Semantic representations during language comprehension are affected by context

Fatma Deniz*[a, b], Christine Tseng*[a], Leila Wehbe[c], Jack L. Gallant[a,d]

[a]Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720, USA
[b]Institute of Software Engineering and Theoretical Computer Science, Technische Universität Berlin, Berlin, Germany
[c]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[d]Department of Psychology, University of California, Berkeley, CA 94720, USA
*all authors contributed equally and are listed alphabetically

Corresponding author: Jack L. Gallant <gallant@berkeley.edu>

Abbreviated title: Context affects SNR and semantic representations

Number of pages: 41

Number of figures: 5

Number of words: Abstract: 250; Significance Statement: 85; Introduction: 648; Discussion: 1437

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

Acknowledgments: CT was supported by the National Science Foundation (NSF Nat-1912373) and UC Berkeley internal funds. FD was supported by the FITweltweit-Program of the German Academic Exchange Service (DAAD), the Moore and Sloan Data Science Environment Postdoctoral Fellowship, and the Federal Ministry of Education and Research (BMBF 01GQ1906). LW was supported by startup funds from the Machine Learning Department at Carnegie Mellon University.

Abstract

The meaning of words in natural language depends crucially on context. However, most neuroimaging studies of word meaning use isolated words and isolated sentences with little context. Because the brain may process natural language differently from how it processes simplified stimuli, there is a pressing need to determine whether prior results on word meaning generalize to natural language. We investigated this issue by directly comparing the brain representation of semantic information across four conditions that vary in context. fMRI was used to record human brain activity while four subjects (two female) read words presented in four different conditions: narratives (Narratives), isolated sentences (Sentences), blocks of semantically similar words (Semantic Blocks), and isolated words (Single Words). Using a voxelwise encoding model approach, we find two clear and consistent effects of increasing context. First, stimuli with more context (Narratives, Sentences) evoke brain responses with substantially higher SNR across bilateral visual, temporal, parietal, and prefrontal cortices compared to stimuli with little context (Semantic Blocks, Single Words). Second, increasing context increases the representation of semantic information across bilateral temporal, parietal, and prefrontal cortices at the group level. However, in individual subjects, only natural language stimuli (Narratives) consistently evoke widespread representation of semantic information across the cortical surface. These results show that context has large effects on both the quality of neuroimaging data and on the representation of meaning in the brain, and they imply that the results of neuroimaging studies that use stimuli with little context may not generalize well to the natural regime.

Significance Statement

Context is an important part of understanding the meaning of natural language, but most neuroimaging studies of meaning use isolated words and isolated sentences with little context. Here we examined whether the results of neuroimaging studies that use out-of-context stimuli generalize to natural language. We find that increasing context improves the quality of neuroimaging data and changes where semantic information is represented in the brain. These results suggest that findings from studies using out-of-context stimuli may not generalize to natural language used in daily life.

Introduction

Language is our main means of communication and an integral part of daily life. Natural language comprehension requires extracting meaning from words that are embedded in context. However, most neuroimaging studies of word meaning use simplified stimuli consisting of isolated words or sentences (Price, 2012). Natural language differs from isolated words and sentences in several ways. Natural language contains phonological and orthographic patterns, lexical semantics, syntactic structure, and compositional- and discourse-level semantics embedded in social context (Hagoort, 2019). In contrast, isolated words and sentences only contain a few of these components (e.g., lexical meaning, local syntactic structure). (For concision, this paper will refer to all differences between natural language and isolated words/sentences as differences in "context.")

Neuroimaging studies that use isolated words and sentences implicitly assume that their results will generalize to natural language. However, because the brain is a highly nonlinear dynamical system (Wu et al., 2006; Breakspear, 2017), the representation of semantic information may change depending on context (Poeppel et al., 2012; Hagoort, 2019; Hamilton and Huth, 2020). Indeed, contextual effects have been demonstrated clearly in other domains. For example, many neurons in the visual system respond differently to simplified stimuli compared to naturalistic stimuli (Simoncelli and Olshausen, 2001; Ringach et al., 2002; David et al., 2004; Touryan et al., 2005). However, few studies have examined whether insights about semantic representation from studies using simplified stimuli will generalize to natural language.

Results from past studies suggest that context has a large effect on semantic representation. Several natural language studies from our lab reported that semantic information is represented in a large, distributed network of brain regions including bilateral temporal, parietal, and prefrontal cortices (Huth et al., 2016; Deniz et al., 2019). In contrast, studies that used isolated words or sentences as stimuli only identified a few brain regions that represent semantic information (left IFG, anterior temporal lobe, inferotemporal cortex, and posterior parietal cortex; for reviews see (Binder et al., 2009; Price, 2010, 2012)).

One way that context might affect neuroimaging results is by affecting the signal-to-noise ratio (SNR) of evoked brain responses. Although no language studies have explicitly looked at evoked BOLD SNR, several converging lines of evidence suggest that context does affect evoked SNR in language studies. (Lerner et al., 2011) examined how language context affects cross-subject correlations in brain responses, and they reported that as the amount of context increased, the number of voxels that were correlated across subjects also increased. In addition, several contrast-based fMRI language studies reported that increasing context evoked larger and more widespread patterns of brain activity (Mazoyer et al., 1993; Xu et al., 2005; Jobard et al., 2007). Finally, most subjects are more attentive when reading natural stories than when reading isolated words, and attention affects BOLD SNR (Bressler and Silver, 2010).

Another more interesting way that context might affect neuroimaging results is by directly changing semantic representations in the brain. Context can change the way that subjects attend to semantic information, and semantic representations in many brain areas shift toward attended semantic categories (Çukur et al., 2013; Sprague et al., 2015; Nastase et al., 2017). Context also changes the statistical structure of language stimuli, and these statistical changes can affect cognitive processes and representations in a variety of ways (Wu et al., 2006; Dahmen et al., 2010; Breakspear, 2017).

To test the hypotheses that context affects evoked SNR and semantic representations, we used fMRI and a voxelwise encoding model approach to directly compare four stimulus conditions that vary in context: Narratives, Sentences, Semantic Blocks, and Single Words (Figure 1). The Narratives condition consisted of four narrative stories used in our previous studies (Huth et al., 2016; Deniz et al., 2019; Popham et al., 2021). The other three conditions used sentences, blocks of semantically similar words, and individual words sampled from the narratives in Huth et al. (2016), Deniz et al. (2019), and Popham et al. (2021).

Materials and Methods

Experimental Design and Statistical Analysis

Subjects. Functional data were collected from two males and two females: S1 (male, age 31), S2 (male, age 24), S3 (female, age 24), S4 (female, age 23). All subjects were healthy and had normal hearing, and normal or corrected-to-normal vision. All subjects were right handed according to the Edinburgh handedness inventory (Oldfield, 1971). Laterality scores were +70 (decile R.3) for S1, +95 (decile R.9) for S2, +90 (decile R.7) for S3, +80 (decile R.5) for S4.

MRI data collection. MRI data were collected on a 3T Siemens TIM Trio scanner with a 32-channel Siemens volume coil, located at the UC Berkeley Brain Imaging Center. Functional scans were collected using gradient echo EPI 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. Thirty axial slices were prescribed to cover the entire cortex and were scanned in interleaved order. A custom-modified bipolar water excitation radiofrequency (RF) pulse was used to avoid signal from fat. Anatomical data were collected using a T1-weighted multi-echo MP-RAGE sequence on the same 3T scanner. Approximately 3.5 hours (214.85 minutes) of fMRI data was collected for each subject.

fMRI data pre-processing. The FMRIB Linear Image Registration Tool (FLIRT) from FSL 5.0 (Jenkinson and Smith, 2001; Jenkinson et al., 2002) was used to motion-correct each functional run. A high-quality template volume was then created for each run by averaging all volumes in the run across time. FLIRT was used to automatically align the template volume for each run to an overall template, which was chosen to be the temporal average of the first functional run for each subject.

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.

A 3rd order Savitsky-Golay filter with a 121-TR window was used to identify low-frequency voxel response drift. This drift was subtracted from the signal before further processing. Responses for each run were z-scored separately before voxelwise modeling. In addition, 10 TRs were discarded from the beginning and the end (20 TRs total) of each run.

Cortical surface reconstruction and visualization. Freesurfer (Dale et al., 1999) was used to generate cortical surface meshes from the T1-weighted anatomical scans. Before surface reconstruction, Blender and pycortex (http://pycortex.org; (Gao et al., 2015)) were used to carefully hand-check and correct anatomical surface segmentations. To aid in cortical flattening, Blender and pycortex were used to remove the surface crossing the corpus callosum and relaxation cuts were made into the surface of each hemisphere. The calcarine sulcus cut was made at the horizontal meridian in V1 as identified from retinotopic mapping data.

Pycortex (Gao et al., 2015) was used to align functional images to the cortical surface. The line-nearest scheme in pycortex was used to project functional data onto the surface for visualization and subsequent analysis. The line-nearest scheme samples the functional data at 64 evenly-spaced intervals between the inner (white matter) and outer (pial) surfaces of the cortex and averages the samples. Samples are taken using nearest-neighbor interpolation, in which each sample is given the value of its enclosing voxel.

Stimuli. Stimuli for all four conditions were generated from ten spoken stories from The Moth Radio

Hour (used previously in (Huth et al., 2016)). In each story, a speaker tells an autobiographical story in front of a live audience. The ten selected stories are 10-15 min long, cover a wide range of topics, and are highly engaging. Transcriptions of these stories were used to generate the stimuli.

Story transcription. 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 transcribed 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 and Liberman, 2008)) was used to automatically align the audio to the transcript. P2FA uses a phonetic hidden Markov model to find the temporal onset and offset of each word and phoneme. The Carnegie Mellon University 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 pronouncing dictionary.

After automatic alignment was complete, Praat (Boersman and Weenink, 2014) was used to manually check and correct each aligned transcript. The corrected, aligned transcript was then spot-checked for accuracy by a different listener. Finally, Praat's TextGrid object was used to convert the aligned transcripts into word representations. The word representation of each story is a list of pairs (W, t), where W is a word and t is the time in seconds.

Stimulus Conditions. To evaluate the effect of context on evoked SNR and semantic representation in the brain, four stimulus conditions with different amounts of context were created. These four conditions were Narratives, Sentences, Semantic Blocks, and Single Words.

The Narratives condition consisted of four narratives from The Moth Radio Hour ("undertheinfluence", "souls", "life", “wheretheressmoke”). These four narratives were chosen from the ten narratives used in (Huth et al., 2016). Each narrative was presented in a separate ~10-minute scanning run. One narrative ("wheretheressmoke") was used as the model validation stimulus, and it was presented twice for each subject.

The Sentences condition consisted of sentences randomly sampled from the ten narratives used in (Huth et al., 2016). Sentence boundaries were marked manually, resulting in 1450 sentences with a median sentence length of 13 words (min=5 words, max=40 words). Sentences were presented in four unique ~10-minute scanning runs. One run was used as the model validation stimulus, and it was presented twice for each subject.

