HHS Public Access

Author manuscript

Neuropsychologia. Author manuscript; available in PMC 2016 September 01.

Published in final edited form as:

Neuropsychologia. 2015 September ; 76: 41–51. doi:10.1016/j.neuropsychologia.2014.11.029.

Semantic Variability Predicts Neural Variability of Object Concepts

Elizabeth Musz and **Sharon L. Thompson-Schill**

Department of Psychology and Center for Cognitive Neuroscience, University of Pennsylvania, Philadelphia PA USA

Abstract

The prevailing approach to the neuroscientific study of concepts is to characterize the neural pattern evoked by a given concept, averaging over any variation that might occur upon multiple retrieval attempts (e.g., across time, tasks, or people). This approach— which diverges substantially from approaches to studying conceptual processing with other methods—treats all variation as noise. Here, our goal is to determine whether variation in neural patterns evoked by semantic retrieval of a given concept is more than just measurement error, and instead reflects variation arising from contextual variability. We measured each concept's diversity of semantic contexts ("SV") by analyzing its word frequency and co-occurrence statistics in large text corpora. To measure neural variability, we conducted an fMRI study and sampled neural activity associated with each concept when it appeared in three separate, randomized contexts. We predicted that concepts with low SV would exhibit uniform activation patterns across stimulus presentations, whereas concepts with high SV would exhibit more dynamic representations over time. We observed that a concept's SV score predicted its corresponding neural variability. This finding supports a flexible, distributed organization of semantic memory, where a concept's meaning and its neural activity patterns both continuously vary across contexts.

Keywords

semantic memory; object concepts; context-dependent representations

1. Introduction

When cognitive psychologists and psycholinguists consider the variability that arises when thinking about concepts, it is often understood to emerge from dynamic interactions between concepts and contexts. When cognitive neuroscientists and neurolinguistics consider this variability, it is usually treated as "noise", and consequently minimized or discarded. For example, efforts to classify multi-voxel patterns activated by thoughts about a chair require

Correspondence should be addressed to: Elizabeth Musz, muszeliz@sas.upenn.edu, Department of Psychology, 3720 Walnut Street, Philadelphia, PA 19104, +1(406)-249-2790.

averaging over many chair-evoked responses, or by limiting analyses to voxels with the most consistent activity patterns. Moreover, experimental subjects are often encouraged to think of the same set of stimulus features upon repeated presentations of the same concept (e.g., Mitchell et al., 2008; Shinkareva et al., 2011). Such methods can decode object-associated patterns with impressive classification accuracy. However, the methods which provide the most predictive power achieve this by collapsing cross-context variations into a single prediction. This implicitly assumes that conceptual representations are situationally invariant.

Rather than being "nuisance noise", neural variation might instead vary across concepts in meaningful, predictable ways. An obvious example of this variation occurs in the case of homonyms (for example, the pattern evoked by "driver" might look more like that evoked by other people or by other tools, depending whether you are thinking about your chauffeur or your golf game). We propose that this is just an extreme case of a more general principle, namely that all concepts exhibit some degree of context-dependent variation in their meaning. In turn, semantic variability should predict the extent of variability in neural signals associated with a concept. Testing this hypothesis requires measuring two characteristics of a given concept: semantic (or contextual) variability and neural variability. We briefly introduce our approach to each of these measures below.

1.1 Semantic Variability

When considering how we might quantify the extent of semantic variability, we consulted a wide body of previous research: Studies have sampled large linguistic corpora to count of the number of unique paragraphs (e.g., Adelman et al., 2006); documents (e.g., Steyvers & Malmberg, 2003) or movie subtitles (e.g., Brysbaert & New, 2009) in which certain concept names (i.e., words) occur. Other work has quantified the similarity of all of the documents in a text corpus that contains a given word, using either Latent Semantic Analysis (e.g., Hoffman et al., 2012) or topic modeling (e.g., Pereira et al., 2011). These methods assume that words are experienced throughout discrete episodic contexts, and these instances are operationalized as the documents in a corpus. Each word receives a quantified description of its entropy over documents, such that "promiscuous" words appearing in many contexts and with many different words are distinguished from "monogamous" words that appear more faithfully in particular contexts (McRae & Jones, 2012). Drawing from these diverse corpora and linguistic methods, we developed a composite measure that reflects the variety of contexts in which each concept occurs, which we henceforth refer to as "semantic variability" (SV).

1.2 Neural Variability

We measured the extent of neural variability by measuring the neural patterns evoked by a particular concept, and computing the correlations between these patterns as the concept's surrounding context varied over time. There are several ways in which we could have experimentally manipulated the variety of contexts in which a given concept appeared. For instance, a concept could be embedded in several different sentence contexts, or it could be probed in various task contexts (e.g., living/non-living or abstract/concrete judgments; for an example, see Hargreaves et al., 2012). However, not all contexts vary in the same ways, and

hence some contexts may be more variable than others. While a central hypothesis of this work is that any concept's representation may be modulated by context, we have no a priori estimates of the magnitude or quality of this effect. For that reason, we have sought to generate contexts without any systematic bias or definition whatsoever. This is best accomplished with a list of random words.

