A Two-Dimensional Space of Linguistic Representations Shared Across Individuals

Greta Tuckute (gretatu@mit.edu)
Department of Brain and Cognitive Sciences & McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139

Elizabeth J. Lee (jelizlee@mit.edu)
Department of Brain and Cognitive Sciences & McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139

Yongtian Ou (ou000036@umn.edu)
Center for Magnetic Resonance Imaging, Department of Radiology, University of Minnesota, Minneapolis, MN 55455

Evelina Fedorenko (evelina9@mit.edu)
Department of Brain and Cognitive Sciences & McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139
Program in Speech and Hearing Bioscience and Technology, Division of Medical Sciences, Harvard University, Boston, MA 02114

Kendrick Kay (kay@umn.edu)
Center for Magnetic Resonance Imaging, Department of Radiology, University of Minnesota, Minneapolis, MN 55455

Abstract

Humans learn and use language in diverse ways, yet all typically developing individuals acquire at least one language and use it to communicate complex ideas. This fundamental ability raises a key question: Which dimensions of language processing are shared across brains, and how are these dimensions organized in the human cortex? To address these questions, we collected ultra-high-field (7T) fMRI data while eight participants listened to 200 linguistically diverse sentences. To identify the main components of variance in the sentence-evoked brain responses, we performed data decomposition and systematically tested which components generalize across individuals. Only two *shared* components emerged robustly, together accounting for about 32% of the explainable variance. Analysis of linguistic feature preferences showed that the first component corresponds to processing difficulty, and the second—to meaning abstractness. Both components are spatially distributed across frontal and temporal areas associated with language processing but, surprisingly, also extended into the ventral visual cortex. These findings reveal a low-dimensional, spatially structured representational basis for language processing shared across humans.

Keywords: language; inter-individual neural organization; spatial topography; fMRI

Introduction

Language enables the transfer of complex ideas across minds—an ability that has laid a critical foundation for human culture. A set of left-lateralized frontal and temporal brain areas—the "language network"—supports language understanding and production, across modalities (spoken, written, and signed; MacSweeney et al. 2002; Deniz et al. 2019; Hu et al. 2022) and across typologically diverse languages (Malik-Moraleda, Ayyash et al. 2022). Although the brain areas that support language are well-established, their internal organization—what dimensions structure their responses and how—remains poorly understood. All frontal and temporal language areas show a similar response profile: they are strongly engaged by structured and meaningful language in controlled paradigms (Rodd et al., 2010; Fedorenko et al., 2020) and track linguistic complexity during naturalistic comprehension (Shain et al. 2020; Wehbe et al. 2021). However, some structure may exist within this network of areas that does not correspond to regional boundaries (e.g., Jain et al., 2020; Regev, Casto et al., 2024). In this study, we used a data-driven structure-discovery approach to explain each voxel's response to diverse sentences as a weighted sum of a smaller number of components. Our goal is to identify the organizing dimensions of language representations that are shared across individuals and to characterize how these dimensions are distributed across the brain.

Methods

We scanned eight proficient English speakers (monolinguals and multilinguals) with 7T fMRI while they listened to 200 spoken sentences (2s each), repeated three times in pseudorandomized order. Sentence-level BOLD responses were estimated using GLMsingle (Prince et al., 2022). We extracted reliable voxel responses from five large anatomical parcels covering the frontal and temporal cortex implicated in language processing (Lipkin et al., 2022), and applied singular value decomposition to the mean-subtracted voxel responses concatenated across participants. This procedure yields the "Sentence PCs", where each Sentence PC denotes how much each sentence drives variance along a principal dimension of voxel activity. To identify Sentence PCs that generalize across individuals, we used a leave-one-participant-out framework: we trained ordinary least squares (OLS) models using Sentence PCs derived from seven participants to predict voxel responses in the held-out participant. We characterized the resulting Sentence PCs using a set of 12 linguistic/semantic properties—combining properties from prior work (Tuckute et al. 2024) with new experiments that directly probe processing difficulty and abstractness of sentence meanings. Finally, we visualized the Sentence PC weights on the cortical surface.

