Bilingual language processing relies on shared semantic representations that are modulated by each language

Catherine Chen[a], Xue L. Gong[b], Christine Tseng[b], Daniel L. Klein[a], Jack L. Gallant[a,b], Fatma Deniz[b,c,d]

[a]Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720, USA
[b]Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720, USA
[c]Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, 10623 Berlin, Germany
[d]Bernstein Center for Computational Neuroscience, 10115 Berlin, Germany
Corresponding author: Fatma Deniz, deniz@tu-berlin.de

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.

Introduction

Over 4 billion people throughout the world are bilingual[1,2] 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[3–9]. 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[10–12], and the perceived emotional intensity or memorability of concepts can change between languages[13,14]. A second group of theories suggests that different languages evoke the same semantic brain representations[4,8,15–20]. These theories are supported by evidence of interference between languages: second language acquisition can increase processing times for false cognates[21–23], and restructure concept categories in one's primary language[24].

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[25–28], we

hypothesized that shared semantic representations are modulated by each language (Figure 1A). For example, language-dependent perceptions of emotional intensity[13,14] 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.

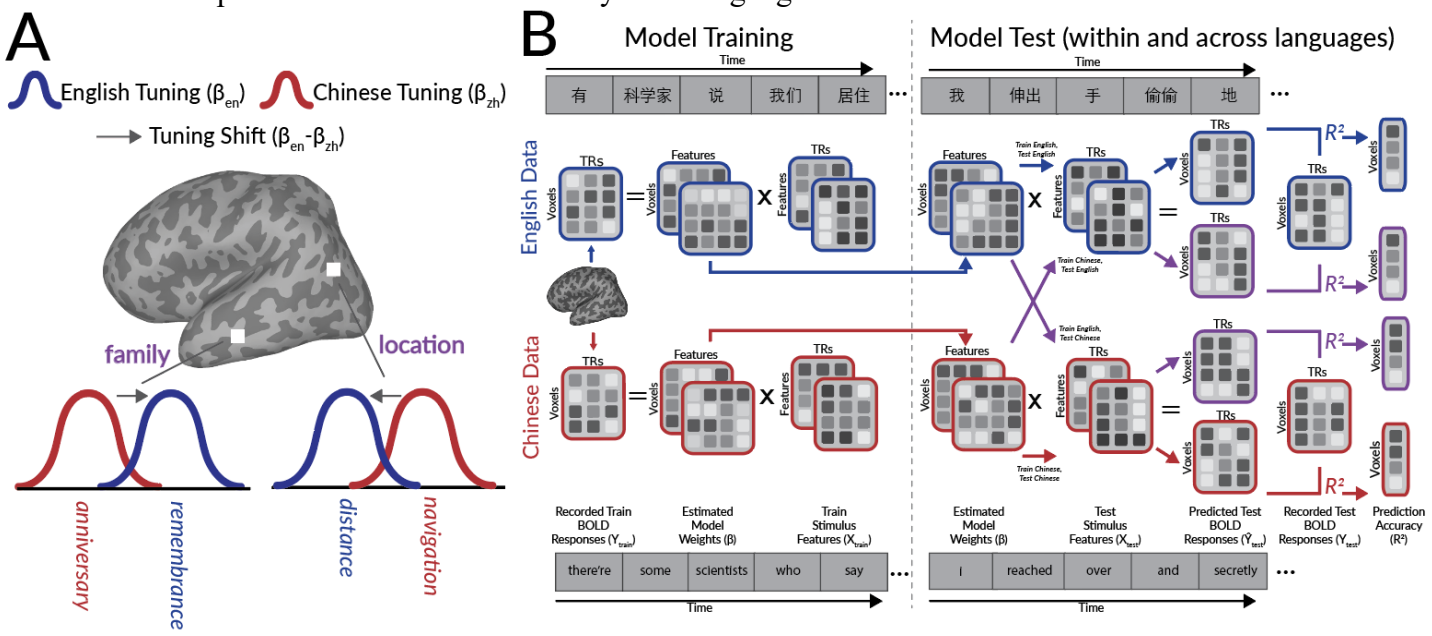

Figure 1. Hypothesis and experimental procedure.
A. Schematic illustrating hypothesized semantic representation shifts. The semantic tuning of a voxel describes its preference for each semantic concept. Shifts in semantic representations were quantified as the change in semantic tuning between languages. 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[29]. 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.

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[29,30]) in which words in different languages that express similar concepts project to nearby vectors. (A separate embedding space[31] produced similar results; Figures S1-S4). Regularized regression was used to estimate voxelwise encoding models separately for each participant and language (Figure 1B). 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.

