Title

Modeling the Structure of the Human Semantic System

Permalink

https://escholarship.org/uc/item/9ck823q4

Author

Chen, Catherine

Publication Date

2024

Modeling the Structure of the Human Semantic System

By

Catherine Chen

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jack Gallant, Co-chair
Professor Daniel Klein, Co-chair
Professor Steven Piantadosi

Summer 2024

Modeling the Structure of the Human Semantic System

Abstract

Modeling the Structure of the Human Semantic System

by

Catherine Chen

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Jack Gallant, Co-chair

Professor Daniel Klein, Co-chair

As humans, we use language throughout our everyday lives. When we use language our brains perform complex processes, which involve perceiving sensory inputs, interpreting linguistic structures, and accessing semantic memory. The human brain has evolved to perform these tasks efficiently and across various contexts: we communicate through different sensory modalities, languages, and levels of abstraction. How does the human brain support language use? Answering this question will deepen our knowledge of human cognition, which in turn can improve diagnoses of language disorders, enhance language education strategies, and inform the development of more flexible artificial language systems.

Prior work suggests that language use engages a network of interacting brain regions. However, it remains unclear how these networks represent the multifaceted aspects of language processing, and how they adapt to the diversity of contexts in which we use language.

This dissertation presents three neuroimaging studies of how the human brain represents language across different contexts. The first experiment (Chapter 2) compares brain representations between two different languages: English and Chinese. This experiment shows that shared semantic representations are systematically modulated by each language to create language-dependent representations. The second experiment (Chapter 3) compares brain representations between different sensory modalities: reading and listening. The results show that representations of language are shared between different sensory modalities, suggesting that pathways for language integration may be shared between modalities. The third experiment (Chapter 4) compares how concepts and relations are represented in the brain, and suggests that the same neural processes may be used to represent both relations and concepts. Together, these three studies show how the human brain encodes language across diverse contexts, and highlight the intricacy, dynamism, and flexibility of the human semantic system.

To my grandparents, whose resilience and curiosity have always inspired me.

Contents

List of Figures

List of Tables

Acknowledgments

The work in this dissertation would not have been possible without the support of many people. I am incredibly fortunate to have worked with people who not only conduct brilliant research, but are also unfailingly humble, curious, and generous with their time. To each of you I am deeply thankful.

To my wonderful advisors Jack Gallant and Dan Klein, for making all of this possible, for providing the freedom to explore interesting research questions, helping me realize what I am capable of, and guiding me along the way. Most of all, thank you for believing in me and giving encouragement when I most needed it. To Fatma Deniz, for helping me to become a better scientist, mentor, and teacher. I am glad that we could work together on so many projects, and I am so excited for you and your new lab. To Sara Popham, for mentoring me throughout my rotation and for many years afterwards, and for introducing me to the world of encoding models, experiment design, and the Gallant Lab. To Tom Dupré la Tour, for your patience with all of my questions, for our collaboration on the timescales project, and for sharing with me your unfailingly clear grasp of signal processing and statistical modeling. To Alicia Zeng, Bram Supriyatno, Cheol Jun Cho, Christine Tseng, Lily Gong, Mathis Lamarre, and Michele Winter, thank you for being fantastic collaborators on projects ranging from bilingual language comprehension to fixing GPUs in lab. To Doug Downey, Gabi Stanovsky, Kyle Lo, and Shannon Shen, for being welcoming and supportive hosts at AI2, and for showing me an intriguing area of research. To Alane Suhr, Steve Piantadosi, and Terry Regier, for sharing your time and ideas, and for your valuable feedback on many different stages of my projects. Thank you to all the other members of the Gallant Lab and Berkeley NLP group who have made this such an enriching and fun environment: Amanda, Carolyn, Emily, Jen, Matteo, Michael, Storm, Tianjiao, Charlie, Daniel, David, Eric, Eve, Jessy, Kayo, Kevin, Mitchell, Nick, Nikita, Rudy, Ruiqi, Sanjay, and Steven. All of this work would not have been possible without the extraordinarily capable people in BAIR and in the EECS department, especially Angie Abbatecola, Roxana Infante, Ami Katagiri, Audrey Sillers, Jean Nguyen, Shirley Salanio, and Susanne Kauer.

To my pre-PhD mentors, thank you for mentoring me and helping me grow as a researcher. To Ken Norman, Chris Baldassano, Qihong Lu, Andre Beukers, and Elad Hazan, thank you for taking me under your wing as an undergraduate, for being incredibly kind mentors, and for inspiring much of my interest in relational representations and NLP. To Moritz Grosse-Wentrup, thank you for welcoming me into your lab, for hosting me for a year in Germany, and for teaching me about causal inference.

Thank you to the friends and family who have supported me throughout graduate school. Thank you especially to Albert, Allison, Alok, Alex, Andrew, Catherina, Chloe, Daniel, Eric, Ivy, Jackey, Jeffrey, Jessica, Heesu, Kevin, Matt, Nate, Nick, Patrick, Rochelle, Roy, Sam, Sally, and Vanessa, whose humor and perspective have been a constant source of joy. To Kevin and Rudy, with whom I have been lucky to share my life both in and out of the lab, thank you for being the best PhD buddies I could have asked for. The Berkeley marina and fire trails (and all those who biked and ran there with me!) have kept me grounded over the past five years. Many of the analyses and sentences in this dissertation were conceptualized either at, or en route to, one of those two places, and the natural beauty of Berkeley has never failed to inspire on even the foggiest day. To Jerry, I am thankful that we finally found each other after many near misses, and to have such a kind and thoughtful partner. Thank you for bringing so much happiness into my life. Finally, to Lucas, Mom, and Dad, thank you for being a constant source of love and support, and for always believing in me. Thank you for making me who I am today.

Chapter 1

Introduction

As humans, we use natural language to understand and communicate about the world. To use language, we perform complex, interacting processes: we transform low-level sensory inputs into word-level representations, integrate information over long strings of words in order to extract higher-level meaning, and connect this meaning to our stored knowledge of the world. Importantly, our ability to use language is highly flexible. We can comprehend language through different modalities, such as in text and speech; communicate in languages with different orthographies and phonologies; and use language to convey different aspects of a situation. How does the human brain support our complex yet flexible ability to use language? In this dissertation, I investigate this question through a series of neuroimaging studies.

Early neuroimaging work about how the human brain processes language often performed controlled contrast studies (Binder et al., 2009). These studies presented participants with sets of carefully controlled stimuli that differed in axes such as the types of concept categories described (Friederici et al., 2000). They recorded brain responses to each condition within the set, and then contrasted the magnitude of recorded responses between sets in order to determine whether certain parts of the brain are more strongly activated by a particular condition.

More recent work has highlighted the importance of using naturalistic experiments in addition to the carefully controlled stimuli of early work. These naturalistic stimuli are more complex and engaging, and elicit brain responses that better reflect how the brain processes language in the real world (Hamilton & Huth, 2020; Nastase et al., 2020). However, naturalistic stimuli pose methodological challenges. These stimuli elicit representations of many different aspects of language processing, and the stimuli vary at a rate that is much faster than the time resolution of many methods for recording brain responses. Thus, methods such as contrast studies are insufficient to disentangle the brain representations that are evoked by naturalistic stimuli. A methodological innovation referred to as the *encoding model framework* allows us to address these challenges. In the encoding model framework, we construct numerical *feature spaces* that reflect aspects of a stimulus that are hypothesized to be represented in the brain. Then we estimate encoding models that predict brain responses from those stimulus features, and use these models to test our hypotheses about where and how each aspect of the stimulus is represented in the brain. This approach provides detailed, ecologically valid descriptions of how the brain represents different aspects of language. Prior studies that used encoding models to analyze brain responses to naturalistic language stimuli have shown that language processing engages a large network of brain regions that span much of the temporal, parietal, and prefrontal cortices (Deniz et al., 2019; Huth et al., 2016; Mitchell et al., 2008; Popham et al., 2021, e.g.,). These studies have produced detailed maps of how semantic information in language is represented in the brain, and how these representations are organized with respect to the representations evoked by other modalities. In the following chapters I build upon this work to provide further insight into the brain representations of language.

Chapter 2 investigates how the brain represents semantic information across different languages. Participants who are fluent in both Chinese and English read the same narratives in both languages while their brain responses were recorded with functional magnetic resonance imaging (fMRI). I modeled brain representations of lexical semantics for each language separately, and developed methods to characterize how these representations shift between the two languages. The results show that lexical semantic representations are largely shared between the two languages, but there are fine-grained shifts that systematically alter semantic representations between the two languages. This study suggests that bilingual language comprehension relies on shared semantic representations that are modulated by each language.

Chapter 3 moves beyond representations of lexical semantic information, and investigates representations of higher-level information that is conveyed across tens, hundreds, and even thousands of words. This study investigates how the brain represents these different levels of information (which we refer to as *timescales*) across different sensory modalities. Participants listened to and read the same narratives while their brain responses were recorded with fMRI. I modeled brain representations of different language timescales, and compared these representations between the two modalities. I find that at the word-level and above, the organization of timescale representations is strikingly consistent between the two modalities. This study suggests that after low-level sensory processing, language is integrated along the same pathways regardless of the sensory modality of the inputs.

Chapter 4 examines how the brain represents not only individual concepts, but also the semantic relations between them. Participants performed an active relation-processing task while brain responses were recorded with fMRI. I modeled brain representations of different relations, and compared these relation representations to concept representations that are evoked during passive language comprehension. I show that representations of relations are organized much like those of concepts: each relation is represented in distinct areas of the cortical surface, forming patterns that are consistent across individuals. This study suggests that the same neural processes may be used for representing both relations and concepts.

Chapter 2

Bilingual Language Processing Uses Shared Semantic Representations that are Modulated by Each Language

2.1 Abstract

Billions of people throughout the world are bilingual and can understand semantic concepts in multiple languages. However, there is little agreement about how the brains of bilinguals represent semantic information from different languages. Some theories suggest that bilingual speakers' brains contain separate representations for semantic information from different languages, while others suggest that different languages evoke the same semantic representations in the brain. To determine how the brains of bilinguals represent semantic information from different languages, we used functional magnetic resonance imaging (fMRI) to record brain responses while participants who are fluent in both English and Chinese read several hours of natural narratives in each language. We then used this data to specifically and comprehensively compare semantic representations between the two languages. We show that while semantic representations are largely shared between languages, these representations undergo fine-grained shifts between languages. These shifts systematically alter how different concept categories are represented in each language. Our results suggest that for bilinguals, semantic brain representations are shared across languages but modulated by each language. These results reconcile competing theories of bilingual language processing.

2.2 Introduction

Over 4 billion people throughout the world are bilingual (Ansaldo et al., 2008; Bot, 2003) and can comprehend semantic concepts in their primary and secondary languages. Brain representations of semantic concepts have been extensively studied for single languages (usually English). However, relatively little is known about how the brains of bilinguals represent semantic concepts across different languages. Some theories suggest that the brains of bilinguals contain separate representations for different languages (de Groot, 1992; Dehaene et al., 1997; Kim et al., 1997; Kroll & De Groot, 2005; MacNamara, 1967; Weinreich, 1986; Xu et al., 2017). These theories are supported by evidence that semantic brain representations are language-dependent: brain lesions can impair concept knowledge in one language but spare others (Gomez-Tortosa et al., 1995; Ku et al., 1996; Paradis, 1985), and the perceived emotional intensity or memorability of concepts can change between languages (Pavlenko, 2002; Schrauf & Rubin, 2004). A second group of theories suggests that different languages evoke the same semantic brain representations (Abutalebi & Green, 2007; Caramazza & Brones, 1980; de Groot, 1992; Grainger et al., 2010; Kroll & Stewart, 1994; Midgley et al., 2008; Potter et al., 1984; Weinreich, 1986). These theories are supported by evidence of interference between languages: second language acquisition can increase processing times for false cognates (Dijkstra et al., 1999; Duyck, 2005; van Hell & Dijkstra, 2002), and restructure concept categories in one’ s primary language (Malik-Moraleda et al., 2023).

A third possibility is that semantic representations for different languages are neither separate nor the same. Based on recent evidence that semantic representations can shift to emphasize task-relevant concepts (Çukur et al., 2013; Deniz et al., 2023; Kiremitçi et al., 2021; Nastase et al., 2017), we hypothesized that shared semantic representations are modulated by each language (Figure 2.2a). For example, language-dependent perceptions of emotional intensity (Pavlenko, 2002; Schrauf & Rubin, 2004) could arise from subtle shifts in brain representations of emotion-related concepts. This hypothesis would reconcile prior contradictory evidence for separate versus the same semantic representations between languages.

To test these three possibilities, we designed a study to compare semantic representations between languages. Six native Chinese speakers who are also fluent in English read natural narratives for over two hours in each language, while functional magnetic resonance imaging (fMRI) was used to record brain responses. Voxelwise modeling was used to estimate semantic brain representations in each language. These brain representations were then compared between the two languages. We found that semantic representations are largely shared between languages, but there are systematic differences between languages. Our results suggest that shared semantic brain representations are modulated by each language.

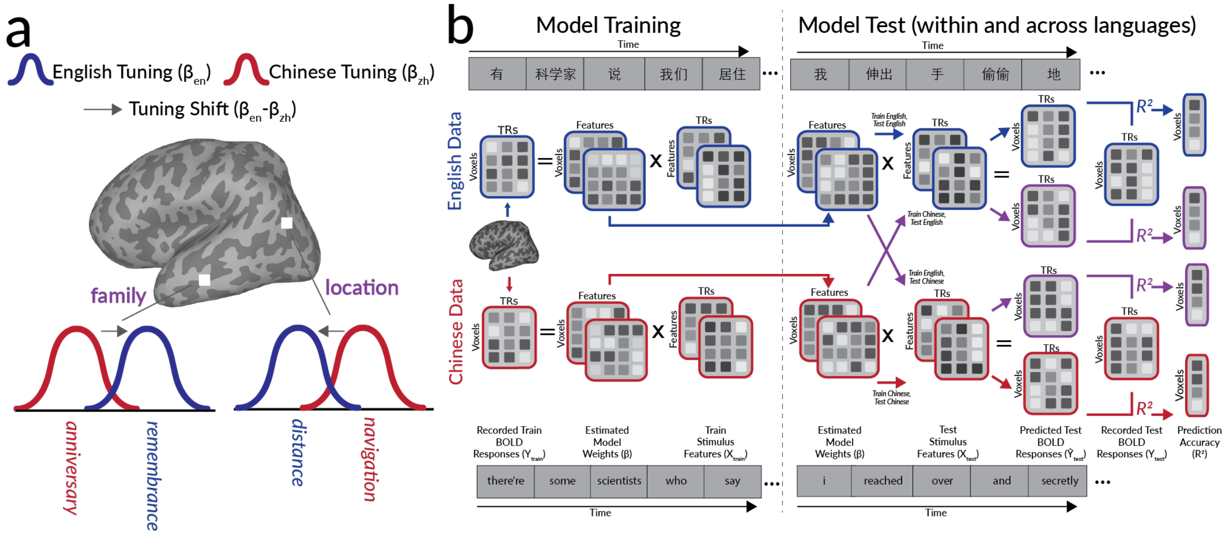
a
English Tuning (β_{en})
Chinese Tuning (β_{zh})
Tuning Shift (β_{en}-β_{zh})
family
location
anniversary
remembrance
distance
navigation
b
Model Training
Model Test (within and across languages)
Time
有
科学家
说
我们
居住
我
伸出
手
偷偷
地
English Data
Chinese Data
TRs
Voxels
Features
Train English, Test English
Train Chinese, Test English
Train English, Test Chinese
Train Chinese, Test Chinese
R^2
Recorded Train BOLD Responses (Y_{train})
Estimated Model Weights (β)
Train Stimulus Features (X_{train})
Estimated Model Weights (β)
Test Stimulus Features (X_{test})
Predicted Test BOLD Responses ($\hat{Y}_{test}$)
Recorded Test BOLD Responses (Y_{test})
Prediction Accuracy (R^2)
there're
some
scientists
who
say
i
reached
over
and
secretly

Figure 2.1: Hypothesis and experimental procedure. **a**. Schematic illustrating hypothesized semantic representation shifts. The semantic tuning of a voxel describes its preference for each semantic concept. Shifts in semantic representations were quantified as the change in semantic tuning between languages. For each voxel, blue and red curves respectively denote semantic tuning in English and Chinese. Arrows represent semantic tuning shifts from Chinese to English. The hypothetical voxel in the parietal cortex represents location-related concepts in both languages, but emphasizes number-related aspects (“distance”) in English, and action-related aspects (“navigation”) in Chinese. The hypothetical voxel in temporal cortex represents family-related concepts in both languages, but emphasizes emotion-related aspects (“remembrance”) in English, and number-related aspects (“anniversary”) in Chinese. **b**. Experiment and modeling procedure. Six fluent Chinese-English bilingual participants read over two hours of narratives in each language while BOLD responses were measured using fMRI. Semantic stimulus features were constructed by projecting each stimulus word into a 300-dimensional embedding space (Bojanowski et al., 2017). Ridge regression was used to estimate encoding models that describe voxelwise semantic tuning for each participant and language. Estimated model weights were used to predict BOLD responses to held-out narratives not used for model estimation. Model weights estimated in one language were used to predict BOLD responses to the same language (red and blue arrows; *within-language*) and to the other language (purple arrows; *across-language*). Prediction accuracy was quantified as the coefficient of determination (CD; R^2 between predicted and recorded BOLD responses.

2.3 Results

Narrative stories were presented to six Chinese-English bilinguals while fMRI was used to record BOLD responses. Each narrative was presented in both English and Chinese as written text. Words were presented one at a time at a natural reading rate. Each participant read narratives for over two hours per language. Semantic stimulus features were extracted by projecting each word into an embedding space (fastText (Bojanowski et al., 2017; Joulin et al., 2018)) in which words in different languages that express similar concepts project to nearby vectors. (A separate embedding space (Devlin et al., 2019) produced similar results; Figures 2.6-2.6). Regularized regression was used to estimate voxelwise encoding models separately for each participant and language (Figure 2.2b). Nuisance features such as word rate and spatiotemporal visual features were regressed out of BOLD responses prior to model estimation. The estimated model weights describe semantic tuning at the highest spatial resolution available in the data. Estimated model weights were compared between English and Chinese to examine whether and how semantic representations differ between languages. To evaluate model accuracy, estimated model weights were used to predict voxel responses to held-out test data. To validate the results and ensure generalization to new participants, data for participants P5 and P6 were not analyzed until the entire analysis pipeline was finalized.

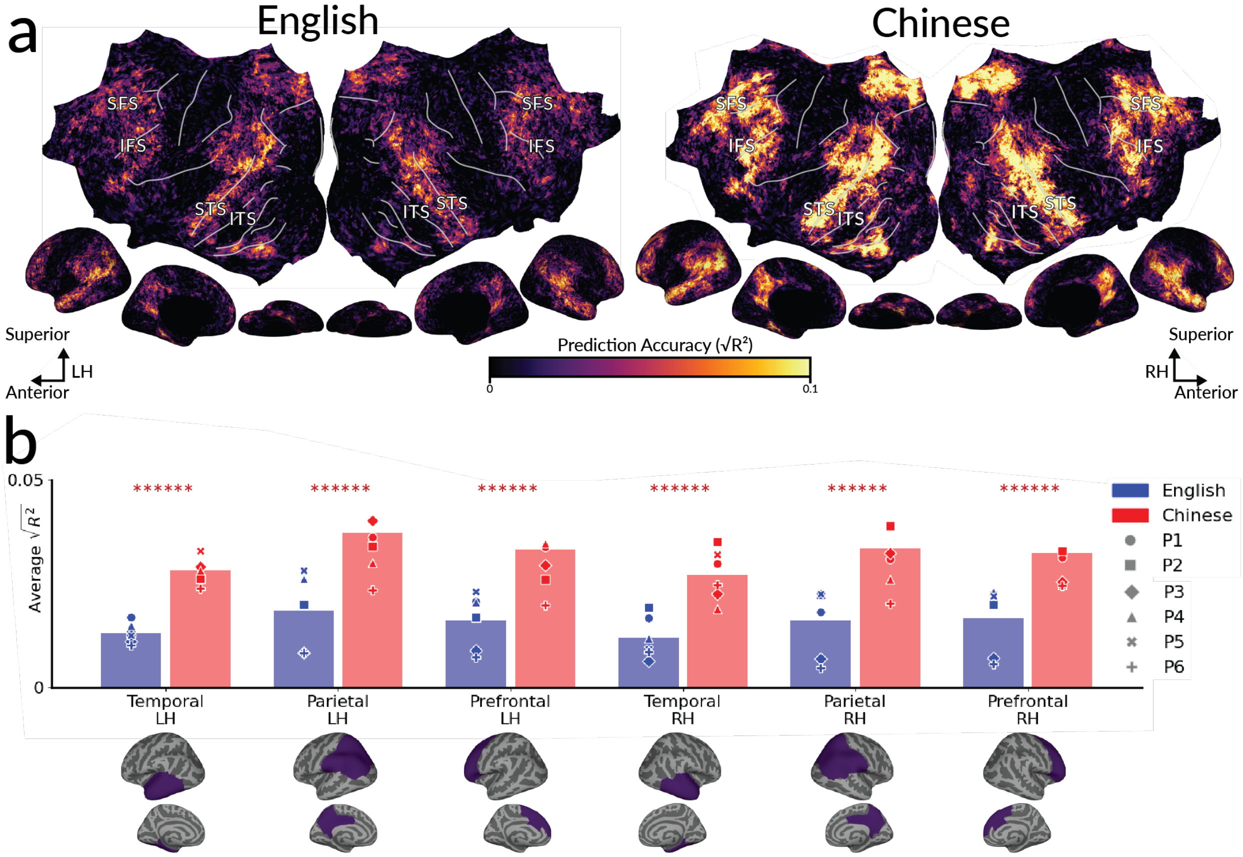

Figure 2.2: Cortical distribution of semantic representations for each language. To determine where semantic information is represented for each language, voxelwise models estimated for each language were used to predict held-out data for the same language. Prediction accuracy was computed as the CD (R^2) between predicted and recorded BOLD responses. a. Group-level prediction accuracy. Results are shown for each language on the flattened corti-

cal surface of the template space. For both languages, prediction accuracy is highest in the bilateral temporal, parietal, and prefrontal cortices. Prediction accuracy in these regions is statistically significant for each participant (Figure 2.6). The same brain regions are well-predicted for both languages. (SFS=superior frontal sulcus; IFS=inferior frontal sulcus; STS=superior temporal sulcus; ITS=inferior temporal sulcus, LH=left hemisphere; RH=right hemisphere) b. Prediction accuracy by cortical region. For each brain region and participant, blue and red markers show the mean prediction accuracy over voxels for English and Chinese respectively. Bars show the mean across participants. Red asterisks denote the number of participants for which prediction accuracy is significantly higher in Chinese than in English (one-sided p<.05 by a permutation test). While the same brain regions are well-predicted for both languages, the semantic model explains more variability in brain responses for Chinese than for English.

Theories that different languages evoke the same or shared semantic brain representations present two predictions. First, semantic information should be represented in the same brain regions across languages. Second, semantic tuning within these regions should be similar between languages. To determine which brain regions represent semantic information for each language, we examined where model weights estimated for each language could predict held-out test data for the same language. Prediction accuracy was computed as the coefficient of determination (CD; R^2) between predicted and recorded BOLD responses for each voxel, participant, and language separately. Group-level results were computed by projecting voxelwise accuracies for each participant to a template space (fsAverage (Fischl et al., 1999)), and then averaging the projected values across participants for each vertex and language separately. Figure 2.3a shows vertexwise group-level prediction accuracy for each language separately. Results for each participant are similar to the group (Figures 2.6 and 2.6). For each language, prediction accuracy is highest within bilateral temporal, parietal, and prefrontal cortices. These regions are sometimes referred to as the *semantic system* (Binder et al., 2009; Huth et al., 2016). These results show that the same brain regions represent semantic information for both languages.

While the same brain regions are well-predicted for both languages, visual inspection of Figure 2.3a indicates that prediction accuracy is overall higher in Chinese than in English. To quantify this difference for each brain region in the semantic system, we used FreeSurfer (Desikan et al., 2006) regions of interest (ROIs) to identify voxels in each region, and then computed average prediction accuracy separately for each language and region. Figure 2.3b shows the average prediction accuracy for each language, region, and participant. Prediction accuracy is significantly greater in Chinese than in English (one-sided p<.05 by a permutation test for all brain regions and participants).

There are two potential explanations for lower English prediction accuracy. First, semantic representations may constitute a lower proportion of overall brain responses to English. For example, non-semantic aspects of language processing (e.g., high-level control) may more strongly influence brain responses to English compared to Chinese. Alternatively, the total amount of explainable signal could be lower in English, such as if participants attended less strongly to English than to Chinese stimuli (Bressler & Silver, 2010). To distinguish between these possibilities, for each language we compute the total amount of explainable signal (*noise-ceiling*), as well as the proportion of total explainable signal that is predicted by the semantic model (*noise-ceiling corrected prediction accuracy*) (Hsu et al., 2004; Sahani & Linden, 2002; Schoppe et al., 2016). Across participants, the noise-ceiling was not lower in English than in Chinese, but the noise-ceiling corrected prediction accuracy was significantly lower in English than in Chinese (Figures 2.6 and 2.6), suggesting that semantic representations constitute a lower proportion of overall brain responses to English.

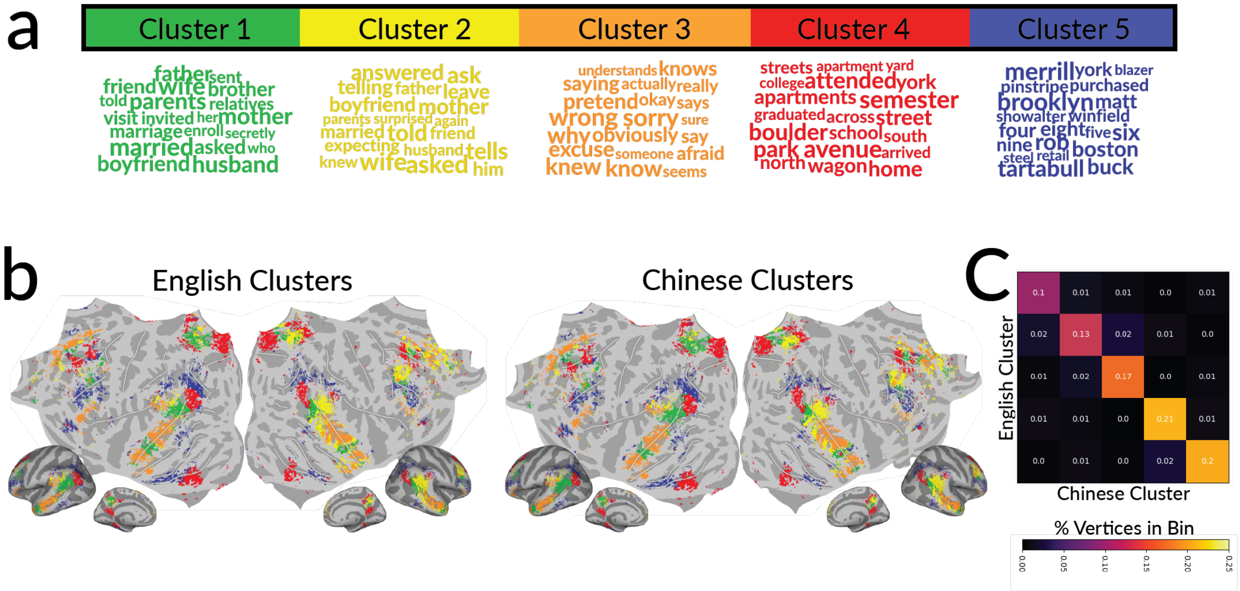

Figure 2.3: Shared semantic representations across languages. Hierarchical clustering on estimated model weights was used to categorize voxels based on semantic tuning in each language. a. Words closest to each cluster. The five clusters represent concepts related to family (Cluster 1, green), communication (Cluster 2, yellow), cognition (Cluster 3, orange), locations (Cluster 4, red), and numbers/names (Cluster 5, blue). b. Cortical distribution of semantic clusters. Group-level results are shown for each language. Vertex color reflects the assigned cluster. Poorly-predicted vertices are shown in grey. C. Confusion matrix of group-level cluster assignments over well-predicted vertices. Cluster assignments match between languages for 81% of vertices (one-sided p<.05, by a permutation test). This shows that semantic representations are largely shared between languages.

Figure 2.3 shows that the same brain regions represent semantic information in both languages. However, within these regions semantic representations could differ between languages. For example, a voxel could activate in response to emotion-related concepts in one language and to location-related concepts in the other. This voxel would represent semantic information in both languages, but the semantic tuning of the voxel would differ between languages. To determine whether semantic tuning is shared between languages, we classify voxels into semantic clusters based on semantic tuning in each language, and then evaluate whether cluster assignments match between languages.

First, we used a cross-validated clustering approach (*model connectivity* (Meschke et al., 2023)) to identify semantic clusters from the estimated model weights. Five clusters best summarized the distribution of model weights across participants and languages (Figure 2.6). To interpret each cluster, we identified the English stimulus words that are closest to each cluster. Distance between a word and a cluster was computed as the Pearson correlation between the word embedding and the cluster centroid. Figure 2.3a lists the closest words to each cluster. The clusters categorize voxels into concepts related to family (Cluster 1), communication (Cluster 2), cognition (Cluster 3), locations (Cluster 4), and numbers/names (Cluster 5).

Then for each participant and language separately, each voxel was assigned to the semantic cluster with the lowest Euclidean distance between the cluster centroid and the voxel' s model weight. To summarize results across participants, we projected model weights for each participant and language to the template space. Then for each vertex of the template space and for each language separately, we computed the mean model weights across participants, and used the mean model weights to assign each vertex to one of the five clusters. Figure 2.3b shows cluster assignments at the group-level. Cluster assignments are shown for vertices that were well-predicted in both languages ($\sqrt{R^2} > 0.1$) in at least one participant. Visual inspection of Figure 2.3b suggests that cluster assignments are consistent between languages. To quantify this consistency, we computed the confusion matrix between cluster assignments in English and Chinese. Figure 2.3c shows the confusion matrix for group-level cluster assignments. For 81% of well-predicted vertices cluster assignments match between languages. For each participant, cluster assignments match between languages for over 70% of well-predicted voxels (Figure 2.6). As a converging test for shared semantic representations between languages, we measured whether model weights estimated in one language could predict voxel responses to the other language (across-language prediction accuracy). Models estimated for English accurately predicted brain responses to Chinese and vice versa throughout bilateral temporal, parietal, and prefrontal cortices (Figures 2.6 and 2.6). Overall, these results show that semantic representations are largely shared between languages.

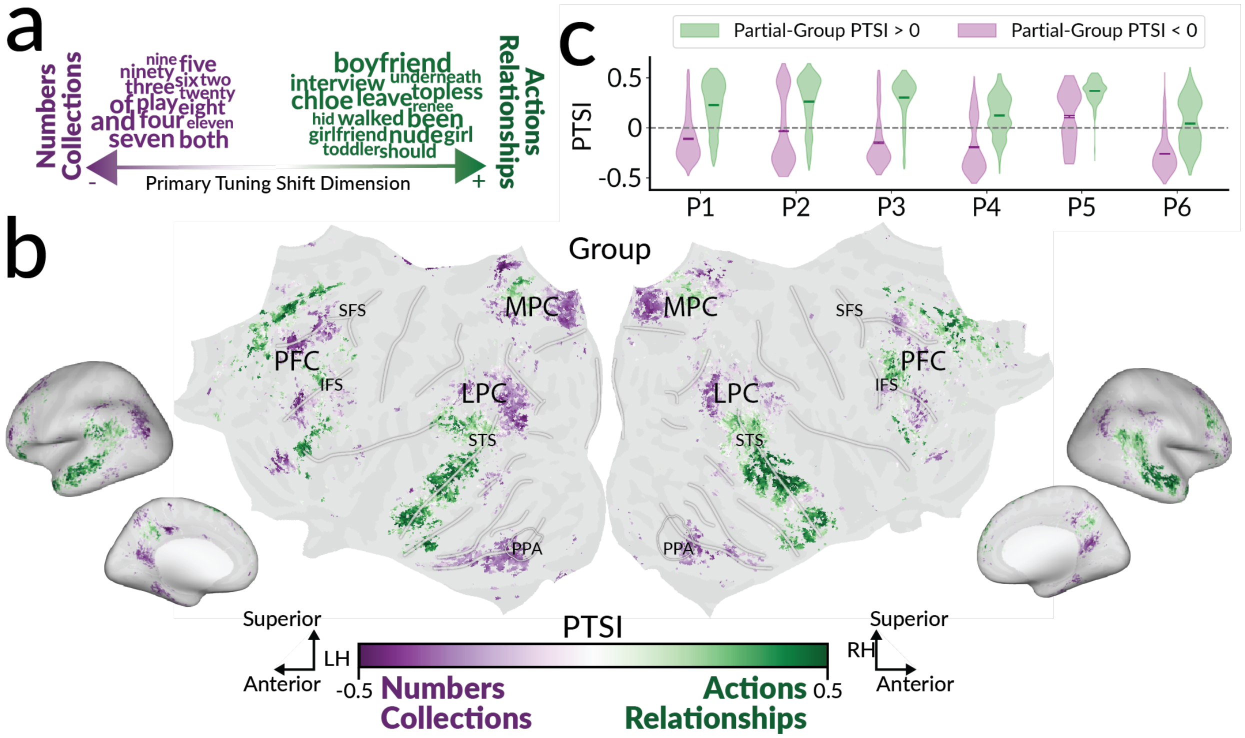

Figure 2.4: Cortical distribution of semantic tuning shifts between languages. Voxelwise *semantic tuning shift* was defined as the change in model weights between English and Chinese: $semantic\_tuning\_shift = \frac{\beta_{en}}{||\beta_{en}||_2} - \frac{\beta_{zh}}{||\beta_{zh}||_2}$. For each voxel, the semantic tuning shift describes which concepts elicit higher BOLD responses in one language relative to the other. The main dimensions of voxelwise semantic tuning shifts were identified using PCA. The first tuning shift PC, which we refer to as the *primary semantic tuning shift dimension* (PSSD), reliably explains variance in semantic tuning shifts across voxels and participants (Figure 2.6). a. Interpretation of the PSSD. Words for which embeddings are most negatively correlated with the PSSD are shown in purple. Words for which embeddings are most positively correlated with the PSSD are shown in green. The negative end of the PSSD emphasizes number/collection-related semantics (purple), while the positive end emphasizes action/relationship-related semantics (green). b. Cortical distribution of semantic tuning shifts. The direction of voxelwise semantic tuning shifts was summarized with the *primary tuning shift index* (PTSI), which is the Pearson correlation between a voxel' s semantic tuning shift and the PSSD. Group-level PTSI is shown on the flattened cortical surface of the template space. Vertices shown in purple shift towards the negative end of the PSSD. These vertices emphasize number/collection-related semantics in English, and emphasize action/relationship-related semantics in Chinese. These vertices are found in bilateral lateral parietal cortex (LPC), fusiform gyrus and near parahippocampal place area (PPA), superior and inferior medial parietal cortex (MPC), and middle frontal cortex (FC). Vertices shown in green shift toward the positive end of the PSSD. These vertices emphasize action/relationship-related semantics in English, and number/collection-related semantics in Chinese. These vertices are found in bilateral STS, LPC, middle MPC, superior frontal gyrus (SFG), and inferior frontal gyrus (IFG). Poorly predicted vertices are shown in grey. c. Consistency of PTSI between participants. For each participant, the other five participants were used

to compute a partial-group estimate of vertexwise PTSI. Violinplots show the distribution of vertexwise PTSI for each participant, separately for vertices in which PTSI is negative (purple violinplots) and positive (green violinplots) in the partial-group. Vertices with positive PTSI in the partial-group also have positive PTSI in the participant ($p<.05$ by a one-sided t-test after Fisher z-transformation), and vertices with negative PTSI in the partial-group also have negative PTSI in the participant ($p<.05$ by a one-sided t-test after Fisher z-transformation; except P5). Thus, the cortical distribution of PTSI is consistent between participants. Overall, there are systematic shifts in semantic tuning throughout the semantic system.

Figures 2.3 and 2.3 suggest that semantic representations are shared between languages. However, given strong behavioral evidence for language-dependent semantic representations (Pavlenko, 2002; Schrauf & Rubin, 2004), we hypothesized that shared semantic representations are modulated by each language (Figure 2.2a). For instance, a voxel might respond to the same concept category (location-related concepts) in both languages, but the semantic tuning of the voxel may subtly shift such that the voxel exhibits greater activation for concepts associated with actions (“navigation”) in English, and for concepts associated with numbers (“distance”) in Chinese.

To determine whether the shared semantic representations shown in Figure 2.3 are modulated by each language, we investigate the change in estimated model weights between languages. We refer to the change in estimated model weights: $\frac{\beta_{en}}{||\beta_{en}||_2} - \frac{\beta_{zh}}{||\beta_{zh}||_2}$ as the voxelwise *semantic tuning shift*. The semantic tuning shift describes which concepts are emphasized in each language relative to the other for each voxel. However, estimated model weights are affected both by the semantic tuning of the voxel and random measurement noise. Thus, for each individual voxel, differences in estimated model weights partially reflect measurement noise. To isolate true shifts in semantic tuning, we focus on the dimensions of estimated semantic tuning shifts that are reliable across voxels and participants.

To identify reliable dimensions of semantic tuning shifts, we used a leave-one-participant-out procedure. For each of the six participants, we concatenated voxelwise semantic tuning shifts from the other five participants and then use principal component analysis (PCA) to obtain 300 orthogonal axes (principal components; PCs) that are sorted by the ratio of variance explained. We evaluated how well each PC explains variance in semantic tuning shifts for the left-out participant. The chance rate was defined as the variance explained by the primary dimensions of semantic tuning within each language. The top semantic tuning shift PC, which we refer to as the *primary semantic tuning shift dimension* (PSSD), reliably explains more variance than chance (Figure 2.6). To obtain an estimate of the PSSD that incorporates data from all six participants, we applied PCA to voxelwise semantic tuning shifts concatenated over all six participants. The PSSD is the main dimension along which semantic representations shift between languages.

To interpret the PSSD, we identified the semantic concepts that correspond to each end of the PSSD. Figure 2.3a shows the English stimulus words for which embeddings are the most positively or negatively correlated with the PSSD. The most negatively correlated words (colored in purple) are related to numbers and collections (e.g., “three” , “both”

). The most positively correlated words (colored in green) are related to actions and human relationships (e.g., “leave” , “boyfriend”). (Interpretation of the PSSD based on Chinese stimulus words is similar and is shown in Figure 2.6.) Thus, semantic tuning shifts towards the negative end of the PSSD emphasize number/collection-related semantics, while shifts towards the positive end emphasize action/relationship-related semantics.

To visualize the cortical distribution of semantic tuning shifts, we examine the direction of semantic tuning shifts along the PSSD. We refer to the Pearson correlation between a voxel’ s semantic tuning shift vector and the PSSD as the *primary tuning shift index* (PTSI). Group-level PTSI was computed by projecting voxelwise PTSI for each participant into the template space and then computing the average over participants for each vertex. Figure 2.3b shows group-level vertexwise PTSI on the flattened surface of the template space. Vertices with negative PTSI (shown in purple) emphasize number/collection-related semantics in English, and action/relationship-related semantics in Chinese. These vertices are found in bilateral lateral parietal cortex (LPC), fusiform gyrus and near parahippocampal place area (PPA), superior and inferior medial parietal cortex (MPC), and middle frontal cortex (FC). Vertices with positive PTSI (shown in green) emphasize action/relationship-related semantics in English, and number/collection-related semantics in Chinese. These vertices are found in bilateral STS, LPC, middle MPC, superior frontal gyrus (SFG), and inferior frontal gyrus (IFG). To quantify the consistency of PTSI across participants, we compared PTSI between participants in the template space. We held out each of the six participants in turn, and used the other five participants to compute a partial-group estimate of PTSI for each vertex. Then we compared vertexwise PTSI between each participant and the partial-group. Figure 2.3c shows the distribution of vertexwise PTSI for each individual participant, separately for vertices in which PTSI is negative and positive in the partial-group. (For each participant we only include vertices that were well-predicted in both languages; $\sqrt{R^2} > 0.1$). Vertices with positive PTSI in the partial-group also have positive PTSI for each individual participant ($p<.05$, by a one-sided t-test after Fisher z-transformation). Vertices with negative PTSI in the partial-group also have negative PTSI for each individual participant ($p<.05$, by a one-sided t-test after Fisher z-transformation; except P6). The distribution of PTSI is consistent between participants.

To ensure that estimated semantic tuning shifts are not biased by idiosyncrasies of the fastText embedding space, we replicated our analyses with a separate semantic embedding space based on a multilingual language model (mBERT (Devlin et al., 2019)) that differs from fastText in its training objectives, training data, and embedding dimensionality (Figures 2.6-2.6).

To ensure that estimated semantic tuning shifts do not merely reflect misalignments in word embeddings between languages, we show that the directions of word embedding misalignment do not explain the estimated semantic tuning shifts (Figure 2.6). Overall, the results in Figure 2.3 show that there are systematic semantic tuning shifts between languages.