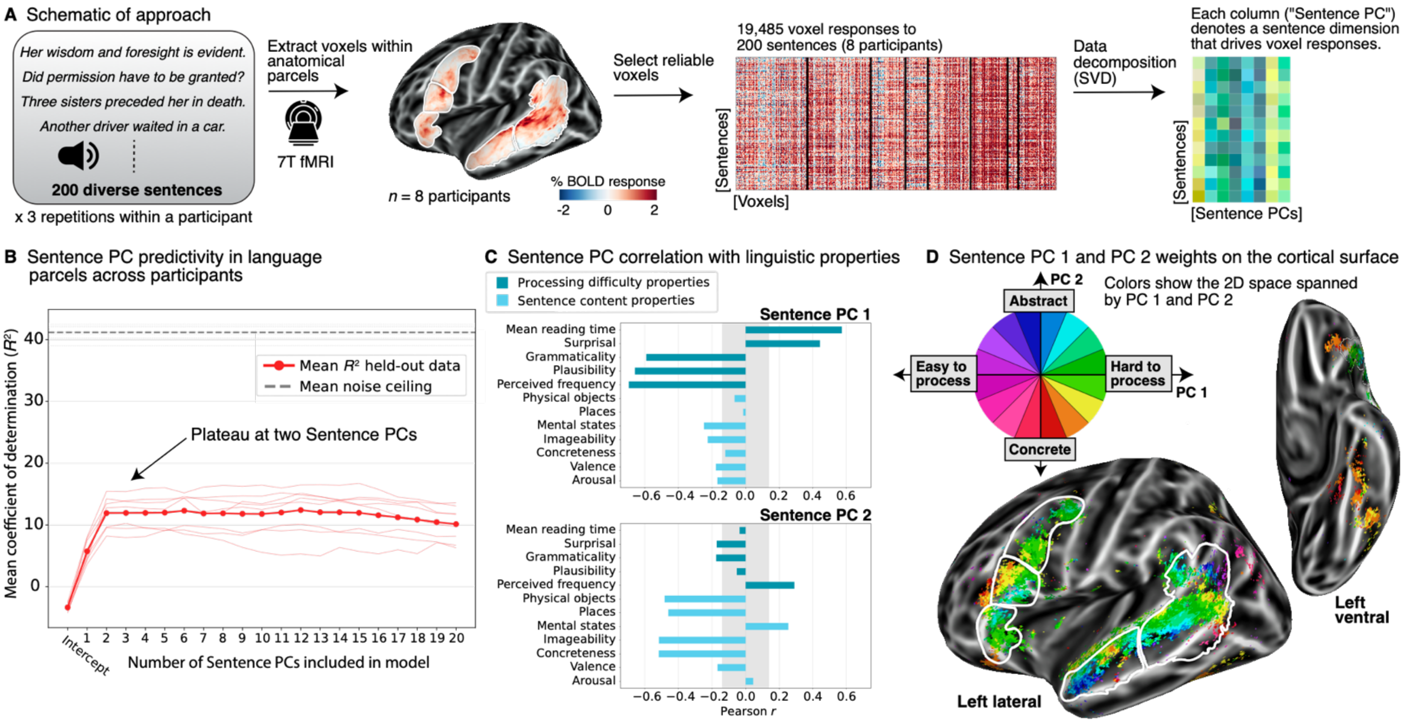

Fig. 1. A. Overview of procedure for deriving components of sentence-evoked brain responses ("Sentence PCs"). **B.** Held-out participant prediction accuracy as a function of the number of Sentence PCs used. **C.** Correlations between Sentence PCs and linguistic features. **D.** Surface maps of Sentence PC weights.

Results

How many distinct components of language are shared across individuals and what characterizes them? To search for shared dimensions of language representations, we identified the principal components ("Sentence PCs") of sentence-evoked brain responses (**Fig. 1A**) and tested their generalizability across individuals. We found that model performance plateaued at two Sentence PCs (**Fig. 1B**); additional components failed to improve prediction accuracy in held-out participants, indicating that PCs beyond two reflect inter-individual neural variability. These two PCs accounted for about 32% of sentence-evoked variance.

Sentence PC 1 was strongly correlated with measures of processing difficulty, such as reading times (Boyce et al., 2020), surprisal, and frequency– altogether capturing a dimension that spans "Easy to process" to "Hard to process" (**Fig. 1C**). Sentence PC 2 correlated most strongly with concreteness and imageability (how much a sentence is tied to perceptual, including visual experience), capturing a dimension from "Concrete" to "Abstract" sentence meanings.