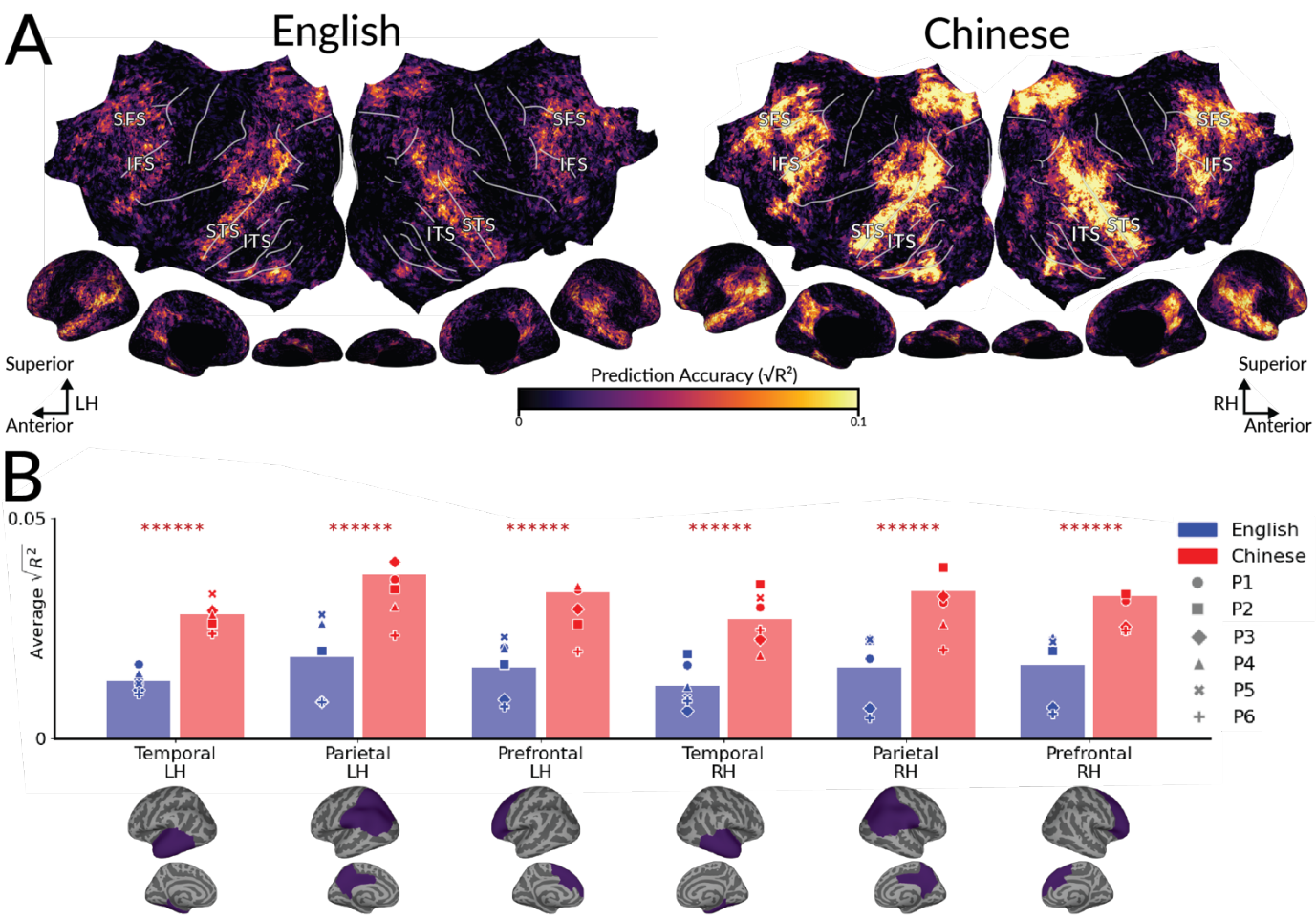

Figure 2. Cortical distribution of semantic representations for each language. To determine where semantic information is represented for each language, voxelwise models estimated for each language were used to predict held-out data for the same language. Prediction accuracy was computed as the CD (R^2) between predicted and recorded BOLD responses. **A.** Group-level prediction accuracy. Results are shown for each language on the flattened cortical 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 S6). 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[32]), and then averaging the projected values across participants for each vertex and language separately. Figure 2A shows vertexwise group-level prediction accuracy for each language separately. Results for each participant are similar to the group (Figure S5 and S6). For each language, prediction accuracy is highest within bilateral temporal, parietal, and prefrontal cortices. These regions are sometimes referred to as the *semantic system*[33,34]. 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 2A 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[35] regions of interest (ROIs) to identify voxels in each region, and then computed average prediction accuracy separately for each language and region. Figure 2B 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[36]. 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*)[37–39]. 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 S7-S8), suggesting that semantic representations constitute a lower proportion of overall brain responses to English.