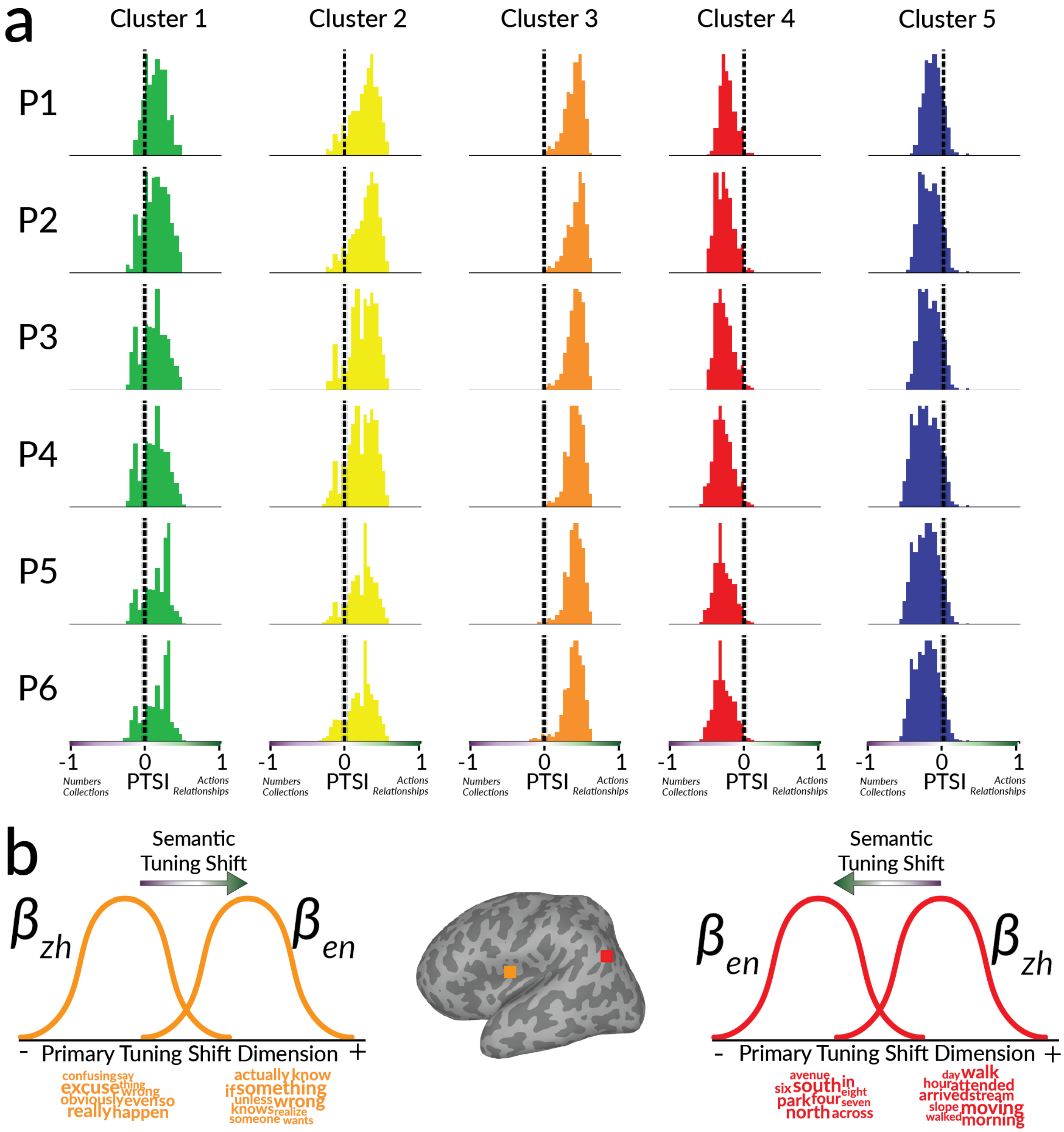

Figure 2.5: Semantic tuning shifts for different semantic clusters. Voxels were categorized into the semantic clusters shown in Figure 2.3. The distribution of semantic tuning shifts was examined for each semantic cluster. a. Semantic tuning shifts for each cluster. Histograms show the distribution of voxelwise PTSI for each cluster and participant separately. Voxels in the family-, communication-, and cognition-related clusters (Clusters 1, 2, and 3) have positive PTSI ($p<.05$ by a two-sided t-test after Fisher z-transformation). Thus, representations of family-,

communication-, and cognition-related concepts shift to emphasize action/relationship-related semantics in English as compared to Chinese. Voxels in the location- and number/name-related clusters (Cluster 4 and 5) have negative PTSI (p<.05 by a two-sided t-test after Fisher z-transformation). Thus, representations of location- and name-number-related concepts shift to emphasize number/collection-related semantics in English as compared to Chinese. b. Semantic tuning shift for two selected voxels. The voxel in frontal cortex (colored in orange) has positive PTSI. This voxel represents concepts related to cognition (Cluster 3) in both languages, but emphasizes action/relationship-related aspects in English (e.g., "know") compared to Chinese (e.g., "really"). The voxel in parietal cortex (colored in red) voxel has negative PTSI. This voxel represents concepts related to locations (Cluster 4) in both languages, but emphasizes number/collection-related aspects in English (e.g., "four") compared to Chinese (e.g., "moving"). These two examples illustrate how semantic tuning is modulated between languages.

Figure 2.3 shows that the semantic tuning of individual voxels shifts between languages. To determine whether semantic tuning shifts systematically modulate concept representations, we tested whether voxels that are tuned towards similar concepts shift in a consistent direction. For each of the clusters shown in Figure 2.3, we examined the direction of semantic tuning shifts for voxels in that semantic cluster. Figure 2.3a shows the PTSI for each cluster and participant, when semantic clusters are identified using the English model weights. Results are consistent when semantic clusters are identified using Chinese model weights (Figure 2.6). For voxels in Clusters 1, 2, and 3 PTSI is positive in all participants (p<.05 for each cluster by a two-sided t-test after Fisher z-transformation). Thus, representations of family, communication, and cognition-related concepts emphasize action/relationship-related aspects in English and number/collection-related aspects in Chinese. For voxels in Clusters 4 and 5 PTSI is negative in all participants (p<.05 for each cluster by a two-sided t-test after Fisher z-transformation). Thus, representations of location- and number/name-related concepts emphasize number/collection-related aspects in English and action/relationship-related aspects in Chinese.

Figure 2.3b illustrates semantic tuning shifts for two selected voxels. One voxel was selected from Cluster 3 and has positive PTSI. This voxel represents cognition-related concepts in both languages, but emphasizes action/relationship-related aspects in English (e.g., "know") compared to Chinese (e.g., "really"). The other voxel was selected from Cluster 4 and has negative PTSI. This voxel represents location-related concepts in both languages, but emphasizes numeric aspects in English (e.g., "four") compared to Chinese (e.g., "moving"). These two examples show how a voxel can represent the same semantic category between languages, but also exhibit semantic tuning shifts between languages. Overall, the results in Figure 2.3 show that each language systematically modulates semantic representations.

2.4 Discussion

This study compares semantic representations in the brain between English and Chinese in fluent Chinese-English bilinguals and provides unique evidence for how the brain represents semantic information in different languages. First, temporal, parietal, and prefrontal cortices represent semantic information in both languages (Figures 2.3, 2.6, 2.6, 2.6, 2.6). Second, semantic representations are largely similar between languages (Figures 2.3, 2.6, 2.6, 2.6). Third, there are systematic shifts in voxelwise semantic tuning between languages (Figures 2.3, 2.3, 2.6, 2.6, 2.6). Our results generalize across participants, including the two held-out participants (P5 and P6). Taken together, these results suggest that in bilinguals shared semantic brain representations are modulated by each language.

Prior neuroimaging work provided mixed evidence as to whether brain responses to different languages are different (Buchweitz, Mason, Hasegawa, & Just, 2009; Dehaene et al., 1997; Honey et al., 2012; Kim et al., 1997) or the same (Buchweitz et al., 2012; Chee et al., 1999; Illes et al., 1999; Klein et al., 1995; Luke et al., 2002; Malik-Moraleda et al., 2022). However, prior studies suffered from three limitations. First, prior studies did not explicitly model semantic representations. Thus, it is unclear whether previously reported results reflect semantic representations or other aspects of language processing. Second, many prior studies involved different participants for different languages (Dunagan et al., 2022; Honey et al., 2012; J. Li et al., 2022; Xu et al., 2017). Thus, individual differences in brain function and anatomy (Fedorenko & Kanwisher, 2009) may have exaggerated reported differences. Third, many prior studies used controlled stimuli. Thus, it was unclear whether reported results generalize to naturalistic settings (Hamilton & Huth, 2020). In this study, we explicitly modeled semantic representations, performed within-participants comparisons, used naturalistic stimuli, and evaluated the generalizability of our results to held-out participants. These contributions enabled us to identify previously unknown shifts in semantic representations between languages.

Our results provide a brain-based explanation that reconciles behavioral evidence of shared and different semantic representations between languages. First, we show that semantic representations in the brain are largely shared. Thus, knowledge of two languages can easily affect each other. This explains how linguistic phenomena such as false cognates (Dijkstra et al., 1999; Duyck, 2005; van Hell & Dijkstra, 2002) can generate interference effects in bilinguals. Second, we show that shifts in semantic tuning systematically modulate brain representations. Thus, the perceived meaning of words can change between languages. This explains behavioral evidence that perceptions of concepts such as numbers and emotions change between languages (Dehaene et al., 1999; Pavlenko, 2002; Schrauf & Rubin, 2004; Spelke & Tsivkin, 2001).

The results and methodology presented here reconciles competing theories of bilingual semantic processing, and will enable future studies of semantic tuning shifts across additional languages as well as over the course of language acquisition.

2.5 Methods

Stimuli

Narrative transcription, translation, and preprocessing

The stimuli consisted of eleven 10- to 15 min narratives from *The Moth Radio Hour*. In each narrative, a speaker tells an autobiographical story in front of a live audience. The selected narratives cover a wide range of topics and have been used in previous studies (de Heer et al., 2017; Deniz et al., 2019; Huth et al., 2016; LeBel et al., 2023). The audio recording of each narrative was manually transcribed, and the written transcription was aligned to the audio recording. (Details of audio transcription and alignment are described in prior work (Deniz et al., 2019)).

The original narratives were performed verbally in English. To construct matched Chinese stimuli, each of the English narratives was translated into Chinese by a professional translation service. To obtain word presentation times that correspond to natural speech, each translated narrative was read aloud by a professional voice actor. Then the written translations were aligned to these audio recordings. Chinese stimuli were presented with simplified Chinese characters (简化字).

Stimulus train and test split

Ten train narratives were used for model estimation, and one held-out test narrative was used for model evaluation. The same test narrative was used for both languages. To obtain noise-ceiling estimates, the test narrative was played to each participant four times in each language. (For P1, the test narrative was played only twice in Chinese due to a change in stimulus design after the first collected sessions.)

Stimulus presentation format

The words of each narrative were presented one-by-one at the center of the screen using a Rapid Serial Visual Presentation (RSVP) procedure (Forster, 1970). Each word was presented for a duration equal to the duration of that word in the spoken version of the narrative.

Each word was presented at the center of the screen in isolation, and a white fixation cross was present at the center of the display throughout the experiment. Participants were asked to fixate on a center cross while reading the narrative. Participants’ eye movements were monitored at 60 Hz throughout the scanning sessions using a custom-built camera system equipped with an infrared source (Avotec) and the View-Point EyeTracker software suite (Arrington Research). The eye tracker was calibrated at the end of each run of data acquisition.

Functional MRI data were collected during four 3-hour scanning sessions that were performed on different days. Each scanning session consisted of seven functional runs. Two of these runs presented the test narrative in a single language. The remaining five runs presented five different training narratives. The language of the training narratives was interleaved across runs.

All participants read all the narratives in both English and Chinese. Narrative presentation order was balanced between languages and across participants.

To verify comprehension and attention, at the end of each session participants were asked outside the scanner to recount the contents of each narrative. All participants were able to accurately summarize the contents of each narrative.

fMRI data acquisition

Whole-brain MRI data were collected on a 3T siemens TIM trio scanner at the UC Berkeley Brain Imaging Center. A 32-channel Siemens volume coil was used. Functional scans were collected using a T2*-weighted gradient-echo EPI with repetition time (TR) 2.0045s, echo time (TE) 35ms, flip angle 74°, voxel size 2.24x2.24x4.1 mm (slice thickness 3.5mm with 18% slice gap), matrix size 100x100, and field of view 224x224 mm. Thirty axial slices were prescribed to cover the entire cortex and were scanned in interleaved order. A custom-modified bipolar water excitation radiofrequency (RF) pulse was used to prevent contamination from fat signals. Anatomical data were collected using a T1-weighted multi-echo MP-RAGE sequence on the same 3T scanner.

To stabilize head motion during scanning sessions, participants wore a personalized head case that precisely fit the shape of each participant’ s head (Gao, 2015; Power et al., 2019).

fMRI data pre-processing

Each functional run was motion-corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FSL (Jenkinson et al., 2012). All volumes in the run were averaged across time to obtain a high quality template volume. FLIRT was used to automatically align the template volume for each run to the overall template, which was chosen to be the temporal average of the first functional run for each participant. These automatic alignments were manually checked and adjusted as necessary to improve accuracy. The cross-run transformation matrix was then concatenated to the motion-correction transformation matrices obtained using MCFLIRT, and the concatenated transformation was used to resample the original data directly into the overall template space. Noise from motion, respiratory, and cardiac signals were removed with a component-based detrending method (CompCor (Behzadi et al., 2007)). Responses were z-scored separately for each voxel and narrative. During z-scoring, the mean response across time was subtracted and the remaining response was scaled to have unit variance. Before data analysis, 10 TRs from the beginning and 10 TRs at the end of each narrative were discarded in order to account for the 10 seconds of silence at the beginning and end of each scan and to account for non-stationarity in brain responses at the beginning and end of each scan.

Cortical surface reconstruction and visualization

Cortical surface meshes were generated from the T1-weighted anatomical scans using FreeSurfer software (Fischl et al., 1999). Before surface reconstruction, anatomical surface segmentations were carefully hand-checked and corrected using Blender software and pycortex (Community, 2018; Gao et al., 2015). Relaxation cuts were made into the surface of each hemisphere. Blender and pycortex were used to remove the surface crossing the corpus callosum. The calcarine sulcus cut was made at the horizontal meridian in V1 using retinotopic mapping data as a guide.

Functional images were aligned to the cortical surface using pycortex. Functional data were projected onto the surface for visualization and analysis using the line-nearest scheme in pycortex. This projection scheme samples the functional data at 32 evenly spaced intervals between the inner (white matter) and outer (pial) surfaces of the cortex and then averages together the samples. Samples are taken using nearest-neighbor interpolation, wherein each sample is given the value of its enclosing voxels.

Cortical parcellation

FreeSurfer ROIs were used to anatomically localize the temporal, parietal, and prefrontal regions for each participant. ROIs were based on the Desikan-Killiany atlas (Desikan et al., 2006). The ROIs used for the temporal region were *"bankssts", "inferiortemporal", "middletemporal", "superiortemporal", "temporalpole", "transversetemporal", "fusiform", "entorhinal", "parahippocampal"*. The ROIs used for the parietal region were *"inferiorparietal", "superiorparietal", "supramarginal", "precuneus", "isthmuscingulate", "posteriorcingulate"*. The ROIs used for the prefrontal region were *"caudalmiddlefrontal", "parsopercularis", "parsorbitalis", "parstriangularis", "rostralmiddlefrontal", "superiorfrontal", "frontalpole", "caudalanteriorcingulate"*.

Localizers for known ROIs

Known ROIs were localized separately in each participant using a visual category localizer and a retinotopic localizer (Hansen et al., 2007; Spiridon et al., 2006). Details of localizer experiments are provided in prior work (Deniz et al., 2019; Huth et al., 2016).

Participants

Functional data were collected from six participants: P1 (29F), P2 (25M), P3 (25F), P4 (25M), P5 (24F), P6 (26M). Participant P1 was an author of this paper. Language proficiency of each participant was evaluated by the Language Experience and Proficiency Questionnaire (LEAP-Q) (Kaushanskaya et al., 2020) and the Language History Questionnaire (LHQ3) (P. Li et al., 2020). All participants were fluent in both Mandarin Chinese (native) and English (non-native). Participants began English language acquisition between the ages of

2 and 11, and spent between 5 and 12 years in a country where English is spoken. At the time of the experiment, participants primarily used Chinese in interactions with family, English in interactions at school/work, and a mix of the two languages in interactions with friends. Please see Supplementary Table 2.1 for additional details of participants’ use of each language. All participants were healthy and had normal or corrected-to-normal vision. All subjects were right handed or ambidextrous according to the Edinburgh handedness inventory (laterality quotient of -100: entirely left-handed, +100: entirely right-handed) (Oldfield, 1971a). Laterality scores were +5, 0, +90, +100, +65, +65 for P1-6 respectively.

Statistical analysis

Voxelwise modeling (VM) was used to model BOLD responses (de Heer et al., 2017; Deniz et al., 2019; Huth et al., 2016; Naselaris et al., 2011; M. C.-K. Wu et al., 2006). In the VM framework, stimulus and task parameters are nonlinearly transformed into sets of features (also called feature spaces) that are hypothesized to be represented in brain responses. Linearized regression is used to estimate a separate encoding model for each voxel and feature space. Each encoding model describes how a feature space is represented in the BOLD response of a voxel. A held-out dataset that was not used for model estimation is used to evaluate model prediction accuracy and to determine the significance of the model prediction accuracy.

All model fitting and analysis was performed using custom software written in Python, making heavy use of NumPy (C. R. Harris et al., 2020)], SciPy (Virtanen et al., 2020), Matplotlib (Hunter, 2007), Himalaya (Dupré la Tour et al., 2022), and Pycortex (Gao et al., 2015).

Construction of semantic feature spaces

To capture the semantic content of the stimulus narratives, each word of the stimulus narrative was projected to a 300-dimensional embedding space (Bojanowski et al., 2017). Embedding spaces were constructed separately for English and for Chinese. The embedding spaces were orthogonally transformed to align the English and Chinese embedding spaces, such that there is a high correlation between embeddings of words in different languages that reflect the same concept (Joulin et al., 2018).

Because of non-isometry between embedding spaces of different languages and limitations to multilingual alignment procedures (Søgaard et al., 2018), embedding spaces are not perfectly aligned between languages. For example two words that are direct translations of each other (e.g., “table” and “桌子” [*table*]) may not project to exactly the same vector. To ensure that results are not specific to the idiosyncrasies of a particular embedding space or a particular imperfection in cross-lingual embedding alignments, all analyses were replicated with a different embedding space, multilingual BERT (mBERT, *bert-base-multilingual-cased* (Devlin et al., 2019)). mBERT is a twelve-layer contextual language model that was jointly trained on text from 104 languages. No explicit cross-lingual alignment object was included during the training of mBERT, but embeddings in some layers are implicitly aligned over the course of training (Pires et al., 2019). To obtain word embeddings from mBERT, each sentence of the stimulus narratives was provided as input to mBERT and then the 768-dimensional activation of layer nine was used as an embedding of each word. Layer 9 was chosen because it produces the best aligned embeddings (Figure 2.6), and because intermediate layers of contextual language models have been shown to produce the most accurate predictions of brain responses (Caucheteux & King, 2022; Chen, Dupré la Tour, et al., 2024; Lamarre et al., 2022; Schrimpf et al., 2021; Toneva & Wehbe, 2019).

Construction of low-level feature spaces

Seven low-level feature spaces were constructed to account for the effects of low-level stimulus information on BOLD responses. For both languages, models were fit with feature spaces that reflect visual spatial and motion features (motion energy) (Adelson & Bergen, 1985; Nishimoto et al., 2011; Watson & Ahumada, 1985), word count, single phonemes, diphones, triphones (Gong et al., 2023), and intermediate level features that capture orthographic similarities by measuring the pixelwise overlap between words (Gong, 2024). For English, an additional low-level feature space reflected letter count. For Chinese, an additional low-level feature space reflected character count.

Stimulus feature space preprocessing

Before voxelwise modeling, each stimulus feature was truncated, downsampled, z-scored, and delayed. Data for the first 10 TRs and the last 10 TRs of each scan were truncated to account for the 10 seconds of silence at the beginning and end of each scan and to account for non-stationarity in brain responses at the beginning and end of each scan. An anti-aliasing, 3-lobe Lanczos filter with cut-off frequency set to the fMRI Nyquist rate (0.25 Hz) was used to resample the stimulus features to match the sampling rate of the fMRI recordings. Then the stimulus features were each z-scored in order to account for z-scoring performed on the MRI data (For details see Section 2.5). In the z-scoring procedure, the value of each feature channel was separately normalized by subtracting the mean value of the feature channel across time and then dividing by the standard deviation of the feature channel across time. Lastly, finite impulse response (FIR) temporal filters were used to delay the features in order to model the hemodynamic response function of each voxel. The FIR filters were implemented by concatenating feature vectors that had been delayed by 2, 4, 6, and 8 seconds (following prior work (Deniz et al., 2019; Huth et al., 2016; P. Li et al., 2020)). A separate FIR filter was fit for each feature, participant, and language.

Voxelwise encoding model fitting

Voxelwise encoding models were estimated in order to determine which features are represented in each voxel. Each model consists of a set of regression weights that describes BOLD responses in a single voxel as a linear combination of the features in a particular feature space. Regression weights were estimated using banded ridge regression (Nunez-Elizalde et al., 2019). Unlike standard ridge regression, which assigns the same regularization parameter to all feature spaces, banded ridge regression assigns a separate regularization hyperparameter to each feature space. Banded ridge regression thereby avoids biases in estimated model weights that could otherwise be caused by differences in feature space distributions. Mathematically, for a train dataset with v voxels and n TRs, the m delayed feature spaces $F_i(X), i \in \{1, ..., m\}$ (each dimension p_i) were concatenated to form a feature matrix $F'(X)$ (dimension $\sum_i^m p_i \times n$). Then banded ridge regression was used to estimate a mapping B (dimension $v \times \sum_i^m p_i$) from $F'(X)$ to the matrix of voxel responses Y (dimension $v \times n$). B is estimated according to $\hat{B} = \arg\min_B ||Y - BF(X)||_2^2 + \lambda||CB||_2^2$. The diagonal matrix C of regularization hyperparameters for each feature space and each voxel is optimized over 10-fold cross-validation. See Section 2.5 for details.

Stepwise regression procedure

To remove confounds from stimulus correlations between semantics and low-level sensory stimulus features, a stepwise regression procedure was used. First banded ridge regression was used to jointly estimate encoding models that predict BOLD responses from the seven low-level stimulus features. Only data from the train narratives was used to estimate models. Then, the low-level models were used to predict BOLD responses to the train and test narratives. The predicted BOLD responses $\hat{Y}_{lowlevel,train}$ and $\hat{Y}_{lowlevel,test}$ were subtracted from the true BOLD responses Y_{train}, Y_{test}. The residual BOLD responses $Y_{train} - \hat{Y}_{lowlevel,train}$, $Y_{test} - Y_{lowlevel,test}$ were zscored. During z-scoring, for each voxel separately the mean response across time was subtracted and the remaining response was scaled to have unit variance. The z-scored residual BOLD responses were used to estimate encoding models that predict BOLD responses from semantic stimulus features.

Regularization hyperparameter selection

Five-fold cross-validation was used to find the optimal regularization hyperparameters for each feature space and voxel. Hyperparameter candidates were chosen with a random search procedure (Bergstra & Bengio, 2012): 1000 normalized hyperparameter candidates were randomly sampled from a dirichlet distribution and were then scaled by 21 log-spaced values ranging from 10^{-10} to 10^{10}. The regularization hyperparameters for each feature space and voxel were selected as the hyperparameters that produced the minimum squared error (L2) loss between the predicted voxel responses and the recorded voxel responses ($\arg\min_{hyperparameters} ||\hat{y} - y||_2^2$). Regularization hyperparameters were chosen separately for each participant and language. Hyperparameter search was performed using the Himalaya Python package (Dupré la Tour et al., 2022).

Model estimation and evaluation

The selected regularization hyperparameters were used to estimate model weights that map from the semantic feature space to voxel BOLD responses. Model weights were estimated separately for each voxel, language, and participant. The model weights for each voxel and language reflect the semantic tuning of the voxel in that language.

The test dataset was not used to select hyperparameters or to estimate regression weights. The prediction accuracy R^2 of the feature spaces was computed per voxel as the coefficient of determination (CD) between the predicted voxel responses and the recorded voxel responses on the test dataset. To determine which voxels represent semantic information in each language, prediction accuracy was computed for within-language predictions (train and test on the same language). To determine how well model weights estimated for one language generalize to the other language, prediction accuracy was also computed for across-language predictions (train on one language, and test on a different language).

A permutation test with 1000 iterations was used to compute the statistical significance of prediction accuracy. In each permutation, the test responses were shuffled in blocks of 10 TRs (Chen, Dupré la Tour, et al., 2024; Deniz et al., 2019; Jain et al., 2020; Lamarre et al., 2022; LeBel et al., 2023; Oota et al., 2023; Reddy & Wehbe, 2021; Tang et al., 2024). Shuffling was performed in blocks of 10 TRs in order to preserve autocorrelations in voxel responses. Then the prediction accuracy (R^2) was computed between the predicted responses and the permuted test responses. The distribution of test accuracies over permutation iterations was used as a null distribution to compute the p-value of prediction accuracy for each voxel. A Benjamini-Hochberg correction for multiple comparisons was applied to the voxelwise p-values (Benjamini & Hochberg, 1995). Permutation tests were performed separately for each voxel, language, and participant.

Noise-ceiling correction was performed by normalizing the prediction accuracy of each voxel R^2 by the maximum possible prediction accuracy (*noise-ceiling*) (Hsu et al., 2004; Sahani & Linden, 2002; Schoppe et al., 2016). To compute the noise-ceiling, first the maximum explainable variance (*EV*, also referred to as *signal power*) is computed for each voxel. EV measures the consistency of measured BOLD responses over repeated stimulus presentations, and reflects the amount of response variance in the test data that could be explained by a perfect model. Formally, for a test dataset with N repeats of a T TR test narrative, and recorded BOLD responses $y_1 ... y_N \in R^T$ for a single voxel, EV is defined as follows (each y_i is first zscored across time):

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$EV = \frac{1}{N-1}(NVar\bar{Y} - \frac{1}{N}\sum_{i=1}^{N} Var(Y_i))$$

The noise-ceiling R^2_{max} is obtained by dividing the EV of each voxel by $Var(\bar{Y})$. The noise-ceiling corrected prediction accuracy R^2_{norm} is then obtained for each voxel by dividing the prediction accuracy R^2 by the noise-ceiling R^2_{max}. For very noisy voxels, the estimated noise-ceiling may be lower than the measured R^2 and therefore lead to divergent estimates of R^2_{norm}. We used a heuristic to correct for this divergence. We identified the set of voxels where R^2 is greater than R^2_{max}, selected the maximum R^2_{max} over these voxels, and then clipped R^2_{max} to be above this maximum value. Note that this heuristic results in a conservative estimate of the noise-ceiling corrected prediction accuracy.

Group-level prediction accuracy

Group-level prediction accuracy was computed by computing prediction accuracy for each participant in the participant's native brain space, and then projecting individual participant results into a template space (fsAverage (Fischl et al., 1999)). Average prediction accuracy across six participants was computed for each fsAverage vertex.

Generalization to new participants

To ensure generalization to new participants, two steps were performed. First, the entire analysis was performed at the individual participant level –group-averaged results are shown only as summary statistics. Second, before the final analyses were performed, two out of the six participants were set aside as held-out participants. The data for these participants were not analyzed until the data analysis and interpretation pipeline was finalized (Popham et al., 2021).

Voxel Selection for Tuning Analyses

To ensure that semantic tuning shift analyses were performed only on voxels that represent semantic information in both English and Chinese, all of the following model weight interpretation analyses were performed only on voxels that were well-predicted ($\sqrt{R^2} > 0.1$) in both English and Chinese.

Semantic Tuning Shifts

The semantic tuning shift of each voxel was used to describe how voxelwise semantic tuning changes between languages. First, the model weights were normalized for each language and voxel by dividing each 300-dimensional vector of model weights by the L2-norm of the vector. Then, the semantic tuning shift of each voxel was computed by subtracting the normalized Chinese model weights from the normalized English model weights. Formally, the semantic tuning shift for a voxel with weights β_{en} and β_{zh} was defined as $semantic\_tuning\_shift = \frac{\beta_{en}}{||\beta_{en}||_2} - \frac{\beta_{zh}}{||\beta_{zh}||_2}$.

The semantic tuning shift vector for each voxel describes how semantic tuning changes from Chinese to English. For example, for a voxel that becomes more tuned towards number-related semantics when the stimulus language changes from Chinese to English, the semantic tuning shift vector would point in the direction of number semantics in the embedding space.

Note that defining the semantic tuning shift as the change in tuning from Chinese to English ($\frac{\beta_{zh}}{||\beta_{zh}||_2} - \frac{\beta_{en}}{||\beta_{en}||_2}$) would change the sign of the semantic tuning shift vector but also swap the languages on each end of the vector. Thus, the choice to define semantic tuning shift as the shift from Chinese to English rather than from English to Chinese does not affect the reported results.

Dimensions of semantic tuning shift

Principal component analysis (PCA) was used to determine the main directions of semantic tuning shifts. Because model weights accurately describe semantic tuning only for well-predicted voxels, we only investigated semantic tuning shifts for voxels that were well-predicted in both languages (see Section 2.5 for details; selecting voxels based on significance instead of prediction accuracy produces similar results as shown in Figure 2.6). To increase the influence of better-predicted voxels on the estimated principal components (PCs), the semantic tuning shift of each voxel was scaled by the voxel’ s mean prediction accuracy across languages. The semantic tuning shift vectors were concatenated across participants and languages. PCA was applied to the concatenated semantic tuning shift vectors to find a set of orthogonal axes that best explain variance in voxelwise semantic tuning shift vectors. The PCs that explain the highest variance in voxelwise semantic tuning shift describe the primary semantic dimensions of semantic tuning shifts. We refer to the PC that explains the most variance in voxelwise semantic tuning shifts as the *primary semantic tuning shift dimension* (PSSD). To examine whether the semantic tuning shift for each voxel is towards the negative or positive end of the PSSD, the Pearson correlation was computed between the semantic tuning shift vector of each voxel and the PSSD. For each voxel we refer to this correlation as the *primary tuning shift index* (PTSI).

Weight Clustering

A clustering approach was used to separate voxels into groups that represent similar concepts (Meschke et al., 2023). Model weights for each participant and language were projected to a standard template space (fsAverage (Fischl et al., 1999)). Each projected model weight $\beta_{vertex,participant,language} \in \mathbb{R}^{300}$ was normalized to have unit L2-norm: $\beta_{normalized} = \frac{\beta}{||\beta||_2}$. The normalized model weights were averaged across participants and languages, and then hierarchical clustering was performed on the normalized group-averaged model weights. Clustering was only performed on vertices that were well-predicted in both languages (group-averaged $\sqrt{R^2_{en}} > 0.05$ and $\sqrt{R^2_{zh}} > 0.05$). In total, there are 12 sets of model weights across the six participants and two languages. The number of clusters was chosen based on a leave-one-out cross-validation procedure. One of the twelve sets of model weights was held out in each cross-validation fold. The remaining eleven sets of model weights were averaged across participants and languages for each vertex of the template space, and hierarchical clustering was used to obtain N group-level clusters, separately for N ranging from 2 through 25. For each number N of clusters, the group-level model weight clusters were used to predict BOLD responses for the held out participant and language in the template space. The cross-validation accuracy was computed for each vertex as the CD (R^2) between the predicted and recorded BOLD responses. Only vertices that were well-predicted in both languages (group-averaged $\sqrt{R^2_{en}} > 0.05$ and $\sqrt{R^2_{zh}} > 0.05$) were included in this analysis. The cross-validation score plateaus around five clusters (Figure 2.6). Thus, we chose to use five clusters for the analyses shown in Figure 2.3 and Figure 2.3. This clustering procedure resulted in five 300-dimensional cluster centroids. These centroids define clusters of voxels that are each tuned towards related semantic concepts. Voxels were assigned to clusters based on the Pearson correlation between voxelwise model weights and each of the 300-dimensional cluster centroids. Cluster assignments were performed separately for each participant and language.

2.6 Supplementary Figures

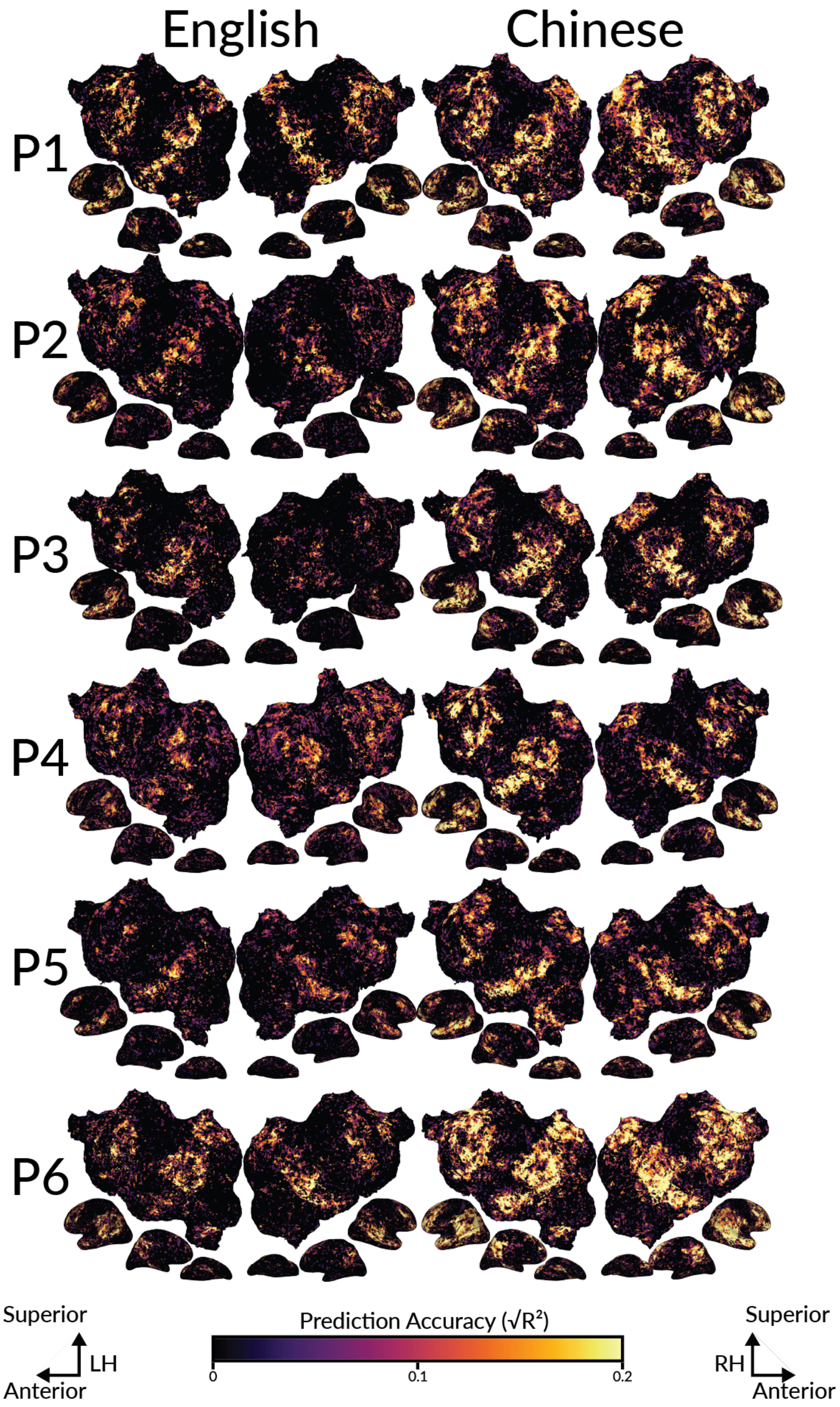

Figure 2.6: Cortical distribution of semantic representations for each language when mBERT is used as the semantic feature space. To validate the results shown in Figure 2.3 multilingual BERT (mBERT) was used instead of fastText as a semantic feature space. VM was used to estimate model weights that map from mBERT features to BOLD responses in each voxel and for each language separately. Estimated model weights for each language were used to predict voxelwise BOLD responses to a held-out dataset in the same language. Prediction accuracy was computed as the CD (R^2) between predicted and recorded BOLD responses. Prediction accuracy for each participant and language is shown on the flattened cortical surface of the participant' s native brain space. For both languages and in each participant the highest prediction accuracy (brightest voxels) is found within the bilateral temporal, parietal, and prefrontal cortices. This suggests that the same brain regions are well-predicted for both languages and that these results are not dependent on the specific semantic feature space that is used.

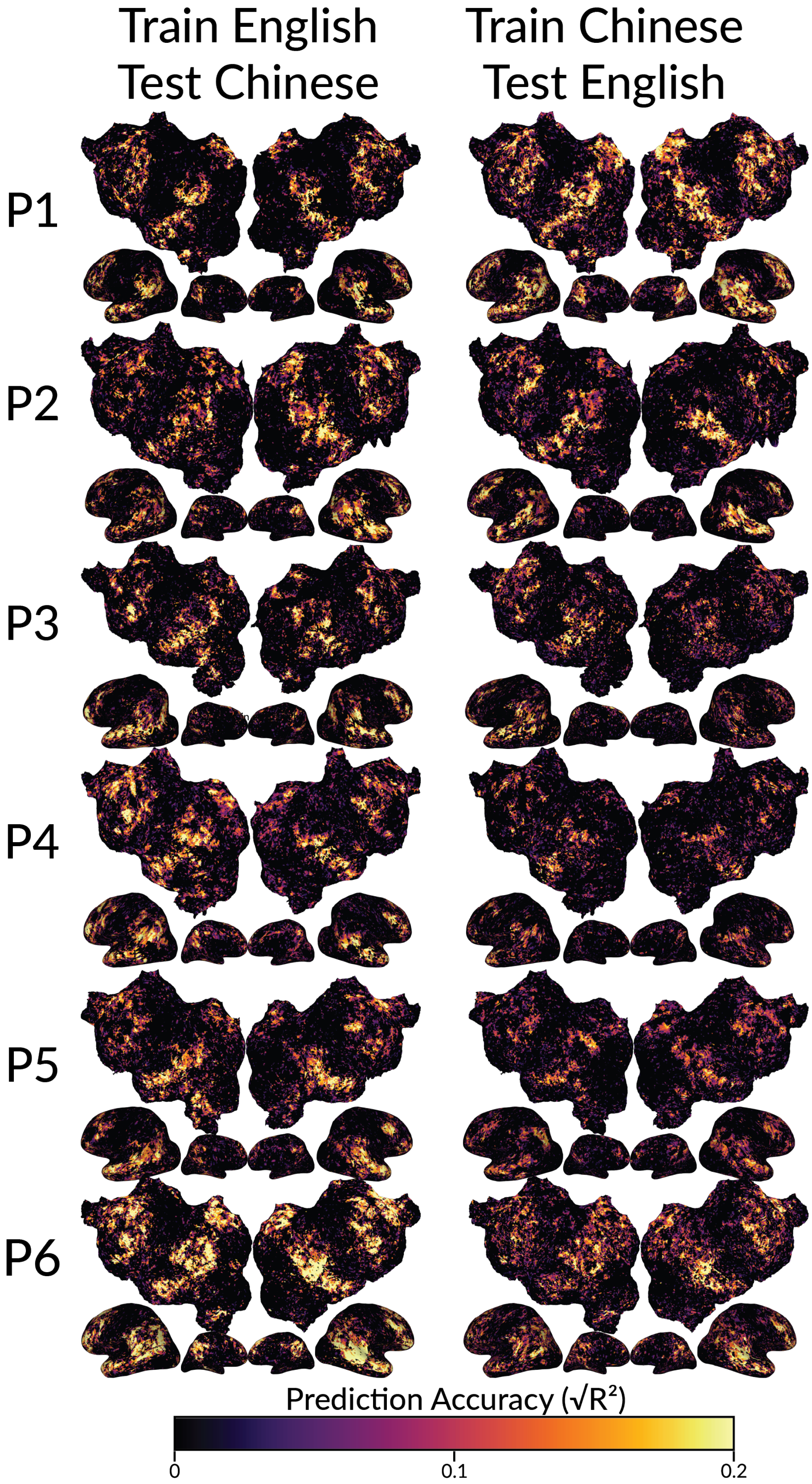
Train English
Test Chinese
Train Chinese
Test English
P1
P2
P3
P4
P5
P6
Prediction Accuracy (√R²)
0
0.1
0.2

Figure 2.7: Shared semantic representations between languages as shown by across-language prediction accuracy when mBERT is used as the semantic feature space. Across-language prediction accuracy is shown for each participant on the flattened cortical surface of the participant' s native brain space. Estimated voxelwise model weights in one language were used to predict the held-out dataset in the other language. Prediction accuracy was computed as the CD (R^2) between predicted and recorded BOLD responses. Prediction accuracy is given by the color scale. Well-predicted voxels appear brighter. In each participant, the semantic model estimated for one language accurately predicts voxel responses to the other language throughout the semantic system. Thus, semantic representations within the semantic system are largely shared between languages.