We measured the variability in neural signals elicited by a given concept as it appeared in three distinct, randomly generated word lists. Here, a concept's context is the items that precede it in a list. Such an approach is common in episodic memory studies: a stimulus item is embedded amongst other words in a sequentially presented list, and the episodic context is thought to gradually drift over time and throughout the list (e.g., the Temporal Context Model; see Polyn, Norman, & Kahana, 2008).

By presenting all concepts in equally random contexts, any given concept's relative semantic variability or stability could spontaneously emerge and manifest in the resulting neural patterns. Insofar as some concepts may have more ambiguous definitions, or stronger dependence on context, this method ensures that we are not simply analyzing the context alone. It trains our focus on the concept itself, without any presupposition about its modulating context.

1.3 Hypotheses

With this measure of neural variability, we could test a few key predictions. Firstly, and in part as a positive control, we compared the neural variability of single-sense nouns to multi-sense nouns. As introduced above, polysemous and homonymous nouns are extreme examples of cross-context variation in meanings, because two or more concepts share a single word form. Under our assumptions, these words should especially exhibit semantic and hence neural variability. While not the main focus of our hypothesis, such a result would validate our metrics of semantic and neural variability.

Secondly, and critically for our overall aims, we predicted a parametric effect of SV among the single-sense nouns. That is, although these "single-sense" nouns would typically be described as referring to a single concept, they nonetheless exhibit a range of SV values, which we hypothesize will be correlated with the extent of neural variability. That is, words with low SV should activate more stable concepts, and thus more stable neural patterns across stimulus presentations, whereas words with high SV should activate more variable concepts, and thus more variable neural patterns.

2. Methods

Subjects

Twenty-one right-handed, native English speakers (13 females; aged 18–26 years) participated in this experiment. Subjects had normal or corrected-to-normal vision and no history of neurological or language disorders. All subjects were recruited from the University of Pennsylvania community and paid $20 per hour for their participation. Subjects gave written informed consent, which was approved by the University of

Pennsylvania Institutional Review Board. Three subjects were replaced for performing below chance on at least one of nine experimental tasks.

2.1 Design overview—We measured neural patterns evoked by three instances of semantic retrieval for each of twenty-five concrete, single-sense nouns (our "target" items), and we calculated neural variability among these three patterns for each word. The procedure was designed both to encourage elaborative episodic encoding of each word and to permit contextual variation to exert an influence on the resulting neural patterns: the task was an intentional episodic encoding paradigm, and the target items were randomly interspersed along a much larger list of stimuli (our "context" items). Details on each follow.

2.2 Materials

2.2.1 Stimuli: The stimulus set comprised 215 concrete, single-sense nouns. These words included both nonliving and living things, from a basic level of semantic categorization (e.g., "dog" instead of "pug" or "animal"). From this larger set, 25 nouns were chosen for target items. These words were pseudo-randomly selected to yield wide range semantic variability values across words. An additional 145 words served as "context" items, in that they appeared in lists with the target items during the episodic encoding task. Finally, 45 nouns served as "lures" in the recognition memory tests that followed. In addition to these single-sense words, we selected 15 polysemous or homonymous nouns (hereafter called "PH words") to serve as our positive control stimuli, based on their use in studies of lexical-semantic ambiguity (e.g., Bedny et al., 2007; Klein & Murphy, 2001).

2.2.2 Semantic variability metric: Drawing from a variety of corpus analysis methods and text databases, we developed a metric of "semantic variability" (SV). SV is composed of seven different variables (Table 1). These variables quantify the magnitude (Variables 1–3) or range (Variables 4–7) of documents in which each word appears.

All target, PH, and context items with scores available for all seven variables were included in the development of SV, resulting in 161 items. To create a composite score for each item, we z-scored each variable to standardize their scales and averaged these z-scores. As a check on the interpretation of this metric, we compared SV scores of the target (single-sense) words and the PH words: As expected, the PH words were consistently assigned higher SV scores than the target words, $t(37.6) = 3.29$, $p = 0.003$ (two-tailed) (Figure 1). Stimulus characteristics for the selected target and PH words are listed in Table 2.

2.2.3 Presentation sequences: As noted in Section 2.1, we sought to elicit conceptual processing associated with each stimulus presentation, while also discouraging any deliberate or specific encoding strategies. Additionally, we sought to create a situation where contextual variability would likely emerge, and where all stimuli were presented in equally random contexts. With these aims in mind, we presented subjects with lists of the stimulus words, where the target items would reappear in separate lists (i.e., among different words). To minimize task constraints, subjects were not given any specific instructions for how to respond during stimulus presentations. However, they were told to remember the words for a subsequent memory test.

Stimuli were assigned to nine lists, where each list consisted of 35 items: five to ten targets, five PH words, and 20–25 context words. Each of the 25 targets and 15 PH words appeared three times in separate, non-adjacent lists. For the context words, 15 of the items on each list were unique (i.e., they appeared in only one list) in order to increase each list's distinctiveness. But, to remove novelty as a cue for task-relevant stimuli, each list (after the first) also included five context items from a previous list. The ordering of each list was completely randomized, with one exception: across its three presentations, a target item never preceded or followed a given context item more than once. New word lists and testing sequences were constructed for each subject.

