JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality

https://doi.org/10.1523/JNEUROSCI.0675-19.2019

Cite as: J. Neurosci 2019; 10.1523/JNEUROSCI.0675-19.2019

Received: 25 March 2019
Revised: 15 July 2019
Accepted: 17 July 2019

The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality

Fatma Deniz[a,b,c], Anwar O. Nunez-Elizalde[a], Alexander G. Huth[a,1], Jack L. Gallant[a,d]

[a] Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720, USA
[b] Berkeley Institute for Data Science, University of California, Berkeley, CA 94720, USA
[c] International Computer Science Institute, Berkeley, CA 94704, USA
[d] Department of Psychology, University of California, Berkeley, CA 94720, USA

[1] Present address: Departments of Computer Science and Neuroscience, University of Texas, Austin, TX 78712, USA

Correspondence should be addressed to:
Jack L. Gallant
2121 Berkeley Way
University of California, Berkeley
Berkeley, CA 94720-1650

E-mail: <gallant@berkeley.edu>.

Running title: *Semantic representations are invariant to modality*

Manuscript summary:
Pages: 23
Words (inc. figure legends, but excluding methods and references): 5,669
Characters (inc. spaces, figure legends, but excluding methods and references): 38,590
Abstract: 167 words, 1166 characters
Figures: 9
References: 79 references, 2,067 words, 13,812 characters

This work was supported by grants the National Science Foundation (IIS1208203), the National Eye Institute (EY019684 and EY022454), the Intelligence Advanced Research Projects Agency (86155-Carnegi-1990360-gallant), and from the Center for Science of Information, an NSF Science and Technology Center (CCF-0939370). FD was supported by the FITweltweit-Program of the German Academic Exchange Service (DAAD) and the Moore and Sloan Data Science Environment Postdoctoral Fellowship. AGH was also supported by the William Orr Dingwall Neurolinguistics Fellowship. We thank Dr. Leila Wehbe and Dr. Mark Lescroart for useful technical discussions, Jasmine Nguyen and Priyanka N. Bhoj for assistance transcribing and aligning stimuli, and Brittany Griffin and Marie-Luise Kieseler for segmenting and flattening cortical surfaces. The authors declare no conflict of interest.

ABSTRACT

An integral part of human language is the capacity to extract meaning from spoken and written words, but the precise relationship between brain representations of information perceived by listening versus reading is unclear. Prior neuroimaging studies have shown that semantic information in spoken language is represented in multiple regions in the human cerebral cortex, while amodal semantic information appears to be represented in a few broad brain regions. However, previous studies were too insensitive to determine whether semantic representations were shared at a fine level of detail rather than merely at a coarse scale. We used fMRI to record brain activity in two separate experiments while participants listened to or read several hours of the same narrative stories, and then created voxelwise encoding models to characterize semantic selectivity in each voxel and in each individual participant. We find that semantic tuning during listening and reading are highly correlated in most semantically-selective regions of cortex, and models estimated using one modality accurately predict voxel responses in the other modality. These results suggest that the representation of language semantics is independent of the sensory modality through which the semantic information is received.

Keywords

fMRI; BOLD; listening; reading; semantics; regression; representation; cross-modal representation

SIGNIFICANCE STATEMENT

Humans can comprehend the meaning of words from both spoken and written language. It is therefore important to understand the relationship between the brain representations of spoken or written text. Here we show that although the representation of semantic information in the human brain is quite complex, the semantic representations evoked by listening versus reading are almost identical. These results suggest that the representation of language semantics is independent of the sensory modality through which the semantic information is received.

INTRODUCTION

Humans have the unique capacity to communicate and extract meaning through both spoken and written language. Although the early sensory processing pathways for listening and reading are distinct, listeners and readers appear to extract very similar information about the meaning of a narrative story (Diakidoy et al., 2005; Rubin et al., 2000). This suggests that the human brain represents semantic information in an amodal form that is independent of input modality (for reviews see (Binder et al., 2009; Price, 2010, 2012; Vigneau et al., 2006)). There is evidence that several cortical regions are activated during both listening and reading (for reviews see (Price, 2010, 2012)). However, the demonstration of some common activation during listening and reading is necessary but not sufficient evidence of a common amodal semantic representation.

A direct and convincing way to determine if listening and reading involve a common underlying semantic representation would be to compare directly the semantic selectivity maps obtained during listening and reading of natural text, in single participants. However, to date no study has performed this crucial comparison. Most imaging studies of the semantic system have examined only one input modality, either spoken or written words (Booth et al., 2002; Démonet et al., 1994, 1992; Devlin et al., 2004; Nakamura et al., 2005; Rissman et al., 2003; Scott et al., 2000; Vandenberghe et al., 1996). Relatively few have studied cross-modal representations by presenting the same stimuli in both modalities (Buchweitz et al., 2009; Chee et al., 1999; Jobard et al., 2007; Liuzzi et al., 2017; Michael et al., 2001; Petersen et al., 1989; Spitsyna et al., 2006). Most of these cross-modality studies observed activity in left lateralized regions such as the left anterior temporal lobe, left superior temporal sulcus (STS), left middle temporal gyrus (MTG) and left inferior frontal gyrus (IFG). Most of these studies used tightly controlled stimuli, such as a set of single isolated words, sentences or curated passages, and an explicit lexical semantic task (Buchweitz et al., 2009; Chee et al., 1999; Liuzzi et al., 2017; Michael et al., 2001). A study that used narrative speech in a listening and reading task demonstrated amodal brain activity in left pSTG, left IFG, bilateral Precuneus, medial prefrontal cortex, and angular gyrus (Regev et al., 2013). However that study did not model semantic information, but only showed that voxel activations in these regions tend to be correlated across these two modalities. Furthermore, previous studies were too coarse grained to determine whether listening and reading shared semantic representations at the level of a single voxel. For example, the semantic representation of listening and reading might have been modal at a fine scale (i.e. single voxel), although amodal at a coarse scale. In sum, the evidence available currently is insufficient to determine whether semantic information obtained during listening and reading are represented in the same way.

To address this issue we used functional magnetic resonance imaging (fMRI) to record blood-oxygen-level dependent (BOLD) activity in human participants while they listened to and read the same narrative stories. We then used voxelwise modeling (VM) combined with banded ridge regression (Nunez-Elizalde et al., 2019) to characterize the semantic selectivity of each voxel in each presentation modality and for each individual participant (see **Materials and Methods** and (Çukur et al., 2013; Huth et al., 2012, 2016; Lescroart et al., 2015; Nishimoto et al., 2011; Stansbury et al., 2013)). Finally, we compared the semantic tuning of each voxel in the two modalities by creating semantic maps (Huth et al., 2016) for both modalities and each individual participant. In addition, we identified modality independent cortical representation of semantic information by predicting voxel responses cross-modally. Comparison of the fit semantic models and semantic maps obtained by listening versus reading provides a sensitive and objective means to determine whether and how semantic selectivity changes depending on the modality with which semantic information is perceived.

MATERIALS and METHODS

Participants

Functional data were collected from six male participants and three female participants: S1 (male, age 31), S2 (male, age 31), S3 (female, age 28), S4 (female, age 25), S5 (male, age 30), S6 (male, age 25), and S7 (male, age 36), S8 (female, age 24), S9 (male, age 24). Two of the participants were authors on the paper (AGH and AONE). All participants listened to and read all the stories. Listening and reading presentations were counterbalanced across participants. All participants were healthy and had normal hearing, and normal or corrected-to-normal vision. One participant was left handed, all other participants were right handed or ambidextrous according to the Edinburgh handedness inventory (Oldfield, 1971) (laterality quotient of -100: entirely left-handed, +100: entirely right-handed). Laterality scores were +90 (decile R.7), +70 (decile R.3), +10 (ambidextrous), +80 (decile R.5), +80 (decile R.5), +80 (decile R.5), -60 (decile L.3), +90 (decile R.7) and +95 (decile R.9) for S1-9, respectively. To stabilize head motion during scanning sessions participants wore a personalized headcase that precisely fit the shape of each participant's head (https://caseforge.co/).

Natural Speech Stimuli

The speech stimuli consisted of ten 10- to 15-minute stories taken from *The Moth Radio Hour* and used previously (Huth et al., 2016). In each story, a speaker tells an autobiographical story in front of a live audience. The ten selected stories cover a wide range of topics and are highly engaging. The model validation dataset consisted of one 10-minute story. This story was played twice for each participant (once during each scanning session), and then the two responses were averaged (see Huth et al. (2016) for more details).

Speech stimuli were played over Sensimetrics S14 in-ear piezoelectric headphones (Sensimetrics, Malden, MA, USA). A Behringer Ultra-Curve Pro hardware parametric equalizer was used to flatten the frequency response of the headphones based on calibration data provided by Sensimetrics. All stimuli were played at 44.1 kHz using the pygame library in Python. All stimuli were normalized to have peak loudness of -1 dB relative to max. However, the stories were performed by different speakers and were not uniformly mastered, so some differences in total loudness remain.

Story transcription and preprocessing

Each story was manually transcribed by one listener, and this transcription was checked by a second listener. Certain sounds (e.g. laughter, lip-smacking and breathing) were also marked in order to improve the accuracy of the automated alignment. The audio of each story was downsampled to 11.5 kHz and the Penn Phonetics Lab Forced Aligner (P2FA; Yuan & Liberman, 2008) was used to automatically align the audio to the transcript. The forced aligner uses a phonetic hidden Markov model to find the temporal onset and offset of each word and phoneme. The Carnegie Mellon University (CMU) pronouncing dictionary was used to guess the pronunciation of each word. The Arpabet phonetic notation was used when necessary to manually add words and word fragments that appeared in the transcript but not in the dictionary.

After automatic alignment was complete, Praat (Boersma and Weenink, 2014) was used to check and correct each aligned transcript manually. The corrected aligned transcript was then spot-checked for accuracy by a different listener.

Finally the aligned transcripts were converted into separate word and phoneme representations using Praat's TextGrid object. The phoneme representation of each story is a list of pairs (P, t), where P is a phoneme and t is the onset time in seconds. Similarly the word representation of each story is a list of pairs (W, t), where W is a word and t is the onset time in seconds.

Natural Reading Stimuli

The same stories from listening sessions were used for reading sessions. Praat's word representation for each story (W, t) was used for generating the reading stimuli. 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 et al., 2009; Forster, 1970). During reading, each word was presented for a duration precisely equal to the duration of that word in the spoken story. RSVP reading is different than natural reading because during RSVP the reader has no control over which word to read at each point in time. Therefore, in order to make listening and reading more comparable we matched the timing of the words presented during RSVP to the rate at which the words occurred during listening.

The pygame library in Python was used to display text on a gray background at 34 horizontal, and 27 vertical degrees of visual angle. Black letters were presented at average 6 (min=1, max=16) horizontal and 3 vertical degrees of visual angle. A white fixation cross was present at the center of the display. Participants were asked to fixate while reading the text. These data were collected during two 3-hour scanning sessions that were performed on different days. Participants' eye movement were monitored at 60 Hz throughout the scanning sessions using a custom-built camera system equipped with an infrared source (Avotec) and the ViewPoint EyeTracker software suite (Arrington Research). The eye tracker was calibrated before the first run of data acquisition. Certain auditory sounds (laughter and applause) were presented as text to provide cues about the ambiance of each story.

Semantic model construction

To account for response variance caused by the semantic content of the story stimuli a 985-parameter semantic feature space based on word co-occurrence statistics in a large corpus of text (Deerwester et al., 1990; Huth et al., 2016; Lund and Burgess, 1996; Mitchell et al., 2008) was used. In short, a word co-occurrence matrix, M, with 985 rows and 10,470 columns was created. The 985 rows describe 985 basic words from Wikipedia's *List of 1000 basic words,* the 10,470 columns are words selected from a very large corpora of 13 transcripts of Moth stories (including the 10 used as stimuli in the experiments described in this paper), 604 popular books available through Project Guttenberg, 2,405,569 Wikipedia pages, 36,333,459 reddit.com user comments (see Huth et al., 2016 for a detailed description).

Iterating through the text corpus, we added 1 to $M_{i,j}$ each time word j appeared within 15 words of basis word i. Once the word co-occurrence matrix was complete, we log-transformed the counts, replacing $M_{i,j}$ with $\log(1 + M_{i,j})$. Next, each row of M was z-scored to correct for differences in basis word frequency, and then each column of M was z-scored to correct for word frequency. Each column of M is now a 985-dimensional semantic vector representing one word in the lexicon.

The semantic model stimulus matrix was then constructed from the stories: for each word-time pair (w, t), within each story the corresponding column of M was selected, creating a new list of semantic vector-time pairs, (M_w, t). These unevenly-sampled lists of vectors were resampled at times corresponding to the fMRI acquisitions using a 3-lobe Lanczos filter with the cutoff frequency set to the Nyquist frequency of the fMRI acquisition (0.249 Hz).