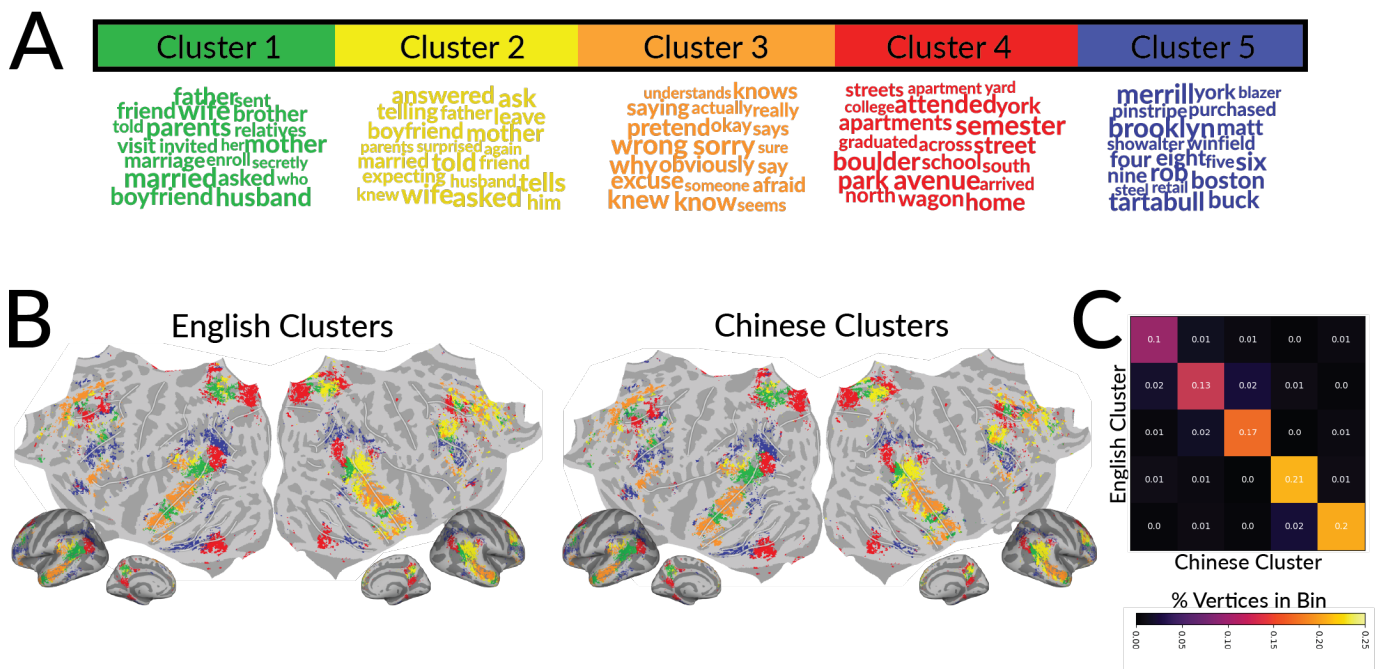

Figure 3. Shared semantic representations across languages. Hierarchical clustering on estimated model weights was used to categorize voxels based on semantic tuning in each language. **A.** Words closest to each cluster. The five clusters represent concepts related to family (Cluster 1, green), communication (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 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*[40]) to identify semantic clusters from the estimated model weights. Five clusters best summarized the distribution of model weights across participants and languages (Figure S9). 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 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 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 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 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 S10). 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 S6, S11). Overall, these results show that semantic representations are largely shared between languages.