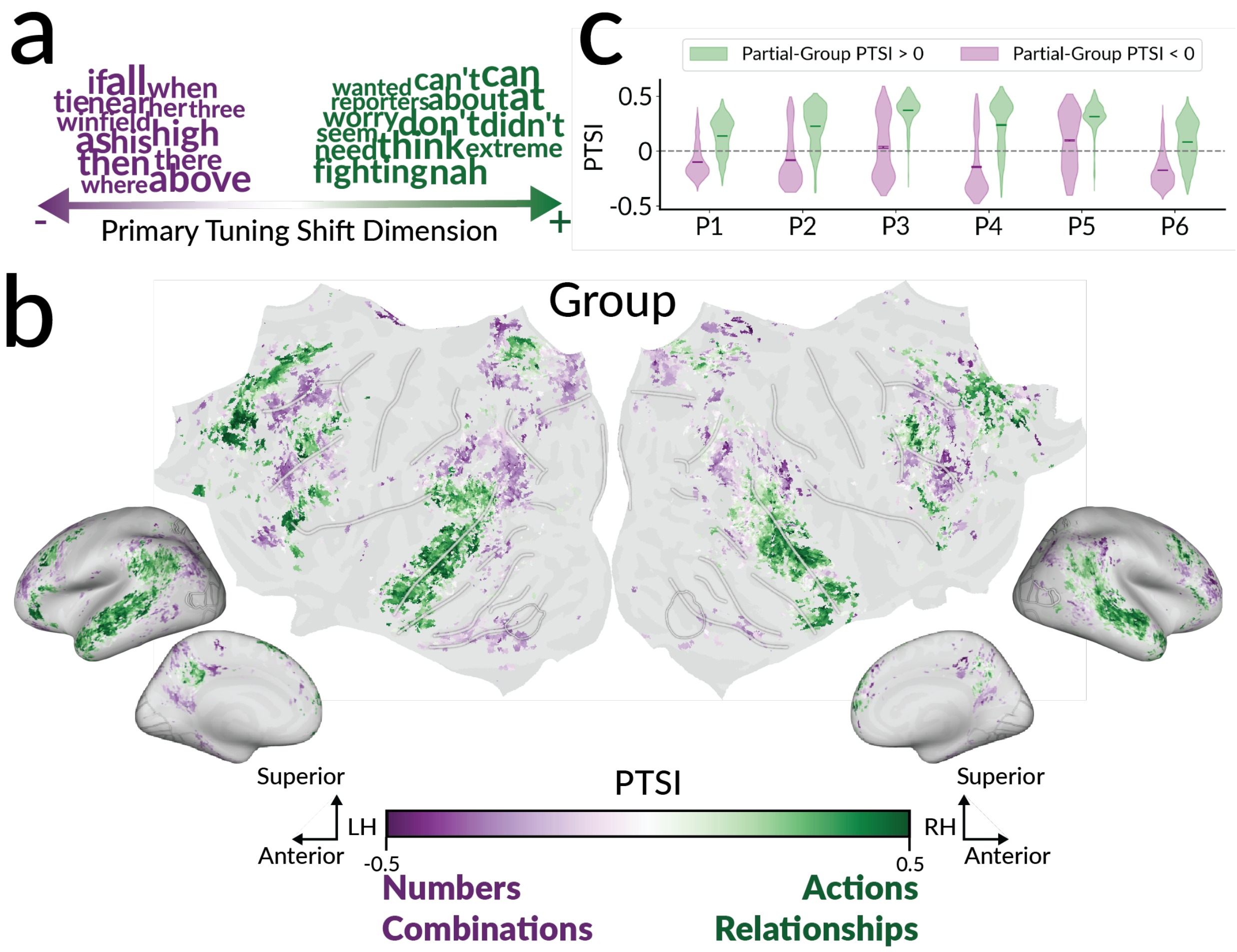

Figure 2.8: Cortical distribution of semantic tuning shifts between languages when mBERT is used as the semantic feature space. To validate the results in Figure 2.3 multilingual BERT (mBERT) was used as the semantic feature space. The semantic tuning shift of each voxel was defined as the change in model weights between English and Chinese ($semantic\_tuning\_shift = \frac{\beta_{en}}{||\beta_{en}||_2} - \frac{\beta_{zh}}{||\beta_{zh}||_2}$). For each voxel the semantic tuning shift describes which concepts elicit higher BOLD responses in one language relative to the other. Principal component analysis (PCA) was used to determine the main dimensions of voxelwise semantic tuning shifts. We refer to the first tuning shift PC as the *primary semantic tuning shift dimension* (PSSD). **a**. Interpretation of the PSSD. To identify words that were closest to each end of the PSSD, we took the contextual embedding of each word in the English stimulus and then identified the embeddings that had the most positive and most negative correlation with the PSSD. Naively comparing the contextual embedding of each stimulus word is heavily biased towards very frequent words, because the same word can have multiple contextual embeddings. Thus the top 1% of most frequent words were removed from this analysis. Words for which embeddings are negatively correlated with the PSSD are shown in purple. Words for which embeddings are positively correlated with the dimension are shown in green. Semantic tuning shifts that are negatively correlated with the PSSD emphasize number/direction-related

semantics (purple), while semantic tuning shifts that are positively correlated with the PSSD emphasize action-related semantics (green). The interpretation of the PSSD is broadly consistent with the results shown in Figure 2.3a for the fastText semantic feature space. **b**. The Pearson correlation between the semantic tuning shift vector and the PSSD was computed. We refer to this correlation as the *primary tuning shift index* (PTSI). PTSI is shown on the flattened cortical surface of the template space. Vertices shown in purple shift towards the negative end of the PSSD. Vertices shown in green shift toward the positive end of the PSSD. Vertices that were not well-predicted in both languages ($\sqrt{R^2} > 0.1$) in at least one participant are shown in grey. The cortical distribution of PTSI is consistent with the results shown in Figure 2.3b for the fastText model. **c**. Consistency in the cortical distribution of PTSI between each participant and the rest of the group. For each participant, the other five participants were used to compute a partial-group estimate of PTSI for each vertex. For each participant, green violin plot depicts the distribution of PTSI over vertices in which partial-group PTSI is positive, and the purple violin plot depicts the distributions PTSI over vertices in which partial-group PTSI is negative. For each participant vertices with positive PTSI tend to also have positive PTSI in the partial-group ($p<.05$ by a one-sided t-test after Fisher z-transformation, except P3 and P5) and vertices with negative PTSI tend to also have negative PTSI in the partial-group ($p<.05$ by a one-sided t-test after Fisher z-transformation). Thus, the cortical distribution of PTSI is consistent between participants. Overall, there are systematic semantic tuning shifts between languages that are consistent across participants. These results suggest that semantic representations are modulated by each language.

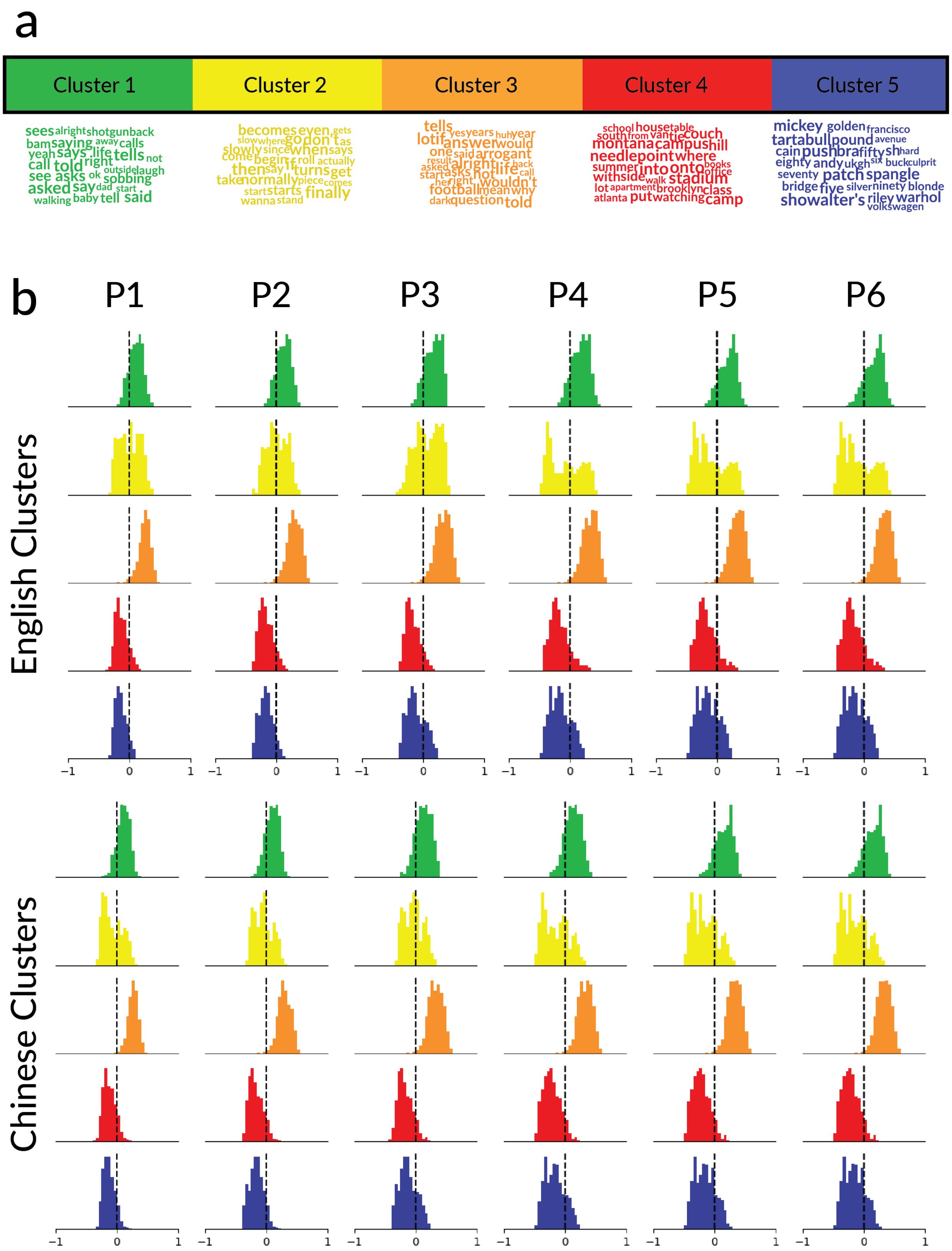
a
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
sees alright shotgun back bam saying away calls yeah says life tells not call told right outside laugh see asks ok sobbing asked say dad start walking baby tell said
becomes even gets slow where go don't as slowly since when says come begin roll actually then say if turns get take normally piece comes start starts finally wanna stand
tells yes years huh year lot if answer would one said arrogant result alright life back asked start asks not call her right wouldn't football mean why dark question told
school house table south from van tie couch montana campus hill needlepoint where summer into onto books office with side walk stadium lot apartment brooklyn class atlanta put watching camp
mickey golden francisco tartabull pound avenue cain push bra fifty sh hard eighty andy ukgh six buck culprit seventy patch spangle bridge five silver ninety blonde showalter's riley warhol volkswagen
b
P1
P2
P3
P4
P5
P6
English Clusters
Chinese Clusters
−1
0
1

Figure 2.9: Semantic tuning shifts per semantic cluster when mBERT is used as the semantic feature space. To validate the results in Figure 2.3 multilingual BERT (mBERT) was used as the semantic feature space in VM. For each participant and language separately, model weights were used to assign each voxel to one of five semantic clusters. a. The meaning of each cluster was determined by finding the stimulus words closest to the cluster centroid. The clusters represent semantics related to communication (Cluster 1, green), cognition (Cluster 3, orange), locations (Cluster 4, red), and numbers/names (Cluster 5, blue). One cluster (Cluster 2, yellow) represents miscellaneous adjectives. Clusters are semantically noisier than for fastText (Figure 2.3a), likely because contextual mBERT embeddings capture more syntactic information than lexical embeddings such as fastText. b. Voxels were assigned to clusters based on English or Chinese mBERT model weights. The direction of semantic tuning shift of each voxel in each cluster was computed using the PTSI metric which is the Pearson correlation between the semantic tuning shift of the voxel and the PSSD (as in Figure 2.3). Histograms indicate the distribution of PTSI values for voxels in each cluster and participant separately. Semantic tuning shifts for voxels in the communication-, and cognition-related clusters (Clusters 1 and 3) have positive PTSI. Thus, representations of communication-, and cognition-related concepts shift to emphasize action/relationship-related semantics in English as compared to Chinese. In contrast, semantic tuning shifts for voxels in the location- and number/name-related clusters (Cluster 4 and 5) have negative PTSI. Semantic tuning shifts for voxels in Cluster 2 (miscellaneous adjectives) are mixed. Overall, representations of location- and number/name-related concepts shift to emphasize number/collection-related semantics in English as compared to Chinese. These results suggest that voxels that represent similar semantic concepts shift in similar directions between languages.

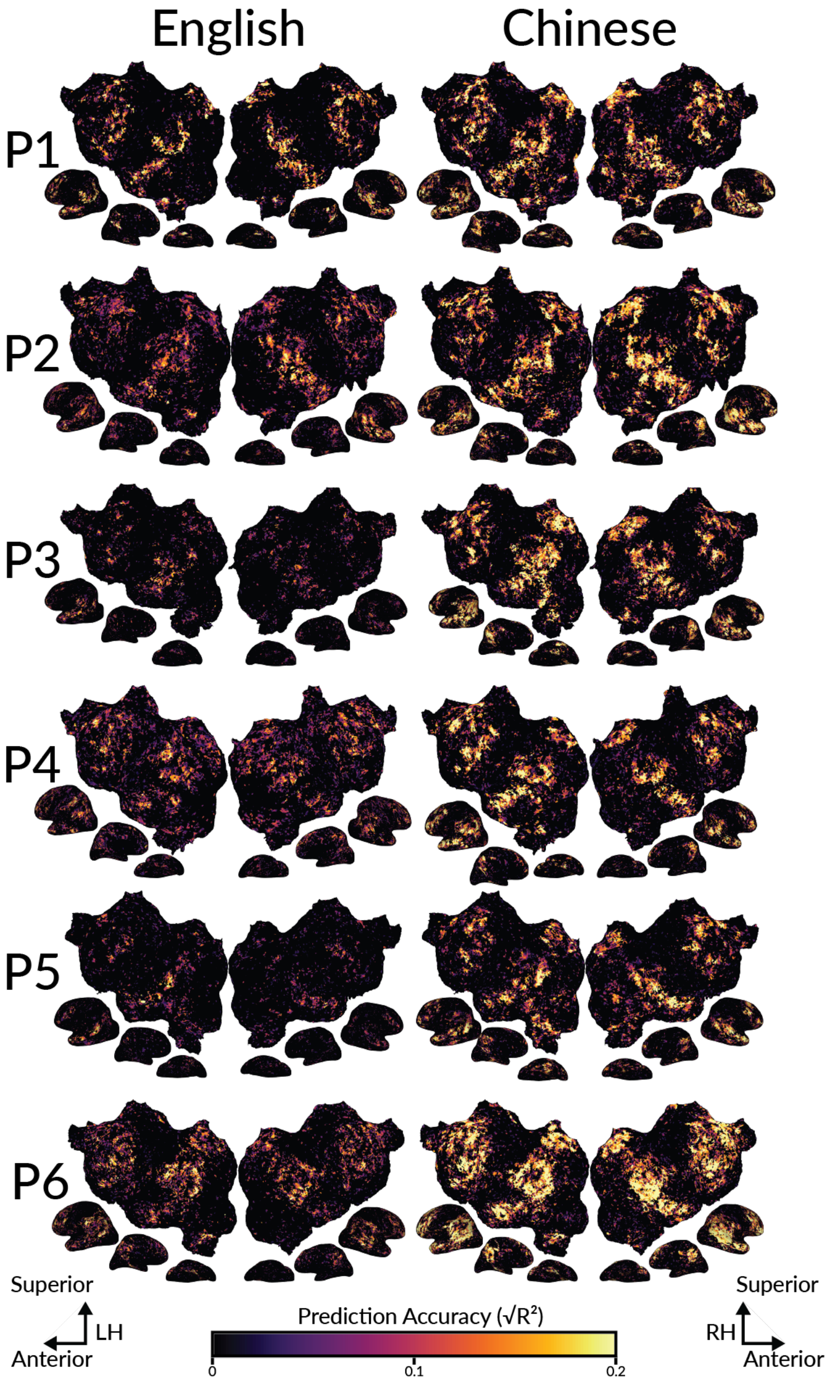

Figure 2.10: Cortical distribution of semantic representations for each language and in each individual participant. To determine where semantic information is represented for each language, voxelwise models estimated for each language were used to predict held-out data for the same language (within-language prediction accuracy). FastText was used as the semantic fea-

ture space. Prediction accuracy was computed as the CD (R^2) between predicted and recorded BOLD responses. Prediction accuracy is shown for each language and participant separately. Results are shown on the flattened cortical surface of each participant’ s native brain space. The color of each vertex indicates prediction accuracy according to the colorbar at the bottom. The magnitude of prediction accuracy differs between participants, partially reflecting individual differences in signal quality (Figure 2.6). For both languages the highest prediction accuracy (brightest voxels) is found within the bilateral temporal, parietal, and prefrontal cortices. Prediction accuracy is significantly positively correlated between languages in each participant (r=0.49, 0.36, 0.28, 0.32, 0.19, 0.46 for S1-S6; one-sided p<.05 for each participant by a permutation test). This suggests that the same brain regions within the semantic system are well-predicted for both languages.

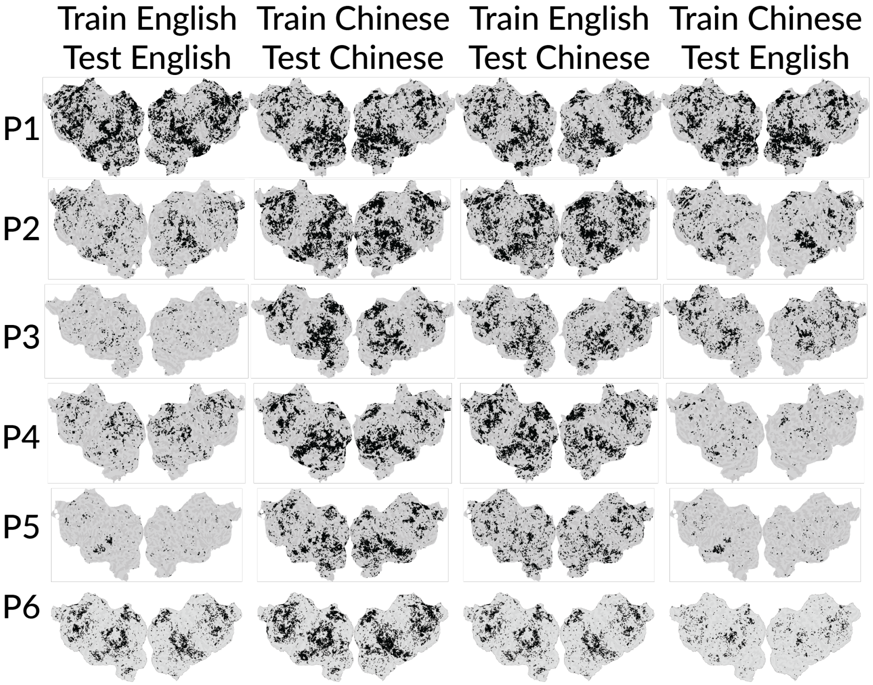

Figure 2.11: Statistical significance of prediction accuracy for each participant and language. Estimated model weights for the fastText semantic feature space were used to predict voxel responses on a held-out dataset. Within-language and across-language prediction accuracy were computed for each voxel as the CD (R^2) between predicted and true BOLD responses on the held-out dataset. The statistical significance of prediction accuracy for each voxel was determined by comparing the prediction accuracy of the estimated models to the prediction accuracy in predicting permuted data. The set of voxels that were significantly well-predicted are shown on the flattened cortical surface of each participant, separately for each language, and separately for within- and across-language testing. Voxels shown in black were significantly well-predicted (one-sided p<.05, FDR corrected with a Benjamini-Hochberg correction for multiple comparisons). The number of significantly well-predicted voxels differs across participants, reflecting individual differences in prediction accuracy shown in Supplementary Figures S5 and S11. Across participants, significantly well-predicted voxels are found within the semantic system, both within- and across-languages.

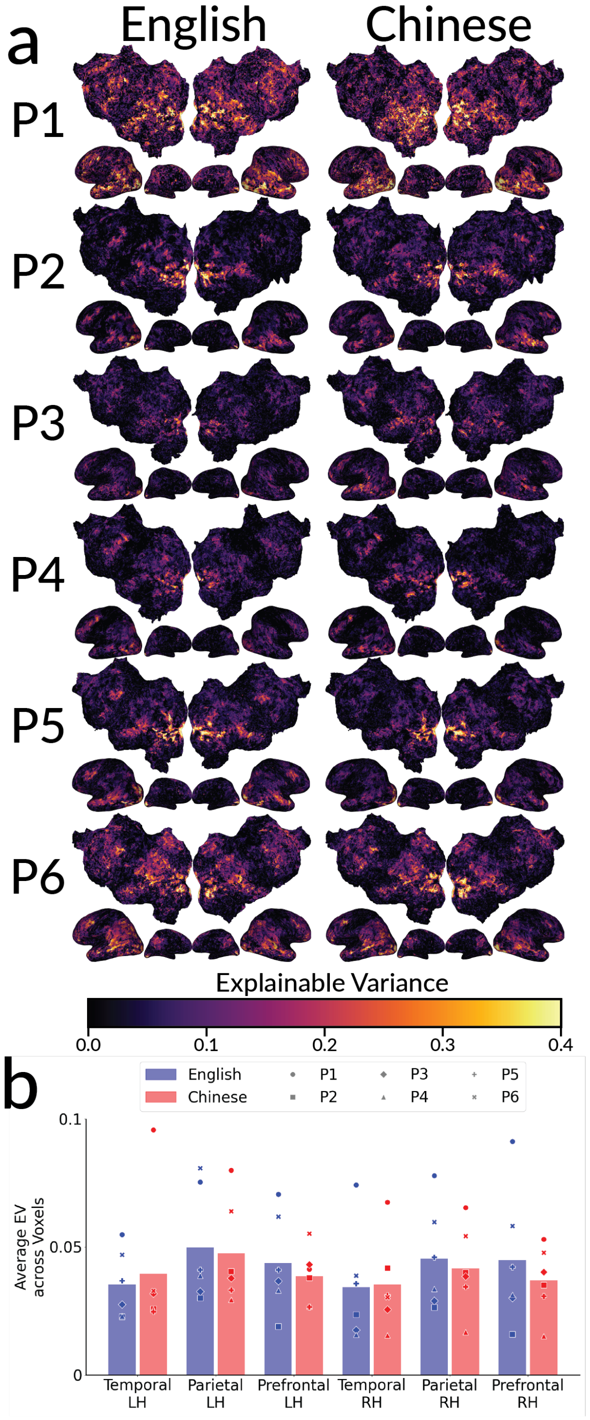

Figure 2.12: Noise-ceiling for each participant and language. Explainable Variance (EV) was computed as a measure of the noise-ceiling for each participant, voxel and language separately. **a**. The explainable variance (EV) of each voxel is shown on the flattened cortical surface of each participant’ s native brain space. EV is shown for English and Chinese separately. The

color of each voxel indicates explainable variance according to the colorbar at the bottom. The magnitude of EV varies across participants, suggesting individual differences in the quality of the recorded BOLD signal. **b**. For each brain region in the semantic system, the average EV across voxels is shown for each participant and language separately. Blue markers indicate EV in English, and red markers indicate EV in Chinese. Bars indicate the mean across participants. EV is not consistently higher in one language than the other.

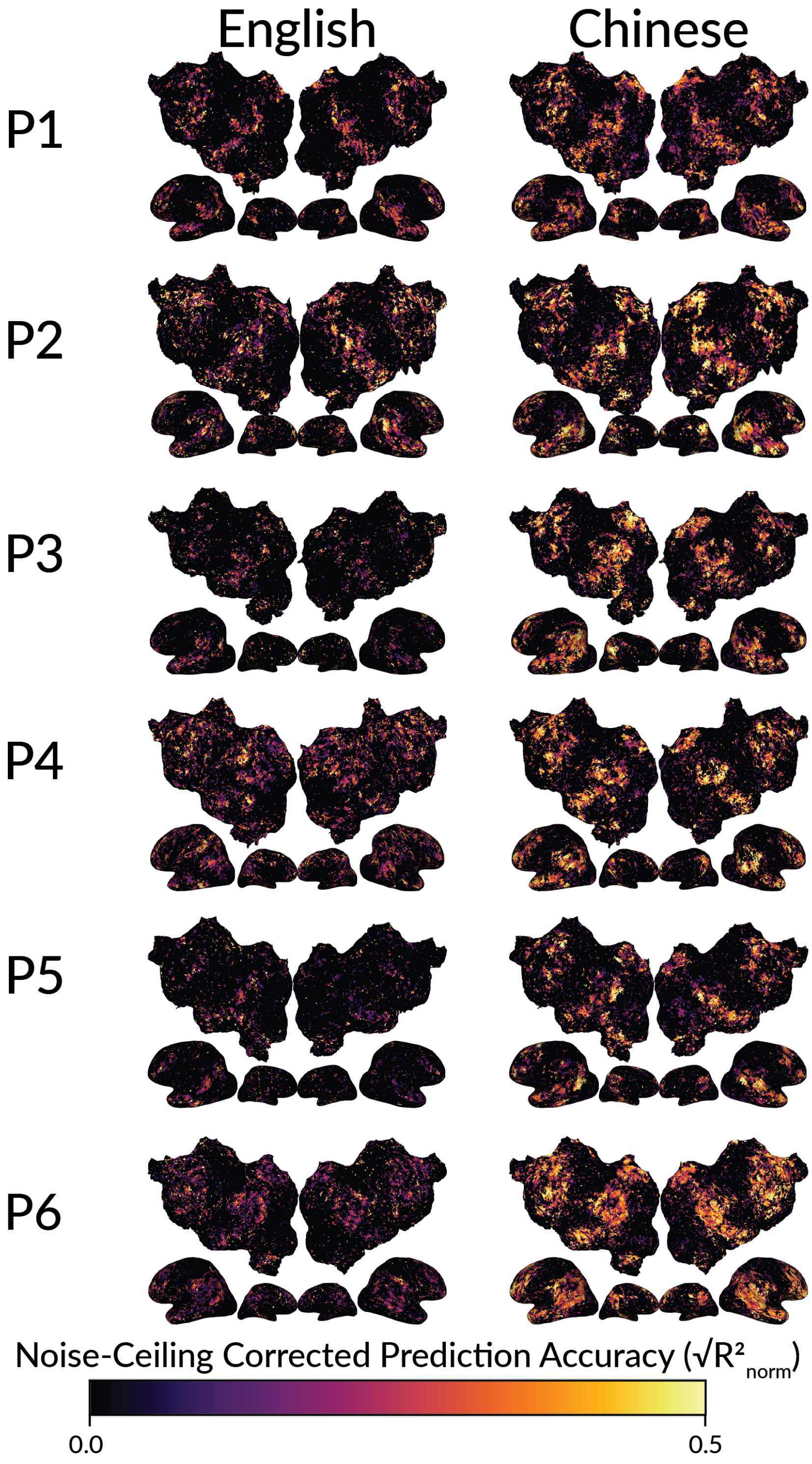
English
Chinese
P1
P2
P3
P4
P5
P6
Noise-Ceiling Corrected Prediction Accuracy ($\sqrt{R^2_{norm}}$)
0.0
0.5

Figure 2.13: Noise-ceiling corrected prediction accuracy for English vs Chinese. To determine whether differences in the noise-ceiling could explain the difference in prediction accuracy between languages, we computed the noise-ceiling corrected prediction accuracy (R^2_{norm}) for each language and participant separately. FastText was used as the semantic feature space. Noise-ceiling corrected prediction accuracy is shown on the flattened cortical surface of each participant’ s native brain space. The color of each voxel indicates prediction accuracy according to the colorbar at the bottom. For each participant, noise-ceiling corrected prediction accuracy is significantly higher in Chinese than in English (one-sided $p<.05$ by a permutation test). This suggests that the semantic model explains a higher proportion of total explainable signal in brain responses in Chinese than in English.

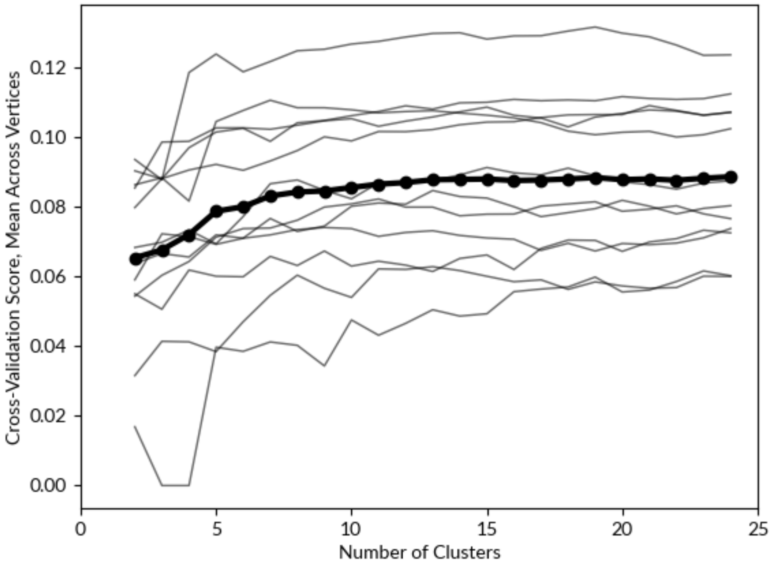

Figure 2.14: Choice of number of clusters to use for semantic clustering. Model weights were estimated for the semantic feature space, separately for each language and participant. A leave-one-out cross-validation approach was used to determine the optimal number of clusters into which to cluster model weights. In total, there are 12 sets of model weights across the six participants and two languages. Each set of model weights was projected to the template space. For each cross-validation fold, one of the twelve sets of model weights was held out. For each vertex the remaining eleven sets of model weights were averaged across participants and languages, and hierarchical clustering was used to obtain N group-level clusters, separately for N ranging from 2 through 25. For each number N of clusters, the group-level model weight clusters were used to predict BOLD responses for the held out participant and language in vertex space. The cross-validation accuracy was computed for each vertex as the CD R^2 between the predicted and recorded BOLD responses. Only vertices that were well-predicted in both languages (group-averaged $\sqrt{R^2_{en}} > 0.05$ and$\sqrt{R^2_{zh}} > 0.05$) were included in this analysis. The mean $\sqrt{R^2}$ over vertices is shown for each number N of clusters. Each thin line indicates the cross-validation scores for one fold. The bolded line indicates the mean across folds. The cross-validation score plateaus around five clusters. Thus, we used five semantic clusters for the analyses in Figure 2.3 and Figure 2.3.

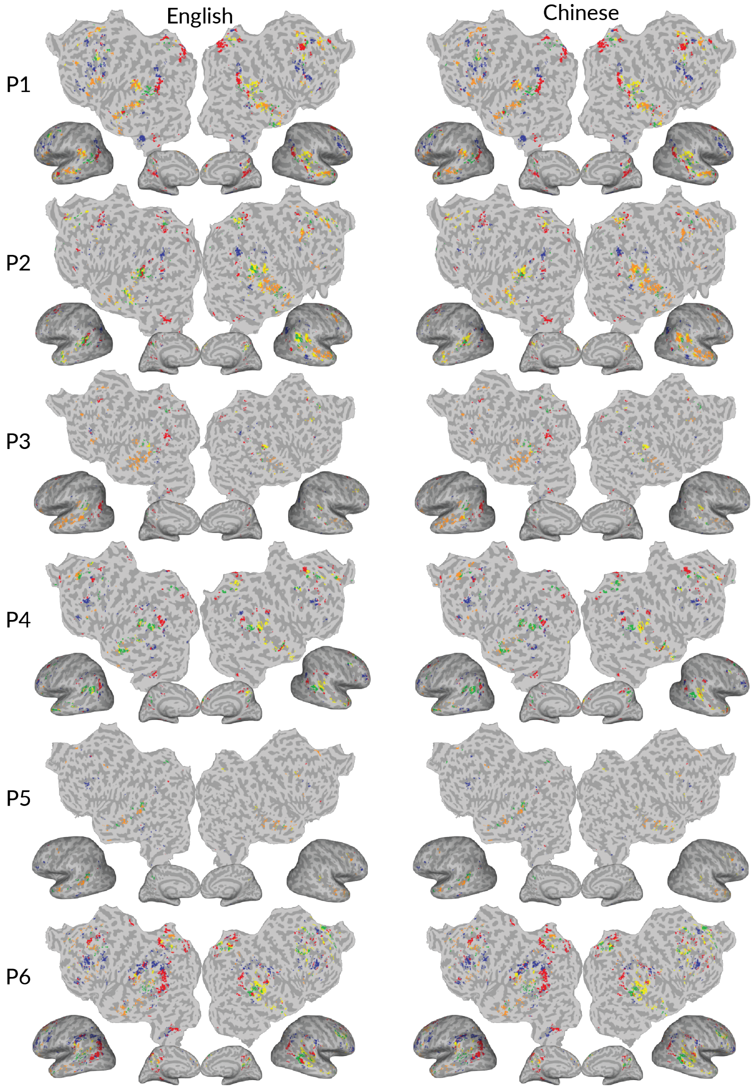

Figure 2.15: Semantic cluster assignments based on semantic model weights in each language and participant. For each participant and language separately, semantic model weights estimated for the fastText semantic feature space were used to assign each voxel to one of the semantic clusters shown in Figure 2.3a. Cluster assignments are shown for each language on the flattened cortical surface of the participant' s native brain space. Each voxel is colored according to the cluster assignment. Voxels that are not well-predicted ($\sqrt{R^2} > 0.1$) are shown in grey. Visual comparison of the flatmaps between languages shows that the cortical distribution of each semantic cluster is similar between English and Chinese. For each participant more than 70% of voxels have the same cluster assignments for both languages (P1: 81%, P3: 71%, P3: 85%, P4: 75%, P5: 87%, P6: 84%). The consistency in semantic cluster assignments between languages reflects that semantic representations are mostly shared between languages.

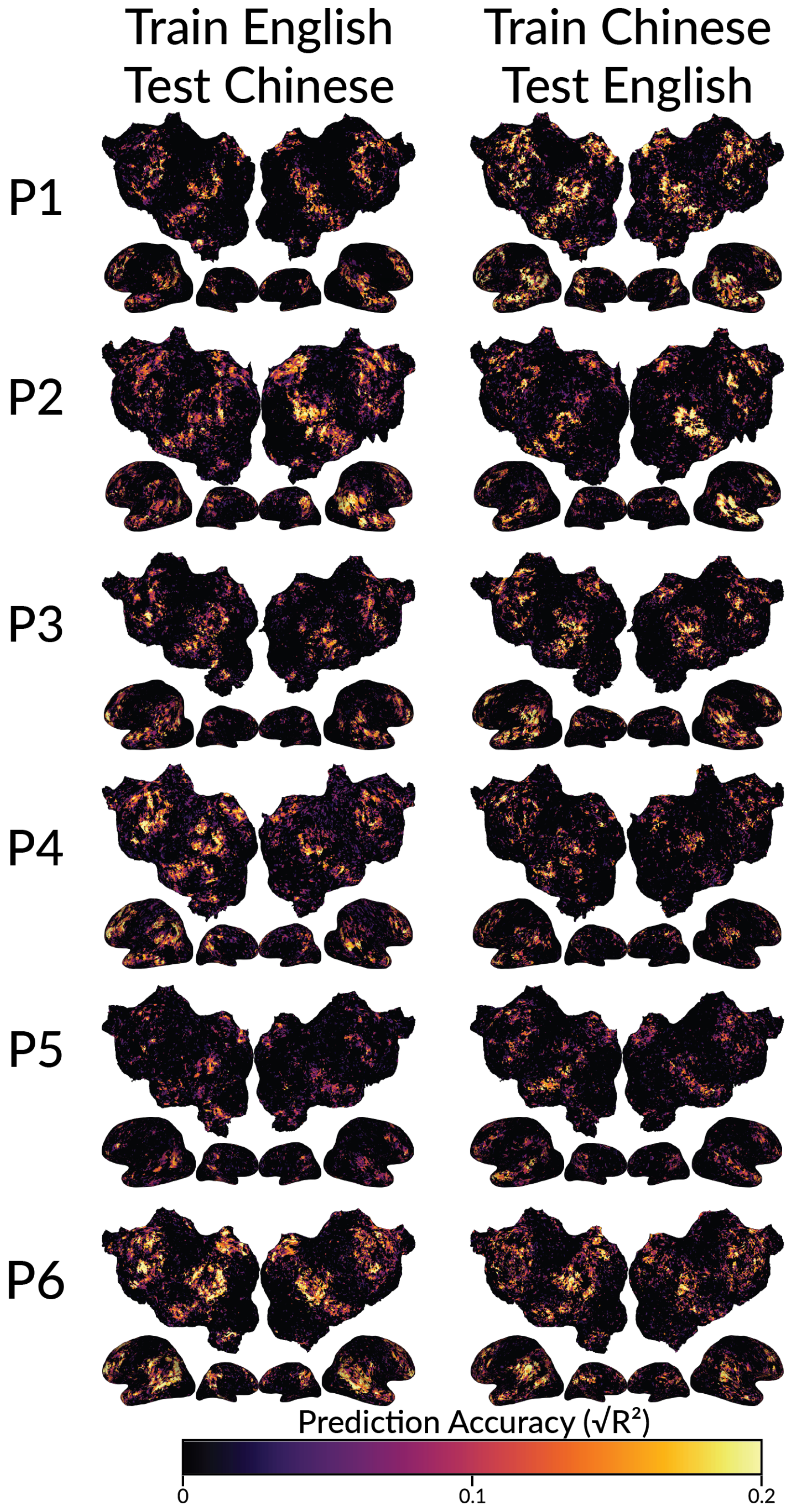
Train English
Test Chinese
Train Chinese
Test English
P1
P2
P3
P4
P5
P6
Prediction Accuracy ($\sqrt{R^2}$)
0
0.1
0.2

Figure 2.16: Shared semantic representations between languages as shown by across-language prediction accuracy. Across-language prediction accuracy is shown for each participant on the flattened cortical surface of the participant' s native brain space. FastText was used as the semantic feature space. Estimated voxelwise model weights in one language were used to predict the held-out dataset in the other language. FastText was used as the semantic feature space. Prediction accuracy was computed as the CD (R^2) between predicted and recorded BOLD responses. Prediction accuracy is given by the color scale. Well-predicted voxels appear brighter. In each participant, the semantic model predicted in one language accurately predicts voxel responses to the other language throughout the semantic system. Thus, semantic representations within the semantic system are largely shared between languages.

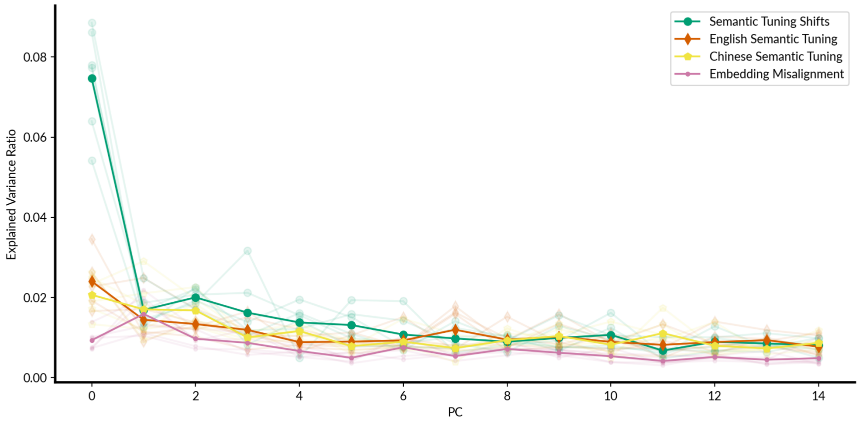

Figure 2.17: Amount of variance explained by semantic tuning shift principal components (PCs). Principal component analysis (PCA) was used to obtain the major dimensions of variation in semantic tuning shifts. To determine how many dimensions reliably capture variation in semantic tuning shifts a leave-one-participant-out procedure was used. At each step, one of the participants was left out and the voxelwise semantic tuning shifts from the other five participants were used to compute the principal components (PCs) of semantic tuning shifts (partial-group semantic tuning shift PCs). The partial-group semantic tuning shift PCs were used to compute the ratio of variance explained in the semantic tuning shifts of the left-out participant (green lines). Semantic tuning shift PCs were compared to three other PCs. First, the English semantic tuning PCs were computed by concatenating across voxels the 300-dimensional vectors of estimated model weights in English and then applying PCA across voxels to the resulting matrix. The English semantic tuning PCs were used to compute the ratio of variance explained in the semantic tuning shifts of each participant (orange lines). Second, the Chinese semantic tuning PCs were computed similarly to the English semantic tuning PCs but using the estimated semantic models weights in Chinese. The English semantic tuning PCs were used to compute the ratio of variance explained in the semantic tuning shifts of each participant (yellow lines). Third, the embedding misalignment PCs were computed by subtracting the embedding of each English word from its Chinese counterpart and then concatenating the 300-dimensional difference vectors across word pairs. Then PCA was applied across voxels to the embedding misalignment matrix. The embedding misalignment PCs were used to compute the ratio of variance explained in the semantic tuning shifts of each participant (pink lines). Transparent lines show variance explained in semantic tuning shifts for each individual participant, and opaque lines show the mean explained variance ratio across all participants. A bootstrapping approach was used to obtain confidence intervals for the variance explained by each PC. The voxel population was resampled 1000 times with replacement, and the tuning shift PCs were recomputed for each bootstrap iteration. 95% confidence intervals are shown by the error bars. The embedding misalignment PCs and the semantic tuning PCs for each language all explain significantly less variance than the PSSD. This suggests that the PSSD reliably captures variation in semantic tuning shifts, and that this dimension of tuning shifts is not merely an artifact of misalignments between embeddings for different languages, or of the primary dimensions of semantic tuning within each language.

-
Primary Tuning Shift Dimension
+

既 - already	路边 - roadside
之 - of	山沟 - ravine
位 - position	砾石 - gravel
以 - by	不速之客 - uninvited guest
更 - more	打电话 - make a phone call
是 - yes	山顶 - hilltop
起 - rise	忍不住 - can't help it
全 - all	阿拉巴马州 - alabama
打 - hit	印第安人 - Indian person
名 - name	德克萨斯 - texas
份 - portion	研究生 - postgraduate
令 - command	所作所为 - behavior
但 - however	基督教 - Christianity
七 - seven	亚特兰大 - Atlanta
和 - together	有生以来 - for one’s whole life

Figure 2.18: Interpretation of the PSSD based on Chinese stimulus words. Chinese stimulus words that best match each end of the PSSD are shown. Words closest to the negative end of the PC are shown in purple. Words closest to the positive end of the PC are shown in green. English translations are listed next to each word. The Chinese words closest to the negative end of the dimension are generally related to numbers. The Chinese words closest to the positive end of the PC are generally related to actions, people, and places. Interpretation of the PSSD is broadly consistent whether interpretations are based on Chinese stimulus words (shown here) or English stimulus words (shown in Figure 2.3a).

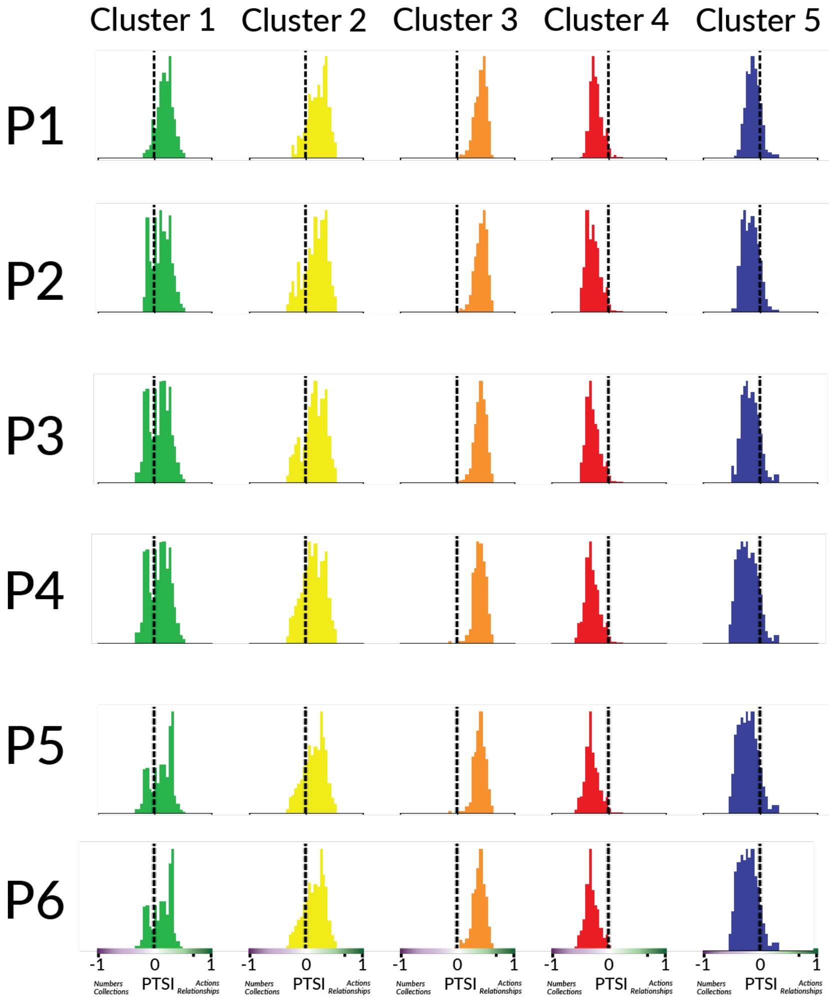

Figure 2.19: Semantic tuning shifts for different clusters when Chinese model weights are used. The semantic tuning shifts for different clusters depicted in Figure 2.3a were computed using English model weights. To test whether these results are independent of the language of model weights used for clustering, Chinese model weights are used to assign each voxel to a semantic cluster. Histograms indicate the distribution of PTSI values for voxels in each cluster, for each participant separately. Semantic tuning shifts for voxels in Clusters 1, 2, and 3 (green, yellow, and orange histograms; except P3 cluster 1 where p=.52) are positively correlated with the PSSD (p<.05 for each cluster by a two-sided t-test after Fisher z-transformation). Semantic tuning shifts for voxels in Clusters 4 and 5 (red and blue histograms) are negatively correlated with the PSSD (p<.05 for each cluster by a two-sided t-test after Fisher z-transformation). These results show that the results shown in Figure 2.3a are consistent whether English or Chinese model weights are used to assign each voxel to a semantic cluster.