Motion-energy model construction

A spatiotemporal Gabor pyramid was used to extract low-level visual features from the sequence of word frames used in the reading experiment (Adelson and Bergen, 1985; Watson and Ahumada, 1985). The word frames were first cropped to 400x400 pixels (14 horizontal, and 14 vertical degrees of visual angle) to include mainly the words and then downsampled to 96x96 pixels to minimize computational cost. The word frames were then converted to the CIE L*A*B* color space (McLaren, 1976) and the color information was discarded. The spatiotemporal Gabor pyramid consisted of a total of 39 three-dimensional Gabor filter pairs of orthogonal quadrature spanning a square grid that covered the screen. The filters consisted of two spatial and one temporal dimension and were created using five spatial frequencies (0, 2, 4, 6, and 8 cycles/image), three temporal frequencies (0, 2 and 4 Hz), and four directions of motion (0, 90, 180, and 270 degrees). Each of the filters was convolved with the sequence of word frames. The resulting filter activations were squared and summed for each quadrature pair, resulting in a 39-dimensional feature vector for each word frame. The output was downsampled to the functional image acquisition rate (2.0045 s) using sinc interpolation (Oliphant, 2007). See (Nishimoto et al., 2011) for more details. However, note that only five spatial frequencies and four directions of motion were used here.

Spectral model construction

A cochleogram model that accounts for the logarithmic filtering of the mammalian coclea described in (de Heer et al., 2017) was used to create the low level auditory features (80 parameters). This model was selected based on an earlier study showing that it outperforms other low level acoustical models (de Heer et al., 2017). The 80 waveforms of the coclear filter bank were between 264 and 7360 Hz, spaced at 25% of the bandwidth. The spectral features were downsampled to the rate of acquisition of the functional images (2.0045 s) using a Lanczos filter.

Syntax model construction

The syntactic properties of each spoken word were labeled. A pre-trained neural network was used to create a parse tree for each sentence of the stories (Andor et al., 2016). Two feature spaces were extracted from the parse trees. The first was constructed from the part-of-speech tags (e.g. noun, verb) by assigning a value of one to each entry in which the part-of-speech tag appeared and all other entries were set to zero (12 parameters). The second feature space captured the word dependencies in the sentence (i.e. direct object, indirect object, etc.) and was constructed by assigning a value of one to each entry in which the word dependency appeared and all other entries were set to zero (44 parameters). For each syntactic feature (e.g. noun), a time course was created with a value of 1 whenever a word was labeled with that feature and 0 otherwise. The syntactic features were then downsampled to the rate of acquisition of the functional images (2.0045 s) using a Lanczos filter.

Phoneme model construction

To account for response variance caused by the low-level phonemic content of the stories, a 39-parameter model that captures how often each of the 39 phonemes in English was spoken over time was constructed. The phoneme representation of the stories were used to construct this model: the lists of phoneme-time pairs (P, t) were re-arranged into 39 lists, each of which contains only the times of a single phoneme. These lists of times were then downsampled to the fMRI acquisition rate (2.0045 s).

Letter model construction

To account for response variance caused by the letters during reading a 26-parameter model that captures how often each of the 26 letters in English was present on screen over time was constructed. This was constructed by counting the number of times a letter was present within a word and then downsampled to the fMRI acquisition rate (2.0045 s).

Word rate, word length variation, phoneme rate, letter rate, and pauses model construction

To account for the highly variable speech rate both within and across stories, single-feature models that simply count the number of words, number of phonemes, number of letters, and number of story speaker's pauses that occurred during the acquisition of each fMRI volume (2.0045 s) were constructed. To account for the variable word lengths during the visual presentation a single-feature word length variation model was constructed by taking the variance of word lengths that occurred during the acquisition of each fMRI volume.

Stimulus downsampling

Before downsampling to the fMRI acquisition rate, the phoneme and semantic models were represented as unevenly-sampled impulse trains. A 3-lobe Lanczos filter with cutoff frequency set to the fMRI Nyquist rate (0.249 Hz) was used to resample these impulse trains at evenly spaced time points corresponding to the middle of each fMRI volume.

Experimental Design and Statistical Analysis

fMRI data acquisition

Each spoken and written story was presented during a separate fMRI scan. The length of each scan was the same as the story. Each scan included 10 seconds (5 TR) of silence both before and after the story. These data were collected during two 3-hour scanning sessions that were performed on different days.

MRI data were collected on a 3T Siemens TIM Trio scanner at the UC Berkeley Brain Imaging Center using a 32-channel Siemens volume coil. Functional scans were collected using gradient echo EPI water excitation pulse sequence with repetition time (TR) = 2.0045s, echo time (TE) = 31ms, flip angle = 70 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. 30 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 avoid signal from fat. Anatomical data were collected using a T1-weighted multi-echo MP-RAGE sequence on the same 3T scanner.

fMRI data pre-processing

Each functional run was motion-corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FSL 5.0 (Jenkinson and Smith, 2001; Jenkinson et al., 2002). All volumes in the run were then averaged across time to obtain a high quality template volume. FLIRT was also 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 temporal average of the cross-modal runs (listening or reading) were also automatically aligned to the same overall template. 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.

Low-frequency voxel response drift was identified using a 3rd order Savitsky-Golay filter with a 120-second window. This drift was subtracted from the signal. Responses of each story were z-scored separately, i.e., the mean response for each voxel was subtracted and the remaining response was scaled to have unit variance. Prior to the voxelwise modeling, 10 TRs from the beginning and 10 TRs at the end of each story were discarded.

Cortical surface reconstruction and visualization

Cortical surface meshes were generated from the T1-weighted anatomical scans using Freesurfer software (Dale et al., 1999). Before surface reconstruction, anatomical surface segmentations were carefully hand-checked and corrected using Blender software and pycortex (Gao et al., 2015) (http://pycortex.org). 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 64 evenly-spaced intervals between the inner (white matter) and outer (pial) surfaces of the cortex, then averages together the samples. Samples are taken using nearest-neighbor interpolation, wherein each sample is given the value of its enclosing voxel.

Localizers for known ROIs

Known regions of interest (ROIs) were localized separately in each participant using standard techniques (Hansen et al., 2007; Spiridon et al., 2006). For all participants ROIs were defined using three experiments: a visual category localizer, an auditory cortex localizer, and a motor localizer. For some participants retinotopic visual ROIs using a retinotopic localizer and area MT+ using an MT localizer were defined.

Visual category localizer. Visual category localizer data were collected in six 4.5-minute scans consisting of 16 blocks, each 16 seconds long. During each block, 20 images of either places, faces, human body parts, non-human animals, household objects, or spatially scrambled household objects were displayed. Each image was displayed for 300 ms followed by a 500 ms blank. Occasionally the same image was displayed twice in a row, in which case the participant was asked to respond with a button press.

The contrast between faces and objects was used to define the fusiform face area (FFA) (Kanwisher et al., 1997) and occipital face area (OFA) (Halgren et al., 1999). The contrast between human body parts and objects was used to define the extrastriate body area (EBA) (Downing et al., 2001). The contrast between places and objects was used to define the parahippocampal place area (PPA) (Epstein and Kanwisher, 1998), occipital place area (OPA) (Nakamura et al., 2000), and retrosplenial cortex (RSC).

Auditory cortex localizer. Auditory cortex localizer data were collected in one 10 minute scan. The participant listened to 10 repeats of a 1-minute auditory stimulus, which consisted of 20-second segments of music (Arcade Fire), speech (Ira Glass), and natural sound (a babbling brook). To determine whether a voxel was responsive to auditory stimuli, the repeatability of the voxel response across the 10 stimulus repeats was calculated using an F-statistic. The F-statistic map was used to define the auditory cortex (AC).

Motor localizer. Motor localizer data were collected during one 10-minute scan. The participant was cued to perform six different motor tasks in a random order in 20-second blocks. For the hand, mouth, foot, speech, and rest blocks the stimulus was simply a word at the center of the screen (e.g. "Hand"). For the saccade block the participant was shown a pattern of saccade targets.

For the “Hand” cue the participant was instructed to make small finger-drumming movements with both hands for as long as the cue remained on the screen. Similarly for the “Foot” cue the participant was instructed to make small toe movements for the duration of the cue. For the “Mouth” cue the participant was instructed to make small mouth movements approximating the nonsense syllables *balabalabala* for the duration of the cue—this requires movement of the lips, tongue, and jaw. For the “Speak” cue the participant was instructed to continuously subvocalize self-generated sentences for the duration of the cue. For the saccade condition the written cue was replaced with a fixed pattern of twelve saccade targets, and the participant was instructed to make frequent saccades between the targets. A linear model was used to find the change in BOLD response of each voxel in each condition relative to the mean BOLD response.

Weight maps for the foot, hand, and mouth responses were used to define primary motor and somatosensory areas for the feet (M1F, S1F), hands (M1H, S1H), and mouth (M1M, S1M); supplementary motor areas for the feet (SMFA) and hands (SMHA); secondary somatosensory area for the feet (S2F) and, in some participants, the hands (S2H); and, in some participants, the ventral premotor hand area (PMVH) (Penfield and Boldrey, 1937). The weight map for saccade responses was used to define the frontal eye field (FEF) (Paus, 1996), frontal operculum eye movement area (FO) (Corbetta et al., 1998), intraparietal sulcus visual areas (IPS), and, in some participants, the supplementary eye field (SEF) (Grosbras et al.). The weight map for speech production responses was used to define Broca's area (BA) (Amunts et al., 2010; Zilles and Amunts, 2018) and the superior ventral premotor speech area (sPMv).

Retinotopic localizer. Retinotopic mapping data were collected in four 9-minute scans. Two scans used clockwise and counterclockwise rotating polar wedges, and two used expanding and contracting rings. Visual angle and eccentricity maps were used to define visual areas V1, V2, V3, V4, LO, V3A, V3B, and V7 (Hansen et al., 2007).

Area MT+ localizer. Area MT+ localizer data were collected in four 90-second scans consisting of alternating 16-second blocks of continuous and temporally scrambled natural movies. The contrast between continuous and temporally scrambled natural movies was used to define visual motion area MT+ (Tootell et al., 1995).

Voxelwise model fitting

A single joint model that included all feature spaces was estimated for each voxel in each dataset (listening and reading) separately using banded ridge regression (see below for details and (Nunez-Elizalde et al., 2019)).

Banded ridge regression assigns a different regularization parameter for every feature space and so reduces bias caused by correlations between feature spaces.

Feature spaces

The feature spaces were motion-energy features (39 parameters), spectral features (80 parameters), word rate (1 parameter), phoneme rate (1 parameter), phonemes (39 parameters), letter rate (1 parameter), letters (26 parameters), word length variation per repetition time (1 parameter), syntactic features (56 parameters), and co-occurence semantics (985 parameters). The motion-energy, spectral, word rate, phoneme rate, phonemes, letter rate, letters, and word length variation features were used to explain away low-level parameters that might otherwise contaminate the semantic model weights.

Before doing regression, each feature channel was z-scored within each story (training and testing features were z-scored independently) by subtracting the mean and dividing by the standard deviation. This was done to match the features to the fMRI responses, which were also z-scored within each story. In addition, 10 TRs from the beginning and 10 TRs at the end of each story were discarded prior to voxelwise modeling.

Banded ridge regression

We combine several feature spaces in the voxelwise modeling approach. In order to assign different levels of regularization to each feature space, we estimate all our models simultaneously using banded ridge regression (Nunez-Elizalde et al., 2019). Under banded ridge regression, brain responses are modeled as a linear combination of all the feature spaces. However, each feature space is assigned a different value of the regularization parameter. Banded ridge regression is a special case of the well-established statistical approach called Tikhonov regression (Tikhonov and Arsenin, 1977). The solution to the Tikhonov regression problem is given by $\hat{\beta} = argmin_{\beta}[\|Y - X\beta\|_2^2 + \|\lambda C\beta\|_2^2]$, where C is the penalty matrix. In case of banded ridge regression, the matrix C is a diagonal matrix whose entries correspond to the regularization levels appropriate for each feature space. To find the optimal regularization parameter for every feature space a wide range of regularization parameters is explored using cross-validation. The regularization parameter is optimized based on prediction accuracy on a held-out data set. Note that in case of $\lambda = 0$ Tikhonov regression reduces to the ordinary least squares and in case of $C = I$ Tikhonov regression reduces to ridge regression.

BOLD responses were modeled as a linear combination of all the feature spaces using linear regression with a non-spherical spatiotemporal multivariate normal prior on the weights (Nunez-Elizalde et al., 2019). This approach allows us to impose different levels of regularization on each feature space within the joint model for each voxel, which is important because of differences in feature space size and signal-to-noise levels. The regularization parameter for each feature space was estimated empirically via cross-validation on a held-out set.

Within the same model, the hemodynamic response function was modeled using a finite impulse response (FIR) filter per voxel and for each subject and modality (listening and reading) separately. This was implemented by modeling the BOLD responses at ten temporal delays corresponding to 0, 2, 4, 6, ..., 16 and 18 seconds. We also imposed a multivariate normal prior on the temporal covariance of the FIR filter. The temporal prior was constructed from a set of HRF basis functions (Penny et al., 2007).

Cross-validation

We used cross-validation to find the optimal regularization parameter for each feature space in the joint model. Because evaluating k regularization parameters for m models leads to k^m combinations, conducting a grid-search in our high-dimensional parameter space is impractical (requiring 10^{10} model fits). To overcome this problem, we used a tree-structured Parzen search (Bergstra et al., 2011). We performed the search 25 times each time using different initialization values and stopped each search after 300 iterations. For every set of regularization parameters tested in each iteration, we performed 5-fold cross-validation twice. We used the coefficient of determination (R^2) between the predicted and the actual voxel responses as our performance metric for each validation fold.