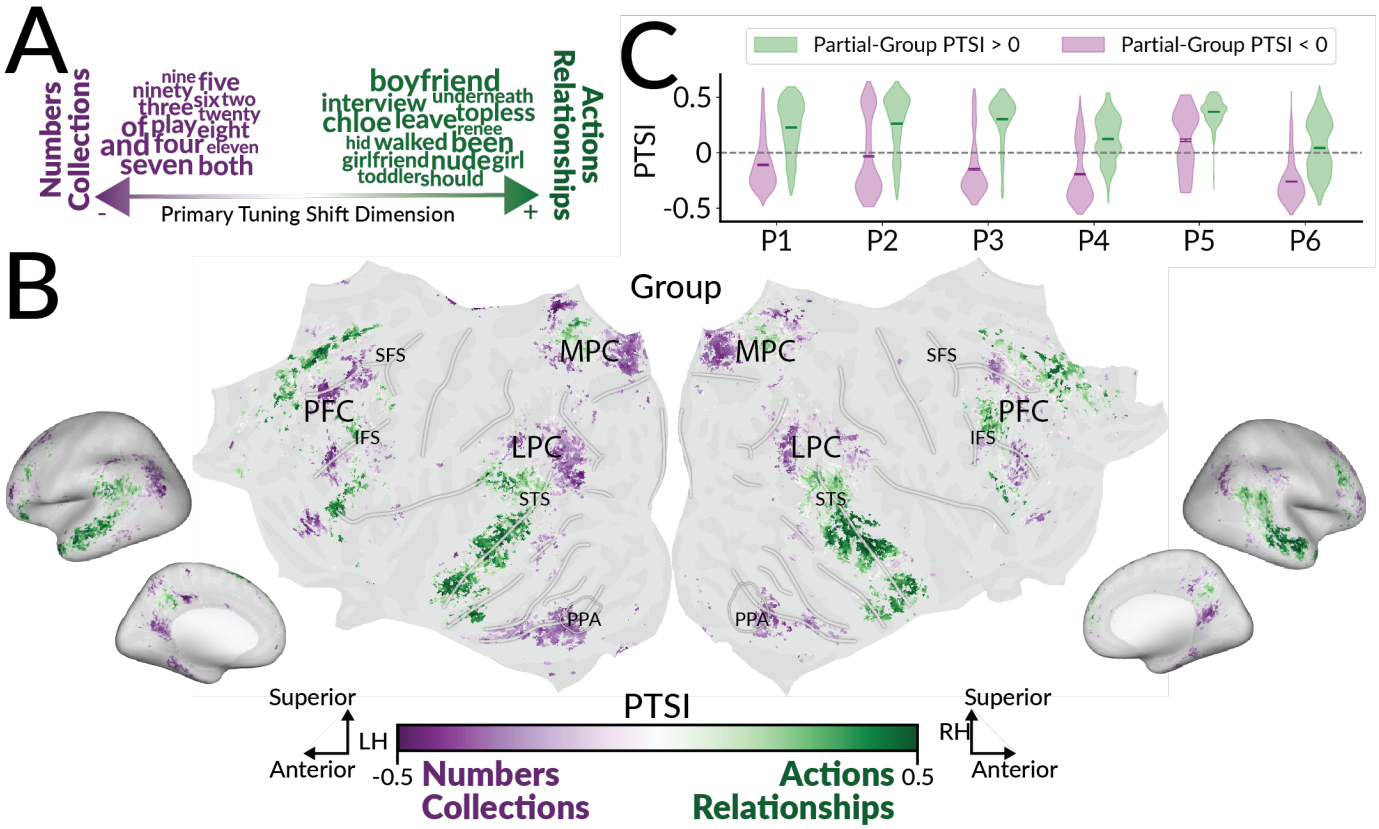

Figure 4. Cortical distribution of semantic tuning shifts between languages. Voxelwise *semantic tuning shift* was defined as the change in model weights between English and Chinese: $semantic_tuning_shift = \frac{\beta_{en}}{||\beta_{en}||_2} - \frac{\beta_{zh}}{||\beta_{zh}||_2}$. For each voxel, the 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 S12). **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 and 3 suggest that semantic representations are shared between languages. However, given strong behavioral evidence for language-dependent semantic representations[13,14], we hypothesized that shared semantic representations are modulated by each language (Figure 1A). 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 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 S12). 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 4A 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 S13.) 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 4B 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 4C 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[31]) that differs from fastText in its training objectives, training data, and embedding dimensionality (Figures S1-S4).

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 S12). Overall, the results in Figure 4 show that there are systematic semantic tuning shifts between languages.