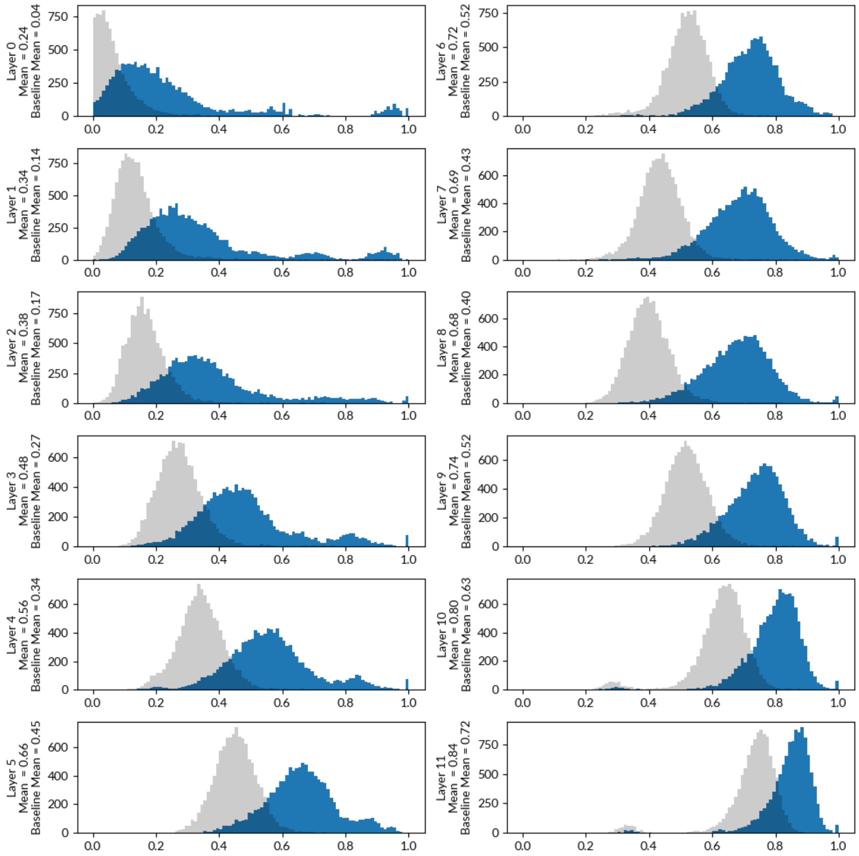

Figure 2.20: Best-aligned layers of multilingual BERT (mBERT). To determine which layer of mBERT produces the best multilingual semantic embedding space, the alignment of each layer of mBERT was measured between English and Chinese. To evaluate alignment between mBERT embeddings of English and Chinese we used the TsinghuaAligner test dataset (Liu & Sun, 2015). This dataset consists of 450 sentences that are provided in both English and Chinese, as well as manually annotated pairs that indicate pairs of English and Chinese words that mean the same thing. For each pair of words, each layer of mBERT was used to obtain two 768-dimensional embeddings: one for the English word, and one for the aligned Chinese word. For each layer separately, the similarity between the respective English and Chinese embeddings was measured as the cosine similarity between the two embeddings. As a baseline comparison, we measured the cosine similarity between the embeddings of randomly selected English-Chinese word pairs that did not mean the same thing. For each layer of mBERT we show the distribution of cosine similarities between aligned English-Chinese word pairs (blue) and the distribution of cosine similarities between randomly selected English-Chinese word pairs (grey). Embeddings for matched pairs of words become more similar in later layers. However, the latest layers produce outliers of poorly aligned word pairs that may bias semantic tuning shift estimates. Layer 9 produces the best aligned embeddings: embeddings of paired words have high cosine similarity relative to embeddings of random word pairs, without substantial outliers. Thus, we use layer 9 embeddings to validate the results from the fastText semantic embedding space (Supplementary Figures S1-S4).

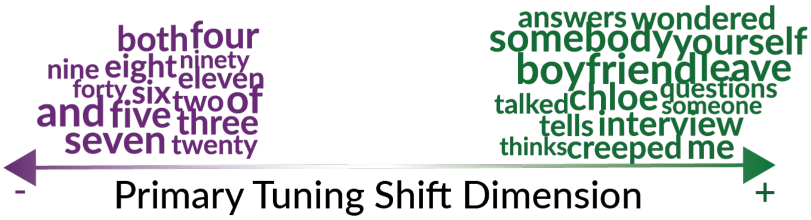

Figure 2.21: PSSD based on significantly well-predicted voxels. In Figure 2.3 the PSSD is estimated using voxels that are well-predicted in both languages ($\sqrt{R^2} > 0.1$ in both languages). To verify that the estimated PSSD is robust to the voxel selection method, we instead estimated the PSSD using voxels that were significantly well-predicted in both languages (one-sided p<.05, after Benjamini-Hochberg correction for multiple comparisons). The words that best match the significance-based estimate of the PSSD are shown. This dimension separates number/collection-related semantics (purple) from action/relationship-related semantics. The Pearson correlation between the significance-based estimate and the prediction accuracy-based estimate of the PSSD is 0.97. Thus, the estimate of the PSSD is robust to the voxel selection method.

Table 2.1: Participant responses to language use questionnaire

	P1	**P2**	**P3**	**P4**	**P5**	**P6**
Age when you began ACQUIRING language 1	1 month	2	since born	Since born	0	Native
Age when you began ACQUIRING language 2	11	7	4	6	2	6
Age when you became FLUENT in language 1	7	7	6	5	5	Native
Age when you became FLUENT in language 2	24	18	20	18	16	19
How many years and months have you spent in a COUNTRY where language 1 is spoken? (E.g. 2 years 3 months)	18 years	18 years	18 years	20 years	18 years	18 years
How many years and months you spent in a COUNTRY where language 2 is spoken? (E.g. 2 years 3 months)	12 years	6 years 1 months	6 years and 9 months	5 years	6 years	6 years 6 months
How many years and months have you spent in a FAMILY where language 1 is spoken? (E.g. 2 years 3 months)	30 years	18 years	18 years	25 years	18 years	18 years
How many years and months you spent in a FAMILY where language 2 is spoken? (E.g. 2 years 3 months)	8 years	0	0	1 month	0	0
How many years and months have you spent in a SCHOOL and/or WORKING environment where language 1 is spoken? (E.g. 2 years 3 months)	18 years	11 years	12 years	20 years	12 years	18 years
How many years and months you spent in a SCHOOL and/or WORKING environment where language 2 is spoken? (E.g. 2 years 3 months)	12 years	6 years 1 months	6 years and 9 months	5 years	6 years	6 years 6 months
Please circle to what extent you are CURRENTLY EXPOSED to language 1 in INTERACTING WITH FRIENDS:	5	4	7	7	5	7
Please circle to what extent you are CURRENTLY EXPOSED to language 2 in INTERACTING WITH FRIENDS:	10	4	5	7	7	5
Please circle to what extent you are CURRENTLY EXPOSED to language 1 in INTERACTING WITH FAMILY:	8	10	3	10	3	10

Please circle to what extent you are CURRENTLY EXPOSED to language 2 in INTERACTING WITH FAMILY:	6	0	0	0	0	0
Please circle to what extent you are CURRENTLY EXPOSED to language 1 at SCHOOL	0	2	0	0	0	0
Please circle to what extent you are CURRENTLY EXPOSED to language 2 at SCHOOL	10	8	10	10	10	9
Please circle to what extent you are CURRENTLY EXPOSED to language 1 at WORK	1	0	0	0	0	2
Please circle to what extent you are CURRENTLY EXPOSED to language 2 at WORK	10	8	10	10	10	9

Chapter 3

The Cortical Representation of Language Timescales is Shared between Reading and Listening

3.1 Abstract

Language comprehension involves integrating low-level sensory inputs into a hierarchy of increasingly high-level features. Prior work studied brain representations of different levels of the language hierarchy, but has not determined whether these brain representations are shared between written and spoken language. To address this issue, we analyze fMRI BOLD data that were recorded while participants read and listened to the same narratives in each modality. Levels of the language hierarchy are operationalized as timescales, where each timescale refers to a set of spectral components of a language stimulus. Voxelwise encoding models are used to determine where different timescales are represented across the cerebral cortex, for each modality separately. These models reveal that between the two modalities timescale representations are organized similarly across the cortical surface. Our results suggest that, after low-level sensory processing, language integration proceeds similarly regardless of stimulus modality.

3.2 Introduction

Humans leverage the structure of natural language to convey complex ideas that unfold over multiple timescales. The structure of natural language contains a hierarchy of components, which range from low-level components such as letterforms or articulatory features, to higher-level components such as sentence-level syntax, paragraph-level semantics, and narrative arc. During human language comprehension, brain representations of low-level components are thought to be incrementally integrated into representations of higher-level components (Christiansen & Chater, 2016). These representations have been shown to form a topographic organization across the surface of the cerebral cortex during spoken language comprehension (Baldassano et al., 2017; Blank & Fedorenko, 2020; Jain & Huth, 2018; Jain et al., 2020; Lerner et al., 2011).

Both written and spoken language consist of a hierarchy of components, but to date it has been unclear to what extent brain representations of these hierarchies are shared between the two modalities of language comprehension. At low levels of the hierarchy, brain representations are known to differ between the two stimulus modalities. For example, visual letterforms in written language are represented in the early visual cortex, whereas articulatory features in spoken language are represented in the early auditory cortex (de Heer et al., 2017; Heilbron et al., 2020). In contrast, many parts of temporal, parietal, and prefrontal cortices process both written and spoken language (e.g., Booth et al., 2002; Buchweitz, Mason, Tomitch, & Just, 2009; Deniz et al., 2019; Liuzzi et al., 2017; Nakai et al., 2021; Regev et al., 2013). It could be the case that in these areas representations of higher-level language components are organized in the same way for both written and spoken language comprehension. On the other hand, these areas could contain overlapping but independent representations for the two modalities. One way to differentiate between these two possibilities would be to directly compare the cortical organization of brain representations across high-level language components between reading and listening. However, prior work has not performed this comparison. Most prior studies of reading and listening have compared brain responses generally, without explicitly describing what stimulus features are represented in each brain area (e.g., Booth et al., 2002; Buchweitz, Mason, Tomitch, & Just, 2009; Liuzzi et al., 2017; Regev et al., 2013). Other studies focused on relatively few components (e.g., low-level sensory features, word-level semantics, and phonemic features), and therefore did not provide a detailed differentiation between different levels of the language hierarchy (Deniz et al., 2019; Nakai et al., 2021). Studies that did differentiate between different levels focused on one modality of language (e.g., Jain & Huth, 2018; Jain et al., 2020; Lerner et al., 2011; Toneva & Wehbe, 2019). Prior studies are therefore insufficient to determine whether brain representations of the language hierarchy are organized similarly between reading and listening.

To address this problem we compared where different levels of the language hierarchy are represented in the brain during reading and listening. Intuitively, levels of processing hierarchy can be considered in terms of numbers of words. For example, low-level sensory components such as visual letterforms in written language and articulatory features in spoken language vary within the course of single words; sentence-level syntax varies over the course of tens of words; paragraph-level semantics varies over the course of hundreds of words. Therefore we operationalize levels of the language hierarchy as language timescales, where a language timescale is defined as the set of spectral components of a language stimulus that vary over a certain number of words. For brevity we refer to "language timescales" simply as timescales.

We analyzed functional magnetic resonance imaging (fMRI) recordings from participants who read and listened to the same set of narratives (Deniz et al., 2019; Huth et al., 2016). The stimulus words were then transformed into features that each reflect a certain timescale of stimulus information: first a language model (BERT) was used to extract contextual embeddings of the narrative stimuli, and then linear filters were used to separate the contextual embeddings into timescale-specific stimulus features. Voxelwise encoding models were used to estimate the average timescale to which each voxel is selective, which we refer to as the "average timescale selectivity". These estimates reveal where different language timescales are represented across the cerebral cortex for reading and listening separately. Finally, the cortical organization of timescale selectivity was compared between reading and listening.

3.3 Results

We compared the organization of timescale representations between written and spoken language comprehension for each participant. First, the set of language-selective voxels for each modality was identified as those for which any of the timescale-specific language feature spaces significantly predicted blood oxygenation level dependent (BOLD) responses (one-sided permutation test, $p < .05$, false discovery rate (FDR) corrected). Then, voxel timescale selectivity was compared between reading and listening across the set of voxels that are language-selective for both modalities. For each participant, voxel timescale selectivity is significantly positively correlated between the two modalities (S1: $r = 0.41$, S2: $r = 0.58$, S3: $r = 0.44$, S4: 0.34, S5: 0.47, S6: 0.35, S7: 0.40, S8: 0.49, S9: 0.52, $p < .001$ for each participant; Figure 3.3a). Visual inspection of voxel timescale selectivity across the cortical surface confirms that the cortical organization of timescale selectivity is similar between reading and listening (Figure 3.3b, Figure 3.3c). For both modalities, timescale selectivity varies along spatial gradients from intermediate timescale selectivity in superior temporal cortex to long timescale selectivity in inferior temporal cortex, and from intermediate timescale selectivity in posterior prefrontal cortex to long timescale selectivity in anterior prefrontal cortex. Medial parietal cortex voxels are selective for long timescales for both modalities. Estimates of timescale selectivity are robust to small differences in feature extraction – results are quantitatively similar when using a fixed rolling context instead of a sentence input context, and when using units from only a single layer of BERT instead of from all layers (Figures S1, S2, S3, S4, and S5). These results suggest that for each individual participant representations of language timescales are organized similarly across the cerebral cortex between reading and listening.

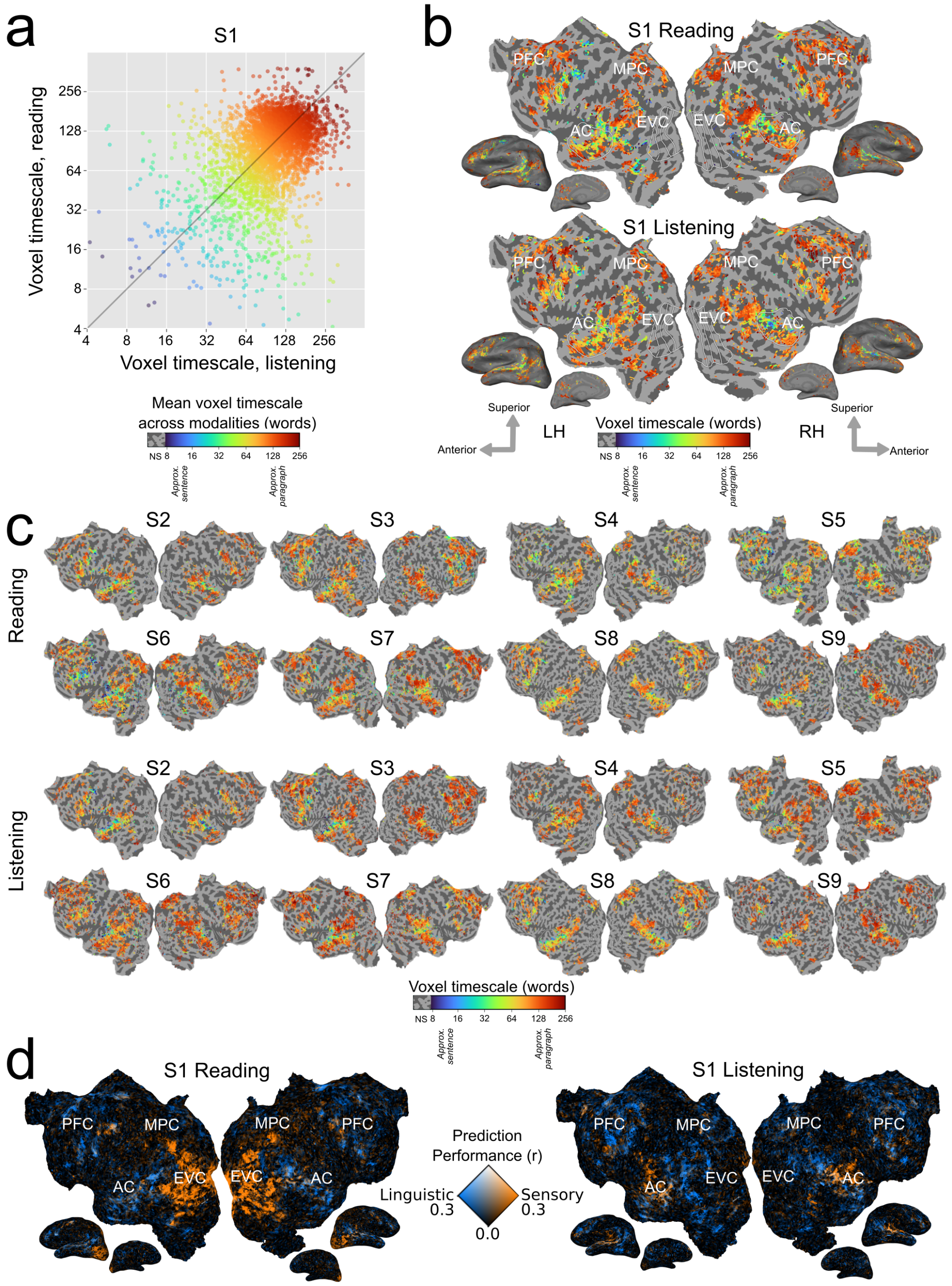

Figure 3.1: Timescale selectivity across the cortical surface.
Voxelwise modeling was used to determine the timescale selectivity of each voxel, for reading and listening separately (See Section 3.5 for details). a. Timescale selectivity during listening (x-axis) vs reading (y-axis) for one representative participant (S1). Each point represents one voxel that was significantly predicted in both modalities. Points are colored according to the mean of the timescale selectivity during reading and listening. Blue denotes selectivity for short

timescales, green denotes selectivity for intermediate timescales, and red denotes selectivity for long timescales. Timescale selectivity is significantly positively correlated between the two modalities ($r = 0.41$, $p < .001$). b. Timescale selectivity during reading and listening on the flattened cortical surface of S1. Timescale selectivity is shown according to the color scale at the bottom (same color scale as in Panel A). Voxels that were not significantly predicted are shown in grey (one-sided permutation test, $p < .05$, FDR corrected; LH, left hemisphere; RH, right hemisphere; NS, not significant; PFC=prefrontal cortex, MPC=medial parietal cortex, EVC=early visual cortex, AC=auditory cortex). For both modalities, temporal cortex contains a spatial gradient from intermediate to long timescale selectivity along the superior to inferior axis, prefrontal cortex (PFC) contains a spatial gradient from intermediate to long timescale selectivity along the posterior to anterior axis, and precuneus is predominantly selective for long timescales. c. Timescale selectivity in eight other participants. The format is the same as in Panel b. d. Prediction performance for linguistic features (i.e., timescale-specific feature spaces) vs. low-level sensory features (i.e., spectrotemporal and motion energy feature spaces) for S1. Orange voxels were well-predicted by low-level sensory features. Blue voxels were well-predicted by linguistic features. White voxels were well-predicted by both sets of features. Low-level sensory features predict well in early visual cortex (EVC) during reading, and in early auditory cortex (AC) during listening. Linguistic features predict well in similar areas for reading and listening. After early sensory processing, cortical timescale representations are consistent between reading and listening across temporal, parietal, and prefrontal cortices.

In contrast to representations of language timescales, low-level sensory features are represented in modality-specific cortical areas. Figure 3.3d shows the prediction performance of linguistic features (i.e., timescale-specific feature spaces), and the prediction performance of low-level sensory features (i.e., spectrotemporal representations of auditory stimuli, and motion energy representations of visual stimuli). Voxels are colored according to the prediction performance of each set of feature spaces: voxels shown in blue are well predicted by the linguistic feature spaces, voxels shown in orange are well predicted by the low-level sensory feature spaces, and voxels shown in white are well predicted by both sets of feature spaces. For both reading and listening, timescale-specific feature spaces predict well broadly across temporal, parietal, and prefrontal cortices. In contrast, low-level stimulus features predict well in early visual cortex (EVC) during reading only, and in auditory cortex (AC) during listening only. These results indicate that during language comprehension, linguistic processing occurs in similar cortical areas between modalities, whereas low-level sensory processing occurs in modality-specific cortical areas.

Within each participant, estimates of timescale selectivity depend not only on the presentation modality, but also on the presentation order. This is because each participant either read all the stories before listening to the stories, or vice versa, and attentional shifts between novel and known stimuli may cause small differences in estimated timescale selectivity. Indeed, activation across higher-level brain regions is often more widespread and consistent for the first presentation modality than for the second presentation modality, indicating that participants attend more strongly to novel stimuli (Figures S6 and S7). In six of the nine participants, timescale selectivity was slightly longer for the first presented modality than for the second presented modality (Figure S8). This change in timescale selectivity between novel and repeated stimuli suggests that the predictability of high-level narrative components in known stimuli may reduce brain responses to longer language timescales. Nevertheless, the overall cortical organization of timescale selectivity was consistent between reading and listening across all nine participants, regardless of whether they first read or listened to the narratives. This consistency indicates that the effects of stimulus repetition on timescale selectivity are small relative to the similarities between timescale selectivity during reading and listening.

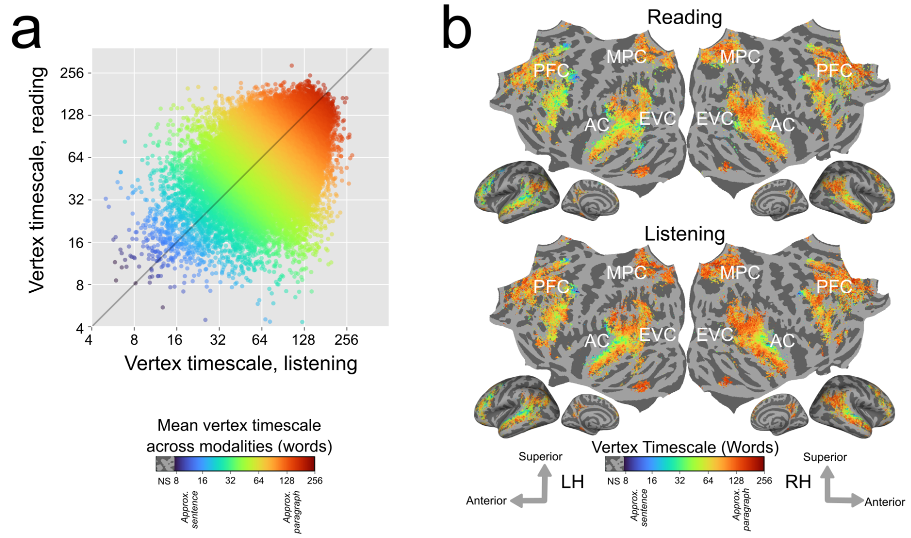

Figure 3.2: Group-level estimates of timescale selectivity in standard brain space. Group-level estimates of timescale selectivity are shown in a standard fsAverage vertex space. The group-level estimate for each vertex was computed by taking the mean over all participants in whom the vertex was language-selective. a. Group-level timescale selectivity during listening (x-axis) vs reading (y-axis). Each point represents one vertex that was significantly predicted in both modalities for at least one-third of the participants. Each point is colored according to the mean of the group-level timescale selectivity during reading and listening. Blue denotes selectivity for short timescales, green denotes selectivity for intermediate timescales, and red denotes selectivity for long timescales. Timescale selectivity is positively correlated between the two modalities (r = 0.48). b. For reading and listening separately group-level timescale selectivity is shown according to the color scale at the bottom (same color scale as in Panel A). Colored vertices were significantly predicted for both modalities in at least one-third of the participants. Vertices that were not significantly predicted are shown in grey (one-sided permutation test, $p < .05$, FDR corrected; NS, not significant; PFC=prefrontal cortex, MPC=medial parietal cortex, EVC=early visual cortex, AC=auditory cortex). Group-averaged measurements of timescale selectivity are consistent with measurements observed in individual participants (Figure 3.3). For both modalities, there are spatial gradients from intermediate to long timescale selectivity along the superior to inferior axis of temporal cortex, and along the posterior to anterior axis of prefrontal cortex (PFC). Precuneus is predominantly selective for long timescales for both modalities. Across participants, the cortical representation of different language timescales is consistent between reading and listening across temporal, parietal, and prefrontal cortices.

In order to consolidate results across all participants, we computed group-level estimates of timescale selectivity. To compute group-level estimates, first the estimates for each individual participant were projected to the standard FreeSurfer fsAverage vertex space (Fischl et al., 1999). Then for each vertex the group-level estimate of timescale selectivity was computed as the mean of the fsAverage-projected values. This mean was computed across the set of participants in whom the vertex was language-selective. Group-level estimates were computed separately for reading and listening. Group-level timescale selectivity was then compared between reading and listening across the set of vertices that were significantly predicted in at least one-third of the participants for both modalities (Figure S9 shows the number of participants for which each vertex was significantly predicted, separately for each modality). This comparison showed that timescale selectivity is highly correlated between reading and listening at the group level ($r = 0.48$; Figure 3.3a). Cortical maps of group-level timescale selectivity (Figure 3.3b) visually highlight that the spatial gradients of timescale selectivity across temporal and prefrontal cortices are highly similar between the two modalities. Gradients of timescale selectivity are also evident within previously proposed anatomical brain networks (Figure S10). Overall, these group-level results show that across participants, the organization of representations of language timescales is consistent between reading and listening.

The results shown in Figure 3.3 and Figure 3.3 indicate that average timescale selectivity is similar between reading and listening. However, average timescale selectivity alone is insufficient for determining whether representations of different timescales are shared between reading and listening – average timescale selectivity could equate voxels with a very peaked selectivity for a single frequency band, and voxels with uniform selectivity for many frequency bands (Figure S11 shows how the uniformity of timescale selectivity varies across voxels). To investigate this possibility we used the timescale selectivity profile, which reflects selectivity for each timescale separately. Although the timescale selectivity profile is a less robust metric than average timescale selectivity (see Section 3.5 for details), the timescale selectivity profile can distinguish between peaked and uniform selectivity profiles.

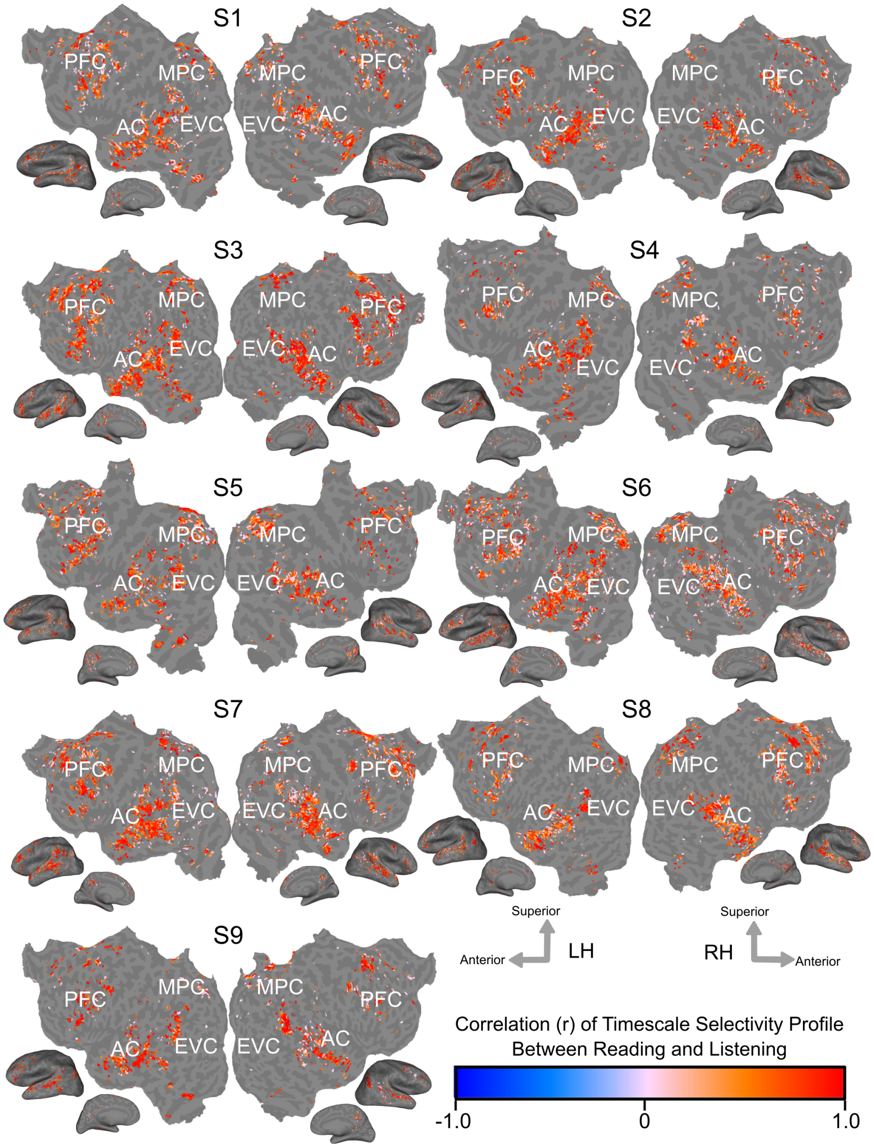
S1
S2
S3
S4
S5
S6
S7
S8
S9
PFC
MPC
AC
EVC
Superior
Anterior
LH
RH
Correlation (r) of Timescale Selectivity Profile
Between Reading and Listening
-1.0
0
1.0

Figure 3.3: Voxelwise similarity of timescale selectivity.
The Pearson correlation coefficient of the timescale selectivity profile between reading and listening is shown on the cortical surfaces of each participant. The correlation coefficient is shown according to the color scale at the bottom. Red voxels have positively correlated timescale selectivity profiles between reading and listening. Blue voxels have negatively correlated timescale selectivity profiles between reading and listening. Voxels that were not significantly predicted in both modalities are shown in grey (one-sided permutation test, $p < .05$, FDR corrected). In areas that are language-selective in both modalities, the timescale selectivity profile is highly correlated across voxels.

Figure 3.3 shows the Pearson correlation coefficient between the timescale selectivity profile in reading and in listening on the flattened cortical surface of each participant. The timescale selectivity profile is highly correlated between reading and listening, across voxels that are language-selective in both modalities.

To further demonstrate the shared organization of cortical timescale selectivity, we compared the cortical distribution of selectivity for each of the eight timescales. For each of the eight timescales, we computed the correlation between selectivity for that timescale during reading and listening across the set of voxels that are language-selective in both modalities. The correlations for each timescale and participant are shown in Figure 3.3a. A full table of correlations and statistical significance is shown in Table S1. Selectivity for each timescale was positively correlated between reading and listening for each timescale and in each individual participant. Most of these correlations were statistically significant (one-sided permutation test, $p < .05$, FDR-corrected). Note that comparing the timescale selectivity metric is more robust to noise in the data than comparing selectivity for each timescale separately (see Section 3.5 for details). Therefore correlations between selectivity for each individual timescale are less consistent across participants than correlation between timescale selectivity.

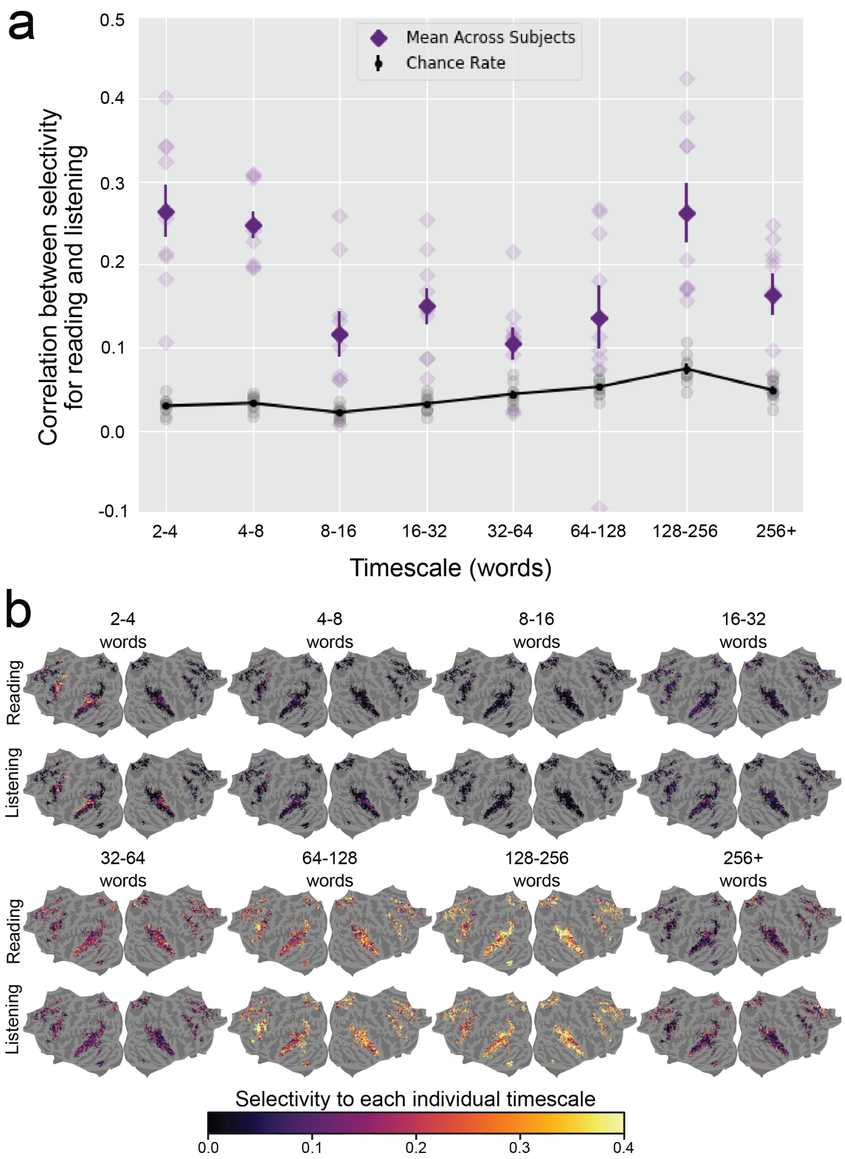

Figure 3.4: Similarity of selectivity for each timescale between reading and listening. Selectivity for each individual timescale was compared between reading and listening across the cerebral cortex. For each voxel, selectivity for each individual timescale describes the extent to which the corresponding timescale-specific feature space explains variation in BOLD responses, relative to the other timescale-specific feature spaces (see Section 3.5 for details). a. For each timescale the Pearson correlation coefficient was computed between selectivity for that timescale during reading and listening, across all voxels that were significantly predicted for both modalities. For each timescale, the mean true correlation across participants is indicated by dark purple diamonds. The mean chance correlation across participants is indicated by black dots (for clarity, these black dots are connected by a black line). Vertical lines through purple diamonds and through black dots are error bars that indicate the standard error of

the mean (SEM) across participants for the respective value. True and chance correlations for each individual participant are respectively indicated by light purple diamonds and grey dots. The true correlation is significantly higher than chance in most individual participants and timescales; see Table S1 for details. b. The group-level selectivity of each vertex to each timescale is shown in fsAverage space for reading and listening separately. Vertices that were not language-selective in both modalities are shown in grey. Outside of primary sensory areas, selectivity for each timescale is distributed similarly across the cortical surface between both modalities. Among voxels that are language-selective in both modalities, each language timescale is represented in similar areas between reading and listening. These results further indicate that there is a shared organization of representations of language timescales between reading and listening.

The cortical distribution of selectivity for each timescale is shown for reading and listening separately in Figure 3.3b. For concision these results are shown at the group-level. Visual inspection of 3.3b shows that for both reading and listening, short timescales (2-4 words, 4-8 words, 8-16 words) are represented in posterior prefrontal cortex and superior temporal cortex; intermediate timescales (16-32 words, 32-64 words) are represented broadly across temporal, prefrontal, and medial parietal cortices; and long timescales (64-128 words, 128-256 words, 256+ words) are represented in prefrontal cortex, precuneus, temporal parietal junction, and inferior temporal cortex. The correlations between selectivity for each timescale and qualitative comparisons of the cortical distribution of selectivity for each timescale between reading and listening indicate that representations of language timescales are organized similarly between reading and listening.

3.4 Discussion

This study tested whether representations of language timescales are organized similarly between reading and listening. We used voxelwise encoding models to determine the selectivity of each voxel to different language timescales and then compared the organization of these representations between the two modalities (Figure 3.5). These comparisons show that timescale selectivity is highly correlated between reading and listening across voxels that are language-selective in both modalities. This correlation is evident in individual participants (Figure 3.3) and at the group-level (Figure 3.3). For both modalities, prefrontal and temporal cortices contain spatial gradients from intermediate to long timescale selectivity, and precuneus is selective for long timescales. Comparisons of selectivity for each individual voxel (Figure 3.3), and to each timescale separately (Figure 3.3), show that the cortical representation of each timescale is similar between reading and listening. These results suggest that the topographic organization of language processing timescales is shared across stimulus modalities.

Prior work has studied brain representations of contextualized and non-contextualized language, separately for written (Toneva & Wehbe, 2019) and spoken language comprehension (Jain & Huth, 2018). Those studies showed that areas within medial parietal cortex, prefrontal cortex, and inferior temporal cortex preferentially represent contextualized information; whereas other areas within superior temporal cortex and the temporoparietal junction do not show a preference for contextualized information. Our results build upon these previous findings by directly comparing representations between reading and listening within individual participants, and by examining representations across a finer granularity of timescales. The fine-grained variation in timescale selectivity that we observed within previously proposed cortical networks supports the hypothesis that language processing occurs along a continuous gradient, rather than in distinct, functionally specialized brain networks (Blank & Fedorenko, 2020).

Our study provides new evidence on the similarities in language processing between reading and listening. To compare brain responses between reading and listening, prior work correlated timecourses of brain responses between participants who read and listened to the same stimuli (Regev et al., 2013). That work found similarities in areas such as superior temporal gyrus, inferior frontal gyrus, and precuneus; and differences in early sensory areas as well as in parts of parietal and frontal cortices. However, that work did not specifically model linguistic features. Therefore the differences they observed between modalities in parietal and frontal cortices may indicate differences in non-linguistic processes such as high-level control processes, rather than differences in language representations. By specifically modeling representations of linguistic features, our results suggest that some of the differences observed in (Regev et al., 2013) could indeed be due to non-linguistic processes such as high-level control. A separate study suggesting that brain representations of language differ between modalities compared brain responses to different types of stimuli for reading and listening: the stimuli used for reading experiments consisted of isolated sentences, whereas the stimuli used for listening experiments consisted of full narratives (Oota et al., 2022). This discrepancy perhaps explains why in (Oota et al., 2022), language models trained on higher-level tasks (e.g., summarization, paraphrase detection) were better able to predict listening than reading data. Our study used matched stimuli for reading and listening experiments, and the similarities we observed highlight the importance of using narrative-length, naturalistic stimuli to elicit brain representations of high-level linguistic features (Deniz et al., 2023).

The method for estimating timescale selectivity that we introduced in this work addresses limitations in methods previously used to study language timescales in the brain (Baldassano et al., 2017; Blank & Fedorenko, 2020; Jain & Huth, 2018; Jain et al., 2020; Lerner et al., 2011). Early methods required the use of stimuli that are scrambled at different temporal granularities (Blank & Fedorenko, 2020; Lerner et al., 2011). However, artificially scrambled stimuli may cause attentional shifts, evoking brain responses that are not representative of brain responses to natural stimuli (Deniz et al., 2023; Hamilton & Huth, 2020; Hasson et al., 2010). Other approaches measured the rate of change in patterns of brain responses in order to determine the temporal granularity of representations in each brain region (Baldassano et al., 2017). However, that approach does not provide an explicit stimulus-response model which is needed to determine whether the temporal granularity in each brain region reflects linguistic or non-linguistic brain representations. Our approach uses voxelwise modeling, which allows us to estimate brain representations with ecologically valid stimuli, and obtain an explicit stimulus-response model. Our method uses spectral analysis to extract stimulus features that reflect different language timescales, decoupling the feature extraction process from specific neural network architectures. This decoupling enables the construction of encoding models that are more accurate and that are also interpretable in terms of timescale selectivity. In the future, our method could be used with pretrained audio or visual models (e.g., wav2vec 2.0 (Baevski et al., 2020) or TrOCR (M. Li et al., 2023)) to estimate selectivity for different timescales of low-level auditory and visual features. In sum, the method for estimating timescale selectivity that we developed in this study allowed us to produce more interpretable, accurate, and ecologically valid models of language timescales in the brain than previous methods.

To further inform theories of language integration in the brain, our approach of analyzing language timescales could be combined with approaches that analyze brain representations of specific classical language constructs. Approaches based on classical language constructs such as part-of-speech tags (Wehbe et al., 2014) and hierarchical syntactic constructs (Brennan et al., 2016; Hale et al., 2015) provide intuitive interpretations of cortical representations. However, these language constructs do not encompass all the information that is conveyed in a natural language stimulus. For example, discourse structure and narrative processes are difficult to separate and define. This difficulty is particularly acute for freely produced stimuli, which do not have explicitly marked boundaries between sentences and paragraphs. Instead of classifically defined language constructs, our approach uses spectral analysis to separate language timescales. The resulting models of brain responses can therefore take into account stimulus language information beyond language constructs that can be clearly separated and defined. In the future, evidence from these two approaches could be combined in order to improve our understanding of language processing in the brain. For example, previous studies suggested that hierarchical syntactic structure may be represented in the left temporal lobe, areas in which our analyses identified a spatial gradient from intermediate to long timescale selectivity (Brennan et al., 2016). Evidence derived from both approaches should be further compared in order to inform neurolinguistic theories with a spatially and temporally fine-grained model of voxel representation that can be interpreted in terms of classical language constructs.