Model estimation and evaluation

We computed the mean prediction performance across cross-validation folds per voxel for each of the 7500 (300 x 25) regularization parameter sets tested. The regularization parameters that yielded the maximum cross-validated prediction performance were selected for each voxel. These regularization parameters were then used to estimate the model weights for each of the voxels in each modality independently for each of the nine subjects.

The estimated model weights were then used to predict the voxel responses to the validation story. Model prediction performance was computed per voxel as the Pearson correlation coefficient between predicted and actual responses.

To validate the voxelwise models, estimated model weights were used to predict responses to a validation story that was not used for model estimation. Only the estimated semantic model weights were used for model predictions. Pearson's correlation coefficient was computed between the predicted responses and the mean of the two validation datasets (291 time points).

Statistical significance was computed by a permutation test with 10,000 iterations and comparing estimated correlations to the empirical null distribution of correlations for each participant and modality separately. At each permutation iteration, the time course of the held-out validation dataset was permuted by blockwise shuffling (10 TRs were blocked to account for autocorrelations in voxel responses), and then Pearson's correlation coefficient between the permuted voxel response and the predicted voxel response was computed for each voxel separately. This produced a distribution of 10,000 estimates of correlation coefficients for each voxel, participant, and modality. These 10,000 estimates define an empirical distribution that was used to obtain a p-value. Resulting p-values were corrected for multiple comparisons within each participant using the false discovery rate (FDR) procedure (Benjamini and Hochberg, 1995).

Voxelwise model fitting and analysis was performed using custom software tikreg (Nunez-Elizalde et al., 2019) written in Python, making heavy use of NumPy (Oliphant, 2006) and SciPy (Oliphant, 2007). Analysis and visualizations were developed using iPython (Perez and Granger, 2007) and the interactive programming and visualization environment jupyter notebook (Kluyver et al., 2016).

Semantic PC projections

Listening model weights and reading model weights were projected onto the semantic subspace that was created in a previous study from our laboratory (Huth et al., 2016). That study recovered a low-dimensional semantic subspace from an aggregated set of estimated semantic model weights using principal components analysis. Taking the dot product of the estimated model weights with the low-dimensional semantic subspace revealed for each voxel a projection along the 985 semantic principal components (PCs). To visualize which semantic concepts are represented in each voxel we used an RGB color space to map the first three semantic PC projections onto the cortical surface separately for the two modalities (Huth et al., 2012, 2016).

Correlating the semantic principal components

Pearson’s correlation coefficient was computed between each semantic projection in listening and the corresponding semantic projection in reading. To find out whether the semantic projections could be correlated by chance, a permutation test with 10,000 iterations was performed for each individual participant separately. The correlation was computed for the 10,000 best predicted voxels by the co-occurrence semantics model in both modalities. The best predicted voxels were selected by taking an average of listening and reading model prediction accuracies per voxel and selecting the 10,000 voxels with highest mean predictions. At each permutation iteration (i) the time courses of the feature matrix was permuted (note that the feature matrix is the same for listening and reading sessions), (ii) banded ridge regression was performed between the fMRI responses and this permuted matrix, (iii) the estimated model weights were projected onto the semantic principal component space, (iv) Pearson's correlation coefficient between projections of the listening and reading weights onto the semantic subspace were computed separately for each PC. This results in a distribution of 10,000 estimates of correlation coefficients for each semantic PC and participant. Statistical significance was defined as any correlation coefficient that exceeded 95% of all of the permuted correlations.

Cross-modality voxelwise model fitting

Estimated model weights (see *Voxelwise model fitting*) from one modality (e.g. listening) were used to predict voxel responses in the other modality (e.g. reading). Model prediction accuracy was then computed using Pearson's correlation coefficient between cross-modal prediction responses (e.g. listening model estimates predicting reading responses) and the mean of the two validation responses (e.g. reading responses).

RESULTS

We sought to determine whether and how the cortical representation of semantic information in narrative language might depend on the modality with which it is perceived. Nine participants listened to and read narrative stories while whole-brain BOLD activity were recorded by means of functional MRI (see **Figure 1**). The experimental stimuli consisted of more than two hours of narrative stories from *The Moth Radio Hour*, along with written transcriptions of the same stories. In the reading condition we used a rapid serial visual presentation (RSVP) method (Buchweitz et al., 2009; Forster, 1970) to present the stories at precisely the same rate as they occurred during listening. That is, in the reading condition each word was presented serially at exactly the same time, and for exactly the same duration as when it was spoken. The semantic content of the stories was estimated continuously by projecting the narrative into a word embedding space based on word co-occurrence statistics (Church and Hanks, 1990; Lund and Burgess, 1996; Mitchell et al., 2008; Turney and Pantel, 2010; Wehbe et al., 2014; Xu et al. 2016). We then used voxelwise modeling to estimate a set of weights for each voxel that best characterize the relationship between the semantic features and the recorded BOLD signals separately for each modality. These estimated model weights were then used to predict voxel responses in a held-out validation dataset both within and across modalities. Finally, the semantic tuning of each voxel in the two modalities was compared by projecting the estimated model weights onto the semantic space described in Huth et al. (2016).

Does the cortical distribution of semantically-selective voxels depend on stimulus modality?

We used a voxelwise modeling procedure to determine whether the broad distribution of semantically-selective voxels depends on presentation modality. Semantic features were extracted from the stories and these were used to estimate voxelwise model weights for BOLD signals that were recorded while participants listened to the stories in the training set. These estimated model weights were then used to predict fMRI voxel responses to a separate held-out validation set. We repeated the same procedure for the reading sessions. Several low-level features (low-level visual, spectral, word rate, letter rate, word length variation, phonemes, phoneme rate, and pauses) and syntactic features were included alongside the semantic features as nuisance regressors (see **Materials and Methods**), but these nuisance regressors were discarded after regression and the final model predictions were based only on semantic model weight estimates. The correlation coefficient between the actual responses in the held-out validation dataset and predicted responses were computed to give a measure of model prediction accuracy. These were then mapped onto the cortical surface.

Figure 2 shows voxelwise model prediction accuracy for listening and reading for all voxels in one participant (p<0.05, FDR corrected). **Figure 2a** shows that our semantic model predicts brain activity in a broadly distributed semantic system when participants listen to natural stories, replicating a previous study from our lab (Huth et al., 2016). This system extends across much of lateral temporal cortex (LTC), ventral temporal cortex (VTC), lateral parietal cortex (LPC), medial parietal cortex (MPC), medial prefrontal cortex, superior prefrontal cortex, and inferior prefrontal cortex. **Figure 2b** shows that when participants read natural stories this network of brain regions are similarly well predicted by the semantic model. **Figure 2c** compares prediction accuracy of semantic models fit to listening (depicted on the x-axis) versus reading (depicted on the y-axis). The saturation of each point represents the number of voxels that fall into a given range of prediction accuracy. Most voxels are approximately equally well predicted in both modalities. Overall, the semantic model accurately predicts activity in most of the semantic system independent of the presentation modality.

Figure 3 shows voxelwise model prediction accuracy for listening and reading for all voxels and across nine participants in the standard MNI brain space. **Figure 3a** and **3b** show average prediction accuracy across all participants in listening and reading, respectively. **Figure 3c** and **3d** show for each voxel the number of participants where semantic model prediction accuracy is significant in listening and reading, respectively. These results show that our semantic model predicts brain activity within the semantic system in all participants. However, due to averaging across participants voxel prediction accuracies are lower than in individual participant results (maximum prediction accuracy across all MNI voxels for listening 0.27 ± 0.03, maximum prediction accuracy across all MNI voxels for reading: 0.28 ± 0.03).

Does the representation of semantic information vary with sensory modality?

In order to determine whether semantic representation is modality independent we compared the semantic tuning of each voxel estimated during listening versus reading. The semantic tuning of each voxel is given by a 985-dimensional vector of weights, one weight for each of the 985 semantic features. Because there are ~80,000 cortical voxels in each individual participant and 985 semantic features it is impractical to make comprehensive comparisons for each feature. Therefore, to simplify interpretation the estimated semantic model weights were projected into a low-dimensional semantic subspace that captures most of the information about the semantic selectivity of the voxel population. This semantic subspace was created by applying principal component analysis to an aggregated set of estimated semantic model weights from seven participants included in a previous study from our laboratory (Huth et al., 2016). (Note that three of those who participated in the earlier study are also included in the current study.) The resulting semantic principal components (PCs) are ordered by how much variance they explain across the voxels. By projecting both the listening model weights and the reading model weights separately into these semantic PCs, we ensure that cortical voxels that represent similar concepts will project to nearby points in the semantic space.

To visualize which semantic concepts are represented in each voxel, we mapped the projections of the first three semantic principal components onto each participant's cortical surface separately for the two modalities. Each voxel was then colored according to a simple RGB color scheme, where the color red represents the first semantic PC, the color green represents the second semantic PC, and blue the third semantic PC. Inspection of the listening and reading semantic maps shown in **Figure 4** reveals that the semantic representations in both modalities are very similar. The similarity between the listening and reading semantic maps indicate that individual voxels within the semantic system are tuned for the same semantic concepts regardless of presentation modality.

To quantify the similarity between the semantic maps shown in **Figure 4** for all participants, we correlated the projections of the listening and reading model weights into the semantic PCs across the two modalities (listening and reading). To reduce noise, the 10,000 voxels that were best predicted by the semantic model in the two modalities were selected for this analysis. The listening and reading semantic PC projections were then correlated for each semantic PC separately.

Figure 5 shows these correlation coefficients for the first ten semantic PC projections and for all nine participants. Each colored line shows the correlation between listening and reading semantic projections for one participant. The dotted lines indicate the upper bound of the 95% confidence interval of the correlation value under the null hypothesis. Hence, the dotted lines can be interpreted as a form of statistical significance as estimated by a permutation test (see **Materials and Methods** for details). Inspection of **Figure 5** reveals that the first five semantic PC projections are significantly correlated between listening and reading modalities.

The first three semantic PC projections are those that are mapped onto the cortical surface in **Figure 4**. Correlations of the sixth PC projection and beyond are relatively weaker, but remain above chance level until the seventh PC projection. Taken together, these results indicate that the cortical representation of semantic information is consistent across input modalities.

Is semantic tuning consistent across modalities at the single voxel level?

Here we sought to determine whether all the dimensions of semantic representation depend on input modality at the level of single voxels. To do this the 985 semantic model weights estimated for each voxel during listening were correlated with those semantic model weights estimated during reading.

Figure 6a shows the correlation coefficient between estimated listening and reading model weights for each voxel, mapped onto the cortical surface of one individual participant. Listening and reading model weights are strongly correlated in many regions within the semantic system including bilateral temporal, parietal, and prefrontal cortices (red voxels in **Figure 6a**). These voxels are also significantly well predicted by the semantic model in both modalities. Voxels whose model weights are not correlated are located in few scattered voxels in the bilateral sensory cortex, intraparietal sulcus, and in prefrontal cortex (white voxels). This suggests that voxels that are semantically selective in Listening and Reading modalities (red) represent similar semantic information. **Figure 6b** summarizes the relation between within-modality voxelwise model prediction accuracy and semantic tuning, for each voxel. Each voxel is a single point in the scatterplot, and the correlation between the estimated listening and reading model weights is indicated by the color saturation. Semantic tuning is more similar for voxels that are semantically selective in both modalities (red) than for those that are well predicted in one modality only (blue or green). Negatively correlated voxels are mostly in sensory regions and are not well predicted by the semantic model in either modality. In general, individual voxels located within the semantic system are selective to similar semantic features during both listening and reading.

Can a voxelwise model fit to one modality predict responses to the other modality?

If the semantic representation in most of the semantic system is modality-invariant then voxel models fit to one modality should accurately predict responses in the other modality. **Figure 7** and **Figure 8** show cross-modal predictions for all voxels in all participants. **Figure 7** shows prediction accuracy for a model fit to voxel responses evoked during listening, but predicting responses evoked during reading. **Figure 8** shows prediction accuracy for a model fit to responses evoked during reading, but predicting responses evoked during listening. In both figures voxels whose predictions were not statistically significant are shown in gray (p>0.05, FDR corrected). In both cases, voxels in bilateral temporal, parietal, and prefrontal cortices are well predicted across modalities. Voxels that are not well predicted cross-modally are located in sensory cortices.

Figure 9 shows a summary map of the relationship between cross-modality predictions and within-modality predictions, for each voxel and all participants. Summary statistics for the two cross-modality predictions were computed by taking the average cross-modality prediction accuracy (per voxel average of **Figure 7** and **Figure 8**). Summary statistics for the two within-modality predictions were computed by taking the maximum within-modality prediction accuracy (per voxel maximum of **Figure 2a** and **Figure 2b**). The mean cross-modality prediction accuracy and the maximum within-modality prediction accuracy per voxel were then mapped onto the same participant's flattened cortical surface. Inspection of **Figure 9** allows us to identify voxels that are well predicted both within and across modality. Most voxels that are well predicted within and across modality are located in the semantic system (white voxels in **Figure 9**). Outside the semantic system, some voxels on the border to visual cortex and voxels surrounding the temporal parietal junction are well predicted within modality but not across modality (orange voxels in **Figure 9**). (Note, however, that within-modality data were collected largely within sessions, and between-modality data were collected across sessions. Thus, within-modality prediction accuracy is likely to be somewhat higher than between-modality accuracy for this reason alone.) This result demonstrates that the distribution of semantically selective voxels in most of the semantic system is independent of the modality.