The Semantic Blocks condition consisted of blocks of clustered words from the ten narratives used in (Huth et al., 2016). The word clusters were designed to elicit maximally different voxel responses. To create the clusters, each word was first transformed into its semantic model representation (see Voxelwise model fitting below). The semantic model representation for each word was then projected onto the first ten principal components of the semantic model weights estimated in (Huth et al., 2016). Finally, the projections were clustered with k-means clustering (k=12) to create 12 word clusters. During each scanning run, subjects saw 12 different blocks of 114 words each. The words in each block were sampled from one of the word clusters, and eight different word clusters were sampled in each run. The frequency with which each cluster was sampled was matched to the frequency with which words from that cluster appeared in the ten narratives. Blocks were presented in four unique ~10-minute long runs. One run was used as the model validation stimulus, and it was presented twice for each subject.

The Single Words condition consisted of words randomly sampled without replacement from the ten narratives used in (Huth et al., 2016). There were 21743 appearances of 2868 unique words across the narratives, and each appearance was sampled uniformly. Words were presented in four unique

10-minute scanning runs. One run was used as the model validation stimulus, and it was presented twice for each subject.

For the Sentences, Semantic Blocks, and Single Words conditions, text descriptions of auditory sounds (e.g., laughter and applause) in the ten narratives were removed. In addition, obvious transcription errors were removed from the list of narrative words for the Semantic Blocks and Single Words conditions. Words that did not make sense by themselves (e.g., "tai", "chi") were also removed. There were five such words: "tai", "chi", "deja", "vu", and "sub."

Stimulus presentation. In all conditions, words were presented individually at the center of the screen using Rapid Serial Visual Presentation (RSVP) (Forster, 1970; Buchweitz et al., 2009). Words in the Narratives and Sentences conditions were presented with the same timing and duration as in the original spoken stories. Words in the Semantic Blocks and Single Words conditions were presented for a baseline of 400 ms with an additional 10 ms for every character. For example, the word "apple" would be presented for 400 ms + 10 ms/character * (5 characters) = 450 ms.

The pygame library in Python was used to display black text on a gray background at 34 horizontal and 27 vertical degrees of visual angle. 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. Subjects were asked to fixate while reading the text. Eye movements were monitored at 60 Hz throughout the scanning sessions using a custom-built camera system equipped with an infrared source (Avotec) and the ViewPoint EyeTracker software suite (Arrington Research). The eye tracker was calibrated before each session of data acquisition.

Explainable variance (EV). To measure the functional SNR of each stimulus condition, we computed the explainable variance (EV). EV was computed as the amount of variance in the response of a voxel that can be explained by the mean response of the voxel across multiple repetitions of the same stimulus. Formally, if the responses of a voxel to a repeated stimulus is expressed as a matrix Y with dimensions (# of TRs in each repetition, # of stimulus repetitions), then EV is given by

1 - [variance(Y - mean(Y, axis=1)) / variance(Y)].

Note that this is the same as the coefficient of determination (R2) where the model prediction is the mean response across stimulus repetitions. For each condition, EV was computed from the two repeated validation runs.

Voxelwise model fitting and validation. To identify voxels that represent semantic information, a linearized finite impulse response (FIR) encoding model (Nishimoto et al., 2011; Huth et al., 2012, 2016) was fit to every cortical voxel in each subject's brain. The linearized FIR encoding model consisted of one feature space designed to represent semantic information in the stimuli, and four feature spaces designed to represent low-level linguistic information. In the semantic feature space, the semantic content of each word was represented by the word's co-occurrence statistics with the 985 words in Wikipedia's List of 1000 basic words (Huth et al., 2016). Thus, each word was represented by a 985-long vector in the semantic feature space. The co-occurrence statistics were computed over a large text corpus that included the ten narrative stories used in Huth et al. (2016), several books from Project Gutenberg, a wide variety of Wikipedia pages, and a broad selection of reddit.com user comments (Huth et al., 2016). The four low-level feature spaces were word rate (1 parameter), letter rate (1 parameter), letters (26 parameters), and word length variation per TR (1 parameter). Together, the five feature spaces had 1014 features.

The features passed through three additional preprocessing steps before being fit to BOLD responses. First, to account for the hemodynamic response, a separate linear temporal filter with four delays was fit for each of the 1014 features, resulting in 4056 final features. This was accomplished by concatenating copies of the features delayed by 1, 2, 3, and 4 TRs (approximately 2, 4, 6, and 8 seconds). Taking the dot product of this concatenated feature space with a set of linear weights is functionally equivalent to convolving the undelayed features with a linear temporal kernel that has non-zero entries for 1-, 2-, 3-, and 4-time point delays. Second, 10 TRs were discarded from the beginning and the end (20 TRs total) of each run. Third, each feature was z-scored separately within each run. This was done so that the features would be on the same scale as the BOLD responses, which were also z-scored within each run.

A single joint model consisting of the 4056 features were fit to BOLD responses using banded ridge regression (Nunez-Elizalde et al., 2019) and the himalaya Python package (see Code Accessibility). A separate model was fit for every voxel in every subject and condition. For every model, a regularization parameter was estimated for each of the five feature spaces using a random search. In the random search, 1000 normalized hyperparameter candidates were sampled from a Dirichlet distribution and scaled by 30 log-spaced values ranging from 10^-5 to 10^20. The best normalized hyperparameter candidate and scaling were selected for each feature space for each voxel. Finally, models were fit again on the BOLD responses with the selected hyperparameters.

To validate the models, estimated feature weights were used to predict responses to a separate, held-out validation dataset. Validation stimuli for the Narratives condition consisted of two repeated presentations of the narrative “wheretheressmoke” (Huth et al., 2016). Validation stimuli for the Sentences, Semantic Blocks, and Single Words conditions consisted of two repeated presentations of one run for each condition. Prediction accuracy was then computed as the Pearson’s correlation coefficient between the model-predicted BOLD response and the average BOLD response across the two validation runs. Statistical significance for each condition was computed with permutation testing. A null distribution was generated by permuting 10-TR blocks of the average validation BOLD response 5000 times and computing the prediction accuracy for each permutation. Resulting p values were corrected for multiple comparisons within each subject using the false discovery rate (FDR) procedure (Benjamini and Hochberg, 1995).

All model fitting and analysis was performed using custom software written in Python, making heavy use of NumPy (Oliphant, 2006) and SciPy (Jones et al., 2001). Analysis and visualizations were developed using iPython (Perez and Granger, 2007), and the interactive programming and visualization environment jupyter notebook (Kluyver et al., 2016).

Code Accessibility. The himalaya package is publicly available on GitHub (https://github.com/gallantlab/himalaya).

Results

The goal of this study was to understand whether context affects evoked SNR and semantic representations in the brain. Previous studies suggest that both evoked SNR and semantic representations will differ across the four experimental conditions (Single Words, Semantic Blocks, Sentences, and Narratives). Here, we analyzed evoked SNR and semantic representations for each of the four conditions in individual subjects.

To estimate evoked SNR, we computed the reliability of voxel responses across repetitions of the same stimulus. Several different sources of noise can influence the variability of voxel responses across stimulus repetitions: magnetic inhomogeneity, voxel response variability, and variability in subject attention or vigilance. Because these sources are independent across stimulus repetitions, pooling voxel responses across repetitions averages out the noise and provides a good estimate of the evoked SNR. In this study, we used explainable variance (EV) as a measure of reliability and computed the EV for two repetitions of one run in each condition to estimate evoked SNR (see Methods).

Figure 3 shows EV for the four conditions in one typical subject (S1) (see Extended Data Figure 3-1 for voxels with significant EV; see Extended Data Figure 3-2 for unthresholded EV for subjects 2-4). In the Single Words condition, appreciable EV is only found in a few scattered voxels located in bilateral primary visual cortex, STS, and IFG (Figure 3a). The number of voxels with significant EV ($p<0.05$, FDR-corrected) in the Single Words condition is 256, 1198, 0, and 0 for subjects 1-4, respectively. A similar pattern is seen in the Semantic Blocks condition, where appreciable EV is only found in a few scattered voxels located in bilateral primary visual cortex, STS, and IFG (Figure 3b). The number of voxels with significant EV ($p<0.05$, FDR-corrected) in the Semantic Blocks condition is 324, 1613, 1201, and 0 for subjects 1-4, respectively. In contrast, both the Sentences and Narratives conditions produce high EV in many voxels located in bilateral visual, parietal, temporal, and prefrontal cortices (Figures 3c and 3d). The number of voxels with significant EV ($p<0.05$, FDR-corrected) in the Sentences condition is 4225, 11697, 2359, and 7251 for subjects 1-4, respectively. The number of voxels with significant EV ($p<0.05$, FDR-corrected) in the Narratives condition is 7622, 8062, 7059, and 2931 for subjects 1-4, respectively. Together, these results show that increasing context increases evoked SNR in bilateral visual, temporal, parietal, and prefrontal cortices.

To quantify semantic representation, we used a voxelwise encoding model (VM) procedure and a semantic feature space to identify voxels that represent semantic information in each condition (Figure 2). We first extracted semantic features from the stimulus words in each condition separately (see Methods). We then used banded ridge regression (Nunez-Elizalde et al., 2019) to fit a separate semantic encoding model for each voxel, subject, and condition. Here we refer to voxels that were predicted significantly by the semantic model (see Methods) as “semantically selective voxels.”

Figure 4 shows semantic model prediction accuracy for semantically selective voxels for the four conditions in one typical subject (S1) (see Extended Data Figure 4-1 for additional subjects; see Extended Data Figure 4-2 for unthresholded semantic model prediction accuracy for all subjects). In the Single Words condition, no voxels are semantically selective in any of the four subjects (Figure 4a, $p<0.05$, FDR corrected). In the Semantic Blocks condition, scattered voxels along the left STS and left IFG are semantically selective (Figure 4b, $p<0.05$, FDR corrected). The number of semantically selective voxels ($p<0.05$, FDR corrected) in the Semantic Blocks condition is 652, 0, 811, and 0 for subjects 1-4, respectively. In the Sentences condition, voxels in the left angular gyrus, left STG, bilateral STS, bilateral ventral precuneus, bilateral ventral premotor speech area (sPMv), bilateral superior frontal sulcus (SFS), and left superior frontal gyrus (SFG) are semantically selective (Figure 4c, $p<0.05$, FDR corrected). The number of semantically selective voxels ($p<0.05$, FDR-corrected) in the Sentences condition is 1626, 3099, 0, and 0 for subjects 1-4, respectively. Finally, in the Narratives condition, voxels in bilateral angular gyrus, bilateral STS, bilateral STG, bilateral temporal parietal junction (TPJ), bilateral sPMv, bilateral ventral precuneus, bilateral SFS, bilateral SFG, bilateral inferior frontal gyrus, left inferior parietal lobule (IPL), and left posterior cingulate gyrus are semantically selective (Figure 4d, $p<0.05$, FDR corrected). The number of semantically selective voxels ($p<0.05$, FDR-corrected) in the Narratives condition is 4505, 7340, 7607, and 1791 for subjects 1-4, respectively. Together, these results suggest that increasing context increases the representation of semantic information in bilateral temporal, parietal, and prefrontal cortices. These results also suggest that this effect is highly variable in individual subjects for non-natural language stimuli (Semantic Blocks, Sentences) but not for natural language stimuli (Narratives).