One limitation of our study comes from the temporal resolution of BOLD data. Because the data used in this study have a repetition time (TR) of 2 seconds, our analysis may be unable to detect very fine-grained distinctions in timescale selectivity. Furthermore, controlling for low-frequency voxel response drift required low-pass filtering the BOLD data during preprocessing. This preprocessing filter may have removed information about brain representations of very long timescales (i.e., timescales above 360 words), thus removing information about these timescales. Future work could apply our method to brain recordings that have more fine-grained temporal resolution (e.g., from electrocorticography (ECoG) or electroencephalography (EEG) recordings) or that do not require low-pass filtering in order to determine whether there are subtle differences in timescale selectivity between modalities. A second limitation arises from the current state of language model embeddings. Although embeddings from language models explain a large proportion of variance in brain responses, these embeddings do not capture all stimulus features (e.g., features that change within single words). In the future, our method can be used with other language models to obtain more accurate estimates of timescale selectivity.

In sum, we developed a sensitive, data-driven method to determine whether language timescales are represented in the same way during reading and listening across cortical areas that represent both written and spoken language. Analyses of timescale selectivity in individual participants and at the group level reveal that the cortical representation of different language timescales is highly similar between reading and listening across temporal, parietal, and prefrontal cortices at the level of individual voxels. The shared organization of cortical language timescale selectivity suggests that a change in stimulus modality alone does not substantially alter the organization of representations of language timescales. A remaining open question is whether a change in the temporal constraints of language processing would alter the organization of representations of language timescales. One interesting direction for future work would be to compare whether a change in the stimulus presentation method (e.g., static text presentation compared to transient rapid serial visual presentation (RSVP)) would alter the organization of language timescale representations.

3.5 Methods

Functional MRI was used to record BOLD responses while human participants read and listened to a set of English narrative stories (Deniz et al., 2019; Huth et al., 2016). The stimulus narratives were transformed into feature spaces that each reflect a particular set of language timescales. Each timescale was defined as the spectral components of the stimulus narrative that vary over a certain number of words. These timescale-specific feature spaces were then used to estimate voxelwise encoding models that describe how different timescales of language are represented in the brain for each modality and participant separately. The voxelwise encoding models were used to determine the language timescale selectivity of each voxel, for each participant and modality separately. The language timescale selectivity of individual voxels was compared between reading and listening. The experimental procedure is summarized in Figure 3.5 and is detailed in the following subsections.

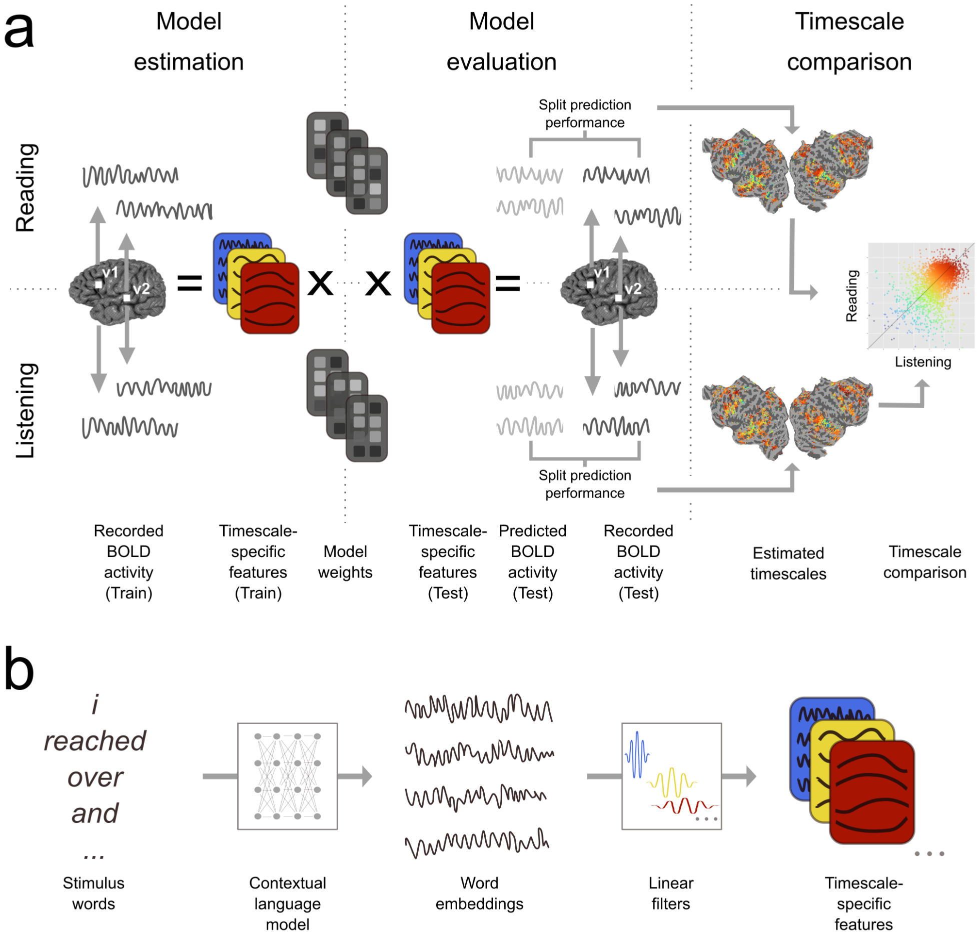
a
Model estimation
Model evaluation
Timescale comparison
Reading
Listening
v1
v2
Split prediction performance
Split prediction performance
Reading
Listening
Recorded BOLD activity (Train)
Timescale-specific features (Train)
Model weights
Timescale-specific features (Test)
Predicted BOLD activity (Test)
Recorded BOLD activity (Test)
Estimated timescales
Timescale comparison
b
i
reached
over
and
...
Stimulus words
Contextual language model
Word embeddings
Linear filters
Timescale-specific features

Figure 3.5: Experimental procedure and voxelwise modeling. The following procedure was used to compare the representation of different language timescales across the cerebral cortex. a. Functional MRI signals were recorded while participants listened to or read narrative stories (Deniz et al., 2019; Huth et al., 2016). Timescale-specific feature spaces were constructed, each of which reflects the components of the stimulus that occur at a specific timescale (See (b) for details). These feature spaces and BOLD responses were used to estimate voxelwise encoding models that indicate how different language timescales modulate the BOLD signal evoked in each voxel, separately for each participant and modality ("Model estimation"). Estimated model weights were used to predict BOLD responses to a separate held-out dataset which was not used for model estimation ("Model evaluation"). Predictions for individual participants were computed separately for listening and reading sessions. Prediction performance was quantified as the correlation between the predicted and recorded BOLD responses to the held-out test dataset. This prediction performance was used to determine the selectivity of each voxel to language structure at each timescale. These estimates were then compared between reading and listening ("Timescale comparison"). b. Timescale-specific feature spaces were constructed from the presented stimuli. A contextual language model (BERT (Devlin et al., 2019)) was used to construct a vector embedding of the stimulus. The resulting stimulus embedding was decomposed into components at specific timescales. To perform this decomposition, the stimulus embedding was convolved across time with each of eight linear filters. Each linear filter was designed to extract components of the stimulus embedding that vary with a specific period. This convolution procedure resulted in eight sets of stimulus embeddings, each of which reflects the components of the stimulus narrative that vary at a specific timescale. These eight sets of stimulus embeddings were used as timescale-specific feature spaces in (a).

MRI data collection

MRI data were collected on a 3T Siemens TIM Trio scanner located at the UC Berkeley Brain Imaging Center. A 32-channel Siemens volume coil was used for data acquisition. Functional scans were collected using gradient echo EPI water excitation pulse sequence with the following parameters: repetition time (TR) 2.0045 s; echo time (TE) 35 ms; flip angle 74 degrees; voxel size 2.24 x 2.24 x 4.1 mm (slice thickness 3.5 mm with 18% slice gap); matrix size 100 x 100; and field of view 224 x 224 mm. To cover the entire cortex, 30 axial slices were prescribed and these were scanned in interleaved order. A custom-modified bipolar water excitation radiofrequency (RF) pulse was used to avoid signal from fat. Anatomical data were collected using a T1-weighted multi-echo MP-RAGE sequence on the same 3T scanner.

To minimize head motion during scanning and to optimize alignment across sessions, each participant wore a customized, 3D-printed or milled head case that matched precisely the shape of each participant’ s head (Gao, 2015; Power et al., 2019). In order to account for inter-run variability, within each run MRI data were z-scored across time for each voxel separately. The data presented here have been presented previously as part of other studies that examined questions unrelated to timescales in language processing (de Heer et al., 2017; Deniz et al., 2019; Huth et al., 2016). Motion correction and automatic alignment were performed on the fMRI data using the FMRIB Linear Image Registration Tool (FLIRT) from FSL 5.0 (Jenkinson et al., 2012). Low-frequency voxel response drift was removed from the data using a third-order Savitzky-Golay filter with a 120s window (for data preprocessing details see (Deniz et al., 2019)).

Participants

Functional data were collected on nine participants (six males and three females) between the ages of 24 and 36. All procedures were approved by the Committee for Protection of Human Subjects at the University of California, Berkeley. All participants gave informed consent. All ethical regulations relevant to human research participants were followed. All participants were healthy, had normal hearing, and had normal or corrected-to-normal vision. The Edinburgh handedness inventory (Oldfield, 1971b) indicated that one participant was left handed. The remaining eight participants were right handed or ambidextrous.

Because the current study used a voxelwise encoding model framework, each participant’s data were analyzed individually, and both statistical significance and out-of-set prediction accuracy (i.e., generalization) are reported for each participant separately. Because each participant provides a complete replication of all hypothesis tests, sample size calculations were neither required nor performed.

Stimuli

Human participants read and listened to a set of English narrative stories while in the fMRI scanner. The same stories were used as stimuli for reading and listening sessions and the same stimuli were presented to all participants. These stories were originally presented at The Moth Radio Hour. In each story, a speaker tells an autobiographical story in front of a live audience. The selected stories cover a wide range of topics and are highly engaging. The stories were separated into a model training dataset and a model test dataset. The model training dataset consisted of ten 10-15 min stories. The model test dataset consisted of one 10 min story. This test story was presented twice in each modality (once during each scanning session). The responses to the test story were averaged within each modality (for details see (Huth et al., 2016) and (Deniz et al., 2019)). Each story was played during a separate fMRI scan. The length of each scan was tailored to the story and included 10s of silence both before and after the story. Listening and reading presentation order was counterbalanced across participants.

During listening sessions the stories were played over Sensimetrics S14 in-ear piezoelectric headphones. During reading sessions the words of each story were presented one-by-one at the center of the screen using a rapid serial visual presentation (RSVP) procedure (Buchweitz, Mason, Tomitch, & Just, 2009; Forster, 1970). Each word was presented for a duration precisely equal to the duration of that word in the spoken story. The stories were shown on a projection screen at 13 x 14 degrees of visual angle. Participants were asked to fixate while reading the text. (For details about the experimental stimuli see (Deniz et al., 2019)).

Voxelwise encoding models

Voxelwise modeling (VM) was used to model BOLD responses (de Heer et al., 2017; Deniz et al., 2019; Huth et al., 2016; Naselaris et al., 2011; M. C.-K. Wu et al., 2006). In the VM framework, stimulus and task parameters are nonlinearly transformed into sets of features (also called “feature spaces”) that are hypothesized to be represented in brain responses. Linearized regression is used to estimate a separate model for each voxel. Each model predicts brain responses from each feature space (a model that predicts brain responses from stimulus features is referred to as an “encoding model”). The encoding model describes how each feature space is represented in the responses of each voxel. A held-out dataset that was not used for model estimation is then used to evaluate model prediction performance on new stimuli and to determine the significance of the model prediction performance.

Construction of timescale-specific feature spaces

To operationalize the notion of language timescales, the language stimulus was treated as a time series and different language timescales were defined as the different frequency components of this time series. Although this operational definition is not explicitly formulated in terms of classic language abstractions such as sentences or narrative chains, the resulting components nonetheless selectively capture information corresponding to the broad timescales of words, sentences, and discourses (Tamkin et al., 2020). To construct timescale-specific feature spaces, first an artificial neural language model ("BERT"(Devlin et al., 2019)) was used to project the stimulus words onto a contextual word embedding space. This projection formed a stimulus embedding that reflects the language content in the stimuli. Then, linear filters were convolved with the stimulus embedding to extract components that each vary at specific timescales. These two steps are detailed in the following two paragraphs.

Embedding extraction An artificial neural network (BERT-BASE-UNCASED (Devlin et al., 2019)) was used to construct the initial stimulus embedding. BERT-base is a contextual language model that contains a 768-unit embedding layer and 12 transformer layers, each with a 768-unit hidden state (for additional details about the BERT-base model see (Devlin et al., 2019)). The w words of each stimulus narrative X were tokenized and then provided one sentence at a time as input to the pretrained BERT-base model (sentence-split inputs were chosen as input context because sentence-level splits mimic the inputs provided to BERT during pretraining). For each stimulus word, the activation of each of the $p = 13 \times 768 = 9984$ units of BERT was used as a p-dimensional embedding of that word. Prior work suggested that language structures with different timescales are preferentially represented in different layers of BERT(Jawahar et al., 2019; Rogers et al., 2021; Tenney et al., 2019) (though some have argued that language timescales are not cleanly separated across different layers of BERT(Niu et al., 2022)). Earlier layers represent lower-level, shorter-timescale information (e.g., word identity and linear word order), whereas later layers represent higher-level, longer-timescale information (e.g., coreference, long-distance dependencies). To include stimulus information at all levels of the language processing hierarchy, activations from all layers of BERT were included in the stimulus embedding. The embeddings of the w stimulus words form a $p \times w$ stimulus embedding $M(X)$. $M(X)$ numerically represents the language content of the stimulus narratives.

Timescale separation The stimulus embedding derived directly from BERT can explain a large proportion of the variance in brain responses to language stimuli (Caucheteux & King, 2022; Lamarre et al., 2022; Schrimpf et al., 2021; Toneva & Wehbe, 2019). However, this stimulus embedding does not distinguish between different language timescales.

In order to distinguish between different language timescales, linear filters were used to decompose the stimulus embedding $M(X)$ into different language timescales. Intuitively, the stimulus embedding consists of components that vary with different periods. Components that vary with different periods can be interpreted in terms of different classical language structures (Tamkin et al., 2020). For example, components that vary with a short period (∼2-4 words) reflect clause-level structures such as syntactic complements, components that vary with an intermediate period (∼16-32 words) reflect sentence-level structures such as constituency parses, and components that vary with a long period (∼128-256 words) reflect paragraph-level structures such as semantic focus. To reflect this intuition, different language timescales were operationalized as the components of $M(X)$ with periods that fall within different ranges. The period ranges were chosen to be small enough to model timescale selectivity at a fine-grained temporal granularity, and large enough to avoid substantial spectral leakage which would contaminate the output of each filter with components outside the specified timescale. The predefined ranges were chosen as: 2-4 words, 4-8 words, 8-16 words, 16-32 words, 32-64 words, 64-128 words, 128-256 words, and 256+ words. To decompose the stimulus embedding into components that fall within these period ranges, eight linear filters b_i $(i \in 1, 2, .., 8)$ were constructed. Each filter b_i was designed to extract components that vary with a period in the predefined range. The window method for filter design was used to construct each filter (F. J. Harris, 1978). Each linear filter was constructed by multiplying a cosine wave with a blackman window (Blackman & Tukey, 1958). The stimulus embedding $M(X)$ was convolved with each of the eight filters separately to produce eight filtered embeddings $M_i(X)$, $i \in 1, 2, ..., 8$, each with dimension $p \times w$. To avoid filter distortions at the beginning and end of the stimulus, a mirrored version of $M(X)$ was concatenated to the beginning and end of $M(X)$ before the filters were applied to $M(X)$. Each filtered embedding $M_i(X)$ contains the components of the stimulus embedding that vary at the timescale extracted by the i-th filter.

Construction of sensory-level feature spaces

Two sensory-level feature spaces were constructed in order to account for the effect of low-level sensory information on BOLD responses. One feature space represents low-level visual information. This feature space was constructed using a spatiotemporal Gabor pyramid that reflects the spatial and motion frequencies of the visual stimulus (for details see (Deniz et al., 2019), (Nishimoto et al., 2011), and (Nakai et al., 2021)). The second feature space represents low-level auditory information. This feature space was constructed using a cochleogram model that reflects the spectral frequencies of the auditory stimulus (for details see (de Heer et al., 2017), (Deniz et al., 2019), and (Nakai et al., 2021)).

Stimulus downsampling

Feature spaces were downsampled in order to match the sampling rate of the fMRI recordings. The eight filtered timescale-specific embeddings $M_i(X)$ contain one sample for each word. Because word presentation rate of the stimuli is not uniform, directly downsampling the timescale-specific embeddings $M_i(X)$ would conflate long-timescale embeddings with the presentation word rate of the stimulus narratives (Jain et al., 2020). To avoid this problem, a Gaussian radial basis function (RBF) kernel was used to interpolate $M_i(X)$ in order to form intermediate signals $M_i'(X)$, following (Jain et al., 2020). Each $M_i'(X)$ has a constant sampling rate of 25 samples per repetition time (TR). After this interpolation step, an anti-aliasing, 3-lobe Lanczos filter with cut-off frequency set to the fMRI Nyquist rate (0.25 Hz) was used to resample the intermediate signals $M_i'(X)$ to the middle timepoints of each of the n fMRI volumes. This procedure produced eight timescale-specific feature spaces $F_i(X)$, each of dimension $p \times n$. Each of these feature spaces contains the components of the stimulus embedding that vary at a specific timescale. These feature spaces are sampled at the sampling rate of the fMRI recordings. The sensory-level feature spaces were not sampled at the word presentation rate. Therefore Gaussian RBF interpolation was not applied to sensory-level feature spaces.

Before voxelwise modeling, each stimulus feature was truncated, z-scored, and delayed. Data for the first 10 TRs and the last 10 TRs of each scan were truncated to account for the 10 seconds of silence at the beginning and end of each scan and to account for non-stationarity in brain responses at the beginning and end of each scan. Then the stimulus features were each z-scored in order to account for z-scoring performed on the MRI data (for details see "MRI data collection"). In the z-scoring procedure, the value of each feature channel was separately normalized by subtracting the mean value of the feature channel across time and then dividing by the standard deviation of the feature channel across time. Note that the resulting feature spaces had low correlation with each other – for each pair of feature spaces, the mean pairwise correlation coefficient between dimensions of the feature spaces was less than 0.1. Lastly, finite impulse response (FIR) temporal filters were used to delay the features in order to model the hemodynamic response function of each voxel. The FIR filters were implemented by concatenating feature vectors that had been delayed by 2, 4, 6, and 8 seconds(Deniz et al., 2019; Huth et al., 2016; Nakai et al., 2021).

Voxelwise encoding model fitting

Voxelwise encoding models were estimated in order to determine which features are represented in each voxel. Each model consists of a set of regression weights that describes BOLD responses in a single voxel as a linear combination of the features in a particular feature space. In order to account for potential complementarity between feature spaces, the models were jointly estimated for all ten feature spaces: the eight timescale-specific feature spaces, and the two sensory-level feature spaces (the two sensory-level feature spaces reflect spectrotemporal features of the auditory stimulus and motion energy features of the visual stimulus) (Dupré la Tour et al., 2022; Nunez-Elizalde et al., 2019).

Regression weights were estimated using banded ridge regression (Nunez-Elizalde et al., 2019). Unlike standard ridge regression, which assigns the same regularization parameter to all feature spaces, banded ridge regression assigns a separate regularization hyperparameter to each feature space. Banded ridge regression thereby avoids biases in estimated model weights that could otherwise be caused by differences in feature space distributions. Mathematically, the m delayed feature spaces $F_i(X), i \in 1, 2, ..., m$ (each of dimension p) were concatenated to form a feature matrix $F'(X)$ (dimension $(m \times p) \times n$). Then banded ridge regression was used to estimate a mapping B (dimension $v \times (\sum_{i=1}^{f} p)$) from $F'(X)$ to the matrix of voxel responses Y (dimension $v \times n$). B is estimated according to $\hat{B} = \arg\min_B ||Y - BF'(X)||_2^2 + \lambda||CB||_2^2$. A separate regularization parameter was fit for each voxel, feature space, and FIR delay. The diagonal matrix C of regularization hyperparameters for each feature space and each voxel is optimized over 10-fold cross-validation. See Section 3.5 for details.

Regularization hyperparameter selection

Data for the ten narratives in the training dataset were used to select regularization hyperparameters for banded ridge regression. 10-fold cross-validation was used to find the optimal regularization hyperparameters for each feature space and each voxel. Regularization hyperparameters were chosen separately for each participant and modality. In each fold, data for nine of the ten narratives were used to estimate an encoding model and the tenth narrative was used to validate the model. The regularization hyperparameters for each feature space and voxel were selected as the hyperparameters that produced the minimum squared error (L2) loss between the predicted voxel responses and the recorded voxel responses ($\arg\min_{hyperparameters} ||\hat{y} - y||_2^2$). Because evaluating k regularization hyperparameters for m feature spaces requires k^m iterations ($10^{10} = 10,000,000,000$ model fits in our case), it would be impractical to conduct a grid search over all possible combinations of hyperparameters. Instead, a computationally efficient two-stage procedure was used to search for hyperparameters (Dupré la Tour et al., 2022). The first stage consisted of 1000 iterations of a random hyperparameter search procedure (Bergstra & Bengio, 2012). 1000 normalized hyperparameter candidates were sampled from a dirichlet distribution and were then scaled by 10 log-spaced values ranging from 10^{-5} to 10^{5}. Then the voxels with the lowest 20% of the cross-validated L2 loss were selected for refinement in the second stage. The second stage consisted of 1000 iterations of hyperparameter gradient descent (Bengio, 2000). This stage was used to refine the hyperparameters selected during the random search stage. This hyperparameter search was performed using the Himalaya Python package (Dupré la Tour et al., 2022). Note that hyperparameter selection in banded ridge regression acts as a feature-selection mechanism that helps account for stimulus feature correlations (Dupré la Tour et al., 2022).

Model estimation and evaluation

The selected regularization hyperparameters were used to estimate regression weights that map from the timescale-specific feature spaces to voxel BOLD responses. Regression weights were estimated separately for each voxel in each modality and participant. The test dataset was not used to select hyperparameters or to estimate regression weights. The joint prediction performance r of the combined feature spaces was computed per voxel as the Pearson correlation coefficient between the predicted voxel responses and the recorded voxel responses. The split-prediction performance $\tilde{r}$ was used to determine how much each feature space contributed to the joint prediction performance r. The split-prediction performance decomposes the joint prediction performance r of all the feature spaces into the contribution $\tilde{r}_i, i \in 1, 2, ...m$ of each feature space. The split-prediction performance is computed as $\tilde{r}_i = \frac{\sum_t \hat{Y}_i[t]Y[t]}{\sqrt{(\sum_t \hat{Y}[t]^2)(\sum_t Y[t]^2)}}$, where t denotes each timepoint (further discussion of this metric can be found in (St-Yves & Naselaris, 2018) and (Dupré la Tour et al., 2022)).

Language-selective voxel identification

The set of "language-selective voxels" was operationally defined as the set of voxels that are accurately predicted by any of the eight timescale-specific feature spaces. To identify this set of voxels, the split-prediction performance was used. The total contribution $\tilde{r}_{all\_timescales}$ of the eight timescale-specific feature spaces to predicting the BOLD responses in each voxel was computed as the sum of the split-prediction performance for each of the eight timescales $\tilde{r}_{all\_timescales} = \sum_{i=1}^{8} \tilde{r}_{timescale_i}$. The significance of $\tilde{r}_{all\_timescales}$ was computed by a permutation test with 1000 iterations. At each permutation iteration, the timecourse of the held-out test dataset was permuted by blockwise shuffling (shuffling was performed in blocks of 10 TRs in order to account for autocorrelations in voxel responses (Deniz et al., 2019; Jain et al., 2020)). The permuted timecourse of voxel responses was used to produce a null estimate of $\tilde{r}_{all\_timescales}$. These permutation iterations produced an empirical distribution of 1000 null estimates of $\tilde{r}_{all\_timescales}$ for each voxel. This distribution of null values was used to obtain the p-value of $\tilde{r}_{all\_timescales}$ for each voxel separately. A false discovery rate (FDR) procedure was used to correct the resulting p-values for multiple comparisons within each participant and modality (Benjamini & Hochberg, 1995). A low p-value indicates that the timescale-specific feature spaces significantly contributed to accurate predictions of BOLD responses in the joint model. Voxels with a one-sided FDR-corrected p-value of less than $p < .05$ were identified as language-selective voxels. The set of language-selective voxels was identified separately for each participant and modality.

Voxel timescale selectivity estimation

The encoding model estimated for each voxel was used to determine voxel timescale selectivity, which reflects the average language timescale for which a voxel is selective. In order to compute timescale selectivity, first the timescale selectivity profile ($\tilde{r}'$) was computed. The timescale selectivity profile reflects the selectivity of each voxel to each of the eight timescale-specific feature spaces. This metric is computed by normalizing the vector of split-prediction performances of the eight timescale-specific feature spaces to form a proper set of proportions: $\tilde{r}'_{timescale_i} = \frac{max(0, \tilde{r}_{timescale_i})}{\sum_{j=1}^{8} max(0, \tilde{r}_{timescale_j})}$.

Comparing each index of the timescale selectivity profile separately cannot distinguish between cases in which a voxel represents similar timescales between reading and listening (e.g., 2-4 words for reading and 4-8 words for listening) and cases in which a voxel represents very different timescales between the two modalities (e.g., 2-4 words for reading and 128-256 words for listening). Therefore, we computed the timescale selectivity $\bar{\mathbf{T}}$ for each voxel, which reflects the average timescale of language to which a voxel is selective (we use the weighted average instead of simply taking the maximum selectivity across timescales, in order to prevent small changes in prediction accuracy from producing large changes in estimated timescale selectivity). To compute voxel timescale selectivity, first the timescale t_i of each feature space $F_i(X)$ was defined as the center of the period range of the respective filter b_i: $t_i = \frac{p_{i,low}+p_{i,high}}{2}$, where $(p_{i,low}, p_{i,high})$ indicates the upper and lower end of the period range for filter i. Then, timescale selectivity was defined as a weighted sum of each feature space log-timescale: $\bar{\mathbf{T}} = 2\widehat{}(\sum_{i=1}^{8}(\tilde{r}_i' \log_2(t_i)))$. Timescale selectivity was computed separately for each voxel, participant, and modality.

Voxel timescale comparison

To compare timescale selectivity between modalities, the Pearson correlation coefficient was computed between timescale selectivity during reading and listening across the set of voxels that are language-selective in both modalities. The significance of this correlation was determined by a permutation test with 1000 iterations. At each iteration and for each modality separately, the timecourse of recorded voxel responses was shuffled. The timecourses were shuffled in blocks of 10 TRs in order to account for autocorrelations in voxel responses. The shuffled timecourses of recorded voxel responses were used to compute a null value for the timescale selectivity of each voxel for each modality separately. The null values of timescale selectivity were correlated between reading and listening to form an empirical null distribution. This null distribution was used to determine the p-value of the observed correlation between timescale selectivity during reading and listening. Significance was computed for each participant separately.

In addition, for each of the eight timescales separately the Pearson correlation coefficient was computed between selectivity for that timescale during reading and listening. This correlation was performed across the set of voxels that are language-selective in both modalities. The significance of the observed correlations were computed by a permutation test. At each of 1000 iterations the timecourse of recorded voxel responses was shuffled and then the shuffled voxel responses were used to compute null values of the timescale selectivity profile. For each timescale-specific feature space separately, the null values of the timescale selectivity profile were used to compute an empirical null distribution for the correlation between selectivity for that feature space during reading and listening. These null distributions were used to determine the p-value of the observed correlations. Significance was computed for each participant and for each timescale-specific feature space separately.

3.6 Supplementary Figures

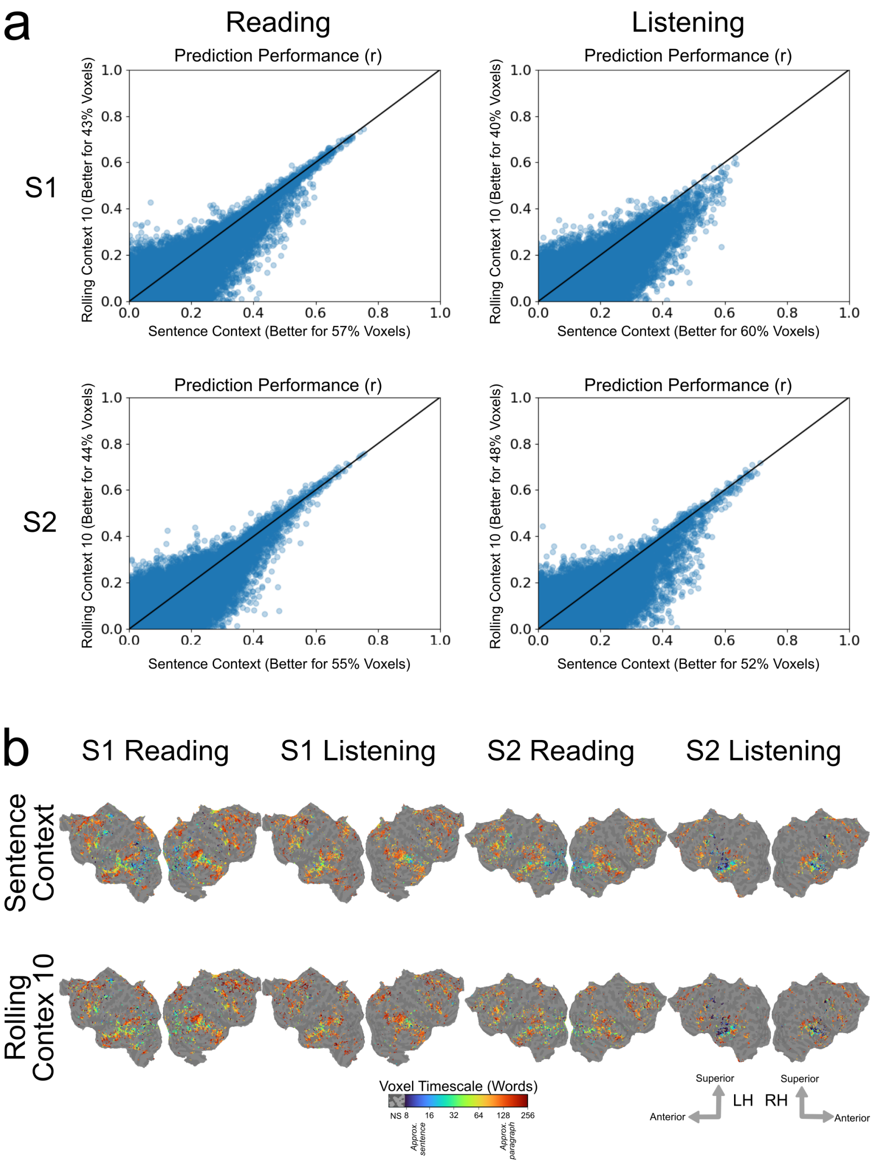

Figure 3.6: Comparison of embedding extraction methods, sentence input context vs fixed rolling input context. a. Encoding model prediction performance (r) obtained from a sentence-split input context (x-axis), and from a rolling input context of 10 words (y-axis). Each point represents one voxel. Axis labels indicate the percentage of voxels for which the respective embedding extraction method produces better performance (some voxels are predicted similarly well with both embedding extraction methods; therefore percentages may not sum to 100). A sentence-split input context produces more accurate predictions of brain responses than a rolling input context of 10 words. b. Timescale selectivity is shown for two representative participants (S1 and S2) and for reading and listening separately. Timescale selectivity is

shown according to the color scale at the bottom. Voxels that were not significantly predicted are shown in grey (one-sided permutation test, $p < .05$, FDR corrected). For both embedding extraction methods, temporal cortex contains a spatial gradient from intermediate to long timescale selectivity along the superior to inferior axis, prefrontal cortex (PFC) contains a spatial gradient from intermediate to long timescale selectivity along the posterior to anterior axis, and precuneus is predominantly selective for long timescales. While Panel A shows that a sentence-split input context length produces more accurate models of brain responses, estimates of timescale selectivity are robust to the input context method.

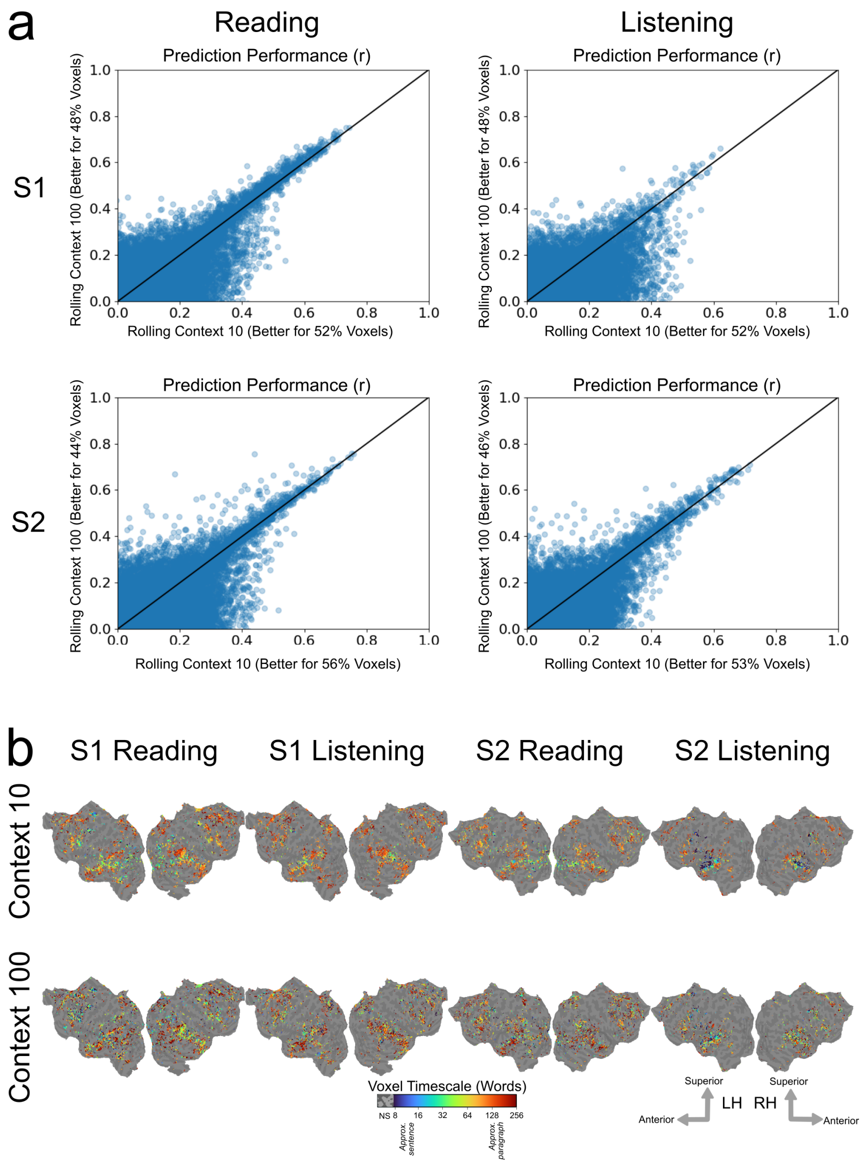

Figure 3.7: Comparison of embedding extraction methods, input context length 10 words vs input context length 100 words. a. Encoding model prediction performance (r) using a rolling input context of 10 words (x-axis) vs a rolling input context of 100 words (y-axis). Each point represents one voxel. Axis labels indicate the number of voxels for which the respective embedding extraction method produces better performance (some voxels are predicted similarly well with both embedding extraction methods; therefore, percentages may not sum to 100.) An input context of 10 words produces more accurate predictions of brain responses than an input context of 100 words. b. Timescale selectivity is shown for two representative participants (S1 and S2), for reading and listening separately. For each significantly predicted voxel, timescale selectivity is shown according to the color scale at the bottom. Voxels that were not significantly predicted are shown in grey (one-sided permutation test, $p < .05$, FDR corrected). For both

embedding extraction methods, temporal cortex contains a spatial gradient from intermediate to long timescale selectivity along the superior to inferior axis, prefrontal cortex (PFC) contains a spatial gradient from intermediate to long timescale selectivity along the posterior to anterior axis, and precuneus is predominantly selective for long timescales. These results suggest using a shorter input context produces more accurate models of brain responses, and that estimates of timescale selectivity are similar between different input context lengths.

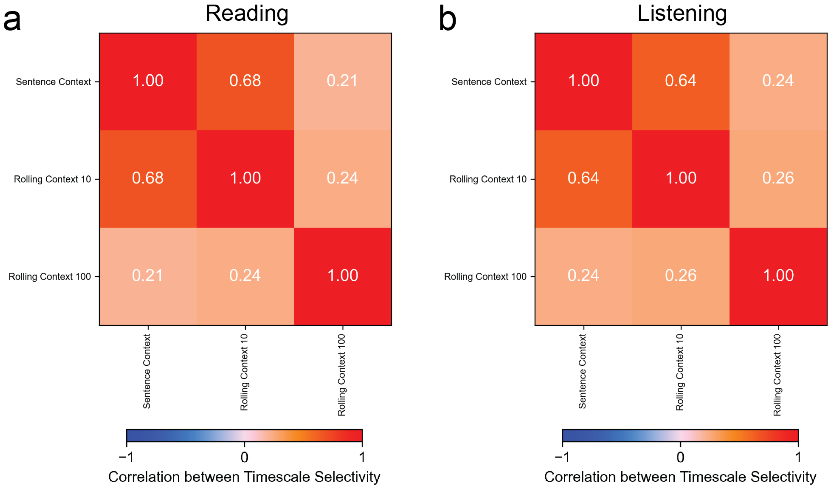

Figure 3.8: Comparison of estimated timescale selectivity between different input context lengths. Timescale selectivity was estimated separately with a sentence-length stimulus input context, a rolling 10-word input context, and a rolling 100-word input context. For each pair of input contexts, group-averaged voxelwise spatial correlation between estimated timescale selectivity is shown, for reading (a) and listening (b) separately. The sentence-length context and rolling 10-word context produce correlated estimates of timescale selectivity. The rolling 100-word context produces estimates of timescale selectivity that are less similar, but still positively correlated with estimates from the other two embedding methods. Furthermore, the similarity between sentence-length and rolling input contexts suggests that the inclusion of future context in sentence-length input contexts does not qualitatively change estimates of timescale selectivity.

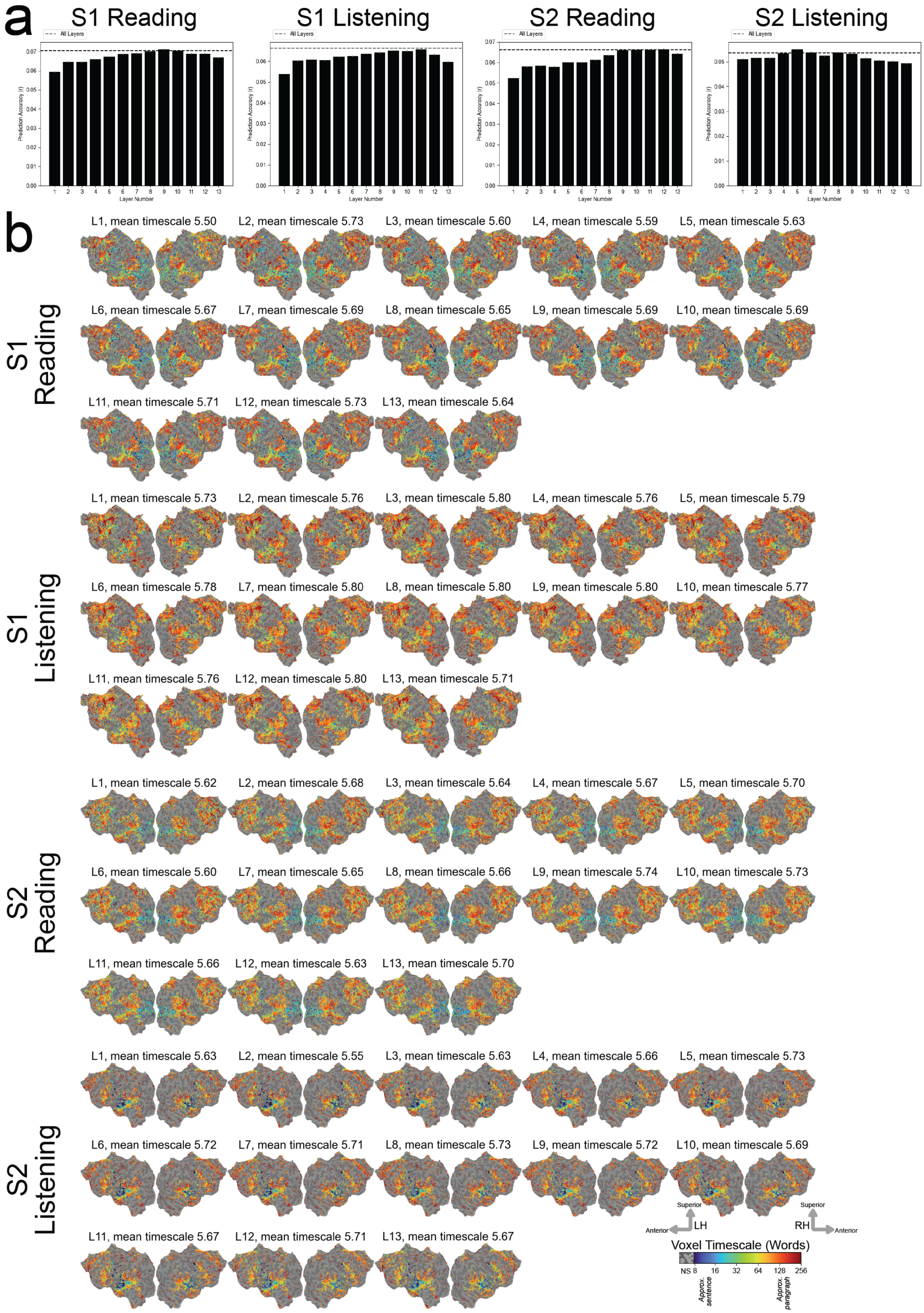
a
S1 Reading
S1 Listening
S2 Reading
S2 Listening
All Layers
Prediction Accuracy (r)
Layer Number
b
S1 Reading
L1, mean timescale 5.50
L2, mean timescale 5.73
L3, mean timescale 5.60
L4, mean timescale 5.59
L5, mean timescale 5.63
L6, mean timescale 5.67
L7, mean timescale 5.69
L8, mean timescale 5.65
L9, mean timescale 5.69
L10, mean timescale 5.69
L11, mean timescale 5.71
L12, mean timescale 5.73
L13, mean timescale 5.64
S1 Listening
L1, mean timescale 5.73
L2, mean timescale 5.76
L3, mean timescale 5.80
L4, mean timescale 5.76
L5, mean timescale 5.79
L6, mean timescale 5.78
L7, mean timescale 5.80
L8, mean timescale 5.80
L9, mean timescale 5.80
L10, mean timescale 5.77
L11, mean timescale 5.76
L12, mean timescale 5.80
L13, mean timescale 5.71
S2 Reading
L1, mean timescale 5.62
L2, mean timescale 5.68
L3, mean timescale 5.64
L4, mean timescale 5.67
L5, mean timescale 5.70
L6, mean timescale 5.60
L7, mean timescale 5.65
L8, mean timescale 5.66
L9, mean timescale 5.74
L10, mean timescale 5.73
L11, mean timescale 5.66
L12, mean timescale 5.63
L13, mean timescale 5.70
S2 Listening
L1, mean timescale 5.63
L2, mean timescale 5.55
L3, mean timescale 5.63
L4, mean timescale 5.66
L5, mean timescale 5.73
L6, mean timescale 5.72
L7, mean timescale 5.71
L8, mean timescale 5.73
L9, mean timescale 5.72
L10, mean timescale 5.69
L11, mean timescale 5.67
L12, mean timescale 5.71
L13, mean timescale 5.67
Superior
Anterior
LH
RH
Voxel Timescale (Words)
NS 8 16 32 64 128 256
Approx. sentence
Approx. paragraph

Figure 3.9: Comparison of embedding extraction methods, all layers of BERT vs only a single layer. Timescale selectivity was estimated with embeddings from each layer of BERT. Results are shown for two representative participants (S1 and S2). a. Bars denote number of significantly predicted voxels for each layer. Errorbars denote standard error. The dashed line denotes the number of significantly predicted voxels obtained from using all layers of BERT together. Including all layers of BERT generally produces better predictions of brain responses than only using a single layer. b. Timescale selectivity estimated with embeddings from each layer separately. For each significantly predicted voxel, timescale selectivity is shown according to the color scale at the bottom. Voxels that were not significantly predicted are shown in grey (one-sided permutation test, $p < .05$, FDR corrected). Estimates of timescale selectivity are similar between the different embedding methods –spatial gradients from intermediate to long timescale selectivity are found along the superior to inferior axis of temporal cortex and along the posterior to anterior axis of prefrontal cortex (PFC), and precuneus is predominantly selective for long timescales. Embeddings from only a single layer of BERT often result in slightly worse prediction performance, but the choice of layer does not substantially affect estimates of timescale selectivity.