DISCUSSION

The experiments presented here were designed to determine whether semantic information obtained during listening and reading are represented within a common underlying semantic system. In separate fMRI sessions participants listened to a spoken story and read a stream of words visually (RSVP using time-locked transcripts of spoken stories). We used voxelwise modeling to estimate semantic selectivity across the entire cerebral cortex, in individual participants, in each voxel separately and in two different presentation modalities (listening and reading).

Our experiments provide three lines of evidence in support of the hypothesis that semantic representations throughout most of the semantic system are invariant to presentation modality. First, voxels in most of the semantic system (temporal, parietal, and prefrontal cortices) are well predicted by the semantic model in each modality independently (**Figure 2-3**). Second, the estimated model weights and the semantic maps are similar between listening and reading (**Figure 4-6**). Third, voxelwise models estimated from one modality (e.g. listening) accurately predict responses in the other modality (e.g. reading) throughout most of the semantic system (**Figure 7-9**).

Our results demonstrate in a single study that semantically amodal voxels span most of the bilateral semantic system. It has been previously proposed that subsequent to early sensory processing, the pathways for processing information by listening or reading converge in semantically-selective regions (Booth et al., 2002; Buchweitz et al., 2009; Carpentier et al., 2001; Chee et al., 1999; Cohen et al., 2004; Constable et al., 2004; Jobard et al., 2007; Liuzzi et al., 2017; Patterson et al., 2007; Spitsyna et al., 2006). Different studies have emphasized different brain regions such as the left anterior temporal lobe, left ventral angular gyrus, left inferotemporal cortex, a region left lateral to the visual word form area (VWFA), left MTG, and the left IFG. Bilateral activations have been reported previously in epileptic patients (Carpentier et al., 2001) or when complex stimuli such as narrative has been used (Jobard et al., 2007; Regev et al., 2013; Spitsyna et al., 2006). Our study shows that semantically amodal voxels are bilaterally distributed across many regions of the temporal, parietal and prefrontal cortices (**Figure 3-6**). Specifically, we show amodal semantic representation in bilateral precuneus, temporal parietal junction (TPJ), angular gyrus (AG), anterior to posterior superior temporal sulcus (STS), superior ventral premotor cortex (sPMv), Broca's area and inferior frontal gyrus (IFG).

One previous report noted that listening and reading evoke different levels of brain activity in anterior and posterior left DLPFC (Regev et al., 2013). However, Regev et al. (2013) did not model linguistic features directly. Therefore, it is unclear whether the differences they identified within left DLPFC are due to differences in semantic representation or some other aspect of linguistic information (e.g. syntax). In our study, we focused solely on semantic representations and our results suggest that semantic representations do not differ between listening and reading in left DLPFC. However, it is possible that this structure may represent other types of linguistic information differently during listening and reading.

One striking difference between our results and those reported in earlier studies is that we find a large network of semantically-selective regions that are independent of the presentation modality, whereas previous studies reported a few amodal semantic regions located mostly in the left hemisphere (Buchweitz et al., 2009; Chee et al., 1999; Jobard et al., 2007; Liuzzi et al., 2017; Petersen et al., 1989). There are three possible factors that contribute to this discrepancy. First, we used rich narrative language as stimuli to study cross-modal semantic representation (**Figure 1**). Previous studies have shown that complex linguistic stimuli such as narrative stories activate many more brain regions than single words or short sentences (Jobard et al., 2007; Lerner et al., 2011; Mazoyer et al., 1993; Xu et al., 2005). Hence, differences in signal-to-noise ratio can account for fewer number of amodal regions identified in previous cross-modality studies that use single words or short sentences.

Second, our voxelwise modeling approach used explicit semantic features, which allowed us to identify brain regions that consistently respond to specific semantic information across different modalities (**Figure 1-6**). To our knowledge, only one previous study of cross-modal representation has used explicit semantic features to model brain activity patterns related to semantics (Liuzzi et al., 2017). That study used as stimuli twenty-four single words derived from only six animate categories, and showed cross-modal representations within left pars triangularis. However, the most likely reason that the Luizzi et al. (2017) study only identified one region as semantically amodal is that single word presentations elicit little brain activity.

Third, the present study is the first that reveals the amodal representation of semantic information during listening and reading in single participants (**Figure 1**). In contrast, most previous neuroimaging studies of language perform comparisons at the group level after transforming individual participant data into a standardized brain space (e.g. MNI or Talairach space). However, the anatomical normalization procedures used in these studies tend to smooth and mask the substantial individual variability in language processing (Caramazza, 1986; Fedorenko and Kanwisher, 2009; Steinmetz and Seitz, 1991). Therefore, studies performing inter-subject averaging might average away meaningful signal and fail to find significant relationships (Fedorenko and Kanwisher, 2009). Indeed, projecting our results into a standard brain space and averaging across individuals reduces prediction performance within modality across much of the brain (compare **Figure 3a** and **Figure 2**). This result already demonstrate that it is important to study cross-modal language representations in individual participants.

Our naturalistic experiment and voxelwise modeling provides a powerful and efficient method for identifying amodal representations in individual human brains. However, the semantic feature space that we used here is only one possible way of representing semantics (Huth et al., 2016; Mitchell et al., 2008; Pereira et al., 2018), and it has some limitations. For example, when people listen to or read a story they likely employ conceptual knowledge at long time scales beyond those using for computing semantic features (Yeshurun et al., 2017). Furthermore, semantic comprehension involves metaphors, humor, sarcasm and narrative information that is not reflected in the current semantic model. It is possible that these unmodeled properties of natural language might have different, modality-specific representations in the brain.

In sum, we demonstrate modality-independent semantic selectivity in most of the bilateral semantic system. The semantic maps recovered in this study show that semantic tuning in individual participants is very similar across the two modalities. Our findings are consistent with the view that sensory regions process unimodal information related to low-level processing of spoken or written language, whereas high-level regions process modality invariant semantic information. Furthermore, our results reveal that modality invariant semantic representations are not isolated in a few left-lateralized regions, but are instead present in many bilaterally distributed regions of the semantic system.

References

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

Amunts, K., Lenzen, M., Friederici, A.D., Schleicher, A., Morosan, P., Palomero-Gallagher, N., and Zilles, K. (2010). Broca's Region: Novel Organizational Principles and Multiple Receptor Mapping. PLoS Biol. *8*, e1000489.

Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., and Collins, M. (2016). Globally Normalized Transition-Based Neural Networks.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B *57*, 289–300.

Bergstra, J.S., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. In Advances in Neural Information Processing Systems 24, J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, eds. (Curran Associates, Inc.), pp. 2546–2554.

Binder, J.R., Desai, R.H., Graves, W.W., and Conant, L.L. (2009). Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. Cereb. Cortex *19*, 2767–2796.

Boersma, P., and Weenink, D. (2014). Praat: doing phonetics by computer.

Booth, J.R., Burman, D.D., Meyer, J.R., Gitelman, D.R., Parrish, T.B., and Mesulam, M.M. (2002). Modality independence of word comprehension. Hum. Brain Mapp. *16*, 251–261.

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

Caramazza, A. (1986). On Drawing Inferences about the Structure of Normal Cognitive Systems from the Analysis of Patterns of Impaired Performance: The Case for Single-Patient Studies.

Carpentier, A., Pugh, K.R., Westerveld, M., Studholme, C., Skrinjar, O., Thompson, J.L., Spencer, D.D., and Constable, R.T. (2001). Functional MRI of language processing: dependence on input modality and temporal lobe epilepsy. Epilepsia *42*, 1241–1254.

Chee, M.W., O'Craven, K.M., Bergida, R., Rosen, B.R., and Savoy, R.L. (1999). Auditory and visual word processing studied with fMRI. Hum. Brain Mapp. *7*, 15–28.

Church, K.W., and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. Comput. Linguist. *16*, 22–29.

Corbetta, M., Akbudak, E., Conturo, T.E., Snyder, A.Z., Ollinger, J.M., Drury, H.A., Linenweber, M.R., Petersen, S.E., Raichle, M.E., Van Essen, D.C., et al. (1998). A common network of functional areas for attention and eye movements. Neuron *21*, 761–773.

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

Dale, A.M., Fischl, B., and Sereno, M.I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. Neuroimage *9*, 179–194.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. (1990). Indexing by Latent

Semantic Analysis. J. Am. Soc. Inf. Sci. *41*, 391–407.

Démonet, J.-F., Price, C., Wise, R., and Frackowiak, R.S.J. (1994). Differential activation of right and left posterior sylvian regions by semantic and phonological tasks: a positron-emission tomography study in normal human subjects. Neurosci. Lett. *182*, 25–28.

Démonet, J.F., Chollet, F., Ramsay, S., Cardebat, D., Nespoulous, J.L., Wise, R., Rascol, A., and Frackowiak, R. (1992). The anatomy of phonological and semantic processing in normal subjects. Brain *115*, 1753–1768.

Devlin, J.T., Jamison, H.L., Matthews, P.M., and Gonnerman, L.M. (2004). Morphology and the internal structure of words. Proc. Natl. Acad. Sci. U. S. A. *101*, 14984–14988.

Diakidoy, I.-A.N., Stylianou, P., Karefillidou, C., and Papageorgiou, P. (2005). The relationship between listening and reading comprehension of different types of text at increasing grade levels. Read. Psychol. *26*, 55–80.

Downing, P.E., Jiang, Y., Shuman, M., and Kanwisher, N. (2001). A Cortical Area Selective for Visual Processing of the Human Body. Science (80-.). *293*, 2470–2473.

Epstein, R., and Kanwisher, N. (1998). A cortical representation of the local visual environment. Nature *392*, 598–601.

Fedorenko, E., and Kanwisher, N. (2009). Neuroimaging of Language: Why Hasn't a Clearer Picture Emerged? Lang. Linguist. Compass *3*, 839–865.

Forster, K.I. (1970). Visual perception of rapidly presented word sequences of varying complexity. Percept. Psychophys. *8*, 215–221.

Gao, J.S., Huth, A.G., Lescroart, M.D., and Gallant, J.L. (2015). Pycortex: an interactive surface visualizer for fMRI. Front. Neuroinform. *9*, 23.

Grosbras, M.H., Lobel, E., Van de Moortele, P.F., LeBihan, D., and Berthoz, A. An anatomical landmark for the supplementary eye fields in human revealed with functional magnetic resonance imaging. Cereb. Cortex *9*, 705–711.

Halgren, E., Dale, A.M., Sereno, M.I., Tootell, R.B.H., Marinkovic, K., and Rosen, B.R. (1999). Location of human face-selective cortex with respect to retinotopic areas. Hum. Brain Mapp. *7*, 29–37.

Hansen, K.A., Kay, K.N., and Gallant, J.L. (2007). Topographic Organization in and near Human Visual Area V4. J. Neurosci. *27*, 11896–11911.

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

Huth, A.G., Nishimoto, S., Vu, A.T., and Gallant, J.L. (2012). A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. Neuron *76*, 1210–1224.

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

Jenkinson, M., and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. Med. Image Anal. *5*, 143–156.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. Neuroimage *17*, 825–841.

Jobard, G., Vigneau, M., Mazoyer, B., and Tzourio-Mazoyer, N. (2007). Impact of modality and linguistic complexity during reading and listening tasks. Neuroimage *34*, 784–800.

Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. J. Neurosci. *17*, 4302–4311.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., et al. (2016). Jupyter Notebooks—a publishing format for reproducible computational workflows. In Positioning and Power in Academic Publishing: Players, Agents and Agendas, pp. 87–90.

Lerner, Y., Honey, C.J., Silbert, L.J., and Hasson, U. (2011). Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. J. Neurosci. *31*.

Lescroart, M.D., Stansbury, D.E., and Gallant, J.L. (2015). Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. Front. Comput. Neurosci. *9*, 135.

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

Lund, K., and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behav. Res. Methods, Instruments, Comput. *28*, 203–208.

Mazoyer, B.M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., Salamon, G., Dehaene, S., Cohen, L., and Mehler, J. (1993). The Cortical Representation of Speech. J. Cogn. Neurosci. *5*, 467–479.

McLaren, K. (1976). XIII-The Development of the CIE 1976 (L* a* b*) Uniform Colour Space and Colour-difference Formula. J. Soc. Dye. Colour. *92*, 338–341.

Michael, E.B., Keller, T.A., Carpenter, P.A., and Just, M.A. (2001). fMRI investigation of sentence comprehension by eye and by ear: modality fingerprints on cognitive processes. Hum. Brain Mapp. *13*, 239–252.

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

Nakamura, K., Kawashima, R., Sato, N., Nakamura, A., Sugiura, M., Kato, T., Hatano, K., Ito, K., Fukuda, H., Schormann, T., et al. (2000). Functional delineation of the human occipito-temporal areas related to face and scene processing. A PET study. Brain *123 (Pt 9)*, 1903–1912.

Nakamura, K., Dehaene, S., Jobert, A., Bihan, D. Le, and Kouider, S. (2005). Subliminal Convergence of Kanji and Kana Words: Further Evidence for Functional Parcellation of the Posterior Temporal Cortex in Visual Word Perception. J. Cogn. Neurosci. *17*, 954–968.

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

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J.L. (2011). Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. Curr. Biol. *21*, 1641–1646.

Nunez-Elizalde, A.O., Huth, A.G., and Gallant, J.L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. Neuroimage.

Oldfield, R.C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia

9, 97–113.