How are the components spatially organized? To understand how the two components are distributed across the brain and whether any systematic patterns exist across individuals, we visualized the PC weights on the cortical surface, averaging across significantly predicted voxels in all eight participants (**Fig. 1D**). Both PCs were present throughout the left fronto-temporal language network (white demarcations), with these areas showing an overall preference for abstract/hard-to-process sentences. Quantifications of these response profiles revealed that PC2—the meaning abstractness component—was more prominently present in the temporal areas compared to the frontal areas ($p<.05$). Surprisingly, we also observed robust prediction in the left ventral visual cortex, typically associated with high-level vision. These voxels were tuned to visualizable, "concrete" sentences, potentially reflecting spontaneous visual imagery or multimodal semantic processing distinct from the computations implemented in the frontal and temporal language areas.

Conclusion

We identify two principal, cross-individual dimensions of language representations in the brain: processing difficulty and meaning abstractness. These findings lay the foundation for developing topographical models of the neural architecture of language.

Acknowledgements

We thank Emily Allen for help with fMRI scanning. We are grateful for feedback on the project from Jacob S. Prince and Nicholas M. Blauch. We thank Selena She, Maya Taliaferro, and Thomas Clark for help with linguistic properties. We thank Jin Li for providing the language VTC parcel. Greta Tuckute was supported by a Friends of McGovern Graduate Fellowship and the K. Lisa Yang ICoN Center Graduate Fellowship. Ev Fedorenko was supported by research funds from the McGovern Institute for Brain Research, the Department of Brain and Cognitive Sciences, MIT's Quest for Intelligence, and a grant from the Simons Foundation to the Simons Center for the Social Brain at MIT. Kendrick Kay acknowledges funding from National Institutes of Health grant (NIH) R01EY034118.

This extended abstract is based on the preprint of Tuckute et al. (2025).

References

Boyce, V., Futrell, R., & Levy, R. P. (2020). Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, *111*, 104082.

Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, *39*(39), 7722-7736.

Fedorenko, E., Blank, I. A., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, *203*, 104348.

Hu, J., Small, H., Kean, H., Takahashi, A., Zekelman, L., Kleinman, D., ... & Fedorenko, E. (2023). Precision fMRI reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cerebral Cortex*, *33*(8), 4384-4404.

Jain, S., Vo, V., Mahto, S., LeBel, A., Turek, J. S., & Huth, A. (2020). Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. *Advances in Neural Information Processing Systems*, *33*, 13738-13749.

Lipkin, B., Tuckute, G., Affourtit, J., Small, H., Mineroff, Z., Kean, H., ... & Fedorenko, E. (2022). Probabilistic atlas for the language network based on precision fMRI data from> 800 individuals. *Scientific data*, *9*(1), 529.

MacSweeney, M., Woll, B., Campbell, R., McGuire, P. K., David, A. S., Williams, S. C., ... & Brammer, M. J. (2002). Neural systems underlying British Sign Language and audio-visual English processing in native users. *Brain*, 125(7), 1583-1593.

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.

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022). Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *Elife*, *11*, e77599.

Regev, T. I., Casto, C., Hosseini, E. A., Adamek, M., Ritaccio, A. L., Willie, J. T., ... & Fedorenko, E. (2024). Neural populations in the language network differ in the size of their temporal receptive windows. *Nature Human Behaviour*, *8*(10), 1924-1942.

Rodd, J. M., Longe, O. A., Randall, B., & Tyler, L. K. (2010). The functional organisation of the fronto-temporal language system: evidence from syntactic and semantic ambiguity. *Neuropsychologia*, *48*(5), 1324-1335.

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.

Tuckute, G., Lee, E. J., Ou, Y., Fedorenko, E., & Kay, K. (2025). A two-dimensional space of linguistic representations shared across individuals. *bioRxiv*, 2025-05.

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

Wehbe, L., Blank, I. A., Shain, C., Futrell, R., Levy, R., von der Malsburg, T., ... & Fedorenko, E. (2021). Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. *Cerebral Cortex*, *31*(9), 4006-4023.