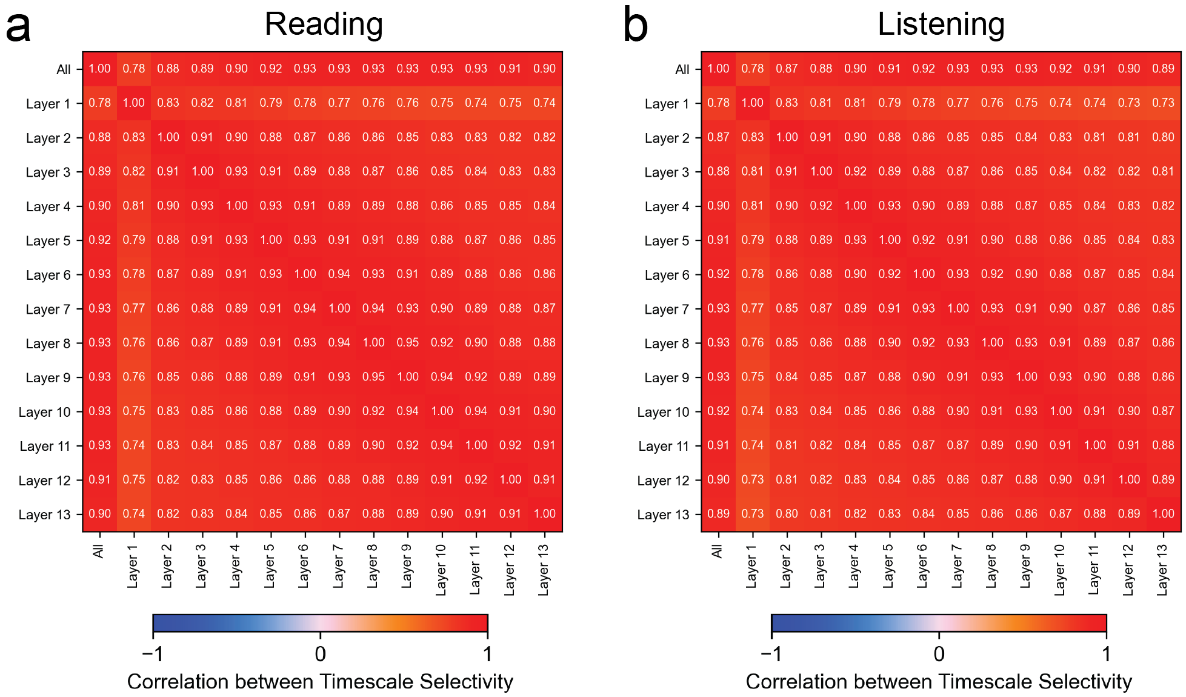

Figure 3.10: Comparison of estimated timescale selectivity between different embedding layers. Timescale selectivity was estimated separately with embeddings from each layer of BERT. Group-averaged voxelwise spatial correlation between timescale selectivity is shown for each pair of layers, for reading (a) and listening (b) separately. Estimates of timescale selectivity are highly correlated across layers. Estimates are more similar for layers that are closer together, suggesting a small effect of stimulus layer on estimates of timescale selectivity. Overall, estimates of timescale selectivity are consistent across different layers.

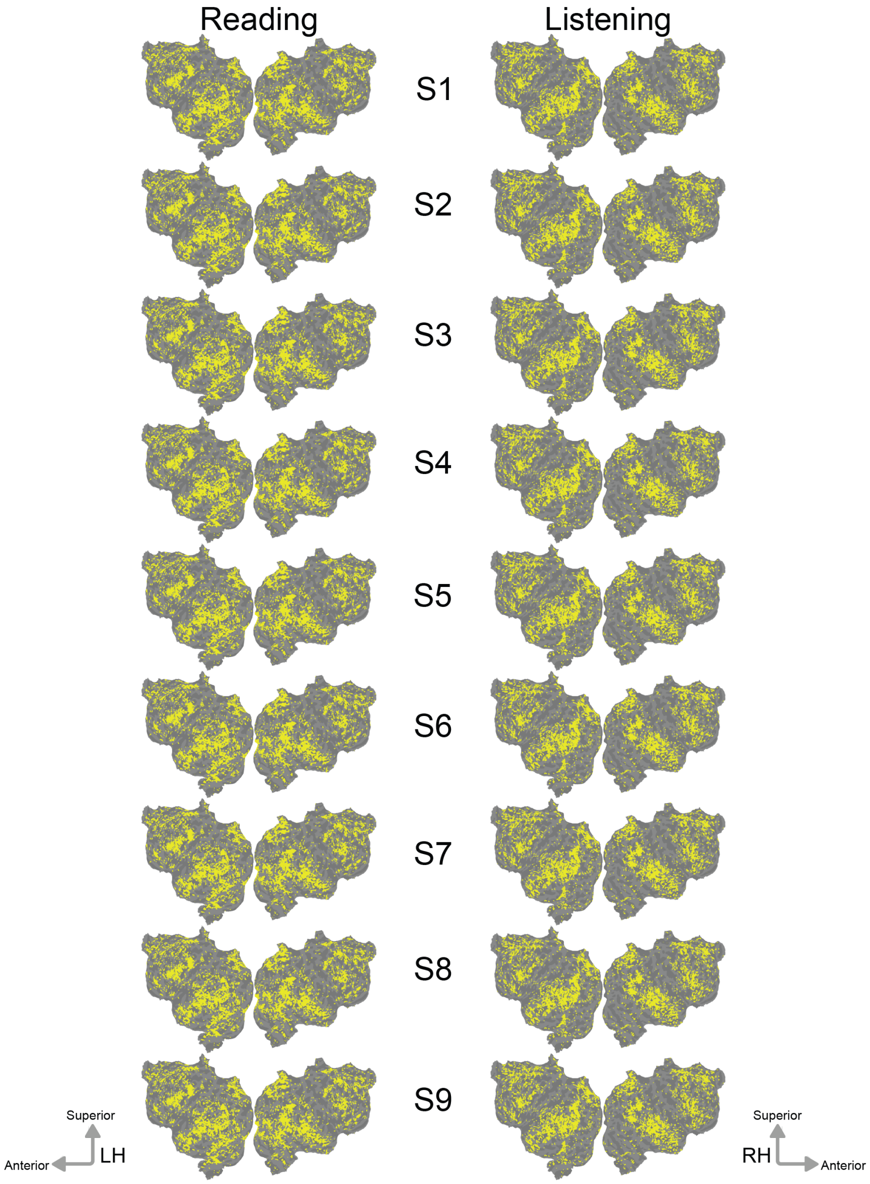

Figure 3.11: Language-selective voxels in each modality. The set of language-selective voxels is shown for reading and listening separately on the flattened cortical surface of each participant. Language-selective voxels are shown in yellow. For both modalities, voxels across temporal, parietal, and prefrontal cortices are language-selective.

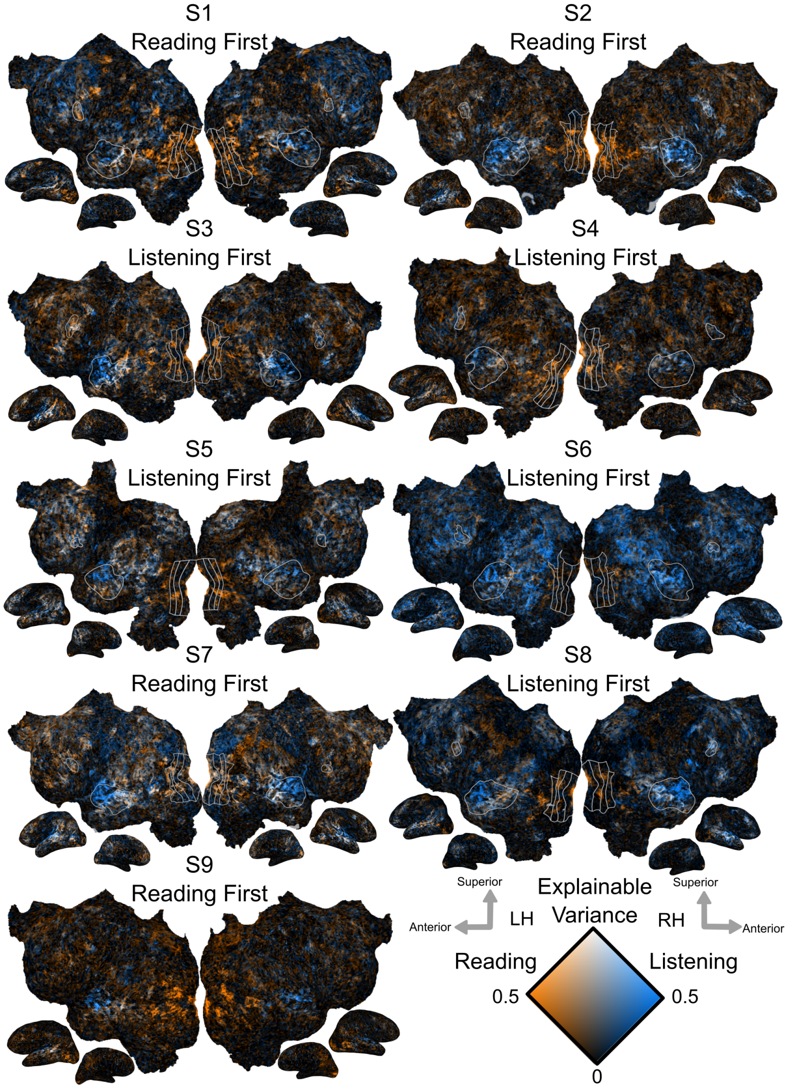

Figure 3.12: Explainable variance in reading vs listening. The explainable variance (EV) for reading and listening is shown on the flattened cortical surface of each participant. The first presented modality for each participant is indicated by the figure subtitles. Orange voxels had high EV for reading. Blue voxels had high EV for listening. White voxels had high EV for both modalities. The stimulus modality with stronger EV varies across participants, possibly due to individual participant preferences for certain modalities or because of small differences in noise across sessions. The EV of a voxel is computed using the measured BOLD response in a voxel over N repetitions of a stimulus with T timepoints $y_1, ...y_N \in \mathbb{R}^T$ as follows (Y must be zscored across time):
$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$
$r_i = y_i - \bar{y}$
$EV = \frac{1}{N}\sum_{i=1}^{N} Var(y_i) - \frac{N}{N-1}\sum_{i=1}^{N} Var(r_i)$

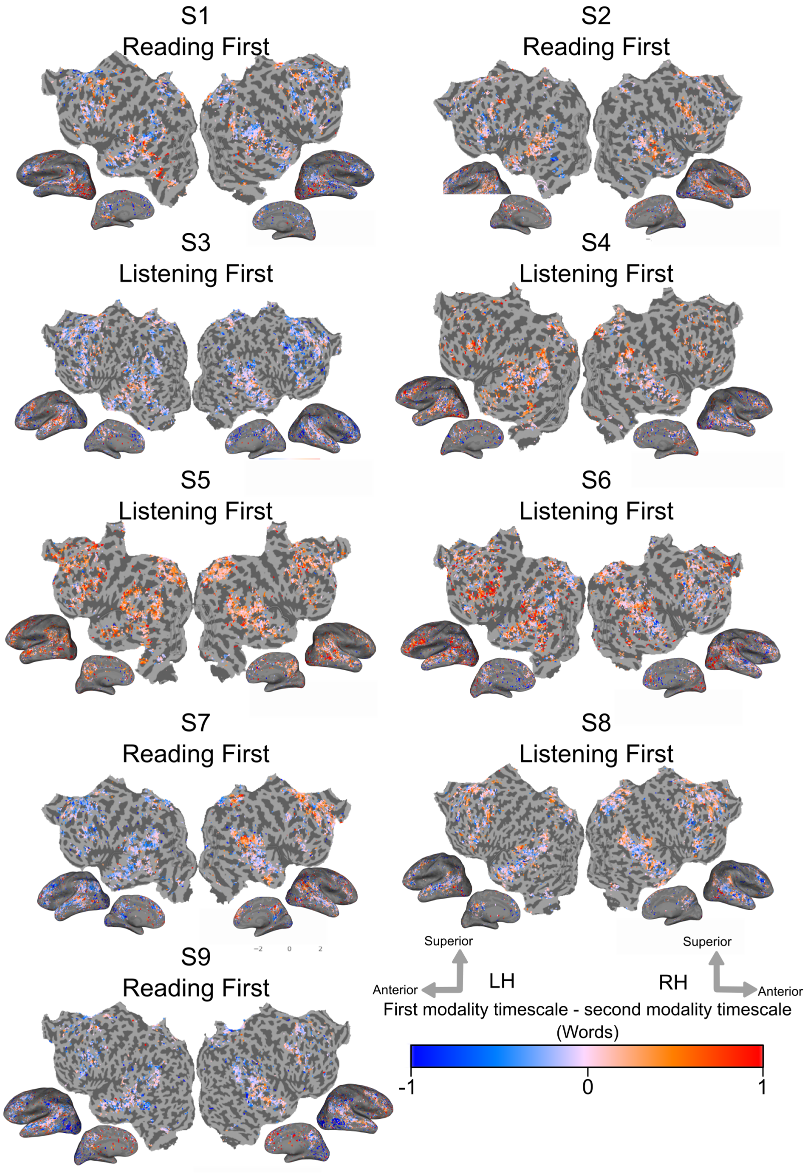

Figure 3.13: Difference between timescale selectivity in first and second presented modalities. The difference between timescale selectivity in the first and second presented modalities is shown on the flattened cortical surface of each participant. The first presented modality for each individual participant is indicated by the figure subtitles. Red voxels have longer timescale selectivity for the first presented modality. Blue voxels have longer timescale selectivity for the second presented modality. Voxels shown in grey were not significantly predicted in both modalities. Timescale selectivity is overall longer in the first presented modality than in the second presented modality in six participants (S1, S2, S4, S5, S6, S8; $p < .05$ by a two-sided t-test for paired samples). Voxel timescale selectivity is on average longer in the second presented modality than in the first presented modality in the other three participants (S3, S7, S9).

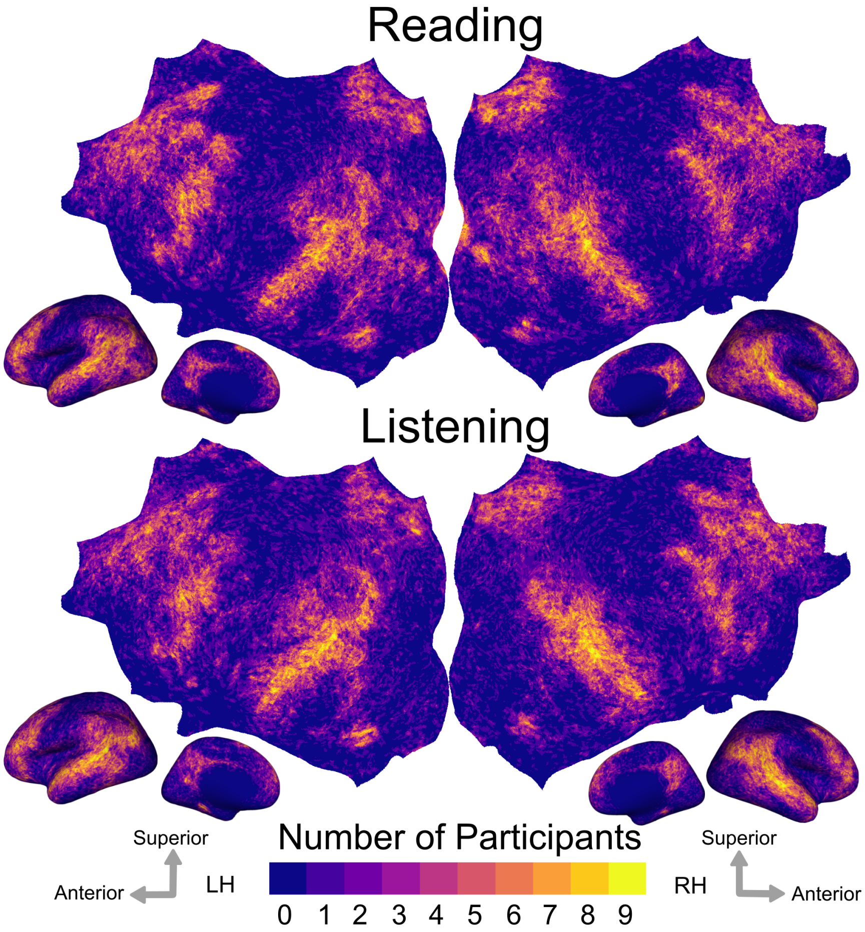

Figure 3.14: Number of participants significantly predicted in each vertex. The set of significantly predicted voxels in each participant was mapped to a standard fsAverage vertex space for reading and listening separately. The number of participants that was significantly predicted for each vertex is shown for reading (a) and listening (b) separately. The number of participants is indicated by the colorbar at the bottom. Brighter vertices are significantly predicted in more participants. Vertices across temporal, parietal, and prefrontal cortices are significantly predicted in most participants for both modalities.

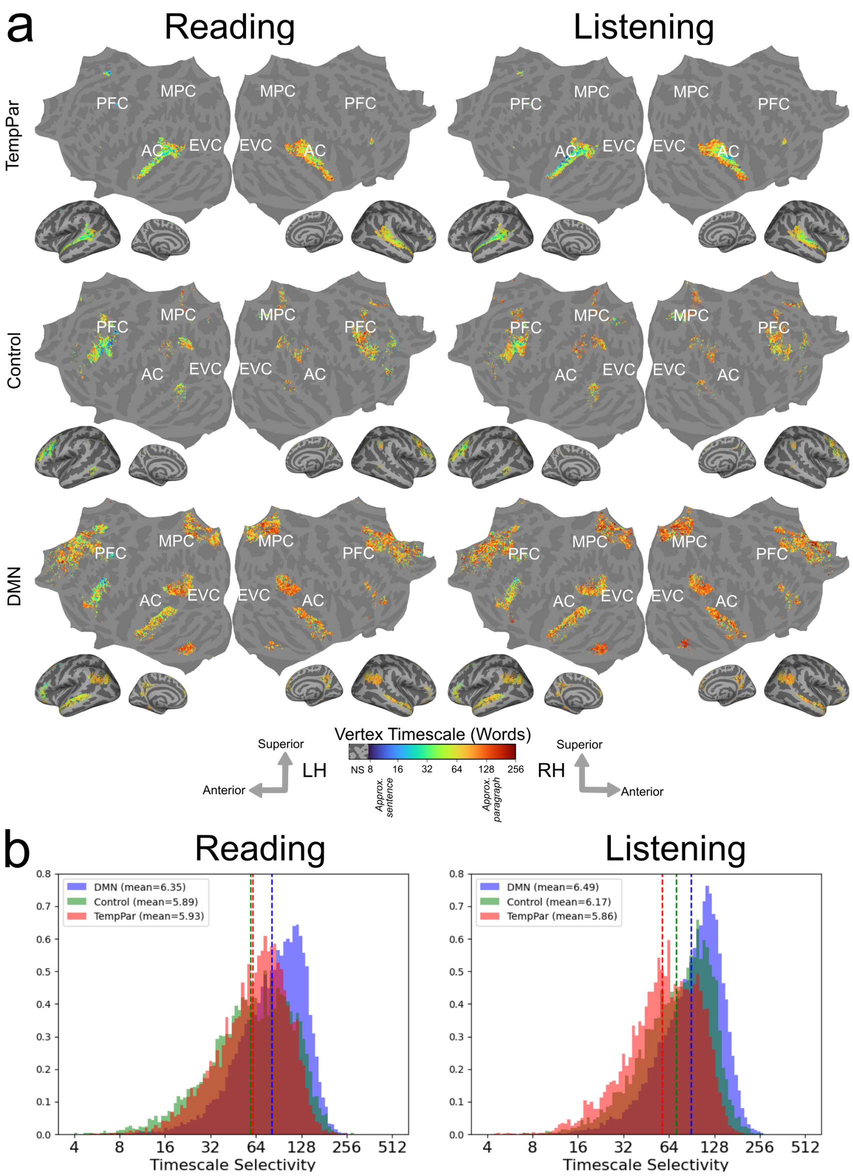

Figure 3.15: Group-level timescale selectivity across previously proposed cortical networks. The Yeo2011 cortical parcellation was used to determine 17 cortical parcels in fsAverage space (Yeo et al., 2011). From this parcellation pre-defined network labels were used to identify three proposed networks (temporo-parietal network (TempPar), cognitive control network (Control), and default mode network (DMN)). a. Group-level timescale selectivity is shown on the flattened cortical surface of the fsAverage template brain, for the three networks separately, and for reading and listening separately. Timescale selectivity is shown according to the color scale at the bottom. Voxels that were not significantly predicted for both modalities in at least three participants are shown in grey (one-sided permutation test, $p < .05$, FDR corrected). The temporo-parietal network (TempPar) contains a gradient from short to long timescale selectivity in superior to inferior temporal cortex. Prefrontal areas of the control network (Control)

contain a mix of timescale selectivity. The default mode network (DMN) contains a gradient from short to long timescale selectivity from posterior to anterior prefrontal cortex. b. The distribution of group-level timescale selectivity for each network is shown for reading and listening separately. Histograms include vertices that were significantly predicted for both modalities in at least three participants. Each network is selective for a range of different timescales. The DMN contains longer timescale selectivity than the other networks. This difference was statistically significant for both reading and listening, at the group level and for eight of the nine individual participants ($p = 0.13$ for DMN vs TempPar S5 reading; $p < .01$ at the group level and for all other participants, modalities, and network pairs). The preference towards longer timescale selectivity within the default mode network is consistent with reports that long-timescale narrative-length information may be processed in this network (Simony et al., 2016).

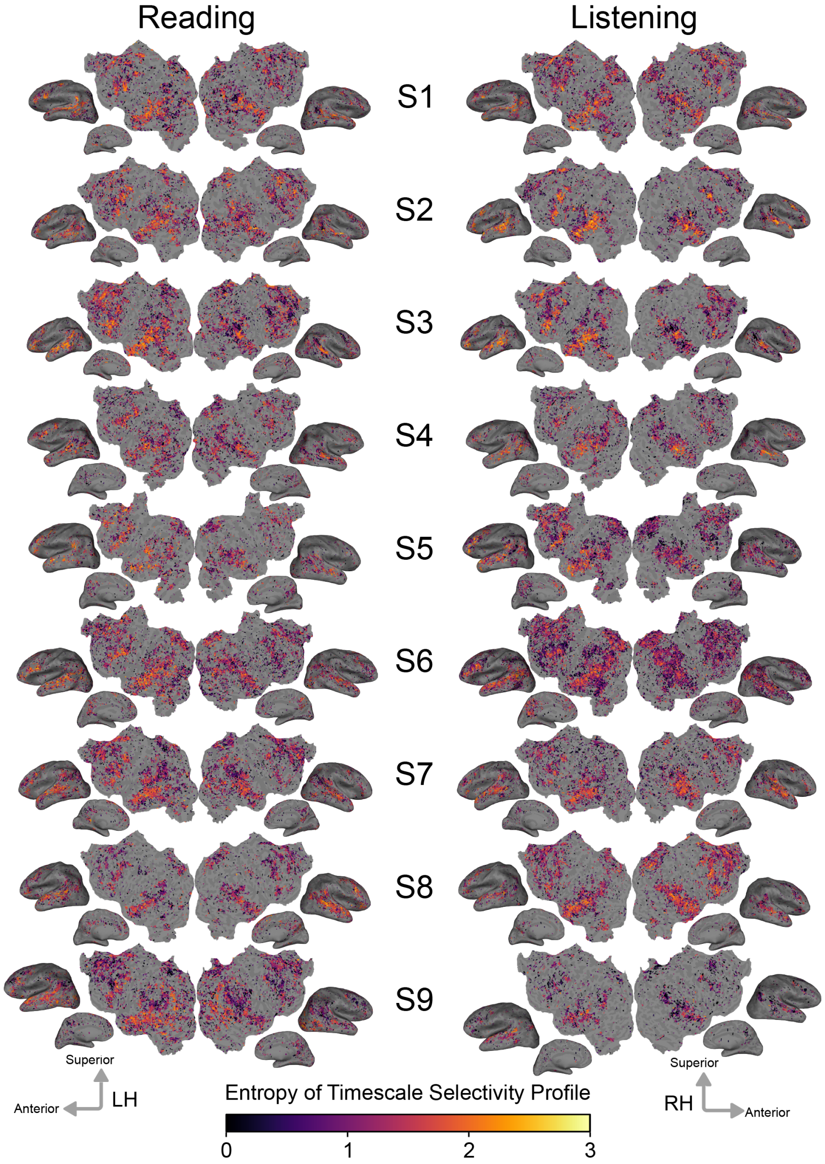

Figure 3.16: Uniformity of timescale selectivity. For each voxel, the uniformity of timescale selectivity was computed as the entropy of the timescale selectivity profile of the voxel. Voxels with more uniform timescale selectivity have a flatter timescale selectivity profile; thus, higher entropy corresponds to more uniform timescale selectivity, whereas lower entropy corresponds to more peaked timescale selectivity. The entropy of the timescale selectivity profile of each voxel is shown according to the color scale at the bottom on the flattened cortical surface of each participant, for reading and listening separately. Brighter voxels have timescale selectivity profiles with higher-entropy. Darker voxels have timescale selectivity profiles with lower-entropy. Voxels that were not significantly predicted are shown in grey (one-sided permutation test, $p < .05$, FDR corrected). Voxel timescale selectivity profiles have higher entropy (i.e., are more uniform) in superior temporal gyrus (STG) and posterior areas of prefrontal cortex (PFC), and

have lower entropy (i.e., are more peaked) in lateral and medial parietal cortex. These results are consistent with (Lerner et al., 2011), which found that areas near STS and posterior PFC displayed a wider range of temporal receptive windows than other brain areas, whereas areas such as precuneus displayed a smaller range of temporal receptive windows.

	2-4 words	4-8 words	8-16 words	16-32 words	32-64 words	64-128 words	128-256 words	256+ words
S1	0.003*	<0.001*	0.004*	<0.001*	0.062	1	0.012*	0.001*
S2	<0.001*	<0.001*	<0.001*	<0.001*	0.048*	0.006*	0.038*	0.008*
S3	<0.001*	<0.001*	0.071	0.004*	0.039*	<0.001*	<0.001*	0.002*
S4	<0.001*	0.001*	0.008*	<0.001*	0.062	0.246	0.021*	0.005*
S5	<0.001*	<0.001*	0.563	0.001*	0.48	0.21	<0.001*	0.425
S6	<0.001*	<0.001*	0.001*	0.01*	0.658	0.043*	0.003*	0.043*
S7	<0.001*	<0.001*	<0.001*	0.092	0.088	0.001*	0.08	0.007*
S8	<0.001*	<0.001*	0.062	0.091	0.018*	0.157	0.003*	0.07
S9	<0.001*	<0.001*	0.001*	<0.001*	0.036*	<0.001*	<0.001*	<0.001*

Table 3.1: **Significance of correlation between selectivity for each timescale during reading and listening.** For each timescale and participant, the Pearson correlation coefficient was computed across voxels between selectivity for that timescale during reading and listening (Figure 4). The p-value of each correlation is shown here. Asterisks indicate p-values that are significant (one-sided permutation test, $p < .05$, FDR corrected with a Benjamini-Hochberg correction for multiple comparisons).

Chapter 4

A Unified Semantic System for Representing Concepts and Relations in the Human Brain

4.1 Introduction

Humans draw upon stored semantic knowledge to communicate and reason about objects in the world. This knowledge consists of concepts (e.g., *bicycle*, *wheel*, *transportation*) and the relations between them (e.g., *is-a*, *has-part*, *used-for*). Prior studies of how the human brain represents semantic knowledge has primarily focused on concepts. These studies have shown that during language comprehension, concepts are represented throughout a network of brain areas in the temporal, parietal, and prefrontal cortices (Binder et al., 2009; Huth et al., 2016). Each area represents specific concepts, forming patterns that are consistent across individuals, modalities, and languages (Chen, Gong, et al., 2024; Deniz et al., 2019; Huth et al., 2016). While prior work has focused on concepts, the ability to represent relations is a fundamental part of human cognition: this ability enables humans to form generalizations, make inferences, and engage in analogical reasoning (Bejar et al., 1991; Chaffin, 1988; Hofstadter & Sander, 2013; Holyoak & Lu, 2021; Unger & Fisher, 2021). Thus, understanding how the brain represents relations, not just individual concepts, is crucial for understanding the neural basis of cognition.

Theoretical models suggest different possibilities about how relations could be represented in the brain. One group of models suggests that representations of relations are embedded within the concepts they connect. For example, in models of semantic memory, relations are represented as labeled links between pairs of concepts, or as features stored within the representation of each concept (Collins & Quillian, 1969; Smith et al., 1974). A second group of models argues that in order to enable flexible reasoning, relations must have their own representations that are abstracted away from the specific concepts involved in each instance of the relation. For example, in models of analogical reasoning, relations are represented as independent units or vectors that are dynamically bound to specific concepts (Doumas & Hummel, 2012; Gentner & Forbus, 2011; Holyoak et al., 2022; Kanerva, 2010).

Based on prior behavioral evidence that humans often treat relations similarly to concepts (Bejar et al., 1991; Chaffin & Herrmann, 1984; Kemp et al., 2018; Popov & Pavlova, 2020), we hypothesized that in the human brain relations have their own representations and, moreover, that relations are represented in the same way as concepts. Under this hypothesis, relations would be organized within the same semantic space as concepts, enabling efficient combination of relations and concepts. This hypothesis suggests three empirical predictions. First, relations should have their own representations: the representation of each relation should be consistent across instances that involve different concepts. Second, the organization of relation representations should match that of concept representations: each area should represent specific relations, forming patterns that are consistent across individuals. Third, there should be a systematic relationship between the cortical organizations of relation and concept representations: two areas that represent the same relation should also represent similar concepts to each other.

Existing evidence provides partial support for this hypothesis. A few neuroimaging studies have reported that representations of relations in the human brain are consistent across different instances of the same relation, suggesting that relation representations are abstracted away from specific concepts (Chiang et al., 2021; Wang et al., 2021). However, these neuroimaging results may have been confounded by the use of different sets of objects for examples of different relations. Furthermore, those studies examined representational similarities and decoding accuracies, rather than explicitly modeling the representation of each relation. Thus, it is unclear where each relation is represented, or how relation representations are organized across the cortical surface: each area could represent specific relations, or broadly represent many different relations. Finally, no prior study has directly compared the organization of relation and concept representations. Thus, it is unclear whether there is a systematic relationship between the concepts and relations that are represented in each area.

We designed a study to test the hypothesis that relations are represented in the same way as other concepts. Six participants each performed a relation-verification experiment. In this experiment, each participant answered over 1000 questions about six semantic relations while functional magnetic resonance imaging (fMRI) was used to record brain responses (Figure 4.1). Voxelwise modeling was used to estimate functional maps that describe selectivity for each of the six relations. Then, a separate narrative comprehension experiment was used to estimate functional maps that describe selectivity for individual concepts (Deniz et al., 2019; Huth et al., 2016). The two sets of maps were used to test our three predictions within each participant. Our results support the hypothesis that relations are represented in the same way as other concepts.

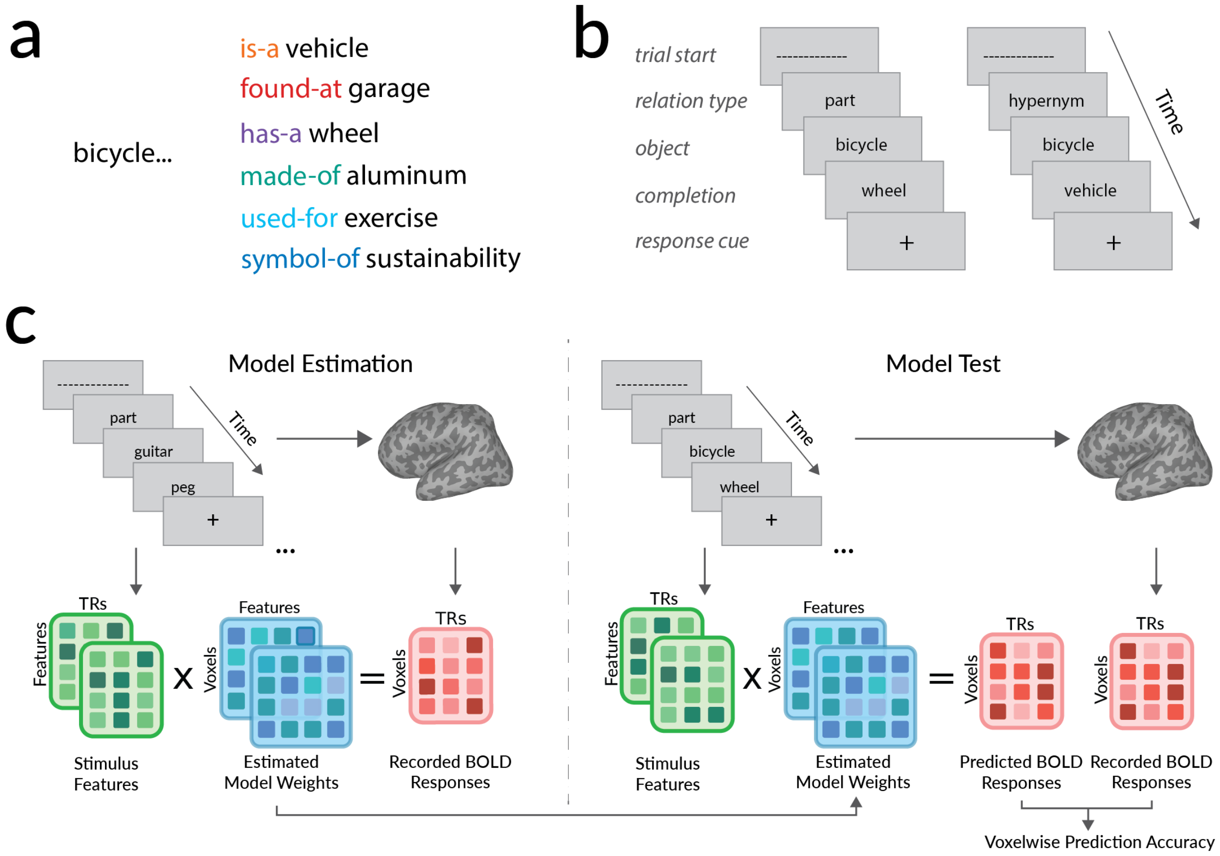

Figure 4.1: Experiment paradigm and voxelwise modeling. **a**. Examples of the six relations in our experiment. Examples are shown for the object "bicycle" . **b**. Experiment paradigm. Participants each performed over 1000 trials of an event-related experiment while fMRI was used to record BOLD responses. Two example trials are shown. In each trial three words were displayed: a relation (e.g., "part"), an object (e.g., "bicycle"), and a potential completion word (e.g., "wheel"). Participants were instructed to press a button to indicate whether the presented instance forms a valid relation. **c**. Modeling framework. For each relation a binary feature space was constructed to describe the times at which participants performed trials for that relation. VM was used to estimate a separate FIR ridge regression model for each feature space, voxel, and participant. Estimated model weights describe how each relation modulates BOLD responses separately in each voxel and for each participant. Estimated model weights were used to predict BOLD responses to a held-out dataset that was not used for model estimation. The held-out dataset used different instances of each relation from the train dataset. Prediction accuracy was quantified as the coefficient of determination (R^2) between predicted and recorded BOLD responses to the held-out dataset.

4.2 Results

A relation-verification experiment was used to model how the brain represents semantic relations. In each trial of this experiment, participants answered questions about one of eight relations. The eight relations include six semantic relations: *is-a* (e.g., bicycle-vehicle), *found-at* (e.g., bicycle-garage), *has-part* (e.g., bicycle-wheel), *made-of* (e.g., bicycle-aluminum), *symbol-of* (e.g., bicycle-freedom), and *used-for* (e.g., bicycle-transportation). These six relations were chosen because they commonly occur in existing studies of semantic relations (Bejar et al., 1991; Jurgens et al., 2012). Furthermore, these relations apply to a wide range of common objects and therefore allowed us to use the same set of objects for each relation in the experiment. The trials also included trials about two non-semantic wordform relations: *alphabetically-before* (e.g., bicycle-stone), and *wordform-match* (e.g., bicycle-bicycle). For brevity, we hereafter refer to semantic relations simply as relations, and non-semantic wordform relations as wordform relations.

To examine how each relation is represented in the brain, we estimated voxelwise models (VMs) that describe how each relation modulates blood oxygen level-dependent (BOLD) responses in the brain. A six-dimensional relation feature space was constructed to describe the relation type of each trial. Each dimension of this feature space corresponds to one of the six relations. Then, banded ridge regression was used to estimate model weights that use the relation feature space to predict BOLD responses in each voxel in each individual participant. Additional feature spaces were included in the regression to account for representations of the two wordform relations, the lexical semantic content of the presented words, visual stimuli, motor reactions, participant reaction times, and participant accuracy. The estimated model weights for each voxel were used to predict BOLD responses to a held-out test set which contains instances of each relation that were not used for model estimation. The product measure was used to compute the contribution $\tilde{R^2}_{featurespace}$ of each feature space to the total prediction accuracy (Hoffman, 1960; Pratt, 1987). The estimated model weights and prediction accuracies reveal which relations are represented in voxel.

First, we tested whether relations have their own representations that are independent of specific pairs of concepts. If this is the case, then the representation of each relation will be consistent across instances that involve different concepts, and therefore the relation feature space will accurately predict brain responses to held-out instances of each relation.

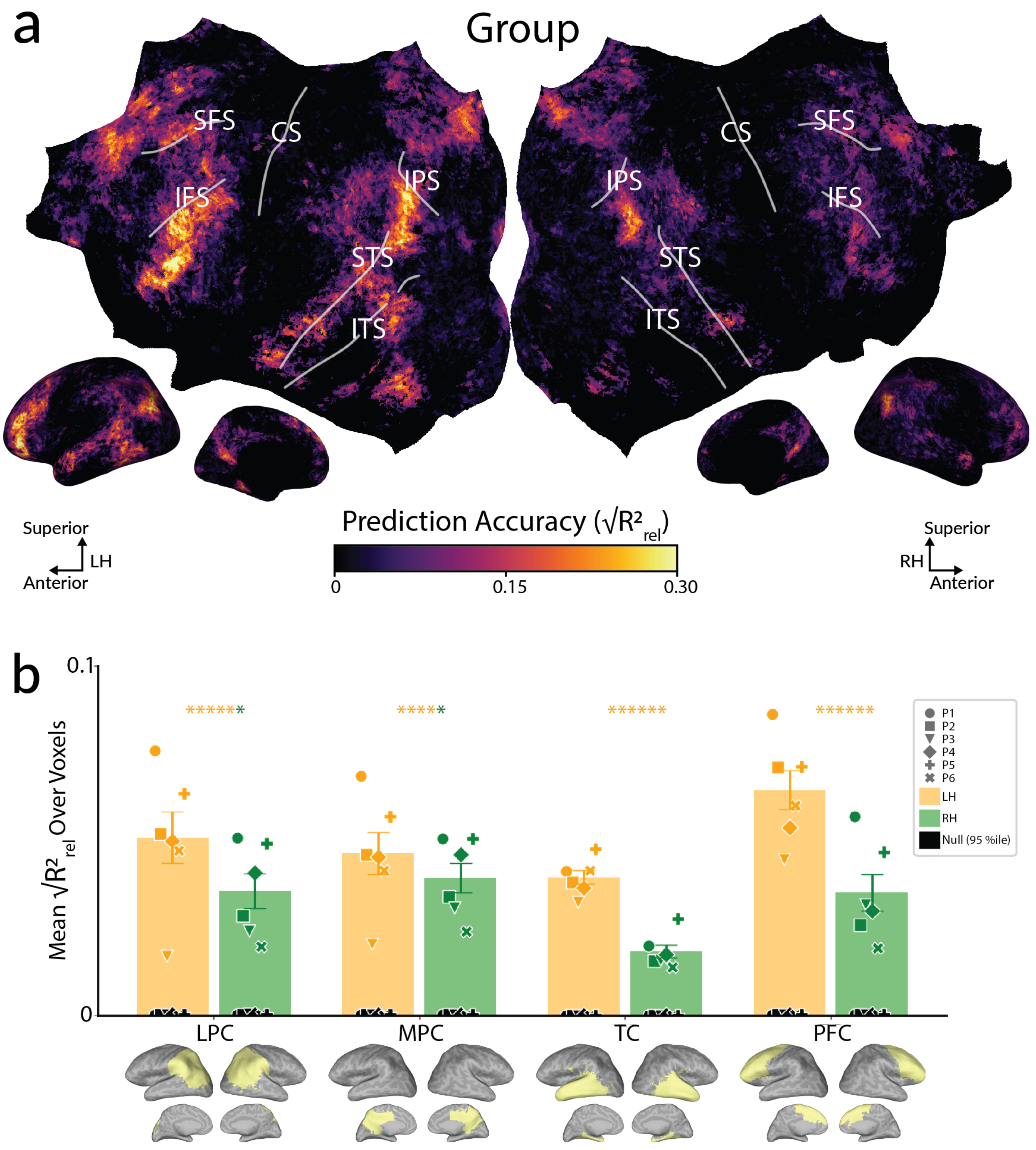

Figure 4.2: Prediction accuracy of the relation feature space. To determine whether relations have their own representations, we tested whether model weights estimated for the relation feature space could accurately predict BOLD responses to held-out instances of each relation. **a**. Group-level prediction accuracy. Results are shown on the flattened cortical surface of

the template space. Vertices that were significantly well-predicted in fewer than one third of participants are shown in black. Prediction accuracy is high throughout bilateral temporal, parietal, and prefrontal cortices. (STS=superior temporal sulcus, ITS=inferior temporal sulcus, IPS=inferior parietal sulcus, SFS=superior frontal sulcus, IFS=inferior frontal sulcus). **b**. Prediction accuracy by cortical region. For each cortical region, yellow and green markers show the mean prediction accuracy over voxels for the left and right hemispheres respectively. Black markers show null prediction accuracy (95th percentile). Bars show the mean across participants. Yellow and green asterisks show the number of participants for which prediction accuracy is significantly higher in the left or right hemisphere. While the temporal, parietal, and prefrontal cortices are significantly well-predicted bilaterally, accuracies are higher in the left than the right hemisphere. These results indicate that relations are represented in a left-lateralized network of cortical regions.