Oliphant, T.E. (2006). Guide to NumPy (Provo, UT: Brigham Young University).

Oliphant, T.E. (2007). SciPy: Open source scientific tools for Python. Comput. Sci. Eng. *9*, 10–20.

Paus, T. (1996). Location and function of the human frontal eye-field: a selective review. Neuropsychologia *34*, 475–483.

Penfield, W., and Boldrey, E. (1937). Somatic Motor and Sensory Representation in the Cerebral Cortex of Man as Studied by Electrical Stimulation. Brain *60*, 389–443.

Penny, W.D., Friston, K., Ashburner, J., Kiebel, S., and Nichols, T. (2007). Statistical parametric mapping : the analysis of funtional brain images (Elsevier/Academic Press).

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S.J., Kanwisher, N., Botvinick, M., and Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. Nat. Commun. *9*, 963.

Petersen, S.E., Fox, P.T., Posner, M.I., Mintun, M., and Raichle, M.E. (1989). Positron Emission Tomographic Studies of the Processing of Singe Words. J. Cogn. Neurosci. *1*, 153–170.

Price, C.J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. Ann. N. Y. Acad. Sci. *1191*, 62–88.

Price, C.J. (2012). A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage *62*, 816–847.

Regev, M., Honey, C.J., Simony, E., and Hasson, U. (2013). Selective and Invariant Neural Responses to Spoken and Written Narratives. J. Neurosci. *33*, 15978–15988.

Rissman, J., Eliassen, J.C., and Blumstein, S.E. (2003). An Event-Related fMRI Investigation of Implicit Semantic Priming. J. Cogn. Neurosci. *15*, 1160–1175.

Rubin, D.L., Hafer, T., and Arata, K. (2000). Reading and listening to oral-based versus literate-based discourse. Commun. Educ. *49*, 121–133.

Scott, S.K., Blank, C.C., Rosen, S., and Wise, R.J.S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. Brain *123*, 2400–2406.

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

Spitsyna, G., Warren, J.E., Scott, S.K., Turkheimer, F.E., and Wise, R.J.S. (2006). Converging Language Streams in the Human Temporal Lobe. J. Neurosci. *26*, 7328–7336.

Stansbury, D.E., Naselaris, T., and Gallant, J.L. (2013). Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex. Neuron *79*, 1025–1034.

Steinmetz, H., and Seitz, R.J. (1991). Functional anatomy of language processing: Neuroimaging and the problem of individual variability. Neuropsychologia *29*, 1149–1161.

Tikhonov, A.N., and Arsenin, V.Y. (1977). Solutions of ill-posed problems (Washington; New York: Winston; distributed solely by Halsted Press).

Tootell, R., Reppas, J., Kwong, K., Malach, R., Born, R., Brady, T., Rosen, B., and Belliveau, J. (1995). Functional analysis of human MT and related visual cortical areas using magnetic resonance imaging. J. Neurosci. *15*, 3215–3230.

Turney, P., and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. J. Artif. Intell. Res. *37*, 141–188.

Vandenberghe, R., Price, C., Wise, R., Josephs, O., and Frackowiak, R.S.J. (1996). Functional anatomy of a common semantic system for words and pictures. Nature *383*, 254–256.

Vigneau, M., Beaucousin, V., Hervé, P.Y., Duffau, H., Crivello, F., Houdé, O., Mazoyer, B., and Tzourio-Mazoyer, N. (2006). Meta-analyzing left hemisphere language areas: Phonology, semantics, and sentence processing. Neuroimage *30*, 1414–1432.

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

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., and Mitchell, T. (2014). Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. PLoS One *9*, e112575.

Xu, J., Kemeny, S., Park, G., Frattali, C., and Braun, A. (2005). Language in context: emergent features of word, sentence, and narrative comprehension. Neuroimage *25*, 1002–1015.

Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C.J., and Hasson, U. (2017). Same Story, Different Story. Psychol. Sci. *28*, 307–319.

Yuan, J., and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. 6–9.

Zilles, K., and Amunts, K. (2018). Cytoarchitectonic and receptorarchitectonic organization in Broca's region and surrounding cortex. Curr. Opin. Behav. Sci. *21*, 93–105.

Figure Legends

Figure 1: Experimental procedure and voxelwise modeling. Nine participants listened to and read over two hours of natural stories in each modality while BOLD responses were measured using fMRI. The presentation time of single words was matched between listening and reading sessions. Semantic features were constructed by projecting each word in the stories into a 985-dimensional word embedding space independently constructed using word co-occurence statistics from a large corpus. These features and BOLD responses were used to estimate a separate finite impulse response (FIR) banded ridge regression model for each voxel in every individual participant. These estimated model weights were used to predict BOLD responses for a separate held-out story that was not used for model estimation. Predictions for individual participants were computed separately for listening and reading sessions. Model performance was quantified as the correlation between the predicted and recorded BOLD responses to this held-out story. Within-modality prediction accuracy was quantified by correlating the predicted responses from one modality (e.g. listening) with the recorded responses to the same modality (e.g. listening). Cross-modality prediction accuracy was quantified by correlating the predicted responses for one modality (e.g. listening) with the recorded responses of the other modality (e.g. reading).

Figure 2: Semantic model prediction accuracy across the cortical surface. Voxelwise modeling was used to estimate semantic model weights in two modalities, listening and reading. Prediction accuracy was computed as the correlation (r) between the participant's recorded BOLD activity to the held-out validation story and the responses predicted by the semantic model. **a.** Accuracy of voxelwise models estimated using listening data and predicting withheld listening data. The flattened cortical surface of one participant is shown. Prediction accuracy is given by the color scale shown at bottom. Voxels that are well predicted appear yellow or white, voxel predictions that are not statistically significant are shown in gray (p>0.05, FDR corrected; LH, left hemisphere; RH, right hemisphere; NS, not significant; EVC, early visual cortex; AC: auditory cortexc; LTC, lateral temporal cortex, VTC, ventral temporal cortex; LPC, lateral parietal cortex; MPC, medial parietal cortex; PFC, prefrontal cortex). **b.** Accuracy of voxelwise models estimated using reading data and predicting withheld reading data. The format is the same as panel **a**. Estimated semantic model weights accurately predict BOLD responses in many brain regions in the semantic system, including lateral temporal cortex (LTC), ventral temporal cortex (VTC), lateral parietal cortex (LPC), medial parietal cortex (MPC), and prefrontal cortex (PFC) in both modalities. In contrast, voxels in the early sensory regions such as the primary auditory cortex and early visual cortex are not well predicted.**c.** Log transformed density plot of the listening (x-axis) versus reading (y-axis) model prediction accuracy. Purple points indicate all voxels. Darker colors indicate a higher number of voxels in the corresponding bin. Voxels with listening prediction accuracy < 0.17 and reading prediction accuracy < 0.19 are not significant. Most voxels are equally well predicted in listening and reading indicating that these voxels represent semantic information independent of the presentation modality.

Figure 3: Semantic model prediction accuracy across all participants in standard brain space. Voxelwise modeling was used to asses semantic model prediction accuracy in the listening and reading modalities for all nine participants as described in **Figure 2a** and **Figure 2b**. Prediction accuracies computed in individual subject's space were then projected into a standard MNI brain space. **a.** Average listening prediction accuracy across nine participants was computed for each MNI voxel in the standard brain space and is mapped onto the cortical surface of the MNI brain. Average prediction accuracy is given by the color scale. Voxels that are well predicted appear brighter. Across all participants the estimated semantic model weights in the listening modality accurately predict BOLD responses in many brain regions in the semantic system, including lateral temporal cortex (LTC), ventral temporal cortex (VTC), lateral parietal cortex (LPC), medial parietal cortex (MPC), and prefrontal cortex (PFC). (LH: Left hemisphere, RH: Right hemisphere, EVC, early visual cortex; AC: auditory cortexc; LTC, lateral temporal cortex, VTC, ventral temporal cortex; LPC, lateral parietal cortex; MPC, medial parietal cortex; PFC, prefrontal cortex) **b.** Average reading prediction accuracy across nine participants was computed for each MNI voxel in the standard brain space and is mapped onto the cortical surface of the MNI brain. The format is the same as in a. Across all participants, estimated semantic model weights in the reading modality accurately predict BOLD responses in the semantic system. **c.** Significant prediction accuracy in each voxel in the listening modality was determined in the subject space and then projected to the standard MNI brain space. The number of subjects with significant semantic model prediction accuracy for a given MNI voxel is then mapped onto the cortical surface of the MNI brain. Number of participants is given by the color scale shown at bottom. Dark red voxels are significantly well predicted in all participants. Dark blue voxels are not significantly predicted in any participant. **d.** Significant prediction accuracy in each voxel in the reading modality was determined in the subject space and then projected to the standard MNI brain space. The number of subjects with significant semantic model prediction accuracy for a given MNI voxel is then mapped onto the cortical surface of the MNI brain. The format is the same as in c. Most of the voxels in the semantic system are significantly predicted by all participants in both modalities.

Figure 4: Semantic tuning maps for listening and reading. The semantic maps for both modalities are displayed on the cortical surface of one participant. **a.** Voxelwise model weights for the listening sessions were projected into a

semantic space created by performing principal component analysis on estimated semantic model weights acquired during a listening experiment published earlier (Huth et al., 2016). Each voxel is colored according to its projection onto the first (red), second (blue) or third (green) semantic PC. The color wheel legend at center indicates the associated semantic concepts. Voxels whose within-modality prediction was not statistically significant are shown in gray (p>0.05, FDR corrected; LH, left hemisphere; RH, right hemisphere; EVC, early visual cortex; AC: auditory cortex; LTC, lateral temporal cortex, VTC, ventral temporal cortex; LPC, lateral parietal cortex; MPC, medial parietal cortex; PFC, prefrontal cortex). **b.** Voxelwise model weights for the reading sessions projected into the semantic space, and colored using the same procedure as in **a.** Comparison of panels a and b reveals that semantically selective voxels are tuned for similar semantic concepts during both listening and reading.

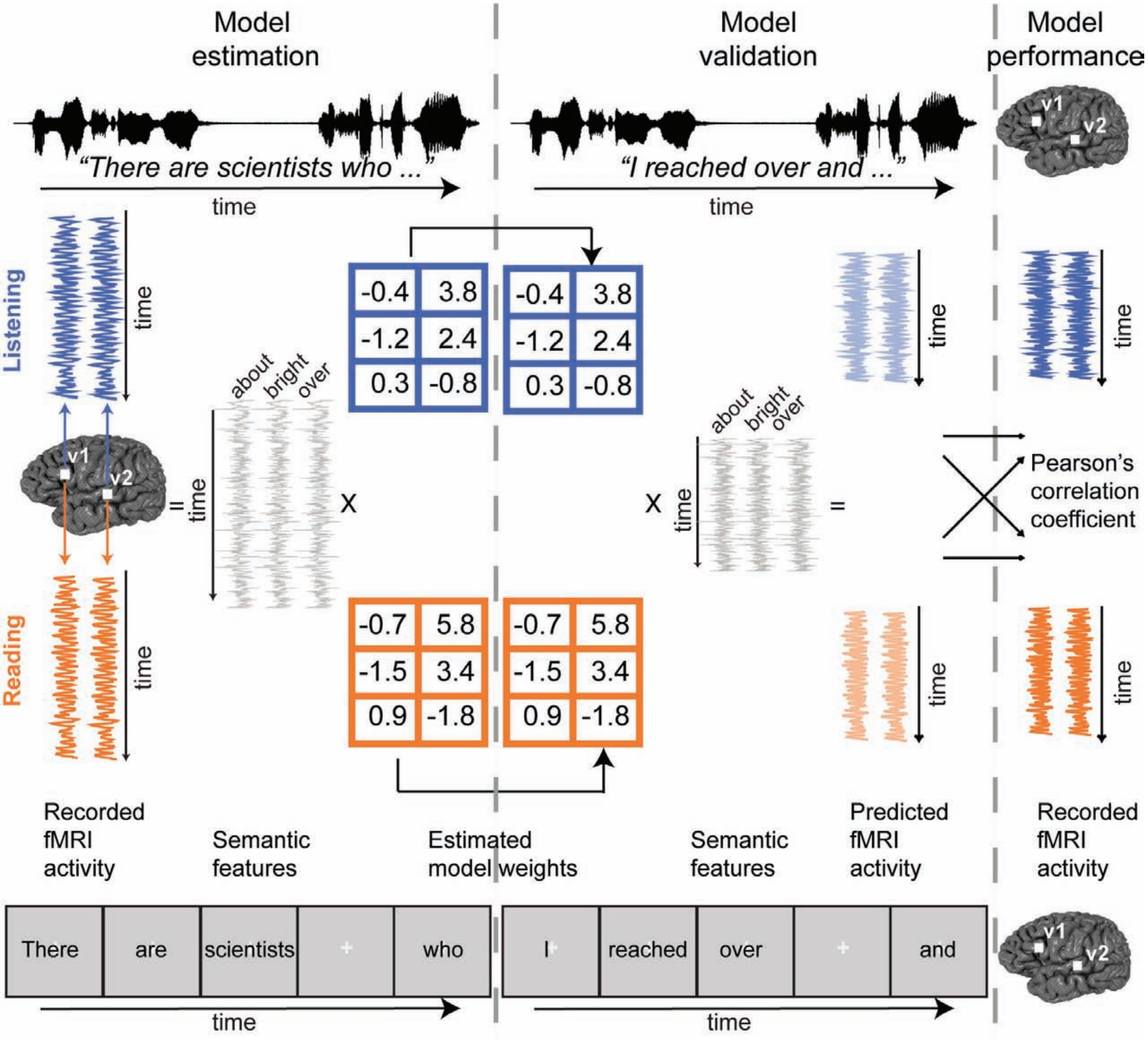
Figure 5: Similarity between listening and reading semantic PC projections. The correlation coefficient between listening and reading semantic PC projections are shown for the first ten semantic PCs and each individual participant separately. Each colored diamond shape indicate one participant and the mean correlation coefficient across participants is indicated by the black solid line. Error bars are standard error of the mean across the correlation coefficients for all participants. The colored dotted lines at the bottom indicate chance level correlation for each semantic PC and participant as computed by a permutation test. At least the first five semantic PC projections are significantly correlated between listening and reading. This shows that the individual dimensions of the semantic maps in **Figure 4** where the first three semantic PCs are displayed are similar across the two modalities.