The results presented in Figure 4 were obtained in each subject’s native brain space. To determine how the representation of semantic information varies across subjects for the four conditions, we transformed the semantic encoding model results obtained for each subject into the standard MNI brain space (Deniz et al., 2019). Figure 5 shows the mean unthresholded model prediction accuracy across subjects (Figure 5a-d) and the number of subjects for which each voxel is semantically selective (Figure 5e-h) for each condition. In the Single Words condition, no voxels are semantically selective in any of the four subjects (Figure 5a and 5e, $p<0.05$, FDR corrected). In the Semantic

Blocks condition, scattered voxels in left STS are semantically selective in two out of four subjects (Figure 5b and 5f, $p<0.05$, FDR corrected). In the Sentences condition, voxels in the bilateral STS, left STG, bilateral ventral precuneus, bilateral angular gyrus, bilateral SFS, and bilateral premotor cortex are semantically selective in two out of four subjects (Figure 5c and 5g, $p<0.05$, FDR corrected). Finally, in the Narratives condition, voxels in bilateral angular gyrus, bilateral STS, right STG, right anterior temporal lobe, bilateral SFS and SFG, left IFG, left IPL, bilateral ventral precuneus, and bilateral posterior cingulate gyrus are semantically selective in all subjects (Figure 5d and 5h, $p<0.05$, FDR corrected), and voxels in left STG and right IFG are semantically selective in three out of four subjects (Figure 5d and 5h, $p<0.05$, FDR corrected). These results are consistent with those in Figure 4, and they suggest that increasing stimulus context increases the representation of semantic information across the cortical surface at the group level. In addition, this effect is inconsistent across individual subjects for non-natural stimuli (Semantic Blocks, Sentences) but not natural stimuli (Narratives).

Because the Narratives condition contains more contextual information than the other three conditions, we hypothesized that we would find more semantically selective voxels in the Narratives condition than in the other three conditions. To test this, we calculated the difference in the number of semantically selective voxels between the Narratives condition and each of the other three conditions. The difference between the Narratives and Single Words conditions is 4505, 7340, 7607, and 1791 voxels for subjects 1-4, respectively ($p<0.05$ for all subjects). The difference between the Narratives and Semantic Blocks conditions is 3853, 7340, 6796, and 1791 voxels for subjects 1-4, respectively ($p<0.05$ for all subjects). Finally, the difference between the Narratives and Sentences conditions is 2879, 4241, 7607, and 1791 voxels for subjects 1-4, respectively ($p<0.05$ for all subjects). The difference between the Narratives and Single Words conditions partly reflects the fact that most voxels have low evoked SNR in the Single Words condition and high evoked SNR in the Narratives condition (Figure 3). Because it is impossible to model noise, differences in evoked SNR across conditions directly affect the number of voxels that achieve a significant model fit. The difference between the Narratives and Semantic Blocks conditions also partly reflects differences in evoked SNR -- for most voxels, evoked SNR is low in the Semantic Blocks condition and high for the Narratives condition (Figure 3). In contrast, the evoked SNR is high for many voxels in both the Narratives and the Sentences conditions (Figure 3), so the difference in the number of semantically selective voxels is unlikely to be due to differences in evoked SNR. Instead, this result suggests that semantic information is represented more widely across the cortical surface in the Narratives condition than in the Sentences condition.

Discussion

The aim of this study was to determine whether and how context affects semantic representations in the human brain. Our results show that both evoked SNR and semantic representations are affected by the amount of context in the stimulus. First, stimuli with relatively more context (Narratives, Sentences) evoke brain responses with higher SNR compared to stimuli with relatively less context (Semantic Blocks, Single Words) (Figure 3). Second, increasing the amount of context increases the representation of semantic information across the cortical surface at the group level (Figures 4, 5). However, in individual subjects, only the Narratives condition consistently increased the representation of semantic information compared to the Single Words condition (Figures 4, 5). These results imply that neuroimaging studies that use isolated words or sentences do not fully map the functional brain representations that underlie natural language comprehension.

Our observations that increasing context increases both the evoked SNR and the cortical representation of semantic information at the group level are fully consistent with results from previous neuroimaging studies. Several previous studies found that stimuli with more context evoke larger, more widespread patterns of brain activity (Mazoyer et al., 1993; Xu et al., 2005; Jobard et al., 2007), that brain activity evoked for individual words is modulated by context (Just et al., 2017), and that brain activity evoked by stimuli with more context are more reliable than those evoked by stimuli with less context (Lerner et al. 2011). Furthermore, previous studies that used narrative stimuli (Wehbe et al., 2014; Huth et al., 2016; Pereira et al., 2018; Deniz et al., 2019; Hsu et al., 2019; Popham et al., 2021) identified many more voxels involved in semantic processing than studies that used isolated words or sentences (for reviews see (Binder et al., 2009; Price, 2010, 2012)).

However, there are several important differences between the results we reported here and those reported in previous neuroimaging studies. First, the 2011 study by Lerner et al. only found voxels with reliable responses in bilateral temporal gyrus and posterior inferior frontal sulcus when using isolated sentences. In contrast, we found many voxels with high EV across bilateral temporal, parietal, and frontal cortices in the Sentences condition (Figure 3). Second, past studies that used isolated sentences found left IFG involved in semantic processing (Constable et al., 2004; Rodd et al., 2005; Humphries et al., 2007). In contrast, we did not find any semantically selective voxels in the Sentences condition for two out of four subjects. Of the remaining two subjects, only one subject had semantically selective voxels in left IFG in the Sentences condition (Figures 4 and 5). Third, past studies that used isolated words found bilateral STS, bilateral lateral sulcus, left IFG, left MTG, and left ITG involved in semantic processing (Mazoyer et al., 1993; Booth et al., 2002; Xu et al., 2005; Jobard et al., 2007; Lerner et al., 2011). In contrast, we did not find any semantically selective voxels in the Single Words condition (Figures 4 and 5). Finally, one previous study looked at brain activity evoked by a stimulus conceptually similar to Semantic Blocks (Mollica et al., 2020). In the study, Mollica et al. (2020) used sentences that were scrambled such that nearby words could be combined into meaningful phrases. They found that the brain activity evoked by scrambled sentences was similar to the brain activity evoked by unscrambled sentences in left IFG, left middle frontal gyrus, left temporal lobe, and left angular gyrus. In contrast, we only found voxels that were semantically selective in both the Semantic Blocks and Sentences conditions in left STS (Figures 4 and 5).

The inconsistencies between this study and past studies most likely stem from four major methodological differences between this study and those earlier studies. First, we avoided smoothing our data before performing analyses. We performed our analyses for each subject in their native brain space, and we did not perform any spatial smoothing across voxels. In contrast, most previous studies performed normalization procedures to transform their data into a standard brain space and applied a spatial smoothing operation across voxels (Lindquist, 2008; Carp, 2012). Spatial smoothing and normalization procedures can incorrectly assign signal to voxels and average away meaningful signal and individual variability in language processing (Steinmetz and Seitz, 1991; Fedorenko and Kanwisher, 2009; Fedorenko et al., 2012; Huth et al., 2016; Deniz et al., 2019). Thus, brain regions identified by past studies may be more relevant at the group level than in individual subjects. These smoothing procedures likely contribute to the inconsistencies observed between past studies and this study.

Second, we used an explicit computational model to identify semantically selective voxels. In contrast, most previous studies identified semantic brain regions by contrasting different experimental conditions (Binder et al., 2008, 2009; Price, 2012). Although past studies designed their experimental conditions to isolate brain activity involved in semantic processing (Binder et al., 2008, 2009), there could be unexpected differences unrelated to semantic processing between the conditions. For example, experiments that contrast a semantic task with a phonological task (Binder et al., 2008, 2009) may have task difficulty as a confound. As a result, it is possible that some semantic brain areas identified by past studies are actually involved in processing the unexpected differences rather than semantics. We would likely not have identified such brain areas in this study, since our semantic model only contains information about semantics.

Third, we evaluated semantic model prediction accuracy on a separate, held-out validation dataset. In contrast, most previous studies drew inferences from analyses performed on only one dataset without a validation dataset (Binder et al., 2009). Performing analyses on only one dataset can lead to inflated results that are overfit to the dataset (Soch et al., 2016). Thus, some semantic brain areas identified by past studies may only be relevant for the specific stimuli, experimental design, or data used in those studies. Such study-specific brain areas would not generalize to other studies, such as this study.

Finally, subjects in our study passively read the stimulus words, which allowed us to directly compare the Narratives condition with the other three conditions. In contrast, many past studies of semantic processing used active tasks involving lexical decisions (Binder et al., 2003), matching (Vandenberghe et al., 1996), or monitoring (Démonet et al., 1992). Active tasks are thought to increase subject engagement, which can increase evoked BOLD SNR. Thus, if we had used an active task, the effect of context on evoked SNR might have been even larger than the differences that we report here. In addition, different active tasks can affect semantic processing differently in the brain (Toneva et al., 2020). Therefore, task effects likely contributed to the inconsistencies observed between past studies and this study.

Our study used only one semantic model, and that model determined which specific voxels were identified as semantically selective. Because this model likely captures some narrative information that is correlated with word-level semantic information, some of the brain activity predicted by our semantic model may actually reflect higher-level linguistic or domain-general representations (Fedorenko et al., 2012; Blank and Fedorenko, 2017). Furthermore, other studies have proposed alternative models that integrate contextual semantic information differently than the model used here (Jain and Huth, 2018; Toneva and Wehbe, 2019), and it is possible that these other models might predict voxel activity better than the semantic model used here. The voxelwise modeling framework provides a straightforward method for evaluating alternative semantic models directly by construction of appropriate feature spaces. Therefore, a valuable direction for future research would be to examine other semantic models, and to include language models that explicitly account for factors such as narrative structure, metaphor, and humor.

In conclusion, our results show that increasing the amount of stimulus context increases both the SNR of evoked brain responses and the representation of semantic information in the brain at the group level. In addition, we find that only natural language stimuli (Narratives) consistently evoke widespread representation of semantic information across the cortical surface in individual subjects. These results imply that neuroimaging studies that use isolated words or sentences to study semantic processing may provide a misleading picture of semantic language comprehension in daily life. Although natural language stimuli are much more complex than isolated words and sentences, the development and validation of the voxelwise encoding model framework for language processing (Huth et al., 2016; de Heer et al., 2017; Deniz et al., 2019; Popham et al., 2021) has made it possible to rigorously test hypotheses about semantic processing with natural language stimuli. To ensure that the results of neuroimaging study generalize to natural language processing, we suggest that future studies of semantic processing should use more naturalistic stimuli.

References

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

Binder JR, Desai RH, Graves WW, Conant LL (2009) Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. Cereb Cortex 19:2767–2796.

Binder JR, McKiernan KA, Parsons ME, Westbury CF, Possing ET, Kaufman JN, Buchanan L (2003) Neural correlates of lexical access during visual word recognition. J Cogn Neurosci 15:372–393.

Binder JR, Swanson SJ, Hammeke TA, Sabsevitz DS (2008) A comparison of five fMRI protocols for mapping speech comprehension systems. Epilepsia 49:1980–1997.