2.3 Procedure—The stimuli were presented in nine scanning runs, with one word list per run, and one testing sequence between each run. Subjects were instructed to pay attention to the words on each list, in order to prepare for a recognition memory test that would immediately follow. Each word was visually presented in the center of the screen for 2,500 ms, with a variable, jittered inter-trial interval (500 ms – 12500 ms), during which a centrally-located fixation cross was present (timings developed using optseq2; http://surfer.nmr.mgh.harvard.edu/optseq/). Word stimuli ranged in size from 3–10 letters, with each letter horizontally subtending approximately 0.5° visual angle. Each word list presentation lasted the entire duration of a single scanner run, approximately 3.5 minutes. The stimulus timing and presentation was controlled by E-prime 2 software (Psychology Software Tools). A schematic of the stimulus display is depicted below (Figure 2).

Immediately after each encoding list, subjects performed a self-paced yes-no recognition memory test. fMRI data were not collected during these tests. Subjects responded via button press whether or not each of the ten words was present in the immediately preceding word list. Each test consisted of five context items and five lure items, in a random order. The context items were randomly selected from any of 20 context items from the immediately preceding word list (that is, either unique or repeated items). The lure items were five unique and novel concrete nouns. Target items never appeared in the recognition memory tests. The next word list presentation, and corresponding scan run, began immediately following the completion of the recognition test.

Across the nine between-list recognition memory tests, subjects successfully responded to 89% of all trials (average hit rate = 84%; correct rejection rate = 94%), with no subjects performing below 50% chance on any of the nine tests.

2.4 fMRI data acquisition—Functional and structural data were collected with a 32-channel array head coil on a 3T Siemens Trio system. The structural data included axial T1-weighted localizer images with 160 slices and 1 mm isotropic voxels (TR = 1620 ms, TE = 3.87, TI = 950 ms). We collected 44 axial slices (3 mm isotropic voxels) of echoplanar fMRI data (TR = 3000 ms, TE = 30 ms). Each of the nine functional scanning sessions lasted 219 seconds. Twelve seconds preceded data acquisition in each functional run to approach steady-state magnetization.

2.5 fMRI preprocessing—Image preprocessing and statistical analyses were performed using the AFNI and SUMA software package (Cox, 1996) and MATLAB (MathWorks).

Before all other analyses, time series data were preprocessed to minimize the effects of noise from various sources, and consequently to provide for a better estimation of the BOLD signal: First, images were corrected for differences in slice acquisition time due to the interleaved slice order within the 3000 ms TR. Next, individual volumes were spatially registered to the last volume of the last functional run in order to correct for head movement, since this was the volume closest in time to the high-resolution anatomical scan. Third, the data were despiked to remove any large values not attributable to physiological processes. For each subject, anatomical gray-matter probabilistic maps were created in Freesurfer (http://surfer.nmr.mgh.harvard.edu/) and applied to the functional data. The volumes were then spatially smoothed using a 3 mm FWHM Gaussian kernel. Finally, the time series data were z-normalized within each run. For the searchlight analysis, these preprocessing steps were repeated, except that subjects' gray matter masks were not applied.

Each stimulus presentation was separately modeled as a three-second boxcar function convolved with a canonical hemodynamic response function. Six motion parameters, which were estimated during the motion-correction step, were also regressed out of the time series data at this step. Beta coefficients were estimated using a modified general linear model that included a restricted maximum likelihood estimation of temporal auto-correlation structure, with a polynomial baseline fit as a covariate of no interest. This GLM analysis yielded a single beta value at each voxel for each stimulus event.

2.6 Neural similarity analysis—For each subject, we selected a set of voxels across which we could compute a measure of neural variability. Voxels were selected using two different methods, each described below. In each subject's voxel set, we extracted three beta values for each of the three item presentations of every target and PH word. Across the selected voxels, we then computed the average pairwise Pearson correlation between the beta values for each item's three separate stimulus presentations. This value served as the metric of neural similarity for a given item.

2.6.1 Whole brain feature selection: For each subject, we selected a set of voxels across which we could compute a measure of neural variability. These voxels were identified in each subject's native space from any voxels labeled as gray matter. We selected voxels with the highest *F*-statistics yielded by the model described above, in which all stimulus events are separately modeled as a single, unique regressor. For a given voxel, the *F*-statistic value reports the variance explained by a model that contrasted (1) words versus fixation and (2) differences across word presentations. Although we did not limit the voxel selection to any specific brain regions, we also added a contiguity constraint: every selected voxel needed to share a face with at least one other selected voxel. We then selected the *n* voxels with the highest *F*-statistic values.

We tested our hypotheses at values of *n* ranging from 25 to 10,000 (following from Hindy et al., 2012). Below, we report detailed analyses for the 500-voxel input; however, the findings we report were robust for *n* of 250 to 1,000 selected features, and up to 2,000 at a trend level. Reports at additional voxel set sizes can be found in Appendix A.