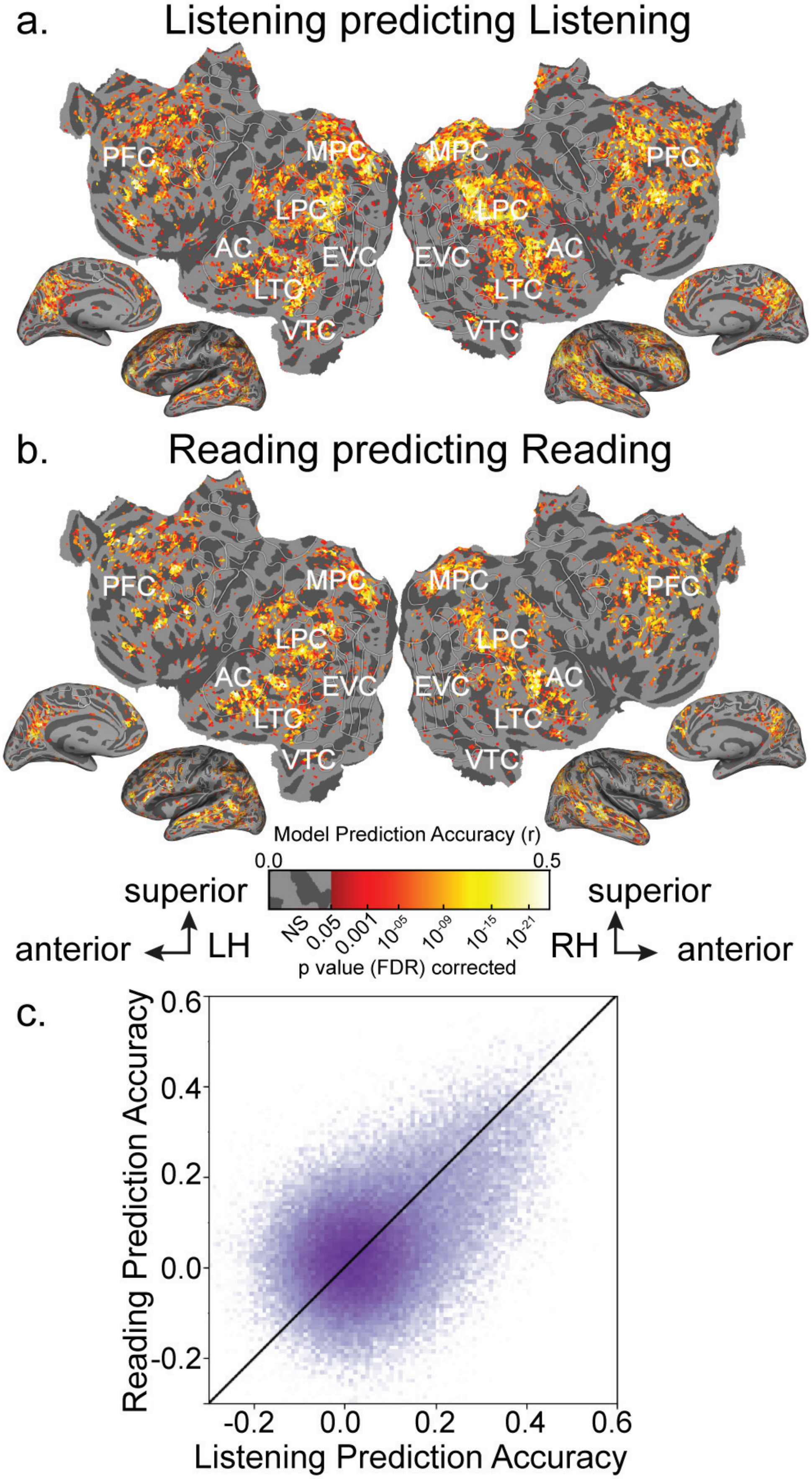
Figure 6: Voxelwise similarity of semantic tuning across listening and reading. Semantic model weights estimated during listening and reading were correlated for each voxel separately. **a.** Correlation coefficient between listening and reading model weights are shown on the flattened cortical surface of one participant. Red voxels are those that are semantically selective in both modalities. Blue voxels are those that are semantically selective in listening, but not reading. Green voxels are those that are semantically selective in reading, but not listening. Gray voxels are not semantically selective in either modality. Color saturation describes the strength of voxel weight correlations. The stronger the color the higher is the correlation between listening and reading model weights. Voxels in the semantic system have similar semantic tuning across all the semantic features. (LH, left hemisphere; RH, right hemisphere; NS, not significant; EVC, early visual cortex; AC: auditory cortex; LTC, lateral temporal cortex, VTC, ventral temporal cortex; LPC, lateral parietal cortex; MPC, medial parietal cortex; PFC, prefrontal cortex) This suggests that across the 985 semantic features semantic information is represented similarly in both modalities in the semantic system. **b.** The relation between within-modality model prediction accuracy and semantic tuning. Listening (x-axis) versus reading (y-axis) prediction accuracy is shown in a scatterplot where each point corresponds to a single voxel in panel **a.** The correlation between the listening and reading model weights is indicated by color saturation and is the same as in panel **a**. Semantic tuning is more similar for voxels that are semantically selective in both modalities (red) than for those that are selective in one modality only (blue and green). Gray voxels are not semantically selective in either modality. This suggests that voxels that are well predicted in both modalities represent similar semantic information.

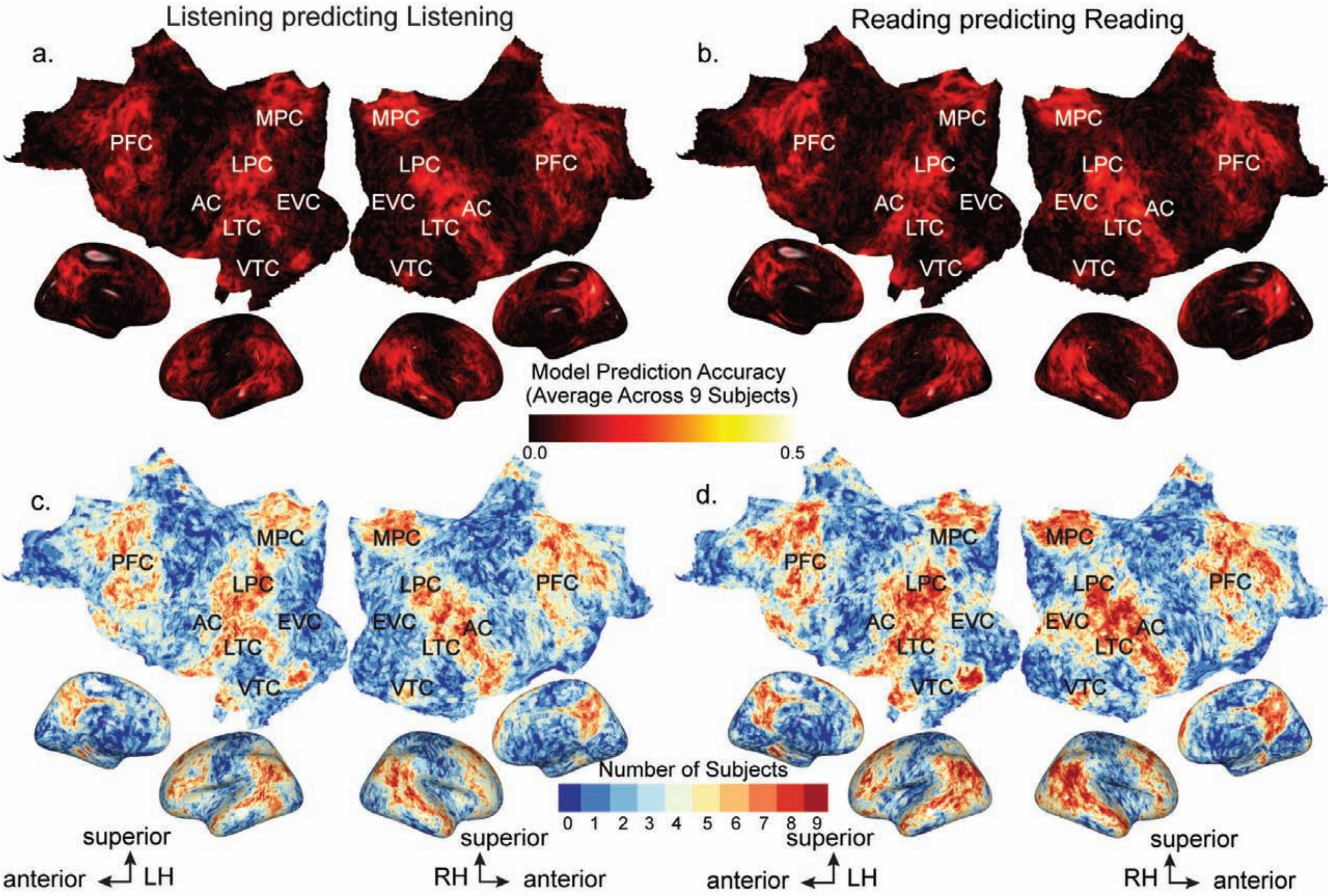
Figure 7: **Semantically amodal voxels as shown by cross-modal predictions (Listening predicting Reading) in all participants.** Estimated semantic model weights in the listening modality were used to predict BOLD activity to the held-out validation story in the reading modality. **a.** Accuracy of voxelwise models estimated during listening predicting reading responses, shown on the same participant's flattened cortical surface as in Figure 2. Prediction accuracy is given by the color scale. Voxels that are well predicted appear yellow or white, voxel predictions that are not statistically significant are shown in gray (p>0.05, FDR corrected; LH, left hemisphere; RH, right hemisphere; NS, not significant; Si: Subject i; EVC, early visual cortex; AC: auditory cortex; LTC, lateral temporal cortex, VTC, ventral temporal cortex; LPC, lateral parietal cortex; MPC, medial parietal cortex; PFC, prefrontal cortex). **b.** Accuracy of voxelwise models estimated during listening predicting reading responses, shown for all other participants. The format is the same as panel **a**. The semantic model estimated in listening accurately predicts voxel responses in reading within the semantic system including bilateral temporal (LTC, VTC), parietal (LPC, MPC), and prefrontal cortices (PFC).