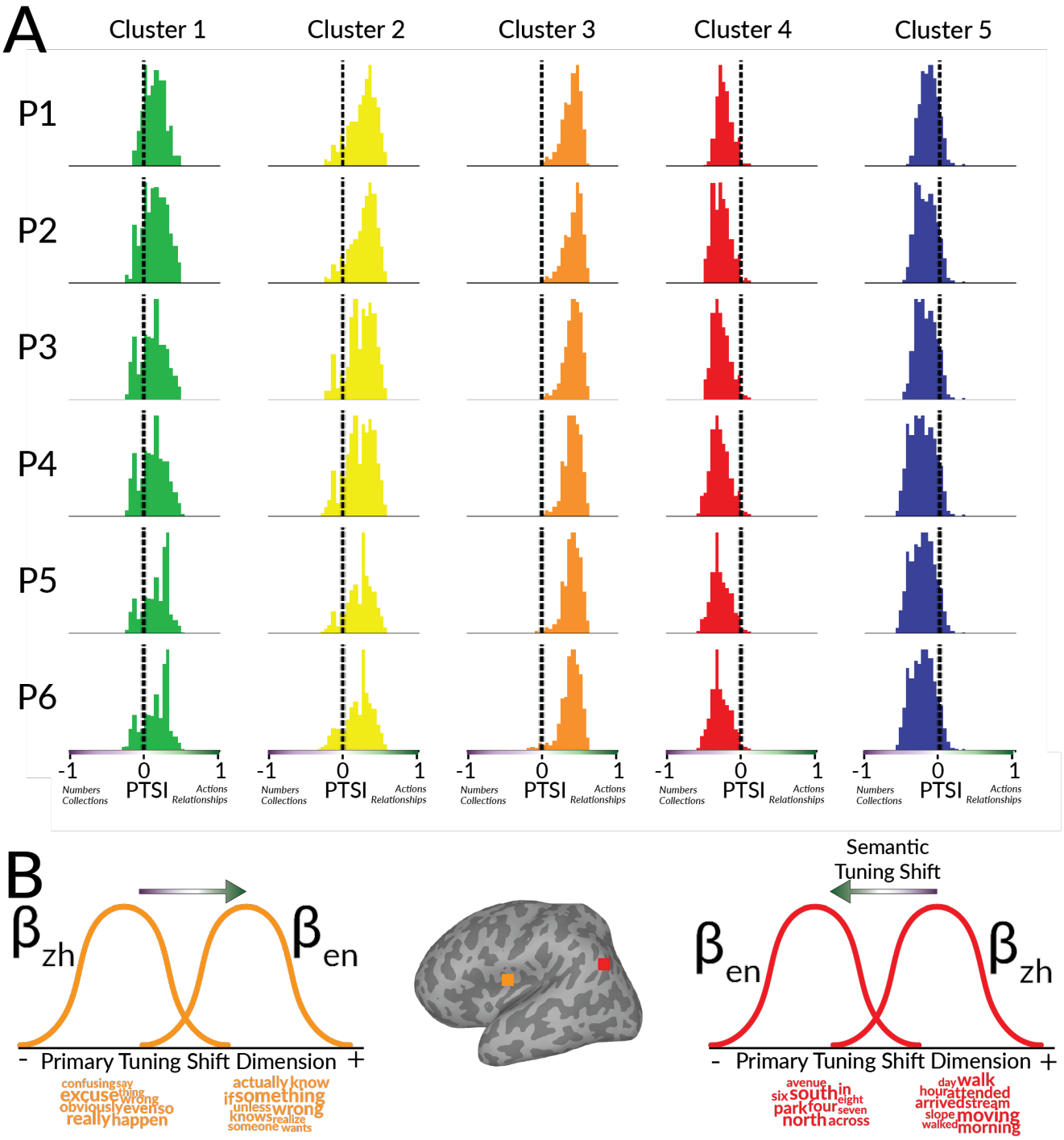

Figure 5. Semantic tuning shifts for different semantic clusters. Voxels were categorized into the semantic clusters shown in Figure 3. The distribution of semantic tuning shifts was examined for each semantic cluster. **A**. Semantic tuning shifts for each cluster. Histograms show the distribution of 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 4 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 3, we examined the direction of semantic tuning shifts for voxels in that semantic cluster. Figure 5A 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 S14). 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 5B 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 5 show that each language systematically modulates semantic representations.

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, S1, S5, S6, S8). Second, semantic representations are largely similar between languages (Figures 3, S2, S10-11). Third, there are systematic shifts in voxelwise semantic tuning between languages (Figures 4-5, S3-4, S14). 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[6,7,41,42] or the same[43–48]. 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[9,41,49,50]. Thus, individual differences in brain function and anatomy[51] may have exaggerated reported differences. Third, many prior studies used controlled stimuli. Thus, it was unclear whether reported results generalize to naturalistic settings[52]. 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[21–23] 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[13,14,53,54].

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.

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[34,55–57]. 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[56]).

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[58]. 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[59,60].

fMRI data pre-processing

Each functional run was motion-corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FSL[61]. 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[62]). 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[32]. Before surface reconstruction, anatomical surface segmentations were carefully hand-checked and corrected using Blender software and pycortex[63,64]. 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[35]. 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[65,66]. Details of localizer experiments are provided in prior work[34,56].

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)[67] and the Language History Questionnaire (LHQ3)[68]. 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 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)[69]. Laterality scores were +5, 0, +90, +100, +65, +65 for P1-6 respectively.

Statistical analysis

Voxelwise modeling (VM) was used to model BOLD responses[34,55,56,70,71]. 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[72], SciPy[73], Matplotlib[74], Himalaya[75], and Pycortex[63].

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[29]. 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[30].

Because of non-isometry between embedding spaces of different languages and limitations to multilingual alignment procedures[76], 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*[31]). 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[77]. 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 (Supplementary Figure S15), and because intermediate layers of contextual language models have been shown to produce the most accurate predictions of brain responses[78–82].

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)[83–85], word count, single phonemes, diphones, triphones[86], and intermediate level features that capture orthographic similarities by measuring the pixelwise overlap between words[87]. 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 *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. 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[34,56,88]). 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[89]. 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 $F_i(X), i \in \{1, \dots, m\}$ (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 $\hat{B} = argmin_B \left|\left|Y - BF'(X)\right|\right|_2^2 + \lambda \left|\left|CB\right|\right|_2^2$. The diagonal matrix C of regularization hyperparameters for each feature space and each voxel is optimized over 10-fold cross-validation. See *Regularization hyperparameter selection* 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 $Y_{train} - \hat{Y}_{lowlevel,train}$, $Y_{test} - \hat{Y}_{lowlevel,test}$ 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[90]: 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 ($argmin_{hyperparameters}||\hat{y} - y||_2^2$). Regularization hyperparameters were chosen separately for each participant and language. Hyperparameter search was performed using the Himalaya Python package[75].

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[56,57,81,82,91–94]. 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[95]. 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*)[37–39]. 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 $y_1 \dots y_N \in R^T$ 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[32]). 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[96].

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 *Voxel Selection for Tuning Analyses* for details; selecting voxels based on significance instead of prediction accuracy produces similar results as shown in Supplementary Figure S16). 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[40]. Model weights for each participant and language were projected to a standard template space (fsAverage[32]). Each projected model weight $\beta_{vertex,participant,language} \in R^{300}$ 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 (Supplementary Figure S9). Thus, we chose to use five clusters for the analyses shown in Figure 3 and Figure 5. 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.

Conflict of interest: The authors declare no competing financial interests.

Acknowledgments: This work was supported by the following grant awards: National Science Foundation (NSF Nat-1912373); German Federal Ministry of Education and Research (BMBF 01GQ1906); Berliner ChancengleichheitsProgramm (BCP) to FD; a National Science Foundation GRFP (DGE 1752814) to CC; an IBM PhD Fellowship to CC; and by internal UC Berkeley funds. We thank Tom Dupré la Tour, Anwar Nunez-Elizalde, Matteo Visconti di Oleggio Castello, and Mathis Lamarre for their help in various stages of this project. We thank Cong Du for their help with stimulus generation.

Author contributions: CC, XG, CT, JG and FD conceived and designed the experiments. CT and XG recruited bilingual participants. CC, CT, and XG collected fMRI data. XG generated the stimuli, preprocessed the fMRI data, and performed surface segmentation. CC performed analyses. CC and FD wrote the manuscript. All authors contributed to editing the manuscript. FD and JG directed and supervised all aspects of this project.