2.6.2 Searchlight analyses: In order to examine whether the putative relation between SV and neural variability was regionally specific, we also conducted a searchlight analysis across the brain. A 3-voxel radius sphere was iteratively centered on each voxel in the brain (Kriegeskorte et al., 2006). This sized sphere included 123 voxels when unrestricted by the brain's boundary, and the diameter of the sphere was 9 mm. For the voxels in a searchlight sphere, we calculated each item's average neural similarity. For each subject, we estimated a linear regression coefficient that used SV values to predict average neural similarity across items. The resulting beta value was then assigned to each searchlight center. Subjects' searchlight maps were then resampled to the functional data resolution, normalized to Talairach coordinates (Talairach & Tournoux, 1998).

We then tested the reliability of the regression coefficient across subjects with a 1-sample *t*-test. To perform this group-level analysis, we first estimated the smoothness of the data in three directions (i.e., xyz coordinates). These estimates were obtained using AFNI's 3dFWHMx on the residual time series data. The average subject-level values were then averaged across subjects (FWHMx = 4.83 mm; FWHMy = 4.85 mm; FWHMz= 3.95 mm). Based on a voxel-level uncorrected alpha of .01 (t=2.84), Monte Carlo simulations (n=50,000) performed with 3dClustSim in AFNI indicated a minimum cluster size of 19 voxels for cluster-level corrected alpha of .05. Although results reported from the searchlight analysis are referred to as clusters of voxels, it is important to point out that such clusters only identify each sphere's center voxel. Some of the sphere's most informative voxels might be located in another region adjacent to the center voxel's region.

3. Results

3.1 Whole-brain distributed patterns

3.1.1 Comparing neural similarity across word types—In each subject's 500 selected voxels, we compared the average within-item neural similarity for single-sense target words versus PH words. Across subjects, the single-sense target words exhibited more within-item neural similarity (mean r = .09) than did the PH words (mean r = .07), $t(20) = 3.03$, $p = 0.006$ (two-tailed) (Figure 3).

3.1.2. Relating semantic variability to neural variability—In each subject, we computed a Pearson correlation between each target item's average neural similarity and its SV score. At the group level, subjects' resulting correlation coefficients were compared to zero in a 1-sample *t*-test. We found a negative relationship between SV and neural similarity, such that items with lower SV scores exhibited greater neural similarity across contexts, and items with higher SV scores had more variability among their cross-context neural patterns, mean r = −.12, $t(20) = -2.89$, $p = 0.009$ (two-tailed) (Figure 4).

3.2 Searchlight Localized patterns

3.2.1. Comparing neural similarity across word types—In each searchlight volume, we computed the average within-item neural similarity for all of the target and PH words. We then computed a mean neural similarity for each word type by averaging across all target items and all PH items. We created two searchlight maps, one in which the average

target neural similarity was assigned to the searchlight center, and one searchlight map with average PH neural similarity at searchlight centers. Across subjects, the two searchlight maps were then submitted to a dependent samples *t*-test to identify searchlight spheres with significant differences between word types. Seven clusters of contiguous searchlight centers emerged as significant (see Table 3).

Three clusters exhibited more neural similarity for target words than PH words, with peak searchlight centers in the right lingual gyrus and extending into the left lingual gyrus (Figure 5) and the superior parietal lobule bilaterally (Figure 6). Four clusters showed the reverse pattern, with peak centers in the left inferior frontal gyrus (pars Triangularis) and right postcentral gyrus (Figure 7), left parahippocampal gyrus, and right superior parietal lobule.

3.2.2. Relating semantic variability to neural variability—In each searchlight volume, we performed an item analysis to test the parametric effect of SV on average within-item neural similarity in the target words. The beta coefficient for SV was then assigned to the searchlight's center. We compared the resulting searchlight maps across subjects in a single-sample *t*-test versus 0 (two-tailed). Four clusters of contiguous searchlight centers emerged as significant. In three left-lateralized clusters, with peak voxels in lingual gyrus, fusiform gyrus (Figure 8), inferior frontal gyrus (par Triangularis) (Figure 9), SV negatively predicted neural similarity. An additional cluster in the right superior medial gyrus showed the opposite effect, such that higher SV scores were associated with greater neural similarity.

Because regions often associated with semantic processing (e.g., the anterior temporal lobes) tend to have poor signal quality, and because no significant clusters emerged in these areas, we checked for signal coverage in these areas. For each subject's wholebrain map, we calculated the temporal signal-to-noise (TSNR) ratio at each voxel by dividing the mean times series data by the standard deviation of the detrended time series data (Murphy et al., 2007). We then normalized the data to a common space and computed a group average map of TSNR values (see Fig. D1). Throughout the bilateral temporal lobes, these values are well above the suggested minimum values for adequate signal detection (e.g., >20; Binder et al., 2011), indicating that TSNR in the temporal lobes was sufficient for detecting fMRI activation.

4. Discussion

The present study aimed to measure and predict neural variation in the conceptual processing of concepts across variations in their semantic contexts. We proposed that concepts with higher semantic variability should have correspondingly larger variations in their cross-context neural representations. We tested this prediction by measuring the similarity of neural activity patterns associated with a given concept, and how these patterns changed across time and context. In agreement with this prediction, significant categorical differences in activation patterns emerged for single- and multi-sense word groups. Additionally, while the neural activity associated with conceptual processing varied across repeated stimulus presentations, this variation was reliably predicted by a stimulus item's

SV score. These findings were observed in subjects' individually selected voxels, well as in group-level whole-brain searchlight analyses.