Blank I, Fedorenko E (2017) Domain-general brain regions do not track linguistic input as closely as language-selective regions. J Neurosci:3642–3616.

Boersman P, Weenink D (2014) Praat: doing phonetics by computer (Version 5.3. 56). Amsterdam:

Praat Available at: https://scholar.google.ca/scholar?cluster=3307900219265089918&hl=en&as_sdt=0,5&sciodt=0,5.

Booth JR, Burman DD, Meyer JR, Gitelman DR, Parrish TB, Mesulam MM (2002) Modality independence of word comprehension. Hum Brain Mapp 16:251–261.

Breakspear M (2017) Dynamic models of large-scale brain activity. Nat Neurosci 20:340–352.

Bressler DW, Silver MA (2010) Spatial attention improves reliability of fMRI retinotopic mapping signals in occipital and parietal cortex. Neuroimage 53:526–533.

Buchweitz A, Mason RA, Tomitch LMB, Just MA (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.

Carp J (2012) The secret lives of experiments: methods reporting in the fMRI literature. Neuroimage 63:289–300.

Constable RT, Pugh KR, Berroya E, Mencl WE, Westerveld M, Ni W, Shankweiler D (2004) Sentence complexity and input modality effects in sentence comprehension: an fMRI study. Neuroimage 22:11–21.

Çukur T, Nishimoto S, Huth AG, Gallant JL (2013) Attention during natural vision warps semantic representation across the human brain. Nat Neurosci 16:763.

Dahmen JC, Keating P, Nodal FR, Schulz AL, King AJ (2010) Adaptation to stimulus statistics in the perception and neural representation of auditory space. Neuron 66:937–948.

Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9:179–194.

David SV, Vinje WE, Gallant JL (2004) Natural stimulus statistics alter the receptive field structure of v1 neurons. J Neurosci 24:6991–7006.

de Heer WA, Huth AG, Griffiths TL, Gallant JL, Theunissen FE (2017) The Hierarchical Cortical Organization of Human Speech Processing. J Neurosci 37:6539–6557.

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

Deniz F, Nunez-Elizalde AO, Huth AG, Gallant JL (2019) The Representation of Semantic Information Across Human Cerebral Cortex During Listening Versus Reading Is Invariant to Stimulus Modality. The Journal of Neuroscience 39:7722–7736 Available at: http://dx.doi.org/10.1523/jneurosci.0675-19.2019.

Fedorenko E, Duncan J, Kanwisher N (2012) Language-Selective and Domain-General Regions Lie Side by Side within Broca's Area. Curr Biol 22:2059–2062.

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

Forster KI (1970) Visual perception of rapidly presented word sequences of varying complexity. Percept Psychophys 8:215–221.

Gao JS, Huth AG, Lescroart MD, Gallant JL (2015) Pycortex: an interactive surface visualizer for fMRI. Front Neuroinform 9:23.

Hagoort P (2019) The neurobiology of language beyond single-word processing. Science 366:55–58.

Hamilton LS, Huth AG (2020) The revolution will not be controlled: natural stimuli in speech neuroscience. Lang Cogn Neurosci 35:573–582.

Hsu C-T, Clariana R, Schloss B, Li P (2019) Neurocognitive Signatures of Naturalistic Reading of Scientific Texts: A Fixation-Related fMRI Study. Sci Rep 9:1–16.

Humphries C, Binder JR, Medler DA, Liebenthal E (2007) Time course of semantic processes during sentence comprehension: an fMRI study. Neuroimage 36:924–932.

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532:453–458.

Huth AG, Nishimoto S, Vu AT, Gallant JL (2012) A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. Neuron 76:1210–1224.

Jain S, Huth AG (2018) Incorporating Context into Language Encoding Models for fMRI. bioRxiv:327601 Available at: https://www.biorxiv.org/content/10.1101/327601v2.

Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825–841.

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

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

Jones E, Oliphant T, Peterson P (2001) SciPy: Open Source Scientific Tools for Python. Available at: https://www.semanticscholar.org/paper/307827ec09187e9c6935e8ff5fd43eeefb901320.

Just MA, Wang J, Cherkassky VL (2017) Neural representations of the concepts in simple sentences: Concept activation prediction and context effects. Neuroimage 157:511–520.

Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick JB, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, Willing C, Jupyter Development Team (2016) Jupyter Notebooks - a publishing format for reproducible computational workflows. ELPUB Available at: https://www.semanticscholar.org/paper/e47868841d87efe261451a43b00d6c81cf7fb7a3.

Lerner Y, Honey CJ, Silbert LJ, Hasson U (2011) Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J Neurosci 31:2906–2915.

Lindquist MA (2008) The Statistical Analysis of fMRI Data. Stat Sci 23:439–464.

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

Mollica F, Siegelman M, Diachek E, Piantadosi ST, Mineroff Z, Futrell R, Kean H, Qian P, Fedorenko

E (2020) Composition is the Core Driver of the Language-selective Network. Neurobiology of Language 1:104–134.

Nastase SA, Connolly AC, Oosterhof NN, Halchenko YO, Guntupalli JS, Visconti di Oleggio Castello M, Gors J, Gobbini MI, Haxby JV (2017) Attention Selectively Reshapes the Geometry of Distributed Semantic Representation. Cereb Cortex 27:4277–4291.

Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL (2011) Reconstructing visual experiences from brain activity evoked by natural movies. Curr Biol 21:1641–1646.

Nunez-Elizalde AO, Huth AG, Gallant JL (2019) Voxelwise encoding models with non-spherical multivariate normal priors. Neuroimage 197:482–492.

Oliphant TE (2006) A guide to NumPy.

Pereira F, Lou B, Pritchett B, Ritter S, Gershman SJ, Kanwisher N, Botvinick M, Fedorenko E (2018) Toward a universal decoder of linguistic meaning from brain activation. Nat Commun 9:1–13.

Perez F, Granger BE (2007) IPython: A System for Interactive Scientific Computing. Computing in Science and Engg 9:21–29.

Poeppel D, Emmorey K, Hickok G, Pylkkänen L (2012) Towards a new neurobiology of language. J Neurosci 32:14125–14131.

Popham SF, Huth AG, Bilenko NY, Deniz F, Gao JS, Nunez-Elizalde AO, Gallant JL (2021) Visual and linguistic semantic representations are aligned at the border of human visual cortex. Nat Neurosci 24:1628–1636.

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

Price CJ (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.

Ringach DL, Hawken MJ, Shapley R (2002) Receptive field structure of neurons in monkey primary visual cortex revealed by stimulation with natural image sequences. J Vis 2:12–24.

Rodd JM, Davis MH, Johnsrude IS (2005) The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. Cereb Cortex 15:1261–1269.

Simoncelli EP, Olshausen BA (2001) Natural image statistics and neural representation. Annu Rev Neurosci 24:1193–1216.

Soch J, Haynes J-D, Allefeld C (2016) How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection. Neuroimage 141:469–489.

Sprague TC, Saproo S, Serences JT (2015) Visual attention mitigates information loss in small- and large-scale neural codes. Trends Cogn Sci 19:215–226.

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

Toneva M, Stretcu O, Póczos B, Wehbe L, Mitchell TM (2020) Modeling Task Effects on Meaning Representation in the Brain via Zero-Shot MEG Prediction. In: Advances in Neural Information

Processing Systems.

Toneva M, Wehbe L (2019) Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In: Advances in Neural Information Processing Systems, pp 14928–14938.

Touryan J, Felsen G, Dan Y (2005) Spatial structure of complex cell receptive fields measured with natural images. Neuron 45:781–791.

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

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

Wu MC-K, David SV, Gallant JL (2006) Complete functional characterization of sensory neurons by system identification. Annu Rev Neurosci 29:477–505.

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

Yuan J, Liberman M (2008) Speaker identification on the SCOTUS corpus. Proceedings of Acoustics.

Figures and Figure legends

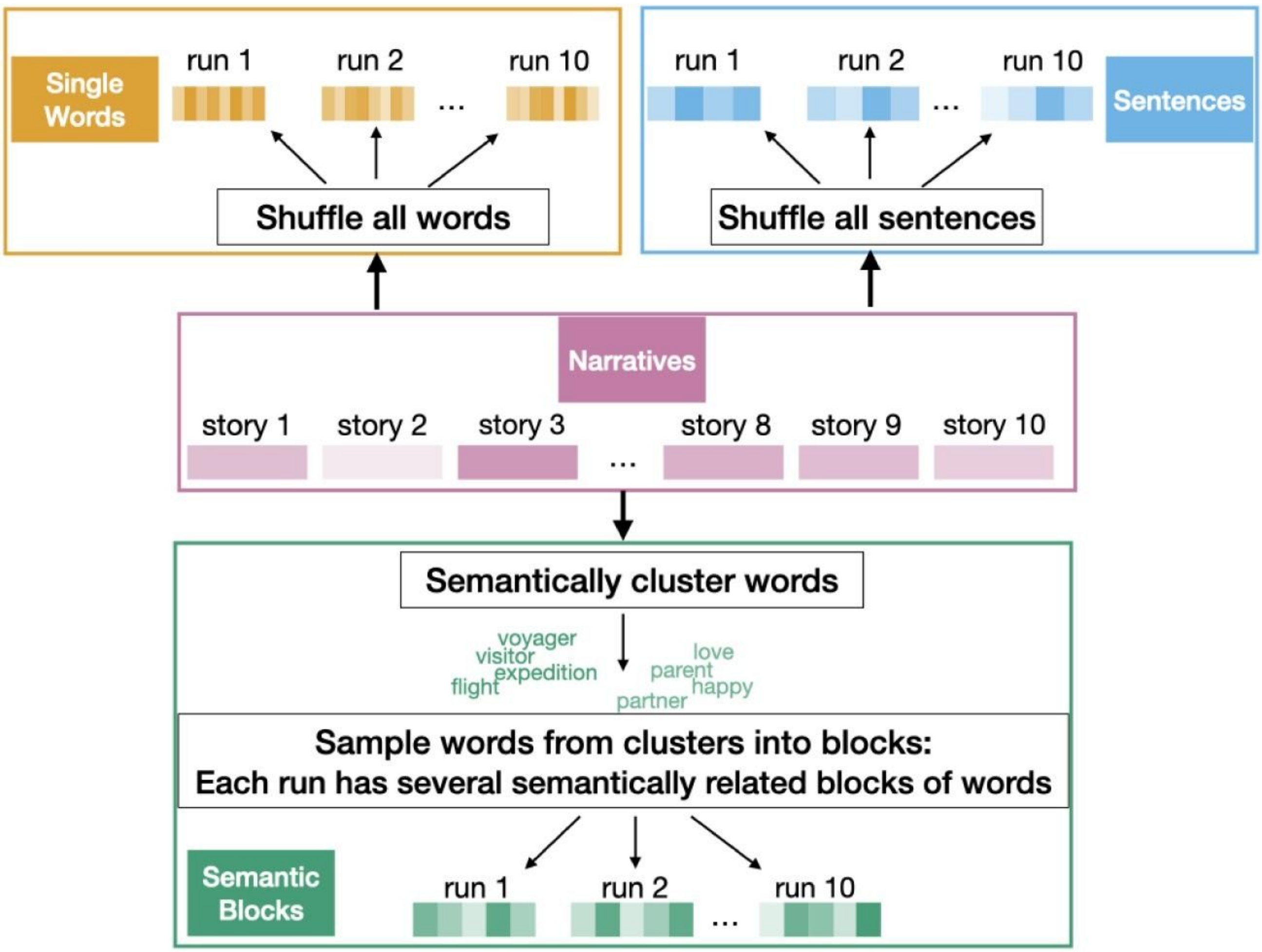


Figure 1: Stimulus conditions. The experiment contained four stimulus conditions that were based on the ten narratives used in Huth et al. (2016). The Single Words condition consisted of words sampled randomly from the ten narratives. The Semantic Blocks condition consisted of blocks of words sampled from clusters of semantically similar words from the ten narratives. There were 12 distinct clusters of semantically similar words, and blocks of words were created by randomly sampling 114 words from one word cluster for each block. The Sentences condition consisted of sentences sampled randomly from the ten narratives. Finally, the Narratives condition consisted of the ten original narratives.