References

1. Ansaldo, A. I., Marcotte, K., Scherer, L. & Raboyeau, G. Language therapy and bilingual aphasia: Clinical implications of psycholinguistic and neuroimaging research. *J. Neurolinguistics* **21**, 539–557 (2008).
2. Bot, K. A bilingual production model: Levelt's 'speaking' model adapted. 399–420 (2003).
3. MacNamara, J. The linguistic independence of bilinguals. *Journal of Verbal Learning & Verbal Behavior* **6**, 729–736 (1967).
4. Weinreich, U. *Languages in Contact: Findings and Problems*. (Mouton, 1968).
5. de Groot, A. M. B. Bilingual lexical representation: A closer look at conceptual representations. *Orthography, phonology, morphology, and meaning.* **435**, 389–412 (1992).
6. Kim, K. H. S., Relkin, N. R., Lee, K.-M. & Hirsch, J. Distinct cortical areas associated with native and second languages. *Nature* **388**, 171–174 (1997).
7. Dehaene, S. *et al.* Anatomical variability in the cortical representation of first and second language. *Neuroreport* **8**, 3809–3815 (1997).
8. *Handbook of Bilingualism: Psycholinguistic Approaches*. vol. 588 (Oxford University Press, New York, NY, US, 2005).
9. Xu, M., Baldauf, D., Chang, C. Q., Desimone, R. & Tan, L. H. Distinct distributed patterns of neural activity are associated with two languages in the bilingual brain. *Sci Adv* **3**, e1603309 (2017).
10. Paradis, M. On the representation of two languages in one brain. *Lang. Sci.* **7**, 1–39 (1985).
11. Gomez-Tortosa, E., Martin, E. M., Gaviria, M., Charbel, F. & Ausman, J. I. Selective deficit of one language in a bilingual patient following surgery in the left perisylvian area. *Brain Lang.* **48**, 320–325 (1995).
12. Ku, A., Lachmann, E. A. & Nagler, W. Selective language aphasia from herpes simplex encephalitis. *Pediatr. Neurol.* **15**, 169–171 (1996).
13. Pavlenko, A. Bilingualism and emotions. *Multilingua* **21**, 45–78 (2002).
14. Schrauf, R. W. & Rubin, D. C. 'language' and 'feel' of bilingual memory. *Socioling. Stud.* **5**, 21–39 (2004).
15. Caramazza, A. & Brones, I. Semantic classification by bilinguals. *Can. J. Exp. Psychol.* **34**, 77–81 (1980).
16. Potter, M. C., So, K.-F., Eckardt, B. V. & Feldman, L. B. Lexical and conceptual representation in beginning and proficient bilinguals. *Journal of Verbal Learning and Verbal Behavior* **23**, 23–38 (1984).
17. Kroll, J. F. & Stewart, E. Category Interference in Translation and Picture Naming: Evidence for Asymmetric Connections Between Bilingual Memory Representations. *J. Mem. Lang.* **33**, 149–174 (1994).
18. Abutalebi, J. & Green, D. Bilingual language production: The neurocognition of language representation and control. *J. Neurolinguistics* **20**, 242–275 (2007).
19. Midgley, K. J., Holcomb, P. J., VanHeuven, W. J. B. & Grainger, J. An electrophysiological investigation of cross-language effects of orthographic neighborhood. *Brain Res.* **1246**, 123–135 (2008).
20. Grainger, J., Midgley, K. & Holcomb, P. J. Chapter 14. Re-thinking the bilingual interactive-activation model from a developmental perspective (BIA-d). in *Language Acquisition and Language Disorders* 267–283 (John Benjamins Publishing Company, Amsterdam, 2010).
21. Duyck, W. Translation and associative priming with cross-lingual pseudohomophones: evidence for nonselective phonological activation in bilinguals. *J. Exp. Psychol. Learn. Mem. Cogn.* **31**, 1340–1359 (2005).
22. van Hell, J. G. & Dijkstra, T. Foreign language knowledge can influence native language performance in exclusively native contexts. *Psychon. Bull. Rev.* **9**, 780–789 (2002).
23. Dijkstra, T., Grainger, J. & van Heuven, W. J. B. Recognition of Cognates and Interlingual Homographs: The Neglected Role of Phonology. *J. Mem. Lang.* **41**, 496–518 (1999).
24. Malik-Moraleda, S., Mahowald, K., Conway, B. R. & Gibson, E. Concepts Are Restructured During Language Contact: The Birth of Blue and Other Color Concepts in Tsimane'-Spanish Bilinguals. *Psychol. Sci.* **34**, 1350–1362 (2023).
25. Çukur, T., Nishimoto, S., Huth, A. G. & Gallant, J. L. Attention during natural vision warps semantic representation across the human brain. *Nat. Neurosci.* **16**, 763–770 (2013).