4.1 Categorical Effects

In support of our hypothesized categorical effect of word type, we observed more neural similarity for target words than PH words. In the group-level searchlight analysis, three brain clusters exhibited this pattern of results. The largest cluster, with a peak searchlight center in the right lingual gyrus, extended bilaterally into the left lingual gyrus and surrounding extrastriate cortex. Two additional searchlight center clusters also exhibited more neural similarity for target words: one in left superior parietal lobule, extending into the inferior parietal lobule, and one in the right superior parietal lobule. While this finding was not the main focus of our study, the result supports our metric of neural similarity. Although both word types exhibited large variation in their neural representations, this variation was reliably greater for PH words than single-sense target words.

Additionally, the searchlight analysis revealed the reverse pattern in four regions: left inferior frontal gyrus (LIFG), right postcentral gyrus, right superior temporal gyrus, and left parahippocampal gyrus. In these searchlight clusters, PH words exhibited greater neural similarity than target words. The LIFG's response is particularly intriguing, since previous work has found that this area is involved in selecting contextually relevant semantic information amidst competition or ambiguity (Thompson-Schill et al., 1997, 1999; Bedny & Thompson-Schill, 2008). We will further discuss the potential functional roles of the LIFG in a following section.

4.2 Parametric Effects

While neural activity patterns associated conceptual processing varied across stimulus presentations, this variation was reliably predicted by the concepts' SV scores. This correlation was observed in each subject's uniquely distributed voxels that had also exhibited a categorical difference of word type. Additionally, this result was observed in a group-level whole-brain searchlight analysis, in local patterns centered in four searchlight clusters. In three left-lateralized clusters centered in the lingual gyrus, fusiform gyrus, and LIFG, higher SV scores inversely predicted neural similarity. These results comport well with our theoretical predictions, whereby variable semantic processing of concepts should in turn evoke more variable neural patterns. Intriguingly, searchlight centers clustered in the right superior medial gyrus showed the reverse result; here, concepts with higher SV scores exhibited greater neural similarity. The direction of this finding is the reverse of what we had predicted, but significance of the result validates our claim that item-wise semantic variability can be used to predict neural similarity.

Additionally, in the whole-brain searchlight analysis, which computed neural similarity in locally distributed multi-voxel patterns, two brain regions exhibited both categorical and parametric differences. In left lingual gyrus, the parametric and categorical effects were observed in overlapping voxels, and both effects were in the predicted direction. In contrast, in LIFG, the searchlight clusters that showed reliable effects did not overlap, and while the parametric effect here matched our hypothesis, the observed categorical difference was

opposite of what we had predicted. Below, we further discuss the findings in these brain areas.

4.3 Early Visual Cortex Findings

In visual cortex, the parametric effect of SV overlapped with searchlights that exhibited the categorical effect of word type: 31 contiguous searchlight spheres exhibited more neural similarity for (1) target words than PH words and (2) target words with low SV than target words with high SV. The center of the overlapping searchlights was located in the left lingual gyrus (Figure 10).

These early visual regions are implicated in studies of object visualization during imagery tasks (Lee et al., 2012) and maintenance of visual representations in working memory (Serences et al., 2009; Harrison & Tong, 2009). Typically, semantic effects in early visual cortex are reported under conditions of explicit mental imagery (e.g., Hindy et al., 2013; Lee et al., 2012). However, additional work has found that early visual areas are recruited even when subjects are not instructed to imagine objects. For example, previous studies from our lab have reported activity in lingual gyrus during retrieval of object shape knowledge (Hsu et al., 2014) and object color knowledge (Hsu et al., 2012). Furthermore, these effects have been found to correlate with subjects' self-reported preference for a visual cognitive style (Hsu et al., 2011).

While we did not explicitly instruct our subjects to imagine the items, and did not debrief them on their encoding strategies, the use of mental imagery might partly explain our findings in these regions. In the context of an explicit episodic encoding paradigm, mental imagery could be an effective strategy for memorizing the presented concepts. One possibility is that subjects engaged in mental imagery while reading the concept names, and that PH and high SV words evoked especially different visualizations—and hence evoked more variable neural patterns—upon their separate presentations. This possibility is supported by recent work by Hindy and colleagues (2013), in which early visual cortex evoked dissimilar patterns when subjects imagined two alternative states of the same object.

Alternatively, although our results indicate that neural variability in these early visual areas is predicted by SV, it is possible that other stimulus characteristics, which correlate with SV, might have contributed to these effects. For instance, amongst our stimulus items, SV is negatively correlated with word length, such that longer words tend to have lower SV values, and words high in SV have fewer letters. Previous studies have indicated that regions of occipital cortex that spatially overlap with our searchlight results are sensitive to letter length, such that there is a positive correlation of BOLD signal with number of letters in early visual regions while subjects read aloud words (e.g., Graves et al., 2010) and pseudowords (e.g., Valdois et al., 2005) and during lexical decision tasks (Schurz et al., 2010). In one study, using word stimuli that matched ours in size, the authors found greater activation while subjects read longer words (7–9 letters long) versus shorter words (4–6 letters long) in regions that overlap with our searchlight results, including left inferior occipital gyrus and left superior parietal gyrus (Church et al., 2011). Greater activation in brain regions associated with visual and attentional processing might reflect longer gaze durations for longer, less frequent words (Rayner, 1998).