Figure 4.2a shows the test prediction accuracy of the relation feature space ($\tilde{R^2}_{relations}$). Prediction accuracy was computed separately for each participant. Group-level accuracies were computed by projecting voxelwise accuracies for each participant to a standard template space (fsAverage (Fischl et al., 1999)), and then averaging the projected accuracies for each vertex in the template space. Group-level results are shown for each vertex of the template space. Results for each participant are similar to the group (Supplementary Figure 4.5). Vertices are significantly well-predicted throughout the bilateral temporal, parietal, and prefrontal cortices (one-sided $p<.05$ by a permutation test, after a Benjamini-Hochberg correction for multiple comparisons (Benjamini & Hochberg, 1995)). These cortical regions are sometimes referred to as the semantic system (Binder et al., 2009; Huth et al., 2016). The results in Figure 4.2a suggest that throughout the semantic system representations of relations are independent of specific pairs of concepts.

Visual inspection of Figure 4.2a suggests that prediction accuracy is higher in the left compared to the right hemisphere. To quantify this difference for each cortical region of the semantic system, we used FreeSurfer regions of interest (ROIs) to identify the set of voxels in the left and right temporal, parietal, and prefrontal cortices. (ROIs were based on the Desikan-Killiany atlas (Desikan et al., 2006).) Then for each cortical region, the average prediction accuracy was computed over voxels in each cortical region. Results were computed for each participant and hemisphere separately. Figure 4.2b shows the average prediction accuracy for each participant, cortical region, and hemisphere. Prediction accuracy is significantly greater in the left than the right hemisphere (one-sided $p<.05$ by a permutation test for all participants and cortical regions, except P3 lateral parietal cortex (LPC), P3 medial parietal cortex (MPC), and P4 MPC). Note that this observed left-lateralization is not merely due to the experiment presentation format: wordform relation trials were presented in the same way as relation trials, but wordform relation representations are right-lateralized (Supplementary Figure 4.5). Overall, Figure 4.2 suggests that relations have their own representations and are represented throughout a left-lateralized network of cortical regions.

Figure 4.2 shows that relations are represented throughout the semantic system. Next, we tested whether these relation representations are organized similarly to concept representations. Prior studies have shown that cortical representations of concepts are organized such that each area represents specific concepts, forming patterns that are consistent across individuals. Thus, if the organization of relation representations matches that of concept representations, then each area should represent specific relations, and the organization of relation representations should be consistent across participants. Alternatively, it could be the case that each area broadly represents many relations, or that the organization of relation representations is highly variable across participants.

To determine where each relation is represented, we computed voxelwise selectivity for each of the six relations. If a voxel is selective for a particular relation, then it will exhibit higher activation during trials of that relation. This means that the relation feature space will accurately predict BOLD responses in that voxel, and the model weights for that relation will be high. Thus, the selectivity S_{R_i} for each relation was operationally defined as the product of of the prediction accuracy of the relation feature space ($\sqrt{\tilde{R}^2}_{relations}$) and the model weight for that relation (β_{R_i}): $S_{R_i} = \sqrt{\tilde{R}^2}_{relations} \times \beta_{R_i}$. Selectivity was computed separately for each relation, participant, and voxel. For each relation, group-level selectivity was computed by projecting voxelwise selectivity for each participant to the template space and then averaging the projected selectivities for each vertex in the template space.

Figure 4.2a shows group-level selectivity for each relation. Results are shown for each vertex of the template space. Each relation appears to selectively activate concentrated patches of voxels. For example, the *found-at* relation is represented in patches in retrosplenial cortex (RSC), anterior to occipital place area (OPA), anterior to parahippocampal place area (PPA), in anterior superior temporal sulcus (STS), and along superior frontal sulcus (SFS); whereas the *has-part* relation is represented in patches superior to OPA, along inferior temporal sulcus (ITS), and along inferior frontal sulcus (IFS). Visual inspection of results for each participant suggests that the areas that represent each relation are consistent across participants (Supplementary Figure 4.5). To quantify this consistency, voxelwise selectivity for each relation and participant was projected to the template space. Then for each relation, the Pearson correlation was used to measure the consistency of vertexwise selectivity between each pair of participants. Figure 4.2b shows these correlations for each relation separately. For each relation and for each pair of participants, the cortical distribution of selectivity is significantly positively correlated between participants ($p<.05$ by a permutation test). These results suggest that each relation is represented in consistent areas across individuals.

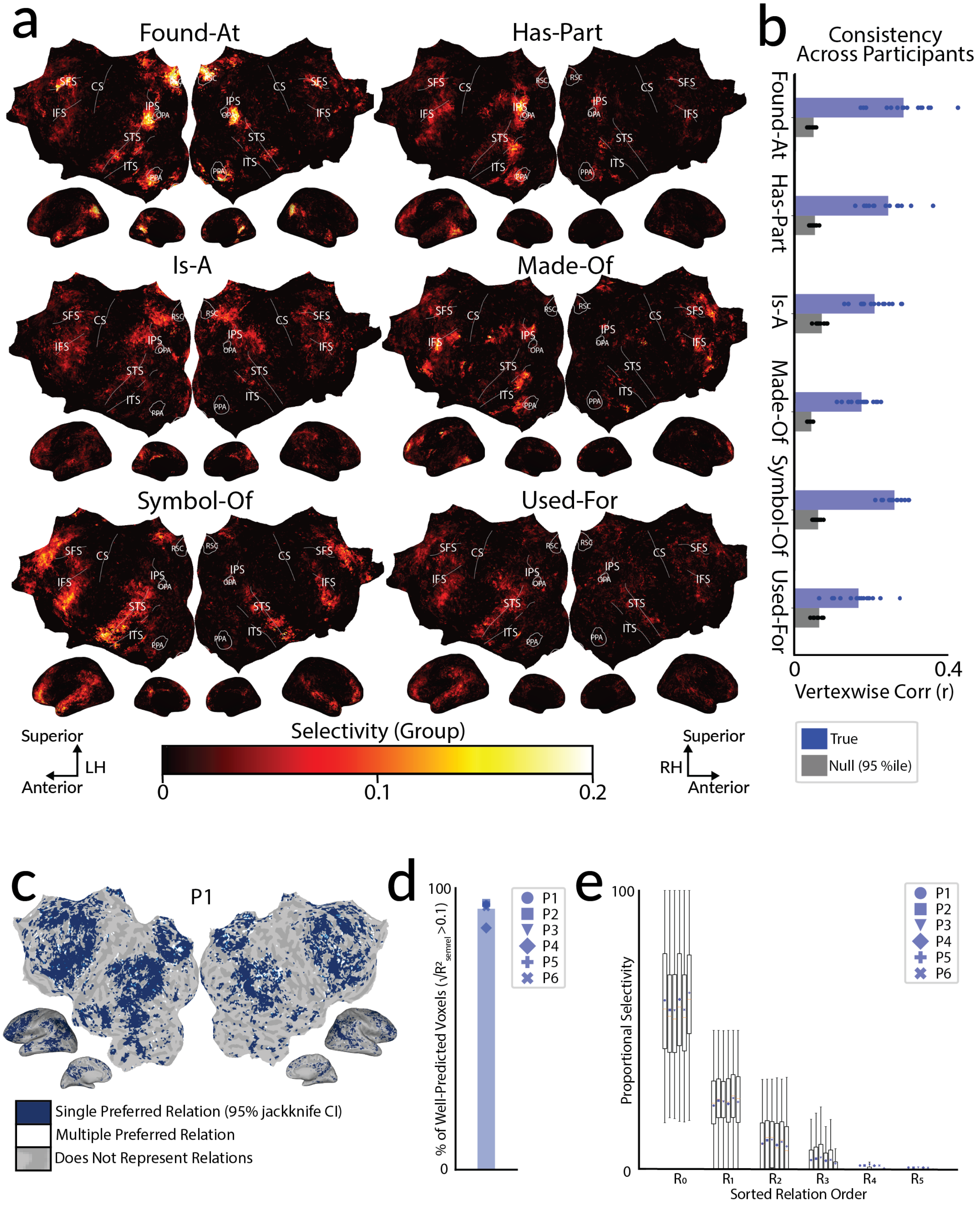
a
Found-At
Has-Part
Is-A
Made-Of
Symbol-Of
Used-For
SFS
CS
IFS
IPS
OPA
STS
ITS
PPA
RSC
Superior
LH
Anterior
Selectivity (Group)
0
0.1
0.2
Superior
RH
Anterior
b
Consistency Across Participants
Found-At
Has-Part
Is-A
Made-Of
Symbol-Of
Used-For
0
0.4
Vertexwise Corr (r)
True
Null (95 %ile)
c
P1
Single Preferred Relation (95% jackknife CI)
Multiple Preferred Relation
Does Not Represent Relations
d
100
0
% of Well-Predicted Voxels ($\sqrt{R^2_{semrel}}$>0.1)
P1
P2
P3
P4
P5
P6
e
100
0
Proportional Selectivity
R_0
R_1
R_2
R_3
R_4
R_5
Sorted Relation Order
P1
P2
P3
P4
P5
P6

Figure 4.3: Cortical distribution of relation representations. To determine how relation representations are organized, we examined voxelwise selectivity for each of the six relations. **a**. Group-level selectivity for each relation. Results are shown on the flattened cortical surface of the template space. Vertex color reflects selectivity for the relation. Each relation appears to be represented in distinct patches throughout the semantic system. **b**. Consistency of relation representations across participants. For each participant and relation, voxelwise selectivity was projected to the template space. Then for each relation, the Pearson correlation was used to measure the consistency in vertexwise selectivity between each pair of participants. Blue markers show true correlations for each pair of participants. Grey markers show null correlations (95th percentile). Blue and grey bars show the mean across participant pairs. For each relation, the cortical distribution of relation selectivity is significantly positively correlated between participants. **c**. Specificity of relation selectivity. To determine whether each voxel is selective for a specific relation, for each voxel the relation with the highest selectivity (R_0) was identified and then a jackknife procedure was used to estimate confidence intervals around the difference between selectivity for R_0 and for the other five relations. A voxel was considered to prefer a single relation if all five confidence intervals were strictly positive. Results are shown for one representative participant (P1) on the flattened surface of the native brain space. Voxels shown in blue prefer a single relation. Voxels shown in white do not prefer a single relation. Voxels shown in grey do not represent relations. Most voxels throughout the semantic system prefer a single relation. **d**. Percentage of voxels that significantly prefer a single relation. Markers show percentages for each participant. The bar shows the mean across participants. In each participant, over 85% of voxels prefer a single relation. **e**. Proportional selectivity for each relation. For each voxel, the six relations were sorted from highest (R_0) to lowest (R_5) selectivity. Boxplots show the distribution over voxels of proportional selectivity for R_0 through R_5. Results are shown separately for each participant. Across voxels and participants, around 60% of selectivity is concentrated on a single relation. These results suggest that most voxels represent specific relations, and that the organization of relation representations is consistent across participants.

To determine whether each voxel represents a specific relation, for each voxel we tested whether selectivity peaks at a specific relation or is evenly spread over multiple relations. This analysis was limited to voxels that represent relations, which were defined according to the prediction accuracy of the relation feature space ($\sqrt{\tilde{R^2}_{relations}} > 0.1$). For each voxel, the relation with the highest selectivity (R_0) was identified, and then a leave-one-run-out jackknife procedure was used to estimate 95% confidence intervals around the difference between selectivity for R_0 and for each of the other five relations (see Methods for details). A voxel was considered to prefer a single relation if all five confidence intervals were strictly positive. Figure 4.2c shows the set of voxels with a single preferred relation. Results are shown for one representative participant (P1). (Results for the other five participants are consistent with P1, as shown in Supplementary Figure 4.5). For each participant, more than 85% of voxels have a single preferred relation (P1: 93%, P2: 95%, P3: 94%, P4: 86%, P5: 93%, P6: 94%; Figure 4.2d). To examine how much of voxelwise relation selectivity is concentrated on a single relation, we computed the proportional selectivity for each of the six relations $\frac{S_{R_i}}{\sum_{j=0}^{5} S_{R_j}}$. Across participants and voxels, around 60% of selectivity is concentrated on one relation (Figure 4.2e). Thus, most voxels represent a specific relation.

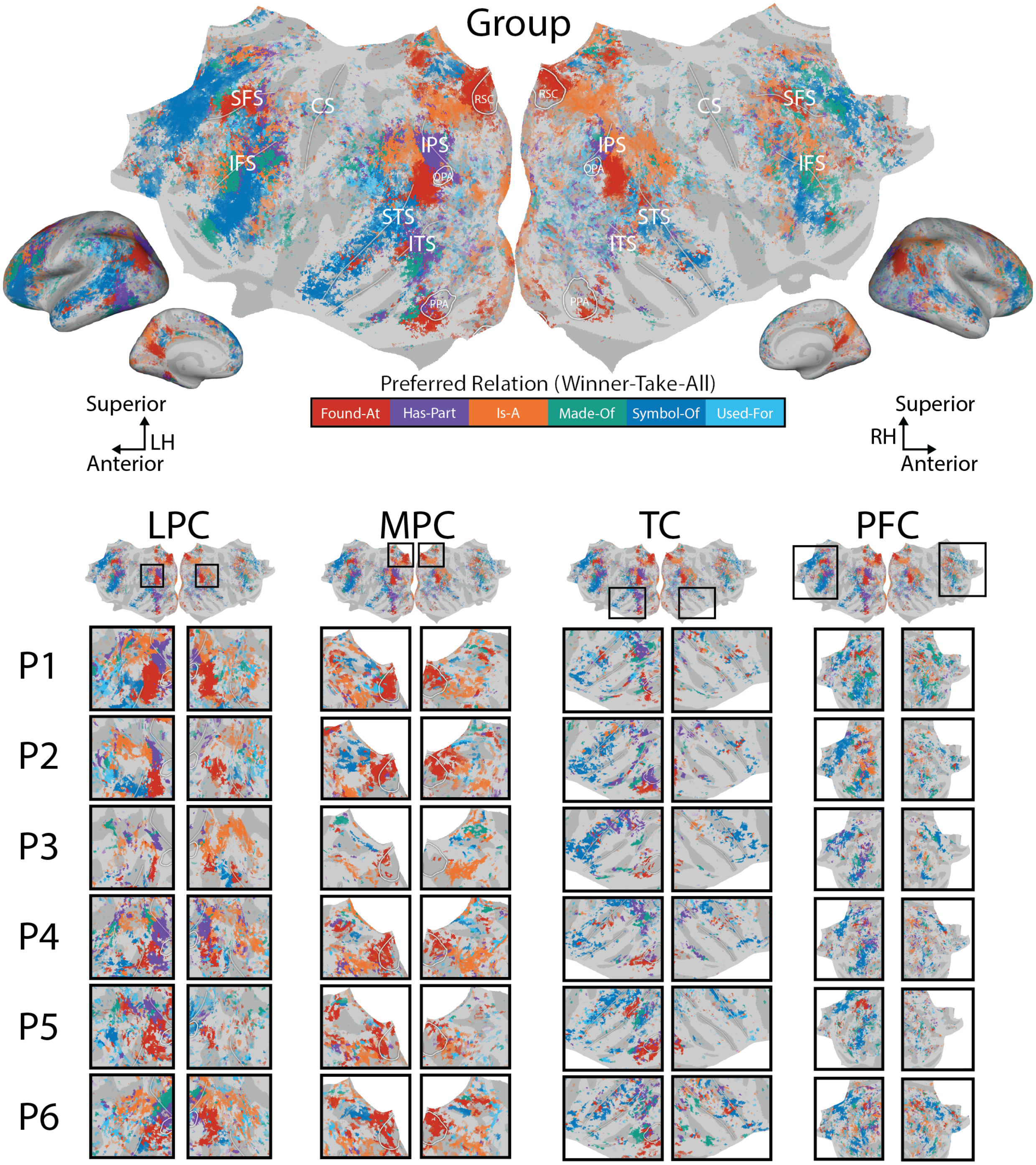

Figure 4.4: Cortical organization of preferred relations. To jointly visualize the arrangement of representations for all six relations, a winner-take-all map was used. Results are shown at the group-level and for each individual participant. All results are shown on the flattened cortical surface of the template space. The color of each vertex reflects the relation with the highest

selectivity (R_0). The opacity of each vertex reflects the prediction accuracy of the relation feature space. The cortical organization of preferred relations appears to be consistent across participants. In LPC *found-at* (red), *has-part* (purple), *is-a* (orange), and *symbol-of* (blue) form a ring of patches around the temporal parietal junction. In MPC *found-at*, *is-a*, and *symbol-of* are respectively shown in, superior to, and anterior to retrosplenial cortex (RSC). In TC *found-at*, *made-of*, and *symbol-of* are respectively shown anterior to parahippocampal place area (PPA), superior to PPA, and along STS. In PFC *symbol-of* is shown superior to SFS and inferior to IFS, and *found-at* is shown along SFS. These results suggest that relation representations form patterns that are consistent across participants. (LPC: lateral parietal cortex, MPC: medial parietal cortex, TC: temporal cortex, PFC: prefrontal cortex). (LPC: lateral parietal cortex, MPC: medial parietal cortex, TC: temporal cortex, PFC: prefrontal cortex).

To visualize how representations of the six relations are arranged across the cortical surface, we used a winner-take-map to jointly show selectivity for all six relations. Figure 4.2 shows results at the group-level and for each individual participant. To facilitate comparisons across participants, all results are shown in the template space. (Results in native participant space are consistent with results in the template space, and are shown in Supplementary Figure 4.5). The arrangement of relation representations is consistent across participants. In lateral parietal cortex (LPC) *found-at* (red), *has-part* (purple), *is-a* (orange), and *symbol-of* (blue) are arranged in a ring of patches around the temporal parietal junction. In medial parietal cortex (MPC) *found-at*, *is-a*, and *symbol-of* are respectively shown in patches in, superior to, and anterior to retrosplenial cortex (RSC). In temporal cortex (TC) *found-at*, *made-of*, and *symbol-of* are respectively shown anterior to PPA, superior to PPA, and along the superior temporal sulcus (STS). In prefrontal cortex (PFC) *symbol-of* is shown superior to superior frontal sulcus (SFS) and inferior to inferior frontal gyrus (IFG), and *found-at* is shown along SFS. These results suggest that relation representations are organized in cortical patterns that are consistent across individuals.

Figures 4.2 and 4.2 suggest that the organization of relation representations matches that of concept representations. It could be the case that relations and concepts are part of the same organization, such that voxels that represent the same relation represent similar concepts to each other. Alternatively, representations of relations and of concepts could form two overlapping but functionally independent networks, such that voxels that represent the same relation do not necessarily represent the same concepts. To distinguish between these two possibilities, we compared voxelwise selectivity for relations and for concepts.

To estimate voxelwise selectivity for concepts, we used a passive language comprehension experiment that has been used in prior work (see Methods for details; (de Heer et al., 2017; Deniz et al., 2019; Huth et al., 2016)). In this passive language comprehension experiment, participants read or listened to narrative stories while their brain responses were recorded with fMRI. Then each word of the narratives was projected to a 300-dimensional word embedding space (word2vec, Mikolov et al., 2013), and VM was used to estimate model weights that reflect voxelwise selectivity for different concepts.

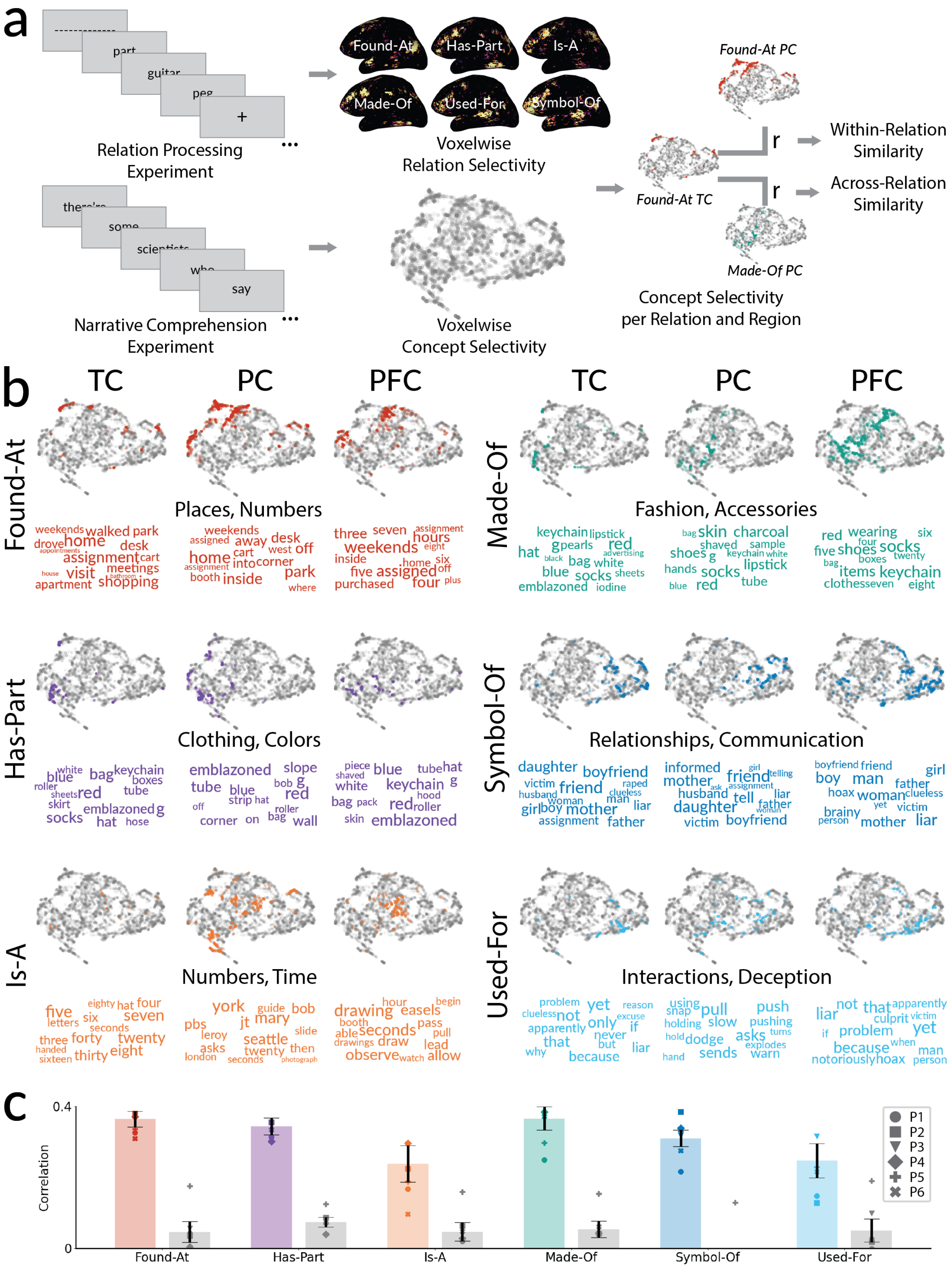
a
part
guitar
peg
+
Relation Processing Experiment
Found-At
Has-Part
Is-A
Made-Of
Used-For
Symbol-Of
Voxelwise Relation Selectivity
there're
some
scientists
who
say
Narrative Comprehension Experiment
Voxelwise Concept Selectivity
Found-At PC
Found-At TC
Made-Of PC
r
r
Within-Relation Similarity
Across-Relation Similarity
Concept Selectivity per Relation and Region
b
TC
PC
PFC
Found-At
Places, Numbers
weekends walked park drove home desk appointments assignment cart house visit meetings bathroom apartment shopping
weekends assigned away desk cart west off home assignment into corner booth inside park where
three seven assignment hours weekends eight inside home six five assigned off purchased four plus
Made-Of
Fashion, Accessories
keychain lipstick hat pearls red black bag white advertising blue socks sheets emblazoned iodine
bag skin charcoal shoes shaved sample keychain white hands socks lipstick blue red tube
red wearing six four five shoes socks bag boxes twenty items keychain clothes seven eight
Has-Part
Clothing, Colors
white blue bag keychain boxes roller sheets red tube skirt emblazoned g socks hat hose
emblazoned slope tube blue bob g strip hat red off roller corner on bag wall
piece shaved blue tube hat white keychain g hood bag pack red roller skin emblazoned
Symbol-Of
Relationships, Communication
daughter boyfriend victim friend raped husband clueless woman man girl boy mother liar assignment father
informed girl friend telling mother ask assignment husband tell liar daughter father woman victim boyfriend
boyfriend friend girl boy man father hoax woman clueless brainy yet victim person mother liar
Is-A
Numbers, Time
eighty hat four five six seven letters seconds three forty twenty handed sixteen thirty eight
york guide bob pbs jt mary slide leroy seattle asks twenty then london seconds photograph
hour begin drawing easels booth pass able seconds pull drawings draw lead observe watch allow
Used-For
Interactions, Deception
problem yet reason clueless not excuse only if apparently never that but liar why because
using snap pull push holding slow pushing hold dodge asks turns explodes hand sends warn
not that apparently liar culprit victim if problem yet because when man notoriously hoax person
c
Correlation
0.4
0
Found-At
Has-Part
Is-A
Made-Of
Symbol-Of
Used-For
P1
P2
P3
P4
P5
P6

Figure 4.5: Comparison of selectivity for relations and for concepts. To determine whether there is a systematic relationship between the organizations of relation and concept representations, we tested whether voxels that represent the same relation also represent similar concepts to each other. **a**. For each participant, the relation-verification experiment was used to estimate voxelwise selectivity for relations (as shown in Figure 4.1), and a separate passive narrative comprehension experiment was used to estimate voxelwise selectivity for concepts (as in Deniz et al., 2019; Huth et al., 2016). For each relation, we identified the voxels that represent the relation ($S_{R_i} > 0.1$) and then tested whether those voxels are selective for similar concepts to each other. **b**. Concept selectivity, shown separately for voxels that are selective for each relation. Voxelwise concept selectivity was projected into a 2D UMAP space. Scatterplots show voxelwise concept selectivity in the reduced 2D space, separately each relation and separately for voxels in the temporal (TC), parietal (PC), and prefrontal cortices (PFC). Results are shown for one representative participant (P1). For each relation and region separately, word clouds show the words that are closest to the mean concept selectivity. Voxels that represent the same relation appear to represent similar concepts. **c**. Similarity in concept selectivity. The Pearson correlation was used to quantify the similarity between the full 300-dimensional vectors of concept selectivity. Similarities were computed for pairs of voxels that both represent the same relation (within-relation), as well as for pairs of voxels that represent different relations (across-relation). For each relation and participant, colored markers show the mean over voxels of within-relation similarity. Grey markers show the mean over voxels of across-relation similarity. Bars and errorbars show the mean and standard error of the mean across participants. Stars indicate the number of participants in which similarities are significantly higher within- than across-relations (by a 95% bootstrap confidence interval). Across relations and participants, within-relation similarity in concept selectivity is significantly higher than across-relation concept selectivity. Overall, these results suggest that there is a systematic relationship between the cortical organizations of relation and concept representations.

Voxelwise selectivity for relations and for concepts was used to test whether voxels that represent the same relation tend to represent concepts in a similar part of the semantic space. For each relation, we took the set of voxels that represent the relation ($S_{R_i} > 0.1$) and examined the concept selectivity of those voxels. Because concept selectivity is a high-dimensional vector and each of the dimensions is not inherently interpretable, it is impractical to directly visualize the selectivity of each voxel for each concept. Therefore we inspected a 2D projection of concept selectivity. The 2D projection was obtained by the UMAP algorithm, which is optimized to preserve the similarity structure of the full 300-dimensional space (McInnes et al., 2018). Figure 4.2b shows voxelwise concept selectivity in this reduced 2D space for one representative participant (P1). To ensure that comparisons are not biased by the spatial autocorrelation of voxel representations, results are shown separately for voxels in the temporal (TC), parietal (PC), and prefrontal (PFC) cortices. In voxels that represent the same relation, concept selectivity appears to be in a consistent part of the semantic space. To interpret the concepts that are represented in each group of voxels, for each relation and cortical region separately we computed the mean concept selectivity over voxels, and then identified the 15 words that are closest to that mean concept selectivity. Similarity between a word and the mean concept similarity was measured as the Pearson correlation coefficient between the 300-dimensional embedding of that word and the 300-dimensional vector of mean concept selectivity. The word clouds in Figure Figure 4.2b show the set of 15 words closest to the mean concept selectivity of each group of voxels. Voxels that represent the same relation tend to be selective for a consistent set of concepts. For example, voxels that represent the *found-at* relation are selective for concepts associated with places and numbers, and voxels that represent the *made-of* relation are selective for concepts associated with fashion and accessories.

To quantify the similarity in concept selectivity between pairs of voxels that represent the same relations, we measured the Pearson correlation between the full 300-dimensional vectors of concept selectivity between pairs of voxels that represent the same relation. For comparison, similarity of concept selectivity was also measured between pairs of voxels in which only one of the voxels represents that relation. As in Figure 4.2b, we only considered pairs of voxels in which each voxel was from a different cortical region. Figure 4.2c shows mean similarity of concept selectivity across pairs of voxels that represent the same relation (within-relation) and that represent different relations (across-relation). Results are shown for each relation and participant separately. Voxel pairs were resampled with replacement to obtain 95% bootstrap confidence intervals around the mean similarity of concept selectivity. These confidence intervals were estimated separately for each participant and relation, and separately for within-relation and across-relation pairs. The confidence intervals for within-relation similarity lie outside that of across-relation similarity for each participant and each relation. These results suggest that voxels that represent the same relation tend to be selective for similar concepts. The results in Figure 4.2 suggest that there is a systematic relationship between the cortical organizations of relation and concept representations.

4.3 Discussion

We investigated how the brain represents relations between concepts, and how these representations compare to those of individual concepts. We show that relations have their own representations throughout the temporal, parietal, and prefrontal cortices (Figure 4.2). The organization of these representations matches that of concept representations: each area represents specific relations, forming patterns that are consistent across participants (Figures 4.2 and 4.2). Finally, there is a systematic relationship between the cortical organizations of concept and relation representations (Figure 4.2). Our results generalize across participants, including the two held-out participants (P5 and P6). Our results support theories that in the human brain relations have their own abstract representations that are not tied to specific pairs of concepts (Doumas & Hummel, 2012; Gentner & Forbus, 2011). Moreover, these results suggest that relations are represented in the same way as concepts.

Prior neuroimaging studies of semantic representations in the brain have primarily investigated representations of concepts during passive language comprehension. Those studies showed how representations of individual concepts are organized throughout the semantic system (Chen, Gong, et al., 2024; Deniz et al., 2019; Huth et al., 2016), and found that this organization implicitly encodes information about relations between concepts (Zhang et al., 2020). However, those studies used passive language comprehension tasks that did not explicitly elicit relation representations. Thus, while they could investigate how relations affect the structure of concept representations in the brain, they could not directly study representations of relations. A few studies have used relation processing experiments to study how the brain represents relations (Chiang et al., 2021; M.-H. Wu et al., 2022). These studies reported that areas throughout the temporal, parietal, and prefrontal cortices encoded information about the type of relation. However, these studies only tested for similarities and differences between brain representations, and did not examine how each relation is represented in the brain. Thus, it was unclear how relation representations are organized in the brain, and whether relation representations are consistently organized across individuals. Moreover, because prior studies did not directly compare representations of relations to those of concepts, it was unclear whether there is a systematic relationship between where relations and concepts are represented in the brain. In this study, we used a relation processing experiment, modeled how each relation is represented in the brain, and directly compared representations of relations and of concepts in order to understand how the brain represents relations between concepts.

One potential criticism of this study is that the estimated representations of relations merely reflect the semantics of the words involved in trials of each relation. For example, because trials of the *found-at* relation involve words related to locations whereas trials of the *symbol-of* relation involve words related to abstract ideas, the estimated representations of relations could reflect the types of words that are uniquely involved in the instances of each relation. However, there are two reasons why we do not think this is the case. First, lexical semantics of each stimulus word was included as a nuisance feature space during model estimation, and banded ridge regression is optimized to select the best-predicting feature space for each voxel (Dupré la Tour et al., 2022). Thus, if representations merely reflected the lexical semantics of the words in the experiment, then we would not have observed the high prediction accuracies attributed to the relation feature space. Second, the areas in which each relation is represented are not merely the areas that represent the words involved in trials of that relation. For instance, Figure 4.2 shows that the *found-at* relation overlaps not only with representations of places, but also of numbers; and the *symbol-of* relation overlaps not with representations of abstract ideas, but of people and relationships. These results suggest that estimated representations of relations do not merely reflect lexical semantic representations of the words involved in trials of each relation.

Our results provide evidence to support the hypothesis that relations are represented in the same way as concepts. These results support theories that relations have their own abstract relations, and are not merely embedded within concept representations (Doumas & Hummel, 2012). Because our study focuses on well-learned relations and common objects, we cannot make conclusions about how these relations originate. However, we speculate that this shared space exists because relations and concepts are jointly learned over the course of semantic acquisition, and co-occur in shared contexts. Indeed, studies of artificial language models have shown that knowledge of structured relations between concepts can be acquired via simple word-level learning objectives (Bouraoui et al., 2019; Chen, Lin, & Klein, 2021;

Hernandez et al., 2024). In the future, the results and methodology presented here can be used to study how the brain represents novel relations, and how representations of relations change over the course of learning.

4.4 Methods

Experimental Stimuli

The experimental stimuli consisted of 1496 trials of a relation-verification experiment. Each trial consisted of three words: a relation (e.g., "hypernym"), an object (e.g., "bicycle"), and a potentially related concept (e.g., "vehicle"). The trials were evenly divided across the eight relations (six semantic relations and two non-semantic wordform relations). Trials for each relation involved the same set of 60 common objects. Half of the trials contained true examples (e.g., "hypernym-bicycle-vehicle") and half contained false examples (e.g., "hypernym-bicycle-clothing").

At the start of each trial a dotted line was presented for 0.4 seconds. Then a rapid serial visual presentation (RSVP) procedure (Forster, 1970) was used to present the three words one-by-one at the center of the screen. The duration of each word was computed based on the length of the word. Each word was presented for a base length of 0.3 seconds, and 0.01 seconds were added for each letter of the word. Each triple of words was followed by an answer period that lasted between 1 and 3 seconds. Answer period durations were jittered to ensure that trial stop times were not time-locked to the onset of each TR. During the answer period participants were asked to press a button to indicate whether the triple of words formed a valid instance of a relation. During the answer period no words were present on the screen.

Trials were presented across 11 unique runs. Each run contained 136 trials. The order of trials was randomized and the relation types, objects, and correct answers (true or false) were balanced across runs. One of the 11 runs was used as a test run. This run contained triples of words that were not used in any of the other ten runs. The test run was performed twice by each participant (once in each session). The two repeats of the test run were used to get an estimate of the noise ceiling (Hsu et al., 2004; Sahani & Linden, 2002; Schoppe et al., 2016).

Participant reaction time was measured as the amount of time between the answer cue and the participants' answer input. Participant accuracy was measured as the percentage of trials in which the participant gave the same answer as the mean of the rest of the group. Note that some trials may have subject answers, and thus this is a conservative estimate of participant accuracy.

fMRI data were collected during two 2-hour scanning sessions that were performed on different days. Each scanning session consisted of six ten-minute long runs (five were train runs and one was the test run).

Participants

Functional data were collected from six participants between the ages of 25-29 (5 female, 1 male). The stimuli were piloted on the first author of this study. Those pilot data are excluded from the study because the author overlearned the stimuli. All participants were right handed according to the edinburgh handedness inventory (Oldfield, 1971a) (laterality quotient of -100: entirely left-handed, +100: entirely right-handed). Laterality scores were 90, 85, 75, 90, 100, 55 for P1-P6 respectively.

To ensure generalization across participants the entire analysis was performed in each individual participant and consistency was measured between each participant and the group. Furthermore, before performing any analyses two of the six participants (P5, P6) were designated as held-out participants (Popham et al., 2021). Data for these two participants were not analyzed until the entire experiment and analysis pipeline was finalized.

fMRI Data Acquisition

Whole-brain MRI data were collected on a 3T siemens TIM trio scanner at the UC Berkeley Brain Imaging Center. A 32-channel Siemens volume coil was used. For participants P1, P2, P3, P5, and P6 functional scans were collected using a T2*-weighted gradient-echo EPI with repetition time (TR) 2.0045s, echo time (TE) 35ms, flip angle 74°, voxel size 2.24x2.24x4.1 mm (slice thickness 3.5mm), matrix size 100x100, and field of view 224x224 mm. Thirty axial slices were prescribed to cover the entire cortex and were scanned in interleaved order. A custom-modified bipolar water excitation radiofrequency (RF) pulse was used to prevent contamination from fat signals. The functional scans for participant P4 were collected using a sequence with multiband acceleration factor 3, repetition time (TR) 1.156s, echo time (TE) 34ms, flip angle 62°, voxel size 2.5x2.5x2.5 mm (slice thickness 2.5mm), matrix size 84x84, and field of view 210x210 mm. Anatomical data were collected using a T1-weighted multi-echo MP-RAGE sequence on the same 3T scanner.

To stabilize head motion during scanning sessions, participants wore a personalized head case that precisely fit the shape of each participant’ s head (Gao, 2015; Power et al., 2019).

fMRI data pre-processing

Each functional run was motion-corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FSL (Jenkinson et al., 2012). All volumes in the run were averaged across time to obtain a high quality template volume. FLIRT was used to automatically align the template volume for each run to the overall template, which was chosen to be the temporal average of the first functional run for each participant. The cross-run transformation matrix was then concatenated to the motion-correction transformation matrices obtained using MCFLIRT, and the concatenated transformation was used to resample the original data directly into the overall template space. Noise from motion, respiratory, and cardiac signals were removed with a component-based detrending method (CompCor; (Behzadi et al., 2007)). Responses of each run were z-scored separately. During z-scoring, for each voxel separately the mean response across time was subtracted and the remaining response was scaled to have unit variance. Before data analysis, 10 TRs from the beginning and 10 TRs at the end of each run were discarded in order to account for the 10 seconds of silence at the beginning and end of each scan and to account for non-stationarity in brain responses at the beginning and end of each scan

Cortical surface reconstruction and visualization

Cortical surface meshes were generated from the T1-weighted anatomical scans using Freesurfer software (Fischl et al., 1999). Before surface reconstruction, anatomical surface segmentations were carefully hand-checked and corrected using Blender software and pycortex (Community, 2018; Gao et al., 2015). Relaxation cuts were made into the surface of each hemisphere. Blender and pycortex were used to remove the surface crossing the corpus callosum. The calcarine sulcus cut was made at the horizontal meridian in V1 using retinotopic mapping data as a guide.

Functional images were aligned to the cortical surface using pycortex. Functional data were projected onto the surface for visualization and analysis using the line-nearest scheme in pycortex. This projection scheme samples the functional data at 32 evenly spaced intervals between the inner (white matter) and outer (pial) surfaces of the cortex and then averages together the samples. Samples are taken using nearest-neighbor interpolation, wherein each sample is given the value of its enclosing voxel.

Statistical Analyses

Voxelwise modeling (VM) was used to model the recorded BOLD responses (Naselaris et al., 2011; M. C.-K. Wu et al., 2006). In the VM framework, stimulus and task parameters are nonlineraly transformed into feature spaces. Each feature space describes an aspect of the experiment that is hypothesized to be represented in brain responses. Linearized regression is used to estimate a separate encoding model for each voxel and feature space. The model weights describe how each feature space modulates the BOLD response of each voxel. A held-out dataset that was not used to estimate model weights is used to evaluate prediction accuracy and to determine the statistical significance of prediction accuracies.

All model fitting and analysis was performed using custom software written in Python, making heavy use of NumPy (C. R. Harris et al., 2020), SciPy (Virtanen et al., 2020), Matplotlib (Hunter, 2007), Himalaya (Dupré la Tour et al., 2022), and Pycortex (Gao et al., 2015).

Stimulus Feature Spaces

A six-dimensional binary feature space was constructed to reflect the timing of the relation trials. Each dimension of this feature space corresponds to one of the six relations in the experiment, and reflects the onset time and duration of trials involving that relation. An analogous two-dimensional binary feature space was constructed to reflect the timing of the wordform relation trials. Fifteen nuisance feature spaces were constructed to reflect visual spatial and motion features (motion energy) (Adelson & Bergen, 1985; Nishimoto et al., 2011; Watson & Ahumada, 1985), the letters in each word, length of each word, standard deviation of word length within each TR, the mean word length per word, the change in mean word length across TRs, word rate, the elapsed time starting from the beginning of each trial, the elapsed time starting from the beginning of each question, the elapsed time starting from the beginning of each answer, the correct answer to the trial, the participant' s typed answer, the accuracy of the participant' s answer, and the lexical semantics of each word.