26. Kiremitçi, I. *et al.* Attentional Modulation of Hierarchical Speech Representations in a Multitalker Environment. *Cereb. Cortex* **31**, 4986–5005 (2021).
27. Deniz, F., Tseng, C., Wehbe, L., Dupré la Tour, T. & Gallant, J. L. Semantic Representations during Language Comprehension Are Affected by Context. *J. Neurosci.* **43**, 3144–3158 (2023).
28. Nastase, S. A. *et al.* Attention Selectively Reshapes the Geometry of Distributed Semantic Representation. *Cereb. Cortex* **27**, 4277–4291 (2017).
29. Bojanowski, P., Grave, E. & Joulin, A. Enriching word vectors with subword information. *Transactions of the association for computational linguistics* 135–146 (2017).
30. Joulin, A., Bojanowski, P., Mikolov, T., Jegou, H. & Grave, E. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* 2979–2984 (2018).
31. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* 4171–4186 (Association for Computational Linguistics, 2019).
32. Fischl, B., Sereno, M. I., Tootell, R. B. & Dale, A. M. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum. Brain Mapp.* **8**, 272–284 (1999).
33. Binder, J. R., Desai, R. H., Graves, W. W. & Conant, L. L. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* **19**, 2767–2796 (2009).
34. Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
35. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968–980 (2006).
36. Bressler, D. W. & Silver, M. A. Spatial attention improves reliability of fMRI retinotopic mapping signals in occipital and parietal cortex. *Neuroimage* **53**, 526–533 (2010).
37. Sahani, M. & Linden, J. How Linear are Auditory Cortical Responses? in *Advances in Neural Information Processing Systems* (eds. Becker, S., Thrun, S. & Obermayer, K.) vol. 15 (MIT Press, 2002).
38. Hsu, A., Borst, A. & Theunissen, F. E. Quantifying variability in neural responses and its application for the validation of model predictions. *Network* **15**, 91–109 (2004).
39. Schoppe, O., Harper, N. S., Willmore, B. D. B., King, A. J. & Schnupp, J. W. H. Measuring the Performance of Neural Models. *Front. Comput. Neurosci.* **10**, 10 (2016).
40. Meschke, E. X., Castello, M. V. di O., la Tour, T. D. & Gallant, J. L. Model connectivity: leveraging the power of encoding models to overcome the limitations of functional connectivity. *bioRxiv* (2023) doi:10.1101/2023.07.17.549356.
41. Honey, C. J., Thompson, C. R., Lerner, Y. & Hasson, U. Not Lost in Translation: Neural Responses Shared Across Languages. *J. Neurosci.* **32**, 15277–15283 (2012).
42. Buchweitz, A., Mason, R. A., Hasegawa, M. & Just, M. A. 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 (2009).
43. Klein, D., Milner, B., Zatorre, R. J., Meyer, E. & Evans, A. C. The neural substrates underlying word generation: a bilingual functional-imaging study. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 2899–2903 (1995).
44. Illes, J. *et al.* Convergent cortical representation of semantic processing in bilinguals. *Brain Lang.* **70**, 347–363 (1999).
45. Chee, M. W. *et al.* Processing of visually presented sentences in Mandarin and English studied with fMRI. *Neuron* **23**, 127–137 (1999).
46. Buchweitz, A., Shinkareva, S. V., Mason, R. A., Mitchell, T. M. & Just, M. A. Identifying bilingual semantic neural representations across languages. *Brain Lang.* **120**, 282–289 (2012).
47. Malik-Moraleda, S. *et al.* An investigation across 45 languages and 12 language families reveals a universal language network. *Nat. Neurosci.* **25**, 1014–1019 (2022).
48. Luke, K.-K., Liu, H.-L., Wai, Y.-Y., Wan, Y.-L. & Tan, L. H. Functional anatomy of syntactic and semantic processing in language comprehension. *Hum. Brain Mapp.* **16**, 133–145 (2002).