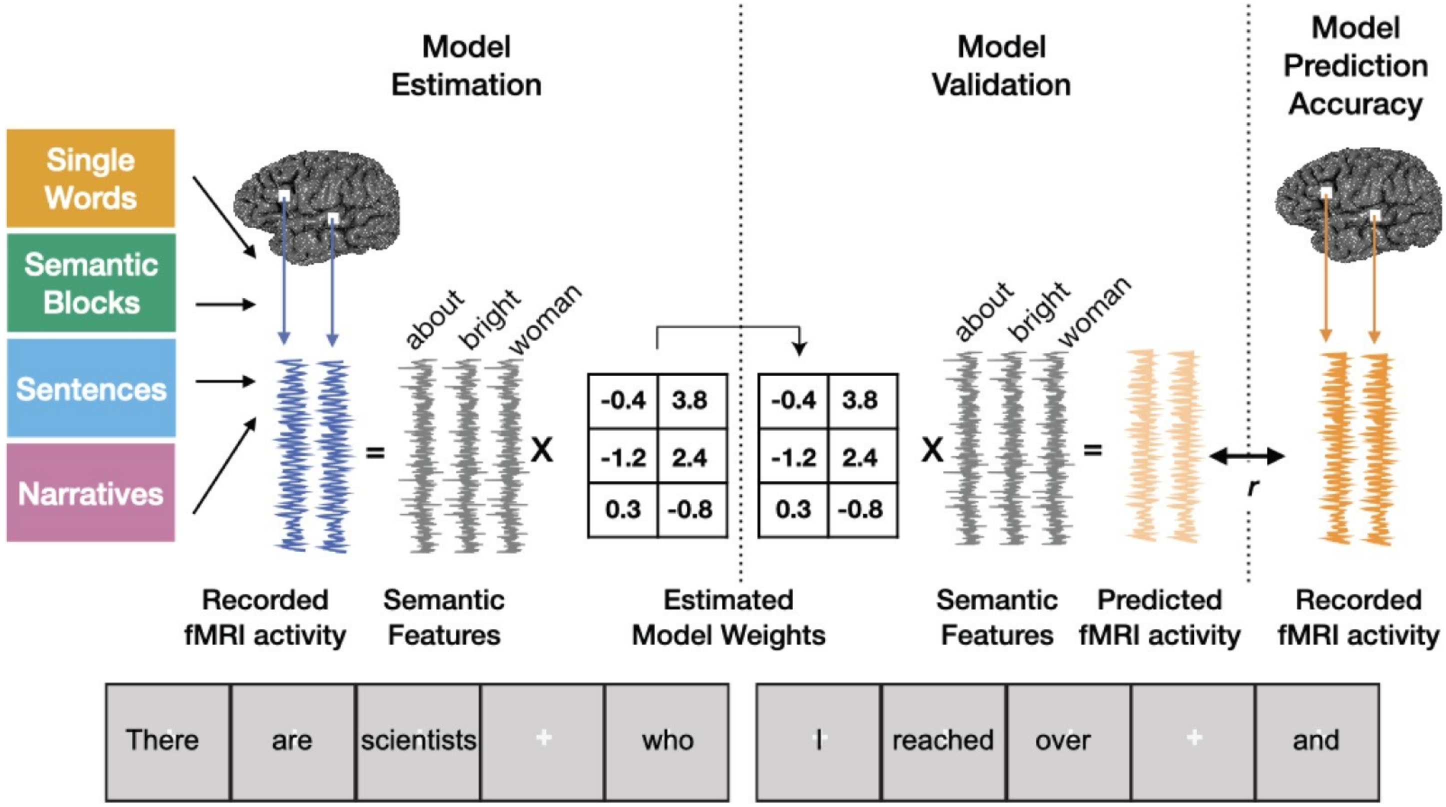


Figure 2: Voxelwise Modeling. Four subjects read words from the four stimulus conditions while BOLD responses were recorded. Each stimulus word was projected into a 985-dimensional word embedding space that was independently constructed using word co-occurrence statistics from a large corpus (Semantic Features). A finite impulse response (FIR) regularized regression model was estimated separately for each voxel in every subject and condition using banded ridge regression (Nunez-Elizalde et al. 2019). The estimated model weights were then used to predict BOLD responses to a separate, held-out validation stimulus. Model prediction accuracy was quantified as the correlation (r) between the predicted and recorded BOLD responses to the validation stimulus.

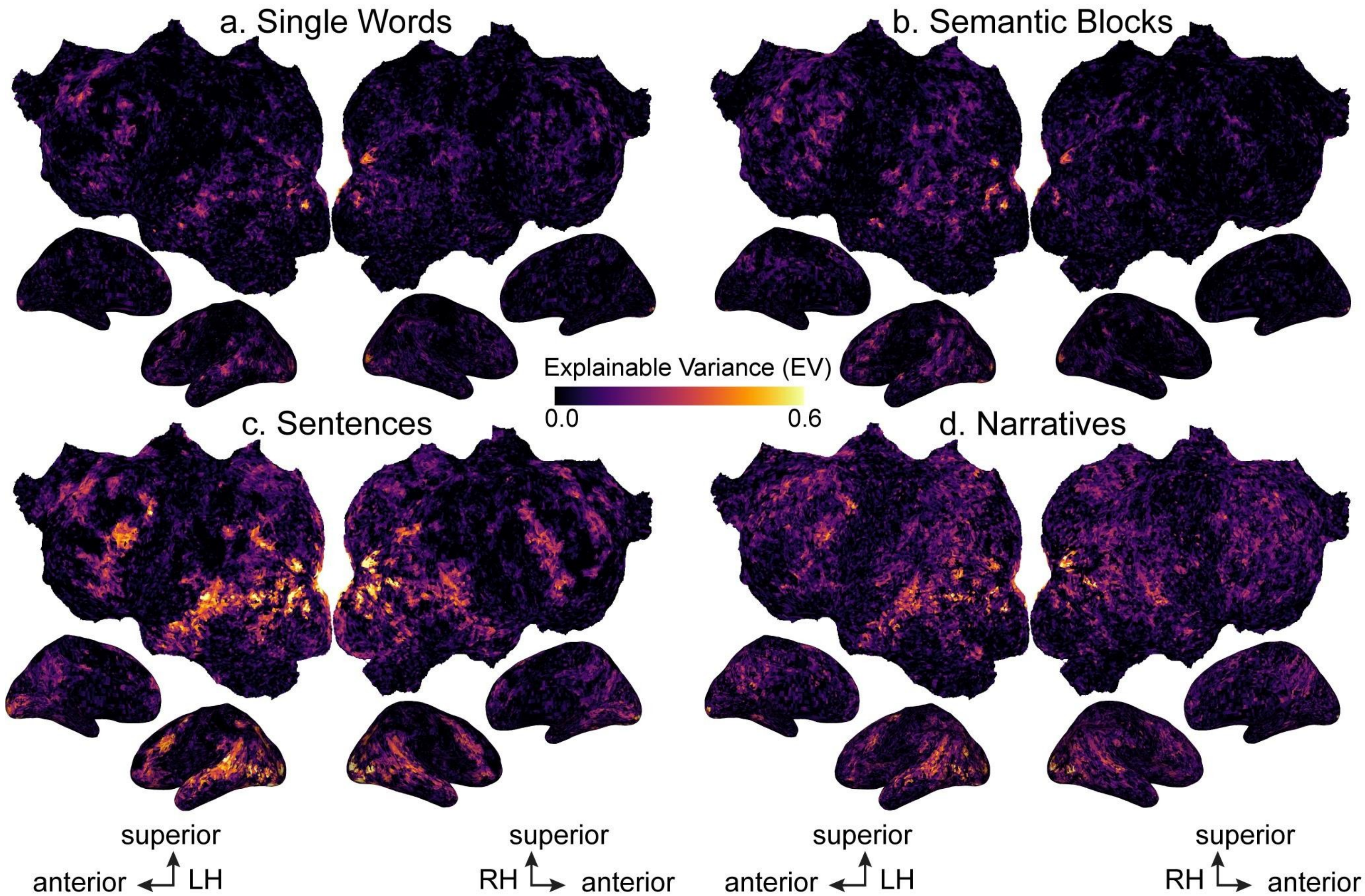


Figure 3. Explainable variance (EV) for the four conditions across the cortical surface. EV for the four conditions is shown for one subject (S1) on the subject's flattened cortical surface. EV was computed as an estimate of the evoked signal-to-noise ratio (SNR). Here EV is given by the color scale shown in the middle, and voxels that have high EV (i.e., high evoked SNR) appear yellow. (LH: Left Hemisphere, RH: Right Hemisphere) The format is the same in all panels. **a.** EV was computed for the Single Words condition and is shown on the flattened cortical surface of subject S1. Scattered voxels in bilateral primary visual cortex, superior temporal sulcus (STS), and inferior frontal gyrus (IFG) have high EV. **b.** EV was computed for the Semantic Blocks condition. Similar to the Single Words condition, scattered voxels in bilateral primary visual cortex, STS, and IFG have high EV. **c.** EV was computed for the Sentences condition. Many voxels in bilateral visual, parietal, temporal, and prefrontal cortices have high EV. **d.** EV was computed for the Narratives condition. Similar to the Sentences condition, voxels in bilateral visual, parietal, temporal, and prefrontal cortices have high EV. Together, these results show that increasing context increases evoked SNR in bilateral visual,

temporal, parietal, and prefrontal cortices. (See Extended Data Figure 3-1 for significant EV voxels for subject S1 and Extended Data Figure 3-2 for EV for all subjects.)