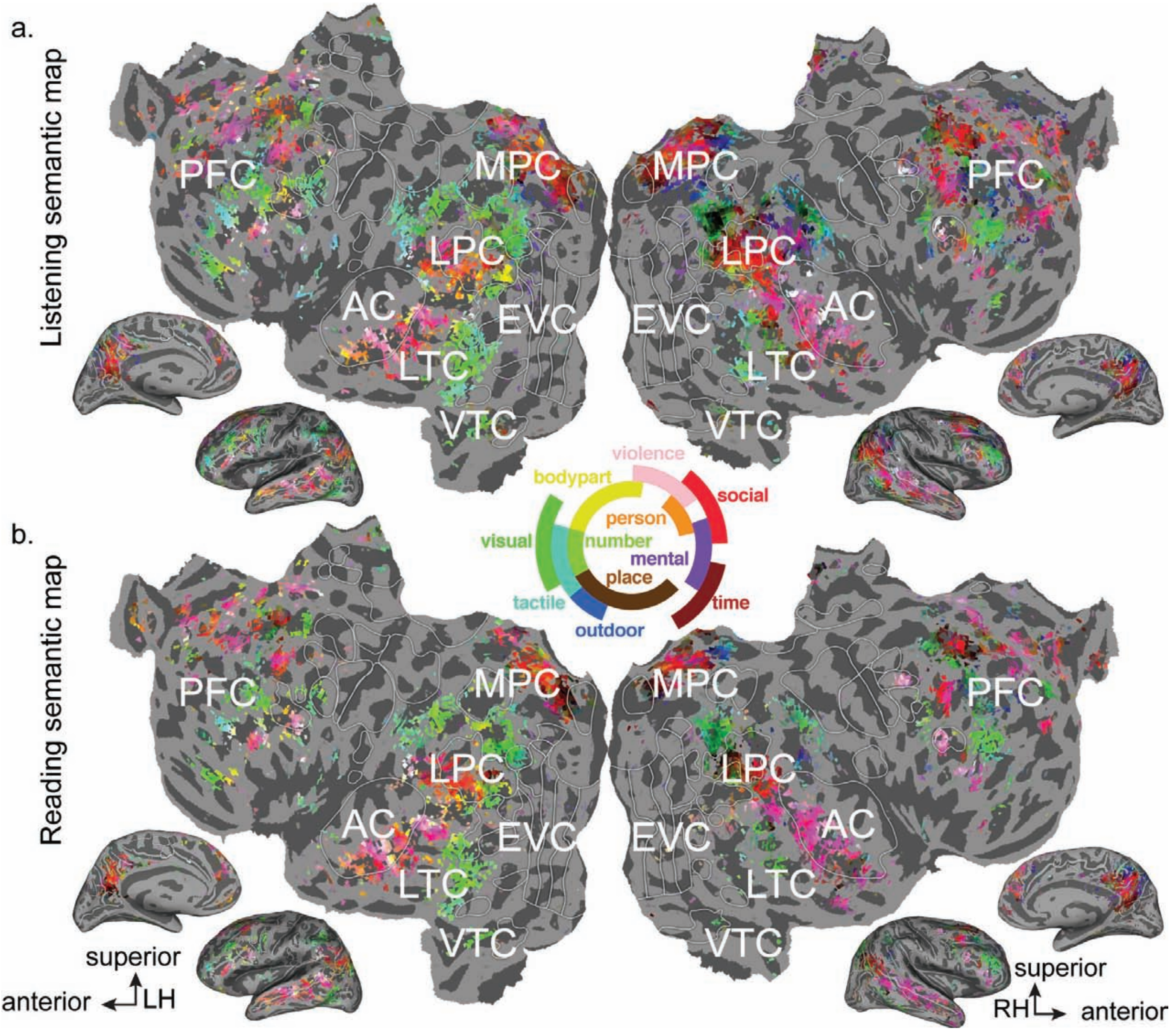
Figure 8: **Semantically amodal voxels as shown by cross-modal predictions (Reading predicting Listening) in all participants.** Estimated semantic model weights in the reading modality were used to predict BOLD activity to the held-out validation story in the listening modality. **a.** Accuracy of voxelwise models estimated during reading predicting listening responses, shown on the same participant's flattened cortical surface as in Figure 2. Prediction accuracy is given by the color scale. Voxels that are well predicted appear yellow or white, voxel predictions that are not statistically significant are shown in gray (p>0.05, FDR corrected; LH, left hemisphere; RH, right hemisphere; NS, not significant; Si: Subject i; EVC, early visual cortex; AC: auditory cortex; LTC, lateral temporal cortex, VTC, ventral temporal cortex; LPC, lateral parietal cortex; MPC, medial parietal cortex; PFC, prefrontal cortex). **b.** Accuracy of voxelwise models estimated during reading predicting listening responses, shown for all other participants. The format is the same as panel **a**. The semantic model estimated in reading accurately predicts voxel responses in listening within the semantic system including bilateral temporal (LTC, VTC), parietal (LPC, MPC), and prefrontal cortices (PFC).

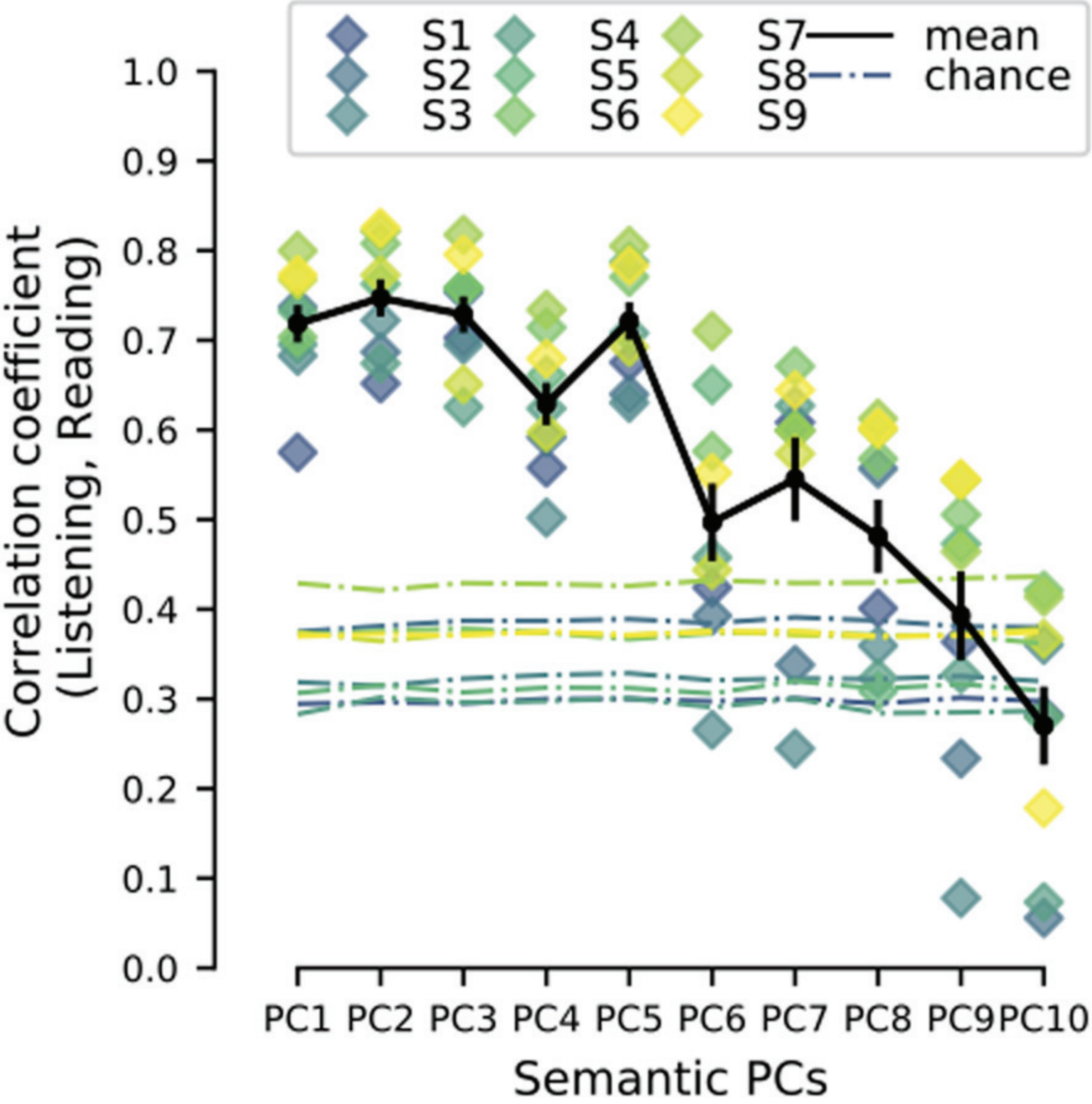
Figure 9: Semantically amodal voxels for all participants. Comparison of voxels that are well predicted across modalities versus within modalities. **a.** The average cross-modality prediction accuracy and the maximum of the within-modality prediction accuracy per voxel are both plotted on the flattened cortical surface of the same participant's flattened cortical surface as in Figure 2. (L2R: Listening predicting Reading, R2L: Reading predicting Listening; L2L: Listening predicting Listening, R2R: Reading predicting Reading; Si: Subject i; LH, left hemisphere; RH, right hemisphere; NS, not significant).Orange voxels are well predicted only within-modality. White voxels are well predicted both within and across modality (in most of the semantic system). Blue voxels are well predicted only across modality. Voxels that are not significant in within- or cross-modality predictions are shown in gray. **b.** The same comparison is plotted for all other participants. The format is the same as panel **a**. Voxels within the semantic system represent semantic information independent of modality.

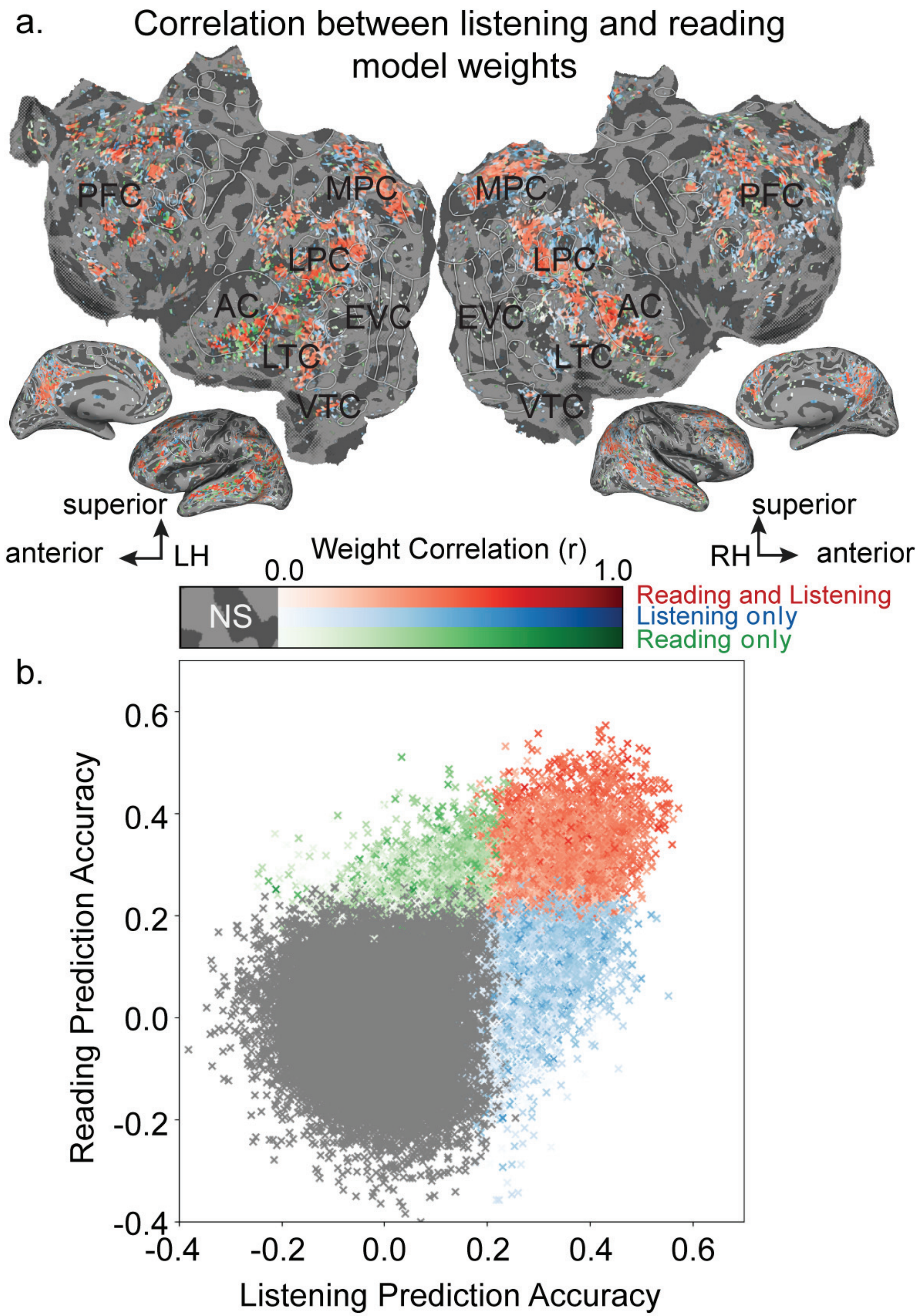

Model
estimation
Model
validation
Model
performance
"There are scientists who ..."
time
"I reached over and ..."
time
v1
v2
Listening
time
Reading
time
=
time
about
bright
over
x
-0.4
3.8
-1.2
2.4
0.3
-0.8
-0.7
5.8
-1.5
3.4
0.9
-1.8
x
time
=
Pearson's
correlation
coefficient
Recorded
fMRI
activity
Semantic
features
Estimated
model weights
Semantic
features
Predicted
fMRI
activity
Recorded
fMRI
activity
There
are
scientists
who
I
reached
over
and
time
time