49. Li, J. *et al.* Le Petit Prince multilingual naturalistic fMRI corpus. *Sci Data* **9**, 530 (2022).
50. Dunagan, D. *et al.* Neural correlates of semantic number: A cross-linguistic investigation. *Brain Lang.* **229**, 105110 (2022).
51. Fedorenko, E. & Kanwisher, N. Neuroimaging of Language: Why Hasn't a Clearer Picture Emerged? *Lang. Linguist. Compass* **3**, 839–865 (2009).
52. Hamilton, L. S. & Huth, A. G. The revolution will not be controlled: natural stimuli in speech neuroscience. *Lang Cogn Neurosci* **35**, 573–582 (2020).
53. Spelke, E. S. & Tsivkin, S. Language and number: a bilingual training study. *Cognition* **78**, 45–88 (2001).
54. Dehaene, S., Spelke, E., Pinel, P., Stanescu, R. & Tsivkin, S. Sources of mathematical thinking: behavioral and brain-imaging evidence. *Science* vol. 284 970–974 (1999).
55. de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L. & Theunissen, F. E. The hierarchical cortical organization of human speech processing. *Journal of Neuroscience* **37**, 6539–6557 (2017).
56. Deniz, F., Nunez-Elizalde, A. O., Huth, A. G. & Gallant, J. L. The Representation of Semantic Information Across Human Cerebral Cortex During Listening Versus Reading Is Invariant to Stimulus Modality. *J. Neurosci.* **39**, 7722–7736 (2019).
57. LeBel, A. *et al.* A natural language fMRI dataset for voxelwise encoding models. *Sci Data* **10**, 555 (2023).
58. Forster, K. I. Visual perception of rapidly presented word sequences of varying complexity. *Percept. Psychophys.* **8**, 215–221 (1970).
59. Gao, J. S. fMRI visualization and methods. (UC Berkeley, 2015).
60. Power, J. D. *et al.* Customized head molds reduce motion during resting state fMRI scans. *Neuroimage* **189**, 141–149 (2019).
61. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825–841 (2002).
62. Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage* **37**, 90–101 (2007).
63. Gao, J. S., Huth, A. G., Lescroart, M. D. & Gallant, J. L. Pycortex: an interactive surface visualizer for fMRI. *Front. Neuroinform.* **9**, 23 (2015).
64. Community, B. O. Blender - a 3D modelling and rendering package. Preprint at http://www.blender.org (2018).
65. Spiridon, M., Fischl, B. & Kanwisher, N. Location and spatial profile of category-specific regions in human extrastriate cortex. *Hum. Brain Mapp.* **27**, 77–89 (2006).
66. Hansen, K. A., Kay, K. N. & Gallant, J. L. Topographic Organization in and near Human Visual Area V4. *J. Neurosci.* **27**, 11896–11911 (2007).
67. Kaushanskaya, M., Blumenfeld, H. K. & Marian, V. The Language Experience and Proficiency Questionnaire (LEAP-Q): Ten years later. *Biling.* **23**, 945–950 (2020).
68. Li, P., Zhang, F., Yu, A. & Zhao, X. Language History Questionnaire (LHQ3): An enhanced tool for assessing multilingual experience. *Bilingualism: Language and Cognition* **23**, 938–944 (2020).
69. Oldfield, R. C. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* **9**, 97–113 (1971).
70. Wu, M. C.-K., David, S. V. & Gallant, J. L. Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.* **29**, 477–505 (2006).
71. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *Neuroimage* **56**, 400–410 (2011).
72. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
73. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
74. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
75. Dupré la Tour, T., Eickenberg, M., Nunez-Elizalde, A. O. & Gallant, J. L. Feature-space selection with banded ridge regression. *Neuroimage* **264**, 119728 (2022).
76. Søgaard, A., Ruder, S. & Vulić, I. On the Limitations of Unsupervised Bilingual Dictionary Induction. in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers) 778–788 (Association for Computational Linguistics, 2018).
77. Pires, T., Schlinger, E. & Garrette, D. How Multilingual is Multilingual BERT? in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* vol. abs/1906.01502 4996–5001 (Association for Computational Linguistics, 2019).
78. Toneva, M. & Wehbe, L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Adv. Neural Inf. Process. Syst.* (2019).
79. Schrimpf, M. *et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* **118**, e2105646118 (2021).
80. Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Communications Biology* **5**, 1–10 (2022).
81. Lamarre, M., Chen, C. & Deniz, F. Attention weights accurately predict language representations in the brain. in *Findings of the Association for Computational Linguistics: EMNLP 2022* (eds. Goldberg, Y., Kozareva, Z. & Zhang, Y.) 4513–4529 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022).
82. Chen, C., Dupré la Tour, T., Gallant, J. L., Klein, D. & Deniz, F. The cortical representation of language timescales is shared between reading and listening. *Communications Biology* **7**, 1–13 (2024).
83. Adelson, E. H. & Bergen, J. R. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* **2**, 284–299 (1985).
84. Watson, A. B. & Ahumada, A. J., Jr. Model of human visual-motion sensing. *J. Opt. Soc. Am. A* **2**, 322–341 (1985).
85. Nishimoto, S. & Gallant, J. L. A Three-Dimensional Spatiotemporal Receptive Field Model Explains Responses of Area MT Neurons to Naturalistic Movies. *J. Neurosci.* **31**, 14551–14564 (2011).
86. Gong, X. L. *et al.* Phonemic segmentation of narrative speech in human cerebral cortex. *Nat. Commun.* **14**, 4309 (2023).
87. Gong, X. L. Language representation in human cerebral cortex. (UC Berkeley, 2024).
88. Nakai, T., Yamaguchi, H. Q. & Nishimoto, S. Convergence of Modality Invariance and Attention Selectivity in the Cortical Semantic Circuit. *Cereb. Cortex* **31**, 4825–4839 (2021).
89. Nunez-Elizalde, A. O., Huth, A. G. & Gallant, J. L. Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage* **197**, 482–492 (2019).
90. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).
91. Jain, S. *et al.* Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. in *Advances in Neural Information Processing Systems* 13738–13749 (2020).
92. Reddy, A. J. & Wehbe, L. Can fMRI reveal the representation of syntactic structure in the brain? in 9843–9856 (Advances in Neural Information Processing Systems, 2021).
93. Oota, S. R., Marreddy, M., Gupta, M. & Bapi, R. How does the brain process syntactic structure while listening? in *Findings of the Association for Computational Linguistics: ACL 2023* (eds. Rogers, A., Boyd-Graber, J. & Okazaki, N.) 6624–6647 (Association for Computational Linguistics, Toronto, Canada, 2023).
94. Tang, J., Du, M., Vo, V. A., Lal, V. & Huth, A. G. Brain encoding models based on multimodal transformers can transfer across language and vision. in *Proceedings of the 37th International Conference on Neural Information Processing Systems* 29654–29666 (Curran Associates Inc., Red Hook, NY, USA, 2024).
95. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
96. Popham, S. F. *et al.* Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nat. Neurosci.* **24**, 1628–1636 (2021).
97. Liu, Y. & Sun, M. Contrastive Unsupervised Word Alignment with Non-Local Features. *AAAI* **29**, (2015).