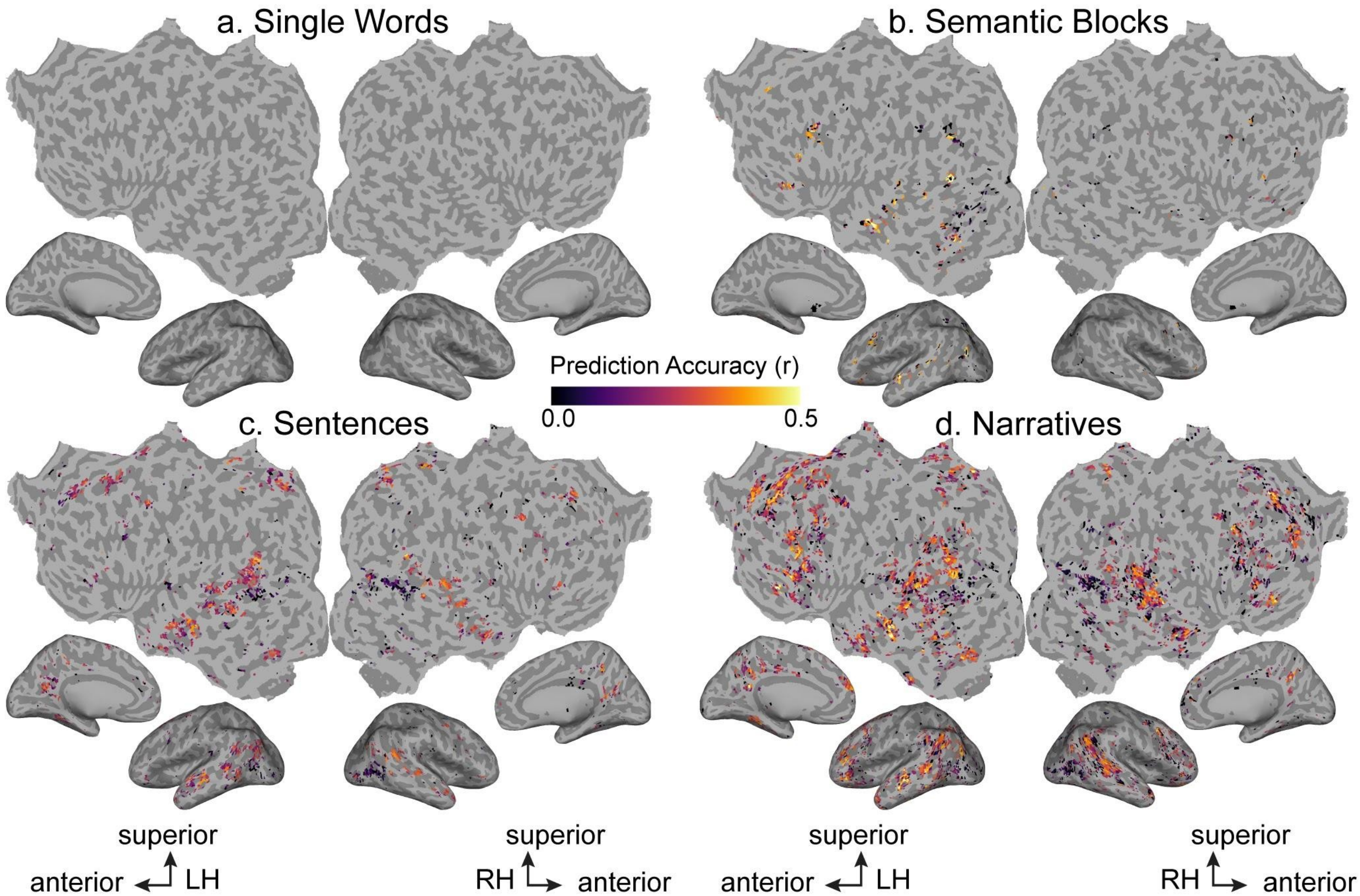


Figure 4. **Semantic model prediction accuracy for the four conditions across the cortical surface.** Semantic model prediction accuracy in the four conditions is shown on the flattened cortical surface of one subject (S1; see Extended Data Figure 4-1 and 4-2 for all subjects). Voxelwise modeling was first used to estimate semantic model weights in the four conditions. Semantic model prediction accuracy was then computed as the correlation (r) between the subject's recorded BOLD activity to the held-out validation stimulus and the BOLD activity predicted by the semantic model. In each panel, only voxels with significant semantic model prediction accuracy ($p<0.05$, FDR corrected) are shown. Prediction accuracy is given by the color scale in the middle, and voxels that have a high prediction accuracy appear yellow. Voxels for which the semantic model prediction accuracy is not statistically significant are shown in gray. (LH: Left Hemisphere, RH: Right Hemisphere) **a.** Semantic model prediction accuracy was computed for the Single Words condition. No voxels are significantly predicted in the Single Words condition. **b.** Semantic model prediction accuracy was computed for the Semantic Blocks condition. The format is the same as panel **a**. Voxels in left STS and IFG are

significantly predicted. **c.** Semantic model prediction accuracy was computed for the Sentences condition. The format is the same as panel **a**. Voxels in left angular gyrus, left STG, bilateral STS, bilateral ventral precuneus, bilateral ventral premotor speech area (sPMv), bilateral superior frontal sulcus (SFS), and left superior frontal gyrus (SFG) are significantly predicted. **d.** Semantic model prediction accuracy was computed for the Narratives condition. The format is the same as panel **a.** Voxels in bilateral angular gyrus, bilateral STS, bilateral STG, bilateral temporal parietal junction (TPJ), bilateral sPMv, bilateral ventral precuneus, bilateral SFS, bilateral SFG, bilateral IFG, left inferior parietal lobule (IPL), and left posterior cingulate gyrus are significantly predicted. Together, these results suggest that increasing context increases the representation of semantic information in bilateral temporal, parietal, and prefrontal cortices.

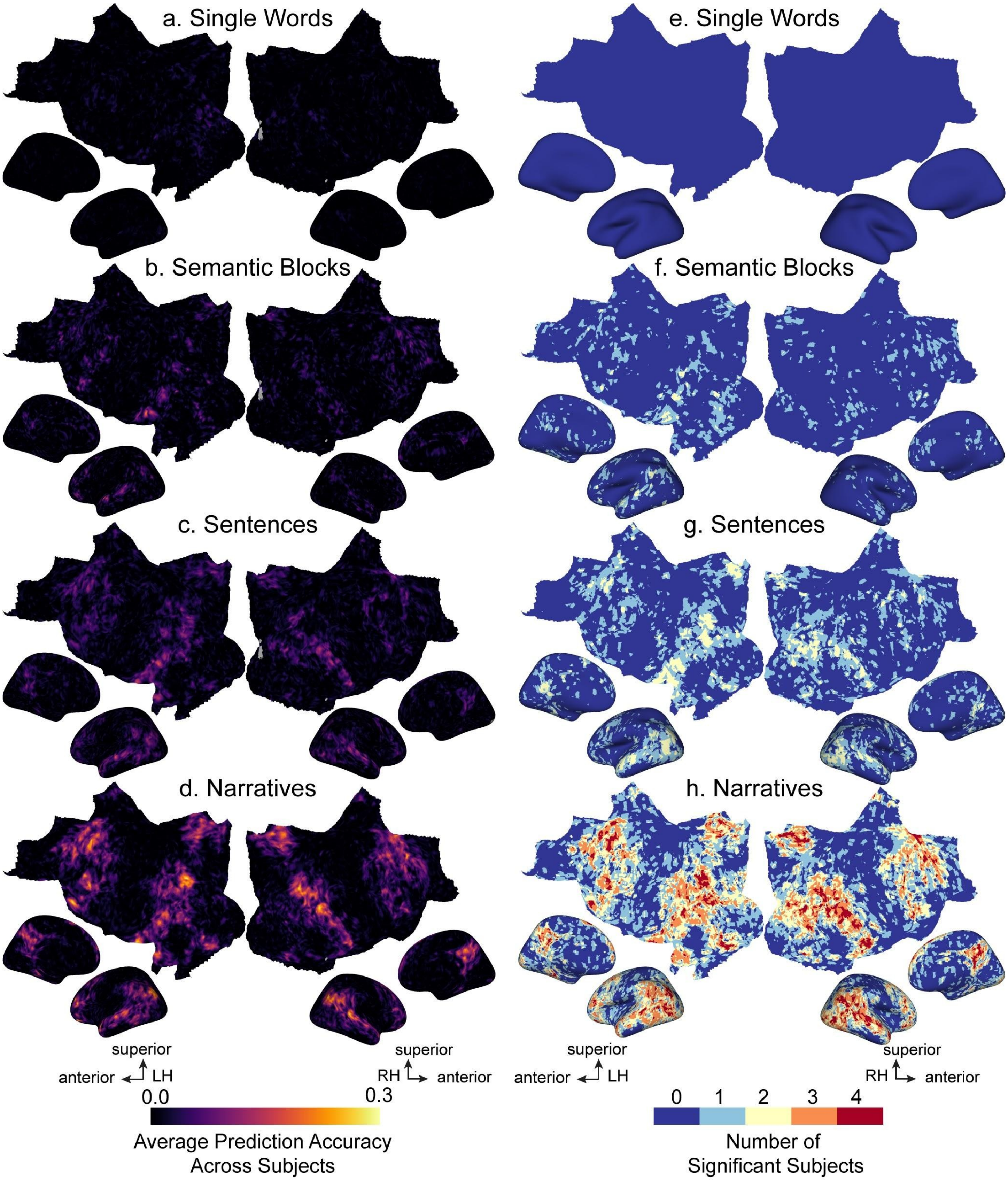


Figure 5. Semantic model prediction accuracy across all subjects for the four conditions in standard brain space. Semantic model prediction accuracy was first computed for each subject and for each condition as described in **Figure 4**. These individualized predictions were then projected into

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.15.472839; this version posted December 16, 2021.

the standard MNI brain space. **a.-d.** Average prediction accuracy across the four subjects is computed for each MNI voxel and shown for each condition on the cortical surface of the MNI brain. Average prediction accuracy is given by the color scale, and voxels with higher prediction accuracy appear brighter. **a.** In the Single Words condition, average prediction accuracy is low across the cortical surface. **b.** In the Semantic Blocks condition, average prediction accuracy is high in voxels in left anterior STS. **c.** In the Sentences condition, average prediction accuracy is high in bilateral STS, STG, anterior temporal lobe, angular gyrus, ventral precuneus, SFS, and SFG. **d.** In the Narratives condition, average prediction accuracy is very high in bilateral STS, STG, MTG, anterior temporal lobe, angular gyrus, IPL, ventral precuneus, posterior cingulate gyrus, Broca's area, IFG, SFS, SFG, and left posterior inferior temporal sulcus. **e.-h.** For each condition, statistical significance of prediction accuracies was determined in each subject's native brain space and then projected into the MNI brain space. The number of subjects with significant prediction accuracy is shown for each voxel on the cortical surface of the MNI brain. The number of significant subjects is given by the color scale shown at bottom. Dark red voxels are significantly predicted in all subjects, and dark blue voxels are not significantly predicted in any subjects. **e.** In the Single Words condition, no voxels are semantically selective for any subjects. **f.** In the Semantic Blocks condition, scattered voxels in left STS are semantically selective in two out of four subjects. **g.** In the Sentences condition, voxels in the bilateral STS, STG, angular gyrus, ventral precuneus, and SFS are semantically selective in two out of four subjects. **h.** In the Narratives condition, voxels in bilateral angular gyrus, bilateral STS, anterior temporal lobe, SFS, SFG, IFG, ventral precuneus, posterior cingulate gyrus, and right STG are semantically selective in all four subjects. The results shown here are consistent with those in **Figure 4**, and they suggest that increasing context increases the representation of semantic information across the cortical surface at the group level but not for individual subjects.