These findings indicate that longer words elicit greater magnitude of BOLD response in early visual regions; however, it is unknown how word length affects the *variability* of multi-voxel patterns evoked by the same word upon repeated presentations, which is the dependent measure in our study. The relationship between univariate BOLD activity and multivoxel neural similarity is not straightforward: an increased BOLD response could be associated with more stable multi-voxel patterns, or it might instead be associated with greater variability in responses. In order to address this possibility, we examined the relation between word length and neural similarity in subject-specific, distributed grey matter voxels; this was marginally significant, $t(20)= 1.98$, $p=.06$.

Because of the high correlation between word length and SV in our stimulus set, we cannot compare the unique variance that each explains. However, there are two reasons to believe that word length is not the *entire* story here. Firstly, neural similarity is inversely predicted by some of the individual measures of semantic variability (that compose our composite measure) that are not correlated with word length (e.g., Variables 5 and 7; see Appendix C). Secondly, prior word length effects on activation are mostly confined to early visual cortex but our correlations with SV are not: We tested whether it was necessary to include early visual regions in order to observe neural variability effects. We transformed anatomical masks of the medial occipital lobes (identified as left and right calcarine sulcus in the SPM Anatomy Toolbox, Eickhoff et al., 2005; see Fig. D1) into each subject's native space. We re-ran our analyses on subjects' whole-brain distributed patterns, now only selecting whole-brain gray matter voxels that were located outside of the calcarine sulci masks. After excluding these regions, the pattern of results was unchanged. Neural similarity was reliably greater for target words (mean $r=.09$) than PH words (mean $r= .07$), $t(20)= 2.54$, $p= .02$. Additionally, SV inversely predicted item-wise neural similarity (mean $r=.11$), $t(20)= -2.98$, $p= .007$. These findings indicate the neural variability effects are also reliably supported in regions outside of early visual cortex. Finally, on this topic, we think it is likely that different stimulus characteristics will contribute to neural variability observed in different brain regions. Even if the effect in early visual cortex is due to a confound with word length, that does not mean this explanation holds across the brain.

4.4 Left Inferior Frontal Gyrus Findings

While the searchlight findings in left lingual gyrus supported our hypotheses and overlapped anatomically, the effects in the LIFG were more varied. In this region, we observed two distinct searchlight clusters which showed divergent effects (Figure 11). In an anterior and medial LIFG cluster, including voxels in the anterior cingulate cortex, SV inversely predicted neural similarity of target items. In line with our predictions, this parametric effect suggests that concept-evoked patterns in anterior regions of LIFG are sensitive to the semantic variability of conceptual representations.

In contrast, in posterior LIFG, the results ran counter to our predictions: PH words exhibited greater neural similarity than target words. One possibility, requiring further investigation, is that the semantically ambiguous PH words evoke a common set of frontally-mediated processes, and hence exhibit more consistent patterns in LIFG. However, such a role may be limited to more posterior regions of LIFG, which do not exhibit sensitivity to continuous

measures of semantic variability of traditionally “single-sense” words. This unexpected finding may also be related to other functional dissociations reported about prefrontal cortex subregions (e.g., Koechlin & Summerfield, 2007; Badre & D’Esposito, 2007), although more work is needed to examine the functional distinctions between posterior and anterior LIFG.

4.5 Characterizing Context

In this study, we observed variation in neural patterns by embedding the target items in randomized word lists. Alternatively, we could have more directly influenced subjects’ interpretations of each item presentation by constructing more item-specific contexts. This could have been accomplished, for example, by hand picking particular words to immediately precede a given target item upon each presentation. For instance, we could have preceded “tulip” by “vase”, “garden”, and “still life”, and we could have preceded “bench” by “park”, “courtroom”, and “ballpark”, in order to manipulate the specific conceptual instantiations of “tulip” and “bench”; however, in part due to the hemodynamic sluggishness of the BOLD signal, we would not be able to discriminate whether greater neural variability for “bench” over “tulip” was due to the variability in the patterns evoked by these two words or due to the variability lingering in the patterns evoked by “park”, “courtroom”, and “ballpark” (compared to “vase”, “garden”, and “still life”).

Instead, by randomly picking the words that preceded each of our target items, we could be sure that our measure of neural variability of the patterns evoked by the target was not unintentionally influenced by the neural variability of the words that preceded it. That is, across subjects (each of whom received a different random list sequence), any differences in the variability of the items that preceded the targets would average out, and so our measure of neural variability can be described as a pure measure of the target concept. With this approach, we observed neural variability that is both robust and reliably predictable by SV.

Amidst the random contexts, object concepts evoked highly variable neural patterns: mean within-item similarity correlations were $r = .07$ for the PH words, and $r = .09$ for target words. These weak correlations indicate that there are several additional sources of neural variability, in addition to the similarity that we have attributed to repeated retrievals of the same concept. For instance, a large portion of the neural variability might be explained by the items that precede a given item in a presentation sequence. Because we deliberately embedded the targeted items in randomized word lists, we are unable model the effects of the preceding items on the resulting neural variability. Future work might find some utility in more explicit manipulations of a concept’s contexts, such that the effects of preceding items on a given item can be accounted for. Such an approach would likely yield stronger correlations of within-item neural similarities.