Stimulus Feature Preprocessing

Before voxelwise modeling, each stimulus feature was truncated, downsampled, z-scored, and delayed. Data for the first 10 TRs and the last 10 TRs of each scan were truncated to account for the 10 seconds of silence at the beginning and end of each scan and to account for non-stationarity in brain responses at the beginning and end of each scan. An anti-aliasing, 3-lobe Lanczos filter with cut-off frequency set to the fMRI Nyquist rate (0.25 Hz) was used to resample the stimulus features to match the sampling rate of the fMRI recordings. Then the stimulus features were each z-scored in order to account for z-scoring performed on the MRI data (For details see Section 4.4). In the z-scoring procedure, the value of each feature channel was separately normalized by subtracting the mean value of the feature channel across time and then dividing by the standard deviation of the feature channel across time. Lastly, finite impulse response (FIR) temporal filters were used to delay the features in order to model the hemodynamic response function of each voxel. The FIR filters were implemented by concatenating feature vectors that had been delayed by 2, 4, 6, and 8 seconds (following e.g., Chen, Dupré la Tour, et al., 2024; Deniz et al., 2019; Huth et al., 2016). A separate FIR filter was fit for each feature, participant, and voxel.

For one participant (P1) the stimuli presented to the participant contained some repeated trials between train and test. To remove this overlap we cleaned the train stimulus features by setting the relation stimulus feature spaces to 0 for any trial that was also included in the test stimuli for participant P1. This train stimulus feature cleaning ensures that the estimated model weights do not rely on trials that are repeated between train and test.

Regularization hyperparameter selection

Five-fold cross-validation was used to find the optimal regularization hyperparameters for each feature space and each voxel. Hyperparameter candidates were chosen with a random search procedure (Bergstra & Bengio, 2012): 1000 normalized hyperparameter candidates were randomly sampled from a dirichlet distribution and were then scaled by 20 log-spaced values ranging from 10^{-10} to 10^{10}. The regularization hyperparameters for each feature space and voxel were selected as the hyperparameters that produced the minimum squared error (L2) loss between the predicted voxel responses and the recorded voxel responses ($\arg\min_{hyperparameters} ||\hat{y} - y||_2^2$). Regularization hyperparameters were chosen separately for each participant and voxel. This hyperparameter search was performed using the Himalaya Python package (Dupré la Tour et al., 2022).

Model estimation and evaluation

The selected regularization hyperparameters were used to estimate model weights that map from the relation feature space to voxel BOLD response. Model weights were estimated separately for each voxel and participant. For each relation, the model weights for the corresponding dimension of the relation feature space reflect voxelwise selectivity for that relation.

The test dataset was not used to select hyperparameters or to estimate regression weights. The prediction accuracy R^2 of the feature spaces was computed per voxel as the coefficient of determination between the predicted voxel responses and the recorded voxel responses.

Group-level prediction accuracy

Group-level prediction accuracy was computed by first computing prediction accuracy for each participant in the participant ' s native brain space, and then projecting individual participant results into a standard fsAverage space. Average prediction accuracy across six participants was computed for each fsAverage vertex.

Voxelwise Relation Selectivity

A voxel was considered to be *selective* for a particular relation if the voxel is more highly activated during trials that involve the relation. If a voxel is highly activated during trials for a particular relation, then the relation feature space will accurately predict BOLD responses in that voxel and the model weights for that relation will be high. Thus, voxelwise selectivity S_{R_i} for each relation was operationally defined as the product of of the prediction accuracy of the relation feature space ($\sqrt{\tilde{R^2}_{relations}}$) and the model weight for that relation (β_{R_i}): $S_{R_i} = \sqrt{\tilde{R^2}_{relations}} \times \beta_{R_i}$. Selectivity was clipped to have a minimum of zero.

Voxelwise Concept Selectivity

To estimate voxelwise selectivity for each concept category we first used an established experiment paradigm to determine the lexical semantic tuning of each voxel. In this experiment participants either read or listen to natural narrative stories while fMRI is used to record BOLD responses in each voxel (See (Deniz et al., 2019; Huth et al., 2016) for more details of the experiment paradigm). Then a feature space was constructed to describe the lexical semantic content of the narrative. To construct this feature space, each word of the narratives was projected to a 300-dimensional word embedding space (word2vec (Mikolov et al., 2013)). Nuisance feature spaces were constructed to reflect the stimulus letters, phonemes, number of phonemes per TR, number of letters per TR, number of words per TR, standard deviation of word length within each TR, spatial and motion frequencies of the visual stimulus (Adelson & Bergen, 1985; Nishimoto et al., 2011; Watson & Ahumada, 1985), and spectral frequencies of the auditory stimulus. To account for the hemodynamic response function of each voxel, FIR filters were used to delay each feature space by 2, 4, 6, and 8 seconds. Banded ridge regression was used to estimate model weights for each of the feature spaces. Model weights were estimated separately for each participant. The model weights for the lexical semantic feature space describe the concept selectivity of each voxel; i.e., how different concepts modulate brain responses in each voxel. To evaluate the prediction accuracy of the estimated model weights, the estimated model weights for each voxel were used to predict BOLD responses to a held-out test set. Total prediction accuracy of all the feature spaces was measured as the coefficient of determination (R^2) between the predicted and observed BOLD responses on this held-out test set. Then the contribution $\tilde{R^2}_{featurespace}$ of each feature space to the total prediction accuracy was computed with the product measure (Hoffman, 1960; Pratt, 1987).

Similarity of Relation Representations between Participants

To test whether the representation of each relation is consistent between participants, selectivity for each relation was compared between participants in the standard template space. For each relation and for each pair of participants, the Pearson correlation was used to measure the similarity between the two participants of vertexwise selectivity for the relation.

A permutation test with 1000 iterations was used to compute the statistical significance of these correlations. This permutation test was performed separately for each relation and for each pair of participants. In each iteration, relation selectivity was shuffled across vertices within each participant. Then the shuffled values were used to compute the correlation between participants of vertexwise selectivity for the relation. Finally, the shuffled correlations were used as a null distribution to compute the p-value of the similarity between participants of vertexwise selectivity for the relation.

Statistical Significance

A permutation test with 1000 iterations was used to compute the significance of prediction accuracy. In each permutation the train semantic features were shuffled in blocks of 10 TRs. Shuffling was performed in blocks of 10 TRs in order to preserve autocorrelations in voxel responses. Then the analysis pipeline was repeated with the shuffled features (included fitting model weights, predicting test responses, and evaluating test prediction accuracy). A fixed regularization parameter was used in each permutation. The distribution of test accuracies over permutation iterations was used as a null distribution to compute the p-value of prediction accuracy for each voxel.

A jackknife procedure (Abdi & Williams, 2022) was used to determine the statistical significance of voxel preference for a specific relation. For each voxel the six relations were sorted from highest to lowest selectivity. The ordered relations were referred to as R_0 (highest selectivity relation) through R_5 (lowest selectivity relation). Then a leave-one-run-out jackknife procedure was used to estimate confidence intervals around the difference between selectivity for R_0 and for each of $[R_1, ..., R_5]$. In this jackknife procedure, each of the 10 train runs was left out in turn, and the remaining 9 runs were used to construct a partial estimate of voxelwise selectivity for each of the relations. These partial estimates were used to estimate confidence intervals around $(S_{R_0} - S_{R_i})$ for $i \in [1, 5]$. A voxel was considered to significantly prefer a specific relation if the 95% bootstrap confidence interval was strictly positive for each of $[R_1, ..., R_5]$.

4.5 Supplementary Figures

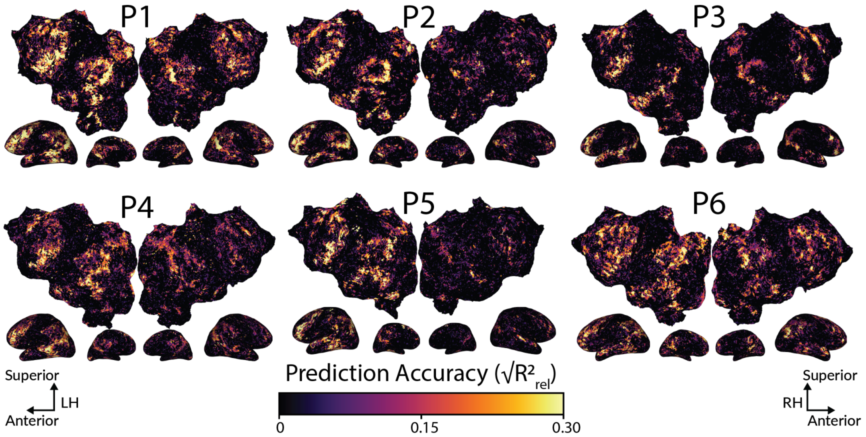

Figure 4.6: Prediction accuracy of the relation feature space, for each participant. Voxelwise prediction accuracy of the relation feature space is shown for each individual participant. Results are shown on the flattened cortical surface of each participant' s native brain space. Voxels that were not significantly well-predicted are shown in black. Prediction accuracy is high throughout bilateral temporal, parietal, and prefrontal cortices. (STS=superior temporal sulcus, ITS=inferior temporal sulcus, IPS=inferior parietal sulcus, SFS=superior frontal sulcus, IFS=inferior frontal sulcus).

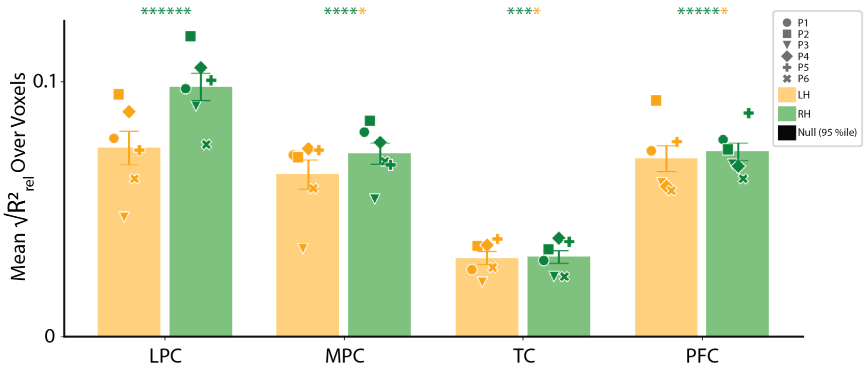

Figure 4.7: Prediction accuracy of the wordform relation feature space. For each cortical region, yellow and green markers show the mean prediction accuracy over voxels for the left and right hemispheres respectively. Black markers show null prediction accuracy (95th percentile). Bars show the mean across participants. Yellow and green asterisks show the number of participants for which prediction accuracy is significantly higher in the left or right hemisphere. Accuracies are higher in the right than the left hemisphere. These results show that representations of wordform relations are right-lateralize, suggesting that the left-lateralization of relation representations shown in Figure 4.2b is not merely due to the experiment presentation format.

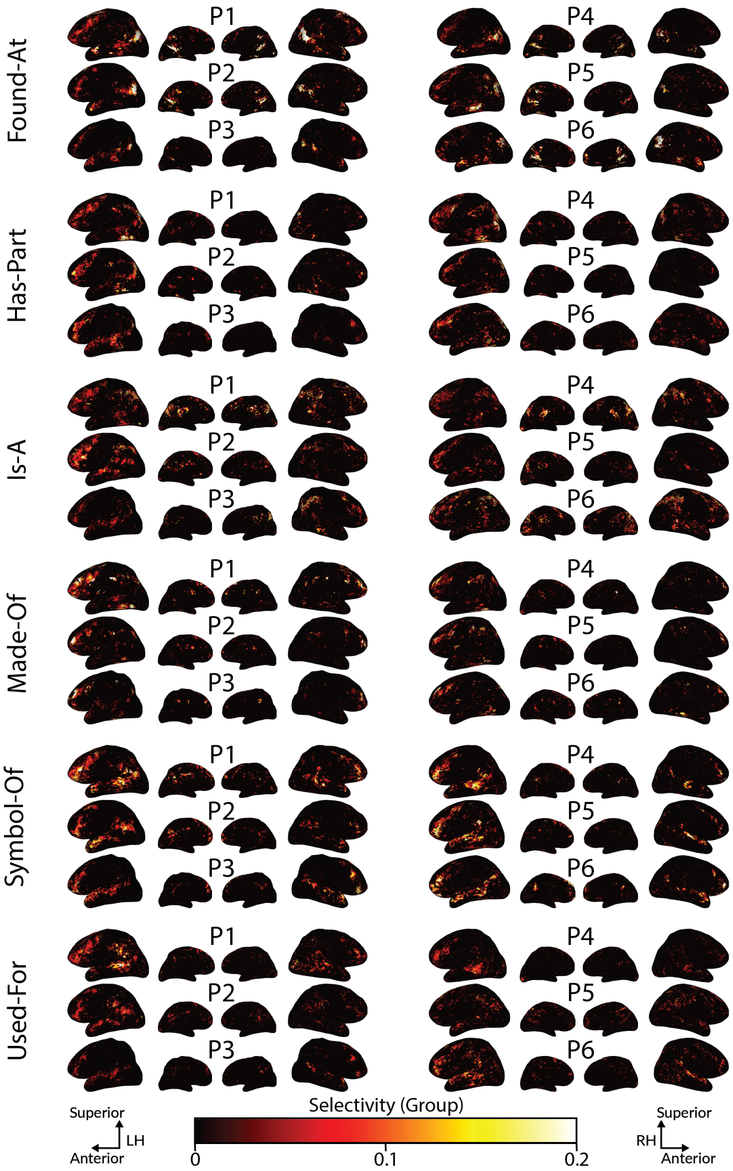

Figure 4.8: Selectivity for each relation for each individual participant. Results are shown on the inflated cortical surface of each participant' s native brain. Voxel color reflects selectivity for the relation. Representations of each relation appear to be consistent across participants.

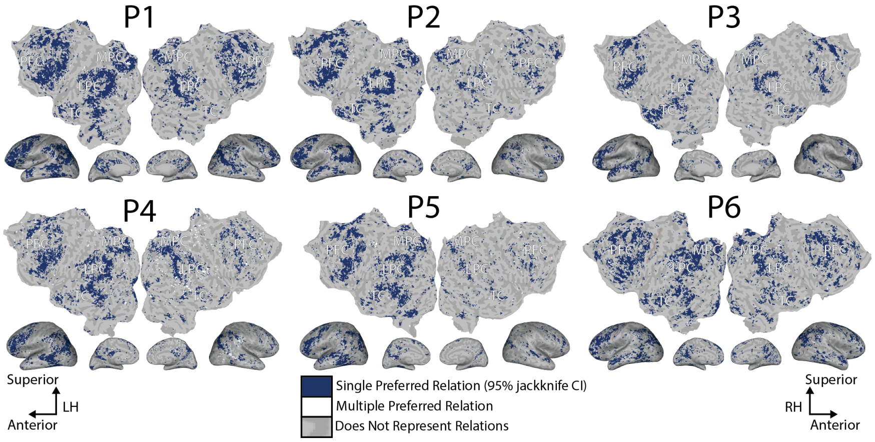
P1
P2
P3
P4
P5
P6
PFC
MPC
LPC
TC
Superior
LH
Anterior
Single Preferred Relation (95% jackknife CI)
Multiple Preferred Relation
Does Not Represent Relations
Superior
RH
Anterior

Figure 4.9: Specificity of relation selectivity, for each participant. To determine whether each voxel is selective for a specific relation, for each voxel the relation with the highest selectivity (R_0) was identified and then a jackknife procedure was used to estimate confidence intervals around the difference between selectivity for R_0 and for the other five relations. A voxel was considered to prefer a single relation if all five confidence intervals were strictly positive. Results are shown for each participant on the flattened surface of the participant' s native brain space. Voxels shown in blue prefer a single relation. Voxels shown in white do not prefer a single relation. Voxels shown in grey do not represent relations. Most voxels throughout the semantic system prefer a single relation.

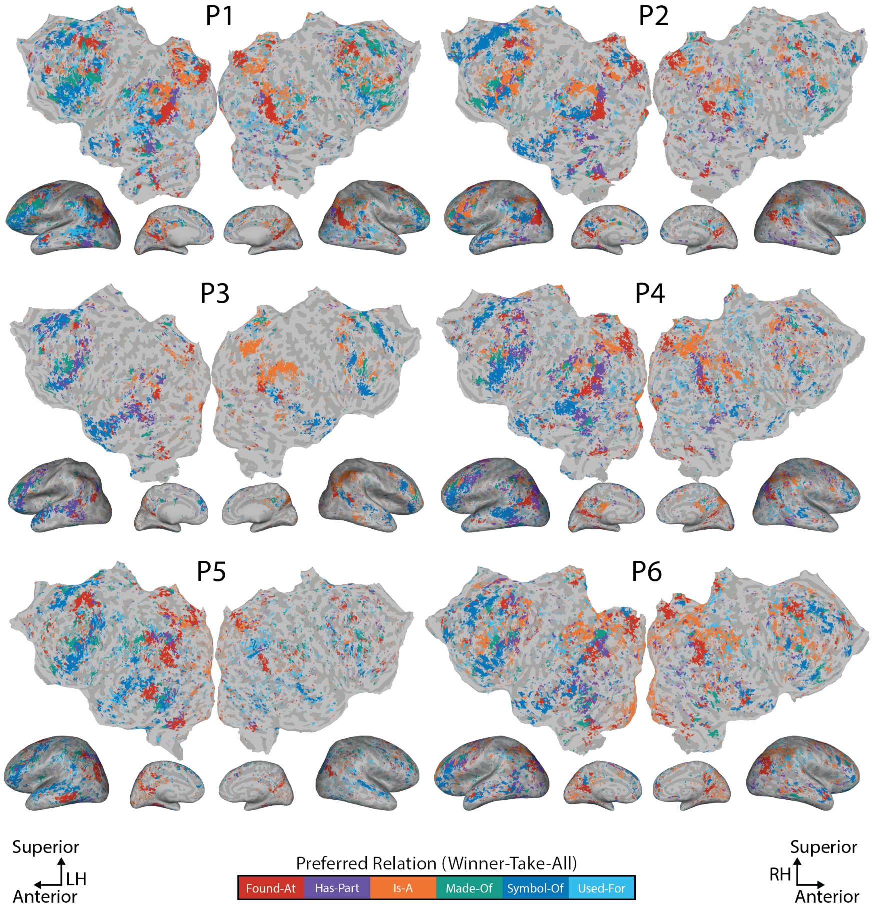

Figure 4.10: Winner-take-all map of relation representations, for each participant in the participant' s native brain space. The color of each voxel reflects the relation with the highest selectivity (R_0). The opacity of each voxel reflects the prediction accuracy of the relation feature space. The cortical organization of preferred relations appears to be consistent across participants.

Chapter 5

Conclusion

The studies presented in this dissertation examine the brain representations that underlie human language processing. Chapter 2 shows that in the brains of bilinguals, semantic representations are shared between languages but are subtly modulated to create language-dependent representations. Chapter 3 shows that representations across different language timescales are shared between sensory modalities. Chapter 4 shows that representations of semantic relations are organized similarly to representations of concepts, suggesting that relations and concepts could be represented in a unified semantic system. The phenomena investigated in these studies represent only a small subset of the processes that underlie language processing, yet they highlight the intricacy, flexibility, and diversity of brain representations in the human semantic system.

These three studies also illustrate the potential for studying language processing in both humans and in artificial models. First, developments in artificial language models (LMs) have provided new tools for understanding how the human brain represents language, and recent work has shown that LMs can be used to produce highly accurate models of brain responses (Caucheteux & King, 2022; Lamarre et al., 2022; Schrimpf et al., 2021; Toneva & Wehbe, 2019). In Chapters 1 and 2, I presented methods that enable us to better interpret these models of brain responses and thereby leverage LMs to better understand how specific aspects of language are represented in the human brain. Second, the human brain presents an opportunity for discovering ways to design better artificial language systems. Despite differences between the low-level mechanisms of the human brain and of artificial systems, the human brain has historically served as inspiration for the design of better artificial systems (e.g., Graves et al., 2016; LeCun et al., 1989). And despite vast advances in LMs, there are still key ways in which they continue to struggle. In work described elsewhere, I found that some artificial models struggle to represent spatial and semantic relations in a way that generalizes to new situations, and that does not require extensive and expensive training and inference (Chen, Lin, & Klein, 2021; Chen, Lu, et al., 2021; Chen et al., 2023). Chapter 3, which investigates brain representations of semantic relations, is in part inspired by that work. I believe that understanding how relations are represented in the human brain can help us devise LMs that can understand relations more flexibly, robustly, and efficiently.

In the future, I am excited to continue exploring language processing in the brain, and to see how these findings can be applied to discover better methods for human language education and rehabilitation, and to more robust and efficient systems for artificial language processing.

Bibliography

Abdi, H., & Williams, L. J. (2022). Jackknife. *The SAGE Encyclopedia of Research Design*. https://api.semanticscholar.org/CorpusID:246899483

Abutalebi, J., & Green, D. (2007). Bilingual language production: The neurocognition of language representation and control. *J. Neurolinguistics*, *20*(3), 242–275.

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, *2*(2), 284–299.

Ansaldo, A. I., Marcotte, K., Scherer, L., & Raboyeau, G. (2008). Language therapy and bilingual aphasia: Clinical implications of psycholinguistic and neuroimaging research. *J. Neurolinguistics*, *21*(6), 539–557.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449–12460.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709–721.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage*, *37*(1), 90–101.

Bejar, I. I., Chaffin, R., & Embretson, S. (1991). *Cognitive and psychometric analysis of analogical problem solving*. Springer Science & Business Media.

Bengio, Y. (2000). Gradient-based optimization of hyperparameters. *Neural computation*, *12*(8), 1889–1900.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2), 281–305.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex*, *19*(12), 2767–2796.

Blackman, R. B., & Tukey, J. W. (1958). The measurement of power spectra from the point of view of communications engineering—part i. *Bell System Technical Journal*, *37*(1), 185–282.

Blank, I., & Fedorenko, E. (2020). No evidence for differences among language regions in their temporal receptive windows. *NeuroImage*, *219*, 116925.

Bojanowski, P., Grave, E., Joulin, A., et al. (2017). Enriching word vectors with subword information. *Transactions of the*.

Booth, J. R., Burman, D. D., Meyer, J. R., Gitelman, D. R., Parrish, T. B., & Mesulam, M. M. (2002). Modality independence of word comprehension. *Human brain mapping*, *16*(4), 251–261.

Bot, K. (2003). A bilingual production model: Levelt's 'speaking' model adapted, 399–420.

Bouraoui, Z., Camacho-Collados, J., & Schockaert, S. (2019). Inducing relational knowledge from bert. *AAAI Conference on Artificial Intelligence*.

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language*, *157*, 81–94.

Bressler, D. W., & Silver, M. A. (2010). Spatial attention improves reliability of fMRI retinotopic mapping signals in occipital and parietal cortex. *Neuroimage*, *53*(2), 526–533.

Buchweitz, A., Mason, R. A., Hasegawa, M., & Just, M. A. (2009). Japanese and english sentence reading comprehension and writing systems: An fMRI study of first and second language effects on brain activation. *Biling.*, *12*, 141–151.

Buchweitz, A., Mason, R. A., Tomitch, L., & Just, M. A. (2009). Brain activation for reading and listening comprehension: An fmri study of modality effects and individual differences in language comprehension. *Psychology & neuroscience*, *2*, 111–123.

Buchweitz, A., Shinkareva, S. V., Mason, R. A., Mitchell, T. M., & Just, M. A. (2012). Identifying bilingual semantic neural representations across languages. *Brain Lang.*, *120*(3), 282–289.

Caramazza, A., & Brones, I. (1980). Semantic classification by bilinguals. *Can. J. Exp. Psychol.*, *34*(1), 77–81.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 1–10.

Chaffin, R. (1988). The nature of semantic relations: A comparison of two approaches. In *Relational models of the lexicon*.

Chaffin, R., & Herrmann, D. J. (1984). The similarity and diversity of semantic relations. *Memory & Cognition*, *12*.

Chee, M. W., Caplan, D., Soon, C. S., Sriram, N., Tan, E. W., Thiel, T., & Weekes, B. (1999). Processing of visually presented sentences in mandarin and english studied with fMRI. *Neuron*, *23*(1), 127–137.

Chen, C., Dupré la Tour, T., Gallant, J. L., Klein, D., & Deniz, F. (2024). The cortical representation of language timescales is shared between reading and listening. *Communications Biology*, *7*(1), 1–13.

Chen, C., Gong, X., Tseng, C., Klein, D., Gallant, J., & Deniz, F. (2024). Bilingual language processing relies on shared semantic representations that are modulated by each language. *bioRxiv*, 2024–06.

Chen, C., Lin, K., & Klein, D. (2021). Constructing taxonomies from pretrained language models. *North American Chapter of the Association for Computational Linguistics*.

Chen, C., Lu, Q., Beukers, A., Baldassano, C., & Norman, K. A. (2021). Learning to perform role-filler binding with schematic knowledge. *PeerJ*, *9*, e11046.

Chen, C., Shen, Z., Klein, D., Stanovsky, G., Downey, D., & Lo, K. (2023). Are layout-infused language models robust to layout distribution shifts? a case study with scientific documents. *Findings of the Association for Computational Linguistics*.

Chiang, J. N., Peng, Y., Lu, H., Holyoak, K. J., & Monti, M. M. (2021). Distributed code for semantic relations predicts neural similarity during analogical reasoning. *Journal of Cognitive Neuroscience*, *33*(3).

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and brain sciences*, *39*. https://doi.org/10.1017/S0140525X1500031X

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, *8*(2), 240–247.

Community, B. O. (2018). Blender - a 3D modelling and rendering package.

Çukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nat. Neurosci.*, *16*(6), 763–770.

de Groot, A. M. B. (1992). Bilingual lexical representation: A closer look at conceptual representations. *Orthography, phonology, morphology, and meaning.*, *435*, 389–412.

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, *37*(27), 6539–6557.

Dehaene, S., Dupoux, E., Mehler, J., Cohen, L., Paulesu, E., Perani, D., van de Moortele, P. F., Lehéricy, S., & Le Bihan, D. (1997). Anatomical variability in the cortical representation of first and second language. *Neuroreport*, *8*(17), 3809–3815.

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science*, *284*(5416), 970–974.

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

Deniz, F., Tseng, C., Wehbe, L., la Tour, T. D., & Gallant, J. L. (2023). Semantic representations during language comprehension are affected by context. *Journal of Neuroscience*, *43*(17), 3144–3158.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, *31*(3), 968–980.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, 4171–4186.

Dijkstra, T., Grainger, J., & van Heuven, W. J. B. (1999). Recognition of cognates and interlingual homographs: The neglected role of phonology. *J. Mem. Lang.*, *41*(4), 496–518.

Doumas, L. A., & Hummel, J. E. (2012). Computational models of higher cognition. *The Oxford handbook of thinking and reasoning*, *19*.

Dunagan, D., Zhang, S., Li, J., Bhattasali, S., Pallier, C., Whitman, J., Yang, Y., & Hale, J. (2022). Neural correlates of semantic number: A cross-linguistic investigation. *Brain Lang.*, *229*, 105110.

Dupré la Tour, T., Eickenberg, M., Nunez-Elizalde, A. O., & Gallant, J. L. (2022). Feature-space selection with banded ridge regression. *NeuroImage*, *264*, 119728. https://doi.org/10.1016/j.neuroimage.2022.119728

Duyck, W. (2005). Translation and associative priming with cross-lingual pseudohomophones: Evidence for nonselective phonological activation in bilinguals. *J. Exp. Psychol. Learn. Mem. Cogn.*, *31*(6), 1340–1359.

Fedorenko, E., & Kanwisher, N. (2009). Neuroimaging of language: Why hasn't a clearer picture emerged? *Lang. Linguist. Compass*, *3*(4), 839–865.

Fischl, B., Sereno, M. I., Tootell, R. B., & Dale, A. M. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human brain mapping*, *8*(4), 272–284.

Forster, K. I. (1970). Visual perception of rapidly presented word sequences of varying complexity. *Perception & psychophysics*, *8*(4), 215–221.

Friederici, A. D., Opitz, B., & Von Cramon, D. Y. (2000). Segregating semantic and syntactic aspects of processing in the human brain: An fmri investigation of different word types. *Cerebral cortex*, *10*(7), 698–705.

Gao, J. S., Huth, A. G., Lescroart, M. D., & Gallant, J. L. (2015). Pycortex: An interactive surface visualizer for fmri. *Frontiers in neuroinformatics*, 23.

Gao, J. S. (2015). *Fmri visualization and methods*. University of California, Berkeley.

Gentner, D., & Forbus, K. D. (2011). Computational models of analogy. *Wiley interdisciplinary reviews: cognitive science*, *2*(3), 266–276.

Gomez-Tortosa, E., Martin, E. M., Gaviria, M., Charbel, F., & Ausman, J. I. (1995). Selective deficit of one language in a bilingual patient following surgery in the left perisylvian area. *Brain Lang.*, *48*(3), 320–325.

Gong, X. L. (2024). *Language representation in human cerebral cortex* [Doctoral dissertation, UC Berkeley].

Gong, X. L., Huth, A. G., Deniz, F., Johnson, K., Gallant, J. L., & Theunissen, F. E. (2023). Phonemic segmentation of narrative speech in human cerebral cortex. *Nat. Commun.*, *14*(1), 4309.

Grainger, J., Midgley, K., & Holcomb, P. J. (2010). Chapter 14. re-thinking the bilingual interactive-activation model from a developmental perspective (BIA-d). In *Language* *acquisition and language disorders* (pp. 267–283). John Benjamins Publishing Company.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., et al. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471–476.

Hale, J., Lutz, D., Luh, W.-M., & Brennan, J. (2015). Modeling fMRI time courses with linguistic structure at various grain sizes. *Proceedings of the 6th workshop on cognitive modeling and computational linguistics*, 89–97.

Hamilton, L. S., & Huth, A. G. (2020). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *35*(5), 573–582.

Hansen, K. A., Kay, K. N., & Gallant, J. L. (2007). Topographic organization in and near human visual area V4. *J. Neurosci.*, *27*(44), 11896–11911.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Harris, F. J. (1978). On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, *66*(1), 51–83. https://doi.org/10.1109/PROC.1978.10837

Hasson, U., Malach, R., & Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in cognitive sciences*, *14*(1), 40–48.

Heilbron, M., Richter, D., Ekman, M., Hagoort, P., & De Lange, F. P. (2020). Word contexts enhance the neural representation of individual letters in early visual cortex. *Nature communications*, *11*(1), 1–11.

Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., & Bau, D. (2024). Linearity of relation decoding in transformer language models. *International Conference on Learning Representations*.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological bulletin*, *57*(2), 116.

Hofstadter, D., & Sander, E. (2013). The forgotten fuel of our minds. *New scientist*, (2915), 30–33.

Holyoak, K. J., Ichien, N., & Lu, H. (2022). From semantic vectors to analogical mapping. *Current Directions in Psychological Science*, *31*(4), 355–361.

Holyoak, K. J., & Lu, H. (2021). Emergence of relational reasoning. *Current Opinion in Behavioral Sciences*, *37*, 118–124.

Honey, C. J., Thompson, C. R., Lerner, Y., & Hasson, U. (2012). Not lost in translation: Neural responses shared across languages. *J. Neurosci.*, *32*(44), 15277–15283.

Hsu, A., Borst, A., & Theunissen, F. E. (2004). Quantifying variability in neural responses and its application for the validation of model predictions. *Network*, *15*(2), 91–109.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, *9*(03), 90–95.

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*, 453–458.

Illes, J., Francis, W. S., Desmond, J. E., Gabrieli, J. D. E., Glover, G. H., Poldrack, R., Lee, C. J., & Wagner, A. D. (1999). Convergent cortical representation of semantic processing in bilinguals. *Brain Lang.*, *70*(3), 347–363.

Jain, S., & Huth, A. (2018). Incorporating context into language encoding models for fMRI. *Advances in Neural Information Processing Systems*, *31*, 6628–6637.

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

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does bert learn about the structure of language? *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/P19-1356

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *Neuroimage*, *62*(2), 782–790.

Joulin, A., Bojanowski, P., Mikolov, T., Jegou, H., & Grave, E. (2018). Loss in translation: Learning bilingual word mapping with a retrieval criterion.

Jurgens, D., Mohammad, S., Turney, P., & Holyoak, K. (2012). Semeval-2012 task 2: Measuring degrees of relational similarity. ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 356–364.

Kanerva, P. (2010). What we mean when we say" what's the dollar of mexico?": Prototypes and mapping in concept space. *2010 AAAI fall symposium series*.

Kaushanskaya, M., Blumenfeld, H. K., & Marian, V. (2020). The language experience and proficiency questionnaire (LEAP-Q): Ten years later. *Biling.*, *23*(5), 945–950.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*(1), 109–128.

Kim, K. H. S., Relkin, N. R., Lee, K.-M., & Hirsch, J. (1997). Distinct cortical areas associated with native and second languages. *Nature*, *388*(6638), 171–174.

Kiremitçi, I., Yilmaz, Ö., Çelik, E., Shahdloo, M., Huth, A. G., & Çukur, T. (2021). Attentional modulation of hierarchical speech representations in a multitalker environment. *Cereb. Cortex*, *31*(11), 4986–5005.

Klein, D., Milner, B., Zatorre, R. J., Meyer, E., & Evans, A. C. (1995). The neural substrates underlying word generation: A bilingual functional-imaging study. *Proc. Natl. Acad. Sci. U. S. A.*, *92*(7), 2899–2903.

Kroll, J. F., & De Groot, A. M. (Eds.). (2005). *Handbook of bilingualism: Psycholinguistic approaches*. Oxford University Press.

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of memory and language*, *33*(2), 149–174.

Ku, A., Lachmann, E. A., & Nagler, W. (1996). Selective language aphasia from herpes simplex encephalitis. *Pediatr. Neurol.*, *15*(2), 169–171.

Lamarre, M., Chen, C., & Deniz, F. (2022). Attention weights accurately predict language representations in the brain. *Findings of the Conference on Empirical Methods in Natural Language Processing.* https://doi.org/10.1101/2022.12.07.519480

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. *Sci Data*, *10*(1), 555.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, *2*.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*(8), 2906–2915.

Li, J., Bhattasali, S., Zhang, S., Franzluebbers, B., Luh, W.-M., Spreng, R. N., Brennan, J. R., Yang, Y., Pallier, C., & Hale, J. (2022). Le petit prince multilingual naturalistic fMRI corpus. *Sci Data*, *9*(1), 530.

Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., & Wei, F. (2023). Trocr: Transformer-based optical character recognition with pre-trained models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 13094–13102.

Li, P., Zhang, F., Yu, A., & Zhao, X. (2020). Language history questionnaire (LHQ3): An enhanced tool for assessing multilingual experience. *Bilingualism: Language and Cognition*, *23*(5), 938–944.

Liu, Y., & Sun, M. (2015). Contrastive unsupervised word alignment with Non-Local features. *AAAI*, *29*(1).

Liuzzi, A. G., Bruffaerts, R., Peeters, R., Adamczuk, K., Keuleers, E., De Deyne, S., Storms, G., Dupont, P., & Vandenberghe, R. (2017). Cross-modal representation of spoken and written word meaning in left pars triangularis. *Neuroimage*, *150*, 292–307.

Luke, K.-K., Liu, H.-L., Wai, Y.-Y., Wan, Y.-L., & Tan, L. H. (2002). Functional anatomy of syntactic and semantic processing in language comprehension. *Hum. Brain Mapp.*, *16*(3), 133–145.

MacNamara, J. (1967). The linguistic independence of bilinguals. *Journal of Verbal Learning & Verbal Behavior*, *6*(5), 729–736.

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

Malik-Moraleda, S., Mahowald, K., Conway, B. R., & Gibson, E. (2023). Concepts are restructured during language contact: The birth of blue and other color concepts in Tsimane'-Spanish bilinguals. *Psychol. Sci.*, *34*(12), 1350–1362.

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *J. Open Source Softw.*, *3*, 861. https://api.semanticscholar.org/CorpusID:53244226

Meschke, E. X., Visconti di Oleggio Castello, M., Dupré la Tour, T., & Gallant, J. L. (2023). Model connectivity: Leveraging the power of encoding models to overcome the limitations of functional connectivity. *bioRxiv*.

Midgley, K. J., Holcomb, P. J., VanHeuven, W. J. B., & Grainger, J. (2008). An electrophysiological investigation of cross-language effects of orthographic neighborhood. *Brain Res.*, *1246*, 123–135.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *science*, *320*(5880), 1191–1195.

Nakai, T., Yamaguchi, H. Q., & Nishimoto, S. (2021). Convergence of modality invariance and attention selectivity in the cortical semantic circuit. *Cerebral Cortex*, *31*(10), 4825–4839.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage*, *56*(2), 400–410.

Nastase, S. A., Connolly, A. C., Oosterhof, N. N., Halchenko, Y. O., Guntupalli, J. S., Visconti di Oleggio Castello, M., Gors, J., Gobbini, M. I., & Haxby, J. V. (2017). Attention selectively reshapes the geometry of distributed semantic representation. *Cereb. Cortex*, *27*(8), 4277–4291.

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, *222*, 117254.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, *21*(19), 1641–1646.

Niu, J., Lu, W., & Penn, G. (2022). Does bert rediscover a classical nlp pipeline? *Proceedings of the 29th International Conference on Computational Linguistics*, 3143–3153.

Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage*, *197*, 482–492.

Oldfield, R. C. (1971a). The assessment and analysis of handedness: The edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

Oldfield, R. C. (1971b). The assessment and analysis of handedness: The edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

Oota, S. R., Arora, J., Agarwal, V., Marreddy, M., Gupta, M., & Surampudi, B. (2022). Neural language taskonomy: Which NLP tasks are the most predictive of fMRI brain activity? *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3220–3237. https://doi.org/10.18653/v1/2022.naacl-main.235

Oota, S. R., Marreddy, M., Gupta, M., & Bapi, R. (2023, July). How does the brain process syntactic structure while listening? In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: ACL 2023* (pp. 6624–6647). Association for Computational Linguistics.

Paradis, M. (1985). On the representation of two languages in one brain. *Lang. Sci.*, *7*(1), 1–39.

Pavlenko, A. (2002). Bilingualism and emotions. *21*(1), 45–78.

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? *Annu Meet Assoc Comput Linguistics*, *abs/1906.01502*.

Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., & Gallant, J. L. (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nat. Neurosci.*, *24*(11), 1628–1636.

Popov, V., & Pavlova, M. (2020). The internal structure of semantic relations: Effects of relational similarity and typicality.

Potter, M. C., So, K.-F., Von Eckardt, B., & Feldman, L. B. (1984). Lexical and conceptual representation in beginning and proficient bilinguals. *Journal of verbal learning and verbal behavior*, *23*(1), 23–38.

Power, J. D., Silver, B. M., Silverman, M. R., Ajodan, E. L., Bos, D. J., & Jones, R. M. (2019). Customized head molds reduce motion during resting state fmri scans. *Neuroimage*, *189*, 141–149.

Pratt, J. W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. *Proceedings of the second international Tampere conference in statistics, 1987*, 245–260.

Reddy, A. J., & Wehbe, L. (2021). Can fmri reveal the representation of syntactic structure in the brain? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (pp. 9843–9856, Vol. 34). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2021/file/51a472c08e21aef54ed749806e3e6490-Paper.pdf

Regev, M., Honey, C. J., Simony, E., & Hasson, U. (2013). Selective and invariant neural responses to spoken and written narratives. *Journal of Neuroscience*, *33*(40), 15978–15988.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866.

Sahani, M., & Linden, J. (2002). How linear are auditory cortical responses? In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15). MIT Press.

Schoppe, O., Harper, N. S., Willmore, B. D. B., King, A. J., & Schnupp, J. W. H. (2016). Measuring the performance of neural models. *Front. Comput. Neurosci.*, *10*, 10.

Schrauf, R. W., & Rubin, D. C. (2004). ‘language’ and ‘feel’ of bilingual memory. *Socioling. Stud.*, *5*(1), 21–39.

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), e2105646118.

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature communications*, *7*(1), 12141.

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological review*, *81*(3), 214.

Søgaard, A., Ruder, S., & Vulić, I. (2018). On the limitations of unsupervised bilingual dictionary induction.

Spelke, E. S., & Tsivkin, S. (2001). Language and number: A bilingual training study. *Cognition*, *78*(1), 45–88.

Spiridon, M., Fischl, B., & Kanwisher, N. (2006). Location and spatial profile of category-specific regions in human extrastriate cortex. *Hum. Brain Mapp.*, *27*(1), 77–89.

St-Yves, G., & Naselaris, T. (2018). The feature-weighted receptive field: An interpretable encoding model for complex feature spaces. *NeuroImage*, *180*, 188–202.

Tamkin, A., Jurafsky, D., & Goodman, N. D. (2020). Language through a prism: A spectral approach for multiscale language representations. *Advances in Neural Information Processing Systems*, 5492–5504.

Tang, J., Du, M., Vo, V., Lal, V., & Huth, A. (2024). Brain encoding models based on multimodal transformers can transfer across language and vision. *Advances in Neural Information Processing Systems*, *36*.

Tenney, I., Das, D., & Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. *Annual Meeting of the Association for Computational Linguistics*, 4593–4601.

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, 14954–14964.

Unger, L., & Fisher, A. V. (2021). The emergence of richly organized semantic knowledge from simple statistics: A synthetic review. *Developmental Review*, *60*.

van Hell, J. G., & Dijkstra, T. (2002). Foreign language knowledge can influence native language performance in exclusively native contexts. *Psychon. Bull. Rev.*, *9*(4), 780–789.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature methods*, *17*(3), 261–272.

Wang, W.-C., Hsieh, L.-T., Swamy, G., & Bunge, S. A. (2021). Transient neural activation of abstract relations on an incidental analogy task. *Journal of Cognitive Neuroscience*, *33*(1).

Watson, A. B., & Ahumada, A. J., Jr. (1985). Model of human visual-motion sensing. *J. Opt. Soc. Am. A*, *2*(2), 322–341.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, *9*(11), e112575.

Weinreich, U. (1986). *Languages in contact: Findings and problems*. Mouton.

Wu, M.-H., Anderson, A. J., Jacobs, R. A., & Raizada, R. D. (2022). Analogy-related information can be accessed by simple addition and subtraction of fmri activation patterns, without participants performing any analogy task. *Neurobiology of Language*, *3*(1), 1–17.

Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.*, *29*, 477–505.

Xu, M., Baldauf, D., Chang, C. Q., Desimone, R., & Tan, L. H. (2017). Distinct distributed patterns of neural activity are associated with two languages in the bilingual brain. *Sci Adv*, *3*(7), e1603309.

Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*. https://doi.org/10.1152/jn.00338.2011

Zhang, Y., Han, K., Worth, R., & Liu, Z. (2020). Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature communications*, *11*(1), 1877.