Extended Data Figure legends

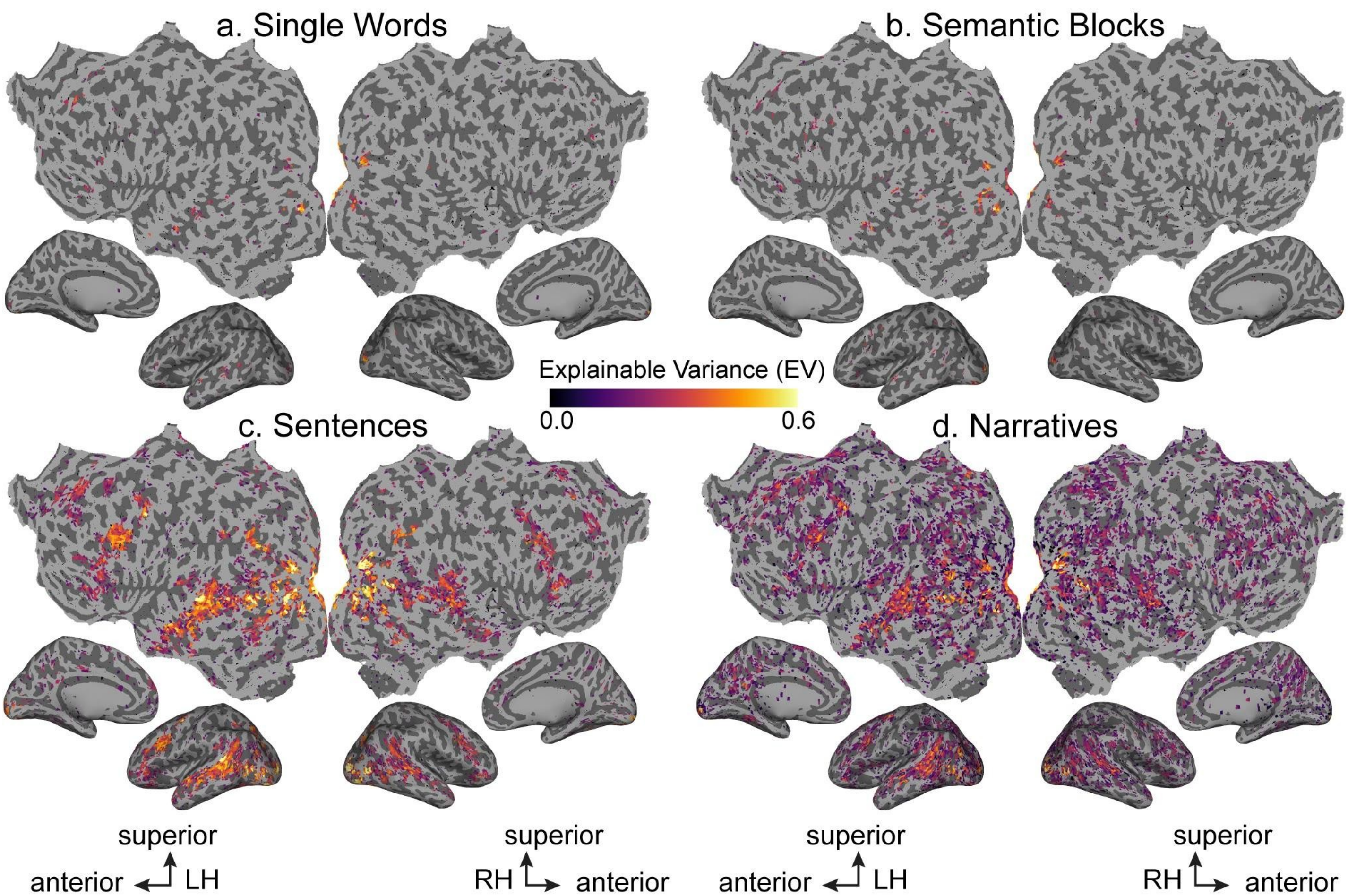
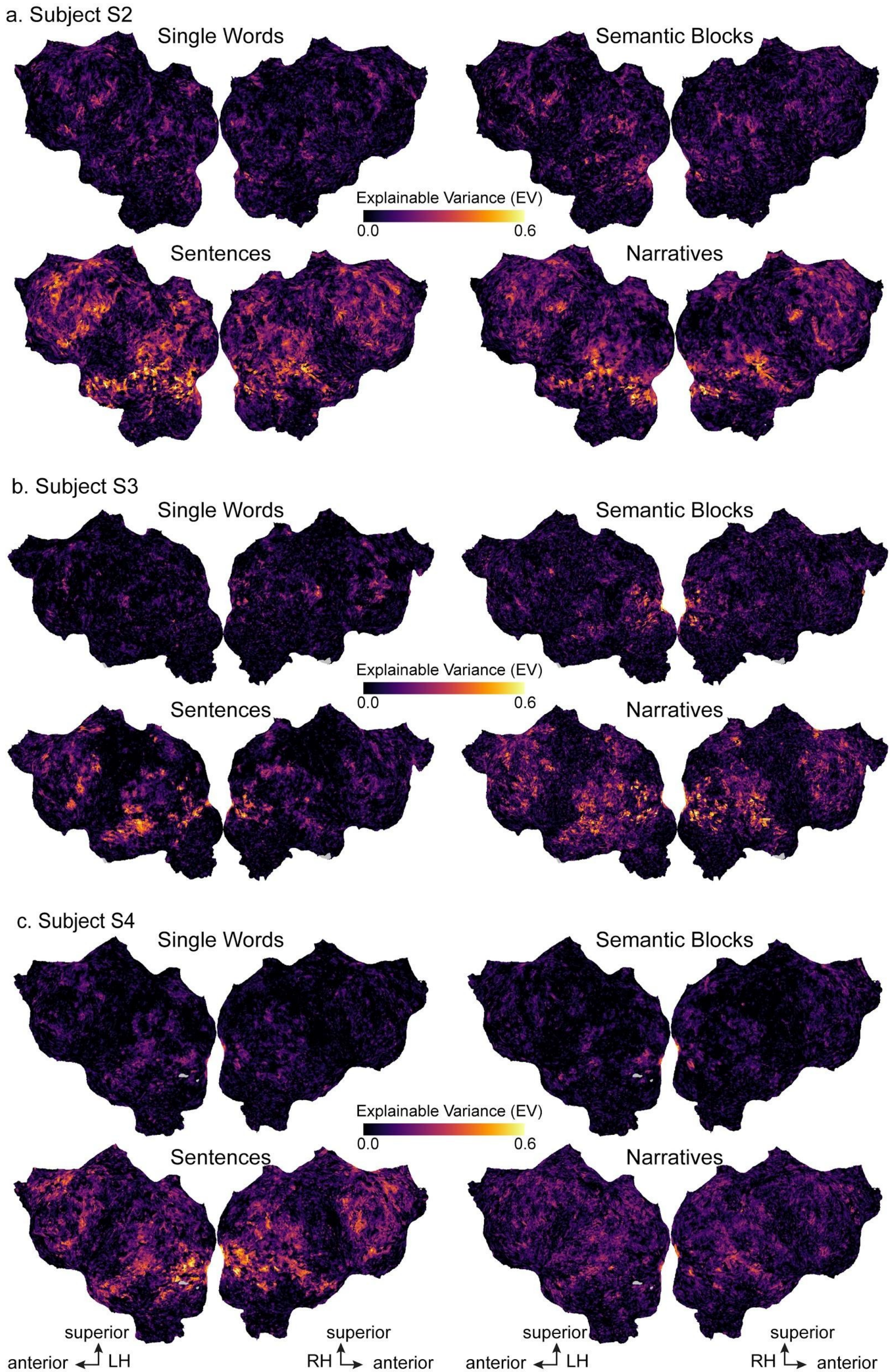


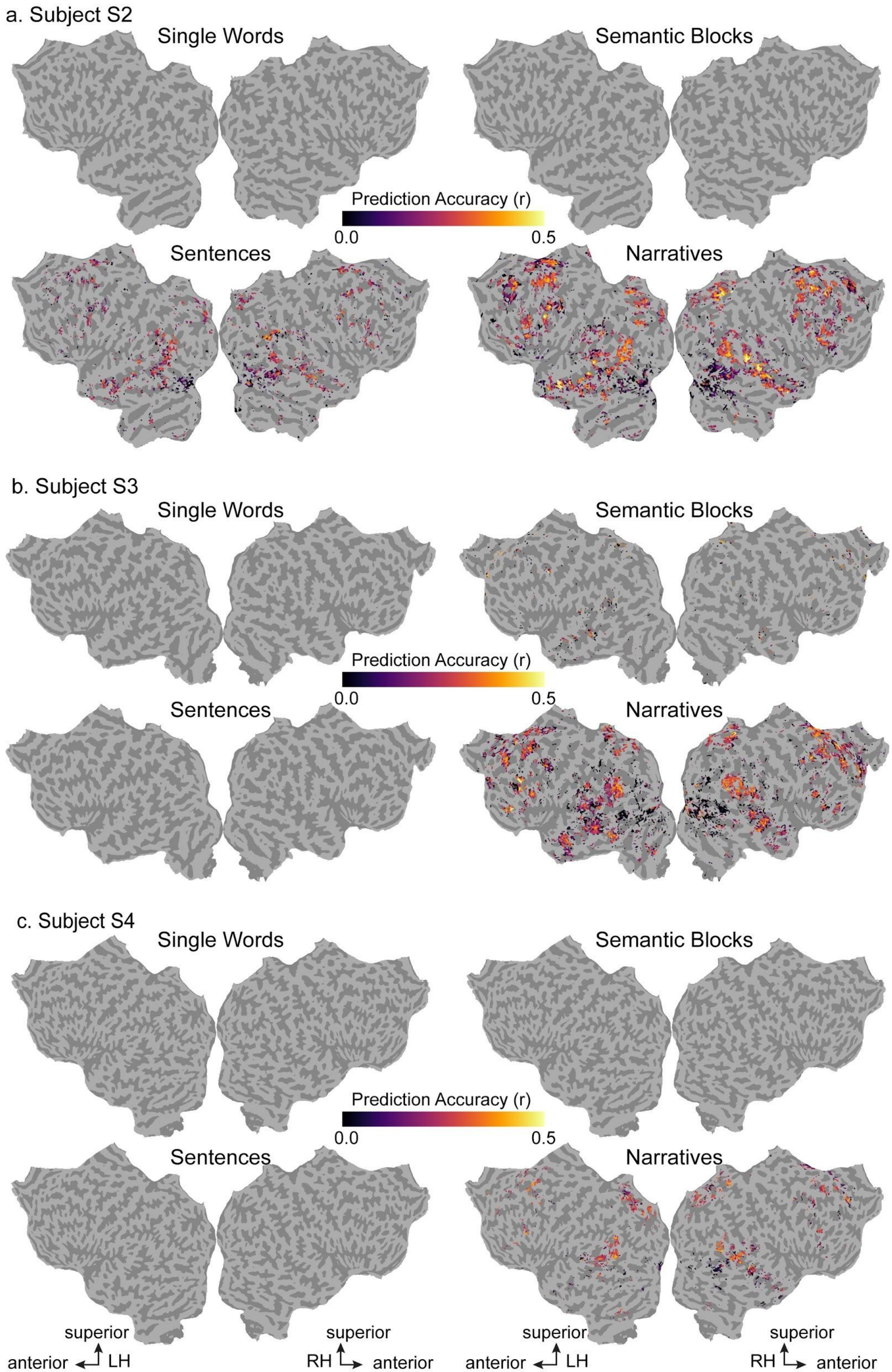
Figure 3-1. Significant explainable variance (EV) for the four conditions across the cortical surface. EV is shown for the four conditions on the flattened cortical surface of one subject (S1). EV was computed as an estimate of the evoked signal-to-noise ratio (SNR). Only voxels with significant EV ($p<0.05$, FDR corrected) are shown. EV is given by the color scale shown in the middle, and voxels that have high EV appear yellow. Voxels with EV values that are not statistically significant are shown in gray. (LH: Left Hemisphere, RH: Right Hemisphere) **a.** EV was computed for the Single Words condition, and significant voxels are shown on the flattened cortical surface of subject S1. Scattered voxels in bilateral primary visual cortex, left STS, and left IFG have significant EV. **b.** Same as panel **a**. but for the Semantic Blocks condition. Similar to the Single Words condition, scattered voxels in bilateral primary visual cortex, left STS, and left IFG have significant EV. **c.** Same as panel **a**. but for the Sentences condition. Many voxels in bilateral visual, parietal, temporal, and prefrontal