In addition to the randomized word lists, context was also defined by the task conditions under which the concepts were retrieved. To encourage variable semantic processing, we used an episodic encoding paradigm. As we describe in Section 4.3, this task context might have encouraged subjects to engage in mental imagery. Such a strategy would activate concepts’ visual properties, relative to more abstract or nonvisual semantic features. In order

to encourage retrieval of a variety of semantic features, future studies might employ tasks that require more explicit retrieval of various kinds of semantic knowledge.

4.6 Predicting Neural Variability

Future studies will benefit from further characterizing the continuous stimulus dimensions that best describe the cross-context variability in multi-voxel patterns. The metric we used to describe neural variation was composed seven separate measures of words' contextual variations, drawn from four different text databases. In addition to our summed z-score version of SV, we also performed a Principal Components Analysis (PCA) in order to reduce the information from the seven original variables into a smaller set of composite dimensions. The first component highly correlated with the SV measure reported above and also reliably predicted the neural data (see Appendix B). However, most of the seven original variables loaded highly on this first component. Moreover, most could predict neural variability independently, without being collapsed into a composite measure (see Appendix C). Future analyses should explore the format and content of text databases from which extracted variables can best explain neural variation.

Furthermore, neural variation might be predicted by additional stimulus properties that are related to a word's breadth of contexts. Concepts high in semantic variability tend to be more frequent and less imageable (Hoffman et al., 2011), and shorter in length and less concrete, relative to concepts that have low semantic variability (see Table 2). The fact that we observe our reported effects when SV correlates with additional these variables suggest that our effects might be in part driven by stimulus characteristics other than SV. Future studies can control for these other stimulus characteristics by minimizing the correlations between them, such that the shared variance can be statistically removed, or through the selection of more controlled experimental stimuli. However, our reported effects are not solely driven by these other variables, because some of the individual measures of semantic variability are not correlated with these additional variables yet they still reliably predict neural variability (see Appendix C).

One could ask, however, whether any of these other variables are in fact producing the observed neural variability in ways in which we had not hypothesized. Perhaps these additional stimulus characteristics jointly or uniquely contribute to neural variability in ways that support additional predictions about semantic representation. Moreover, it is likely the case that different perceptual and psychological factors contribute to the variability in neural patterns observed across different brain regions. This is a potentially interesting, yet currently untested, research topic. But, absent a measure of neural variability, such possibilities could not be further considered. Any of these predictions would be interesting to explore, once one adopts the approach of measuring neural variability, rather than averaging over it.

Additionally, further work is needed to localize the neural activity that best captures this semantic variability. While many studies limit their analyses to voxels with the most stable activation profiles (e.g., Mitchell et al., 2008; Anderson et al., 2014), the present work examines voxels that exhibit maximally different responses across stimulus presentations. In our subjects' gray matter masks, there is only a 0.001% overlap in the top 500 voxels

selected by these two criteria. However, rather than narrowing analyses to either maximally or minimally variable voxels, it is possible that conceptual information is most robustly represented by some combination of both stable and variable patterns of response.

In sum, our results suggest that a concept's meaning varies continuously as a function of its context, such that concepts do not have a fixed, discrete number of senses, but rather a continuous, context-dependent variation in their meaning. Furthermore, neural data that is typically discarded as "noise" might instead represent context-modulated variation in an object's representation. These findings illustrate the possibility of applying a more dynamic view of concepts to investigations of their associated neural patterns.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was supported by NIH grants R0I-DC009209 and R01-EY021717 to S.T.-S. and an NSF graduate fellowship to E.M. The authors wish to thank members of the Thompson-Schill lab for helpful discussions and two anonymous reviewers for their comments on an earlier version of this manuscript.

References

Anderson AJ, Murphy B, Poesio M. Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. Journal of Cognitive Neuroscience. 2014; 26:658–681. [PubMed: 24168217]

Badre B, D'Esposito M. Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. Journal of Cognitive Neuroscience. 2007; 19:2082–2099. [PubMed: 17892391]

Bedny M, Hulbert JC, Thompson-Schill SL. Understanding words in context: the role of Broca's area in word comprehension. Brain Research. 2007; 1146:101–114. [PubMed: 17123486]

Bedny M, McGill M, Thompson-Schill SL. Semantic adaptation and competition during word comprehension. Cerebral Cortex. 2008; 18:2574–2585. [PubMed: 18308708]

Brysbaert M, New B. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods. 2009; 41:977–990. [PubMed: 19897807]

Church JA, Balota DA, Petersen SE, Schlaggar BL. Manipulation of length and lexicality localizes the functional neuroanatomy of phonological processing in adult readers. Journal of Cognitive Neuroscience. 2011; 23:1475–1493. [PubMed: 20433237]

Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Computational Biomedical Research. 1996; 29:162–173.