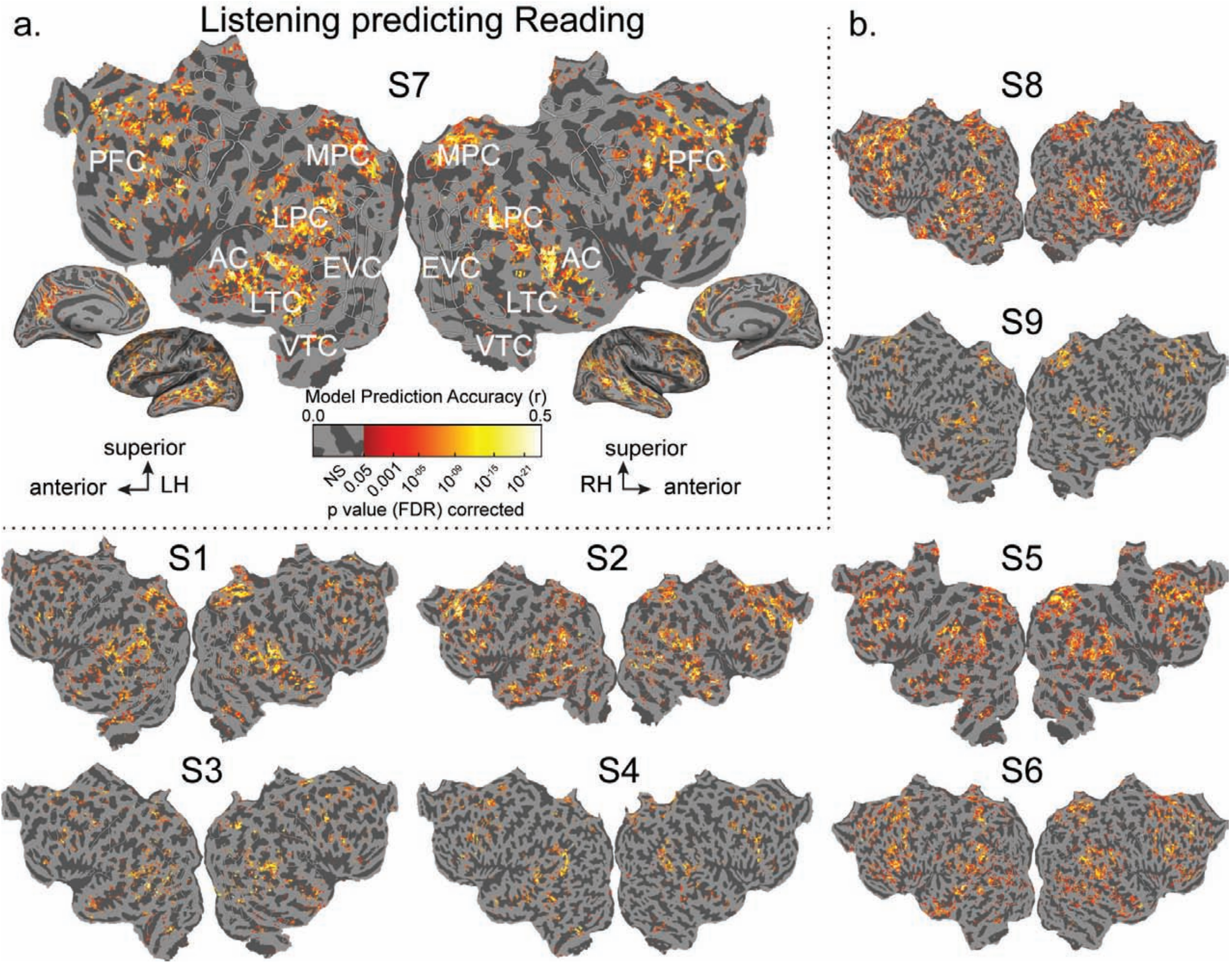

a.
Listening predicting Listening
PFC
MPC
LPC
AC
EVC
LTC
VTC
MPC
PFC
LPC
EVC
AC
LTC
VTC
b.
Reading predicting Reading
PFC
MPC
LPC
AC
EVC
LTC
VTC
MPC
PFC
LPC
EVC
AC
LTC
VTC
Model Prediction Accuracy (r)
0.0
0.5
NS
0.05
0.001
10-05
10-09
10-15
10-21
p value (FDR) corrected
superior
anterior
LH
superior
RH
anterior
c.
Reading Prediction Accuracy
0.6
0.4
0.2
0.0
-0.2
-0.2
0.0
0.2
0.4
0.6
Listening Prediction Accuracy

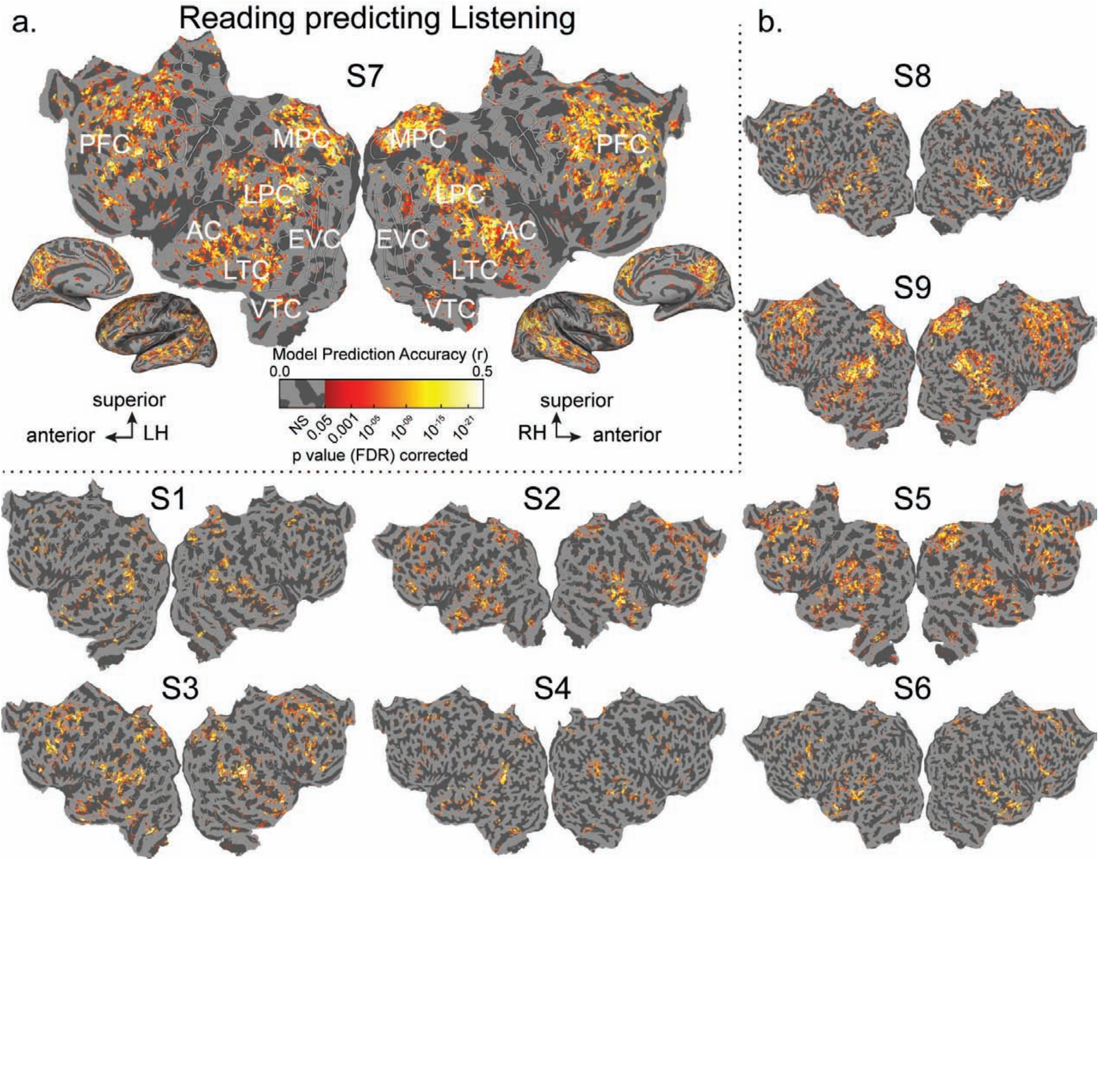

Listening predicting Listening
Reading predicting Reading
a.
b.
c.
d.
PFC
MPC
LPC
AC
EVC
LTC
VTC
Model Prediction Accuracy
(Average Across 9 Subjects)
0.0
0.5
Number of Subjects
0
1
2
3
4
5
6
7
8
9
superior
anterior
LH
RH

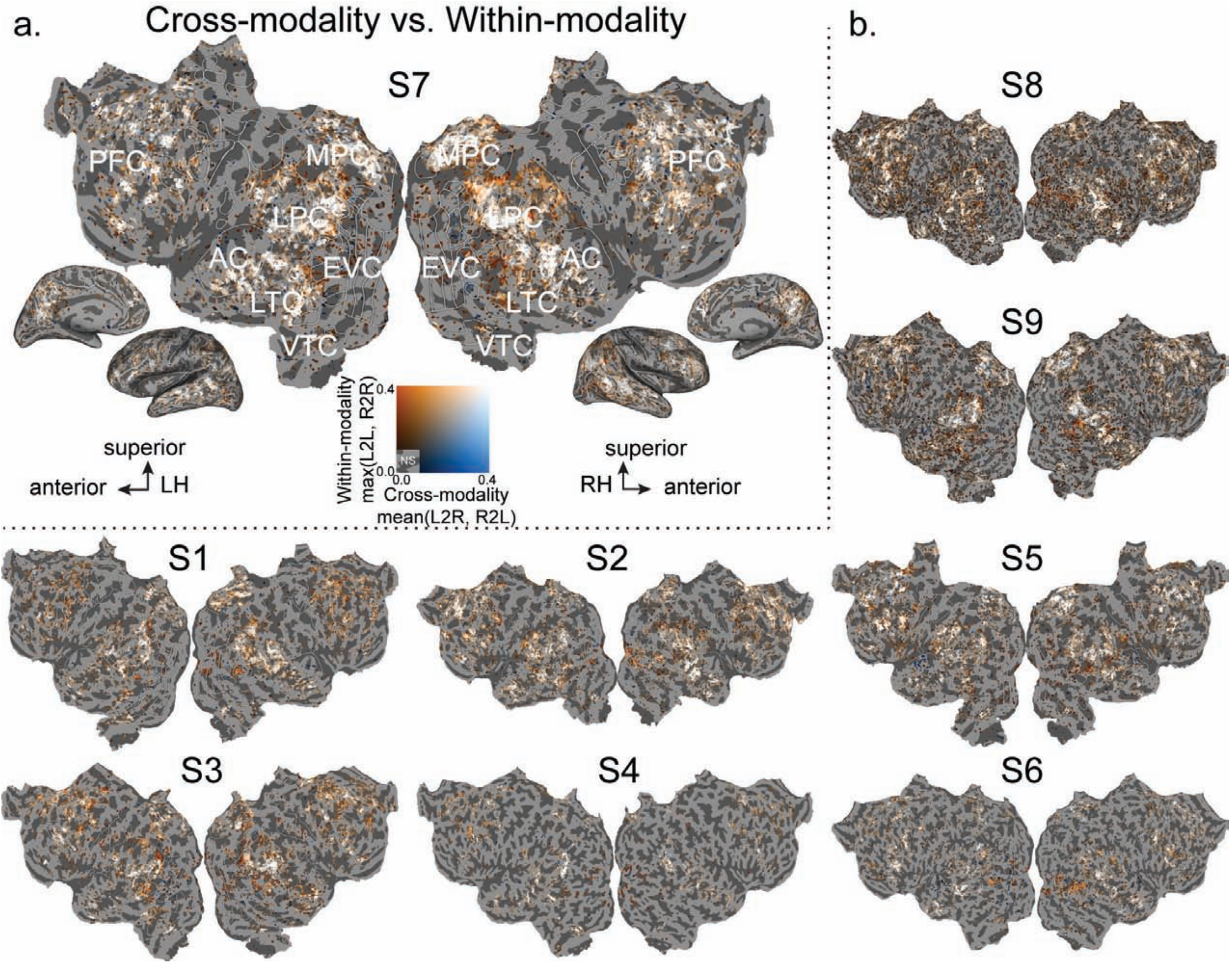

a.
Listening semantic map
PFC
MPC
LPC
AC
EVC
LTC
VTC
MPC
PFC
LPC
EVC
AC
LTC
VTC
violence
bodypart
social
person
visual
number
mental
place
tactile
time
outdoor
b.
Reading semantic map
PFC
MPC
LPC
AC
EVC
LTC
VTC
MPC
PFC
LPC
EVC
AC
LTC
VTC
superior
anterior
LH
superior
RH
anterior


S1
S2
S3
S4
S5
S6
S7
S8
S9
mean
chance
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Correlation coefficient
(Listening, Reading)
PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
PC9
PC10
Semantic PCs


a.
Correlation between listening and reading model weights
PFC
MPC
LPC
AC
EVC
LTC
VTC
MPC
PFC
LPC
EVC
AC
LTC
VTC
superior
anterior
LH
superior
RH
anterior
Weight Correlation (r)
0.0
1.0
NS
Reading and Listening
Listening only
Reading only
b.
Reading Prediction Accuracy
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.4
-0.2
0.0
0.2
0.4
0.6
Listening Prediction Accuracy


a.
Listening predicting Reading
S7
PFC
MPC
LPC
AC
EVC
LTC
VTC
MPC
PFC
LPC
EVC
AC
LTC
VTC
Model Prediction Accuracy (r)
0.0
0.5
NS
0.05
0.001
10-05
10-09
10-15
10-21
p value (FDR) corrected
superior
anterior
LH
superior
RH
anterior
b.
S8
S9
S1
S2
S5
S3
S4
S6


a.
Reading predicting Listening
S7
PFC
MPC
LPC
AC
EVC
LTC
VTC
MPC
PFC
LPC
EVC
AC
LTC
VTC
Model Prediction Accuracy (r)
0.0
0.5
NS
0.05
0.001
10^{-05}
10^{-09}
10^{-15}
10^{-21}
p value (FDR) corrected
superior
anterior
LH
superior
RH
anterior
b.
S8
S9
S1
S2
S5
S3
S4
S6


a.
Cross-modality vs. Within-modality
S7
PFC
MPC
LPC
AC
EVC
LTC
VTC
MPC
PFC
LPC
EVC
AC
LTC
VTC
superior
anterior
LH
Within-modality
max(L2L, R2R)
0.4
0.0
NS
0.0
0.4
Cross-modality
mean(L2R, R2L)
superior
RH
anterior
b.
S8
S9
S1
S2
S5
S3
S4
S6