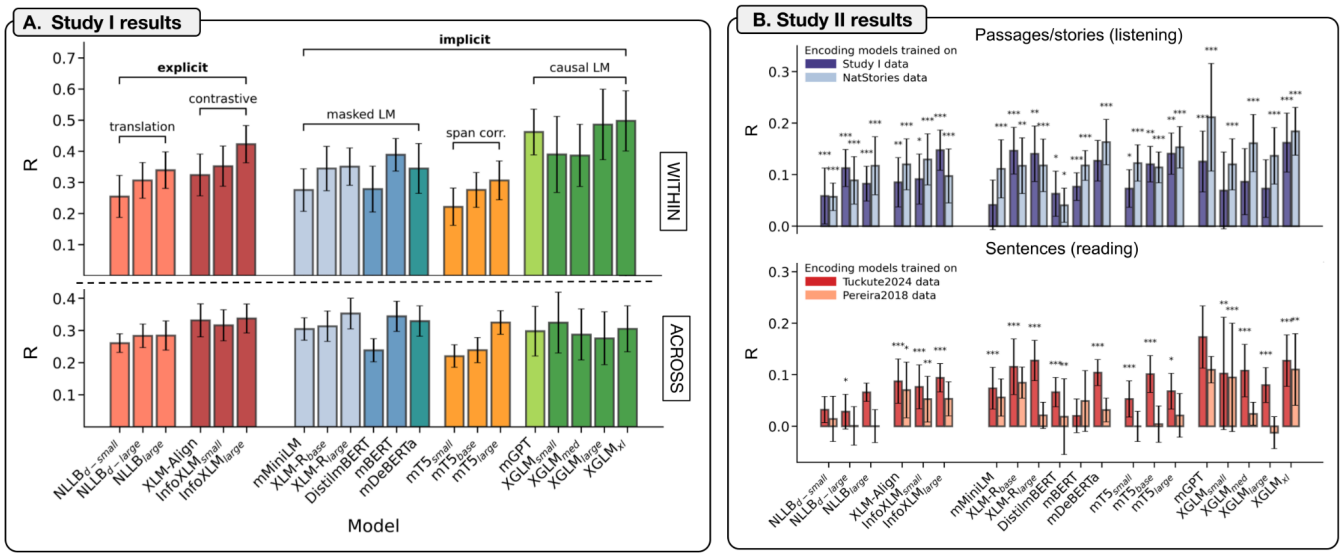

Figure 1. Results of the encoding models across the two experiments. Across subplots A and B, results are averaged across languages. **A. Study I results:** Best-layer WITHIN (top) and ACROSS (bottom) encoding performance by model. The error bars indicate the standard error of the encoding performance across languages. **B. Study II results:** Performance of the models trained on the independent data sources (Study I, *NatStories, Pereira2018, Tuckute2024*) in the transfer to the new data we collected, divided by modality of presentation (top: listening; bottom: reading).

were trained to predict activity in the left-hemisphere language network (functionally defined), averaged across voxels, fROIs, and participants.

Encoding models trained and tested within the same language (WITHIN condition) successfully predicted brain activity above chance (all $p < 0.001$, Figure 1A, top). The strongest performance was obtained by i) causal models (vs. masked or encoder-decoder models), ii) models with more parameters, and iii) models with better next-word prediction abilities. Intermediate-to-deep layers consistently yielded the best predictivity. Critically, encoding models trained in N-1 languages generalized zero-shot to the held-out language (ACROSS condition), even across language families (all $p < 0.001$; Figure 1A, bottom). Cross-lingual encoding transfer was strongest in models with higher cross-lingual embedding alignment (measured by how closely they mapped translation-equivalent sentences across languages).

Study II tested the robustness of cross-lingual transfer under more challenging conditions. We trained encoding models on existing fMRI datasets in English (*Pereira2018*, Pereira et al., 2018; *Tuckute2024*, Tuckute et al., 2024; *NatStories,* Blank et al., 2014) that varied in modality (listening vs. reading) and stimulus type (sentences vs. narratives), as well as on the multilingual data from Study I. These models were then evaluated on newly collected fMRI data from speakers of 9 previously untested languages, using the same passage listening paradigm as in Study I. Encoding models generalized successfully to these new languages, despite differences in language, type of stimuli, and modality (Figure 1B).

Discussion

Our results provide strong evidence for a shared component in how the human brain processes language across typologically diverse languages. Concerning the *kind* of information that is shared, our findings point to semantics (rather than form-level properties): control analyses ruled out shallow form-level predictors (word length and frequency) and found no evidence that formal linguistic similarity (e.g., syntax or phonology) explained generalization. These findings suggest that shared brain responses across languages are primarily driven by linguistic meaning.

References

Antonello, R., & Huth, A. (2024). Predictive coding or just feature discovery? An alternative account of why language models fit brain data. *Neurobiology of Language, 5*(1), 64-79.

Aw, K. L., & Toneva, M. (2023). Training Language Models to Summarize Narratives Improves Brain Alignment. In *Eleventh International Conference on Learning Representations.*

Blank, I., Kanwisher, N., & Fedorenko, E. (2014). A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of neurophysiology, 112*(5), 1105-1118.

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

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, 25*(5), 289-312.

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

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences, 118*(45), e2105646118.

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., ... & Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature neuroscience, 25*(8), 1014-1019.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., ... & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature communications, 9*(1), 963.

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., ... & Fedorenko, E. (2024a). Driving and suppressing the human language network using large language models. *Nature Human Behaviour, 8*(3), 544-561.