Eickhoff SB, Stephan KE, Mohlberg H, Grefkes C, Fink GR, Amunts K, Zilles K. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. NeuroImage. 2005; 25:1325–1335. [PubMed: 15850749]

Hargreaves IS, White M, Pexman PM, Pittman D, Goodyear BG. The question shapes the answer: the neural correlates of task differences reveal dynamic semantic processing. Brain and Language. 2012; 120:73–78. [PubMed: 22078639]

Harrison SA, Tong F. Decoding reveals the contents of visual working memory in early visual areas. Nature. 2009; 458:632–635. [PubMed: 19225460]

Hindy NC, Solomon SH, Altmann GTM, Thompson-Schill SL. A Cortical Network for the Encoding of Object Change. Cerebral Cortex. 2013

Hindy NC, Altmann GTM, Kalenik E, Thompson-Schill SL. The effect of object state-changes on event processing: do objects compete with themselves? Journal of Neuroscience. 2012; 32:5795–5803. [PubMed: 22539841]

Hoffman P, Rogers TT, Ralph MAL. Semantic diversity accounts for the "missing" word frequency effect in stroke aphasia: Insights using a novel method to quantify contextual variability in meaning. Journal of Cognitive Neuroscience. 2011; 23(9):2432–2446. [PubMed: 21254804]

Hoffman P, Lambon Ralph MA, Rogers TT. Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. Behavior Research Methods. 2012; 45:718–730. [PubMed: 23239067]

Hsu NS, Kraemer DJ, Oliver RT, Schlichting ML, Thompson-Schill SL. Color, context, and cognitive style: Variations in color knowledge retrieval as a function of task and subject variables. Journal of Cognitive Neuroscience. 2011; 23(9):2544–2557. [PubMed: 21265605]

Hsu NS, Schlichting ML, Thompson-Schill SL. Feature Diagnosticity Affects Representations of Novel and Familiar Objects. Journal of Cognitive Neuroscience. 2014:1–15. [PubMed: 24047384]

Graves WW, Desai R, Humphries C, Seidenberg MS, Binder JR. Neural Systems for Reading Aloud: A Multiparametric Approach. Cerebral Cortex. 2010; 20(8):1799–1815. [PubMed: 19920057]

Lee S-H, Kravitz DJ, Baker CI. Disentangling visual imagery and perception of real-world objects. NeuroImage. 2012; 59:4064–4073. [PubMed: 22040738]

Klein DE, Murphy GL. The Representation of Polysemous Words. Journal of Memory and Language. 2001; 45:259–282.

Koechlin E, Summerfield C. An information theoretical approach to prefrontal executive function. Trends in Cognitive Sciences. 2007; 11:229–235. [PubMed: 17475536]

Kriegeskorte N, Goebel R, Bandettini P. Information-based functional brain mapping. Proceedings of the National Academy of Science. 2006; 103:3863–3868.

McRae, K.; Jones, M. Semantic Memory. In: Reisberg, D., editor. Oxford Handbook of Cognitive Psychology. New York, NY: Oxford University Press, Inc; 2012.

Mitchell TM, Shinkareva SV, Carlson A, Chang K-M, Malave VL, Mason RA, Just MA. Predicting human brain activity associated with the meanings of nouns. Science. 2008; 320:1191–1195. [PubMed: 18511683]

Murphy K, Bodurka J, Bandettini PA. How long to scan? The relationship between fMRI temporal signal to noise ratio and necessary scan duration. NeuroImage. 2007; 34:565–574. [PubMed: 17126038]

Pereira F, Botvinick M, Detre G. Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. Artificial Intelligence. 2011; 194:240–252. [PubMed: 23243317]

Polyn S, Norman KA, Kahana MJ. A context maintenance and retrieval model of organization processes in free recall. Psychological Review. 2008; 116:129–156. [PubMed: 19159151]

Rayner K. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin. 1998; 124:372–422. [PubMed: 9849112]

Schurz M, Sturm D, Richlan F, Kronbichler M, Ladurner G, Wimmer H. A dual-route perspective on brain activation in response to visual words: Evidence for a length by lexicality interaction in the visual word form area (VWFA). NeuroImage. 2010; 49(3):2649–2661. [PubMed: 19896538]

Serences JT, Ester EF, Vogel EK, Awh E. Stimulus-Specific Delay Activity in Human Primary Visual Cortex. Psychological Science. 2009; 20(2):207–214. [PubMed: 19170936]

Shinkareva SV, Malave VL, Mason RA, Mitchell TM, Just MA. Commonality of neural representations of words and pictures. NeuroImage. 2011; 54:2418–2425. [PubMed: 20974270]

Simmons WK, Reddish M, Bellgowan PSF, Martin A. The selectivity and functional connectivity of the anterior temporal lobes. Cerebral Cortex. 2010; 20:813–825. [PubMed: 19620621]

Steyvers M, Malmberg KJ. The effect of normative context variability on recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2003; 29:760–766.

Talairach, J.; Tournoux, P. Co-planar stereotaxic atlas of the human brain. New York: Thieme; 1988.

Thompson-Schill SL, D'Esposito M, Aguirre GK, Farah MJ. Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. Proceedings of the National Academy of Science. 1997; 94:14792–14797.

Thompson-Schill SL, D'Esposito M, Kan IP. Effects of repetition and competition on activity in left prefrontal cortex during word generation. Neuron. 1999