cortices have significant EV. **d.** Same as panel **a**. but for the Narratives condition. Similar to the Sentences condition, voxels in bilateral visual, parietal, temporal, and prefrontal cortices have high EV.


a. Subject S2
Single Words
Semantic Blocks
Explainable Variance (EV)
0.0
0.6
Sentences
Narratives
b. Subject S3
Single Words
Semantic Blocks
Explainable Variance (EV)
0.0
0.6
Sentences
Narratives
c. Subject S4
Single Words
Semantic Blocks
Explainable Variance (EV)
0.0
0.6
Sentences
Narratives
superior
anterior
LH
superior
RH
anterior
superior
anterior
LH
superior
RH
anterior

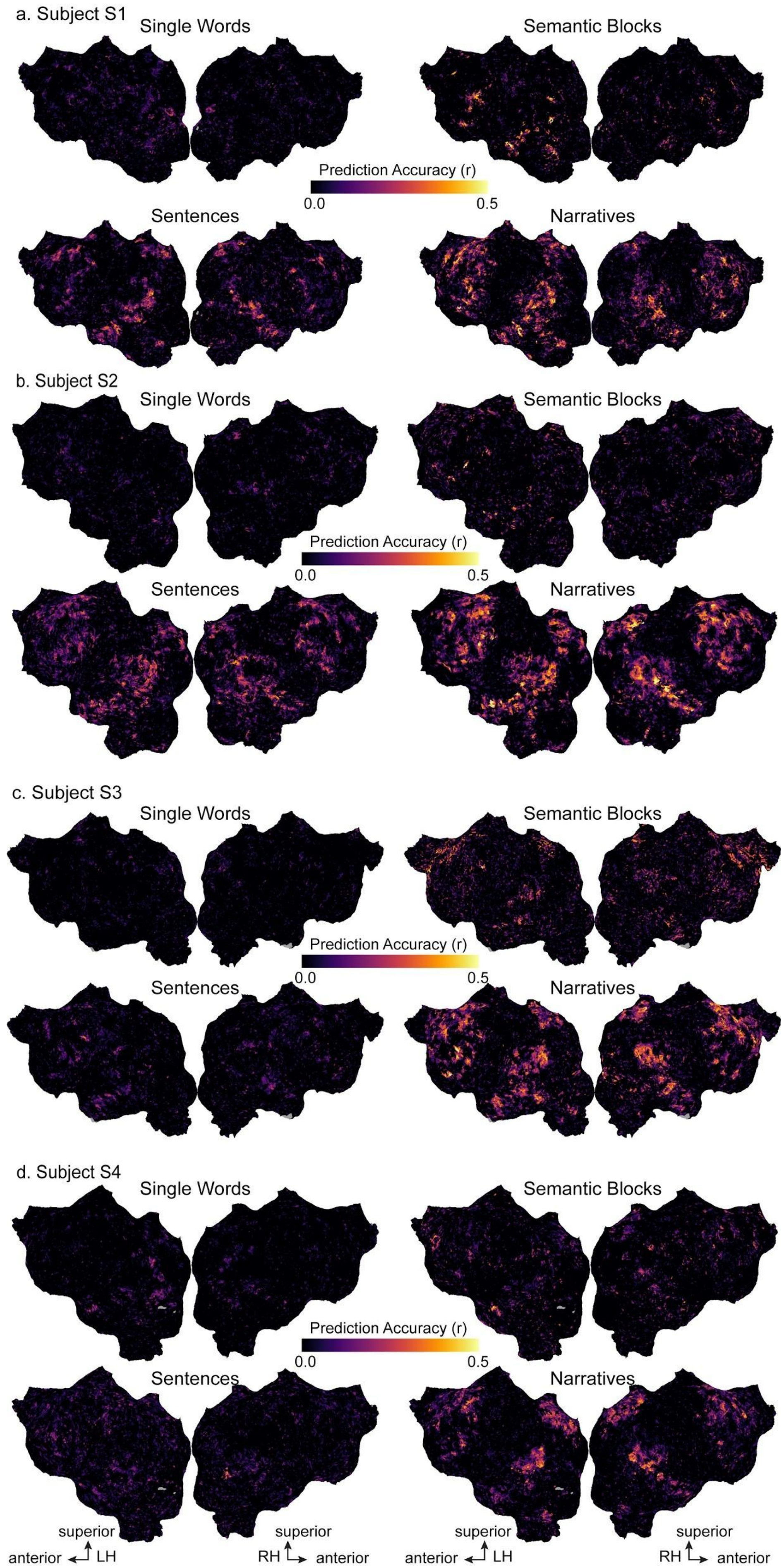
758

Figure 3-2. Explainable variance (EV) for the four conditions across the cortical surface for subjects S2, S3, and S4. EV is shown for the four conditions on the flattened cortical surface of subjects S2, S3 and S4. The format is the same as **Figure 3**. EV was computed as an estimate of the evoked signal-to-noise ratio (SNR). EV is given by the color scale shown in the middle, and voxels that have high EV (i.e., high evoked SNR) appear yellow. (LH: Left Hemisphere, RH: Right Hemisphere) Across all subjects, EV is low across most of the cortical surface in the Single Words and Semantic Blocks conditions. In contrast, EV is high for many voxels in bilateral visual, parietal, temporal, and prefrontal cortices in the Sentences and Narratives conditions.


a. Subject S2
Single Words
Semantic Blocks
Prediction Accuracy (r)
0.0
0.5
Sentences
Narratives
b. Subject S3
Single Words
Semantic Blocks
Prediction Accuracy (r)
0.0
0.5
Sentences
Narratives
c. Subject S4
Single Words
Semantic Blocks
Prediction Accuracy (r)
0.0
0.5
Sentences
Narratives
superior
anterior
LH
superior
RH
anterior
superior
anterior
LH
superior
RH
anterior

767

Figure 4-1. **Semantic model prediction accuracy for the four conditions across the cortical surface for subjects S2, S3, and S4.** Semantic model prediction accuracy in the four conditions is shown on the flattened cortical surface of subjects S2, S3 and S4. The format is the same as **Figure 4**. Voxelwise modeling was first used to estimate semantic model weights in the four conditions. Semantic model prediction accuracy was then computed as the correlation (r) between the subject's recorded BOLD activity to the held-out validation story and the BOLD activity predicted by the semantic model. In each panel, only voxels with significant semantic model prediction accuracy ($p<0.05$, FDR corrected) are shown. Prediction accuracy is given by the color scale in the middle, and voxels that have a high prediction accuracy appear yellow. Voxels with semantic model prediction accuracies that are not statistically significant are shown in gray. (LH: Left Hemisphere, RH: Right Hemisphere) In the Single Words condition, no voxels are significantly predicted in all subjects. In the Semantic Blocks condition, scattered voxels in left STS, left angular gyrus, left sPMv, and bilateral SFS are significantly predicted in subject S3. In the Sentences condition, voxels in bilateral STS, bilateral STG, bilateral angular gyrus, bilateral ventral precuneus, bilateral SFS and SFG, bilateral IFG, and bilateral sPMv are significantly predicted in subject S2. In the Narratives condition, voxels in bilateral angular gyrus, bilateral ventral precuneus, bilateral SFS and SFG, and right STS are significantly predicted in all three subjects. In addition, bilateral STG, left STS, bilateral Broca's area and IFG, and bilateral sPMv are significantly predicted in subjects S2 and S3.


a. Subject S1
Single Words
Semantic Blocks
Prediction Accuracy (r)
0.0
0.5
Sentences
Narratives
b. Subject S2
Single Words
Semantic Blocks
Prediction Accuracy (r)
0.0
0.5
Sentences
Narratives
c. Subject S3
Single Words
Semantic Blocks
Prediction Accuracy (r)
0.0
0.5
Sentences
Narratives
d. Subject S4
Single Words
Semantic Blocks
Prediction Accuracy (r)
0.0
0.5
Sentences
Narratives
superior
anterior LH
superior
RH anterior
superior
anterior LH
superior
RH anterior

786

Figure 4-2. **Un-thresholded semantic model prediction accuracy for the four conditions across the cortical surface for all subjects.** Un-thresholded semantic model prediction accuracy in the four conditions is shown for all subjects on each subject's flattened cortical surface. Voxelwise modeling was first used to estimate semantic model weights in the four conditions. Semantic model prediction accuracy was then computed as the correlation (r) between the subject's recorded BOLD activity to the held-out validation story and the BOLD activity predicted by the semantic model. Prediction accuracy is given by the color scale in the middle, and voxels that have a high prediction accuracy appear yellow. (LH: Left Hemisphere, RH: Right Hemisphere) In the Single Words condition, prediction accuracy is high in scattered voxels in primary visual cortex in subjects S1 and S4. In the Semantic Blocks condition, prediction accuracy is high in voxels in left STS and left angular gyrus in subjects S1 and S3. In addition, prediction accuracy is high in voxels in left Broca's area and IFG in subject S1, and prediction accuracy is high in voxels in bilateral SFS, SFG, and ventral precuneus in subject S3. In the Sentences condition, prediction accuracy is high in voxels in bilateral angular gyrus, STS, STG, MTG, anterior temporal lobe, IFG, sPMv, SFS, SFG, and ventral precuneus in subjects S1 and S2. In the Narratives condition, prediction accuracy is high in voxels in bilateral angular gyrus, STS, STG, MTG, anterior temporal lobe, Broca's area and IFG, sPMv, SFS, SFG, ventral precuneus, and posterior cingulate gyrus in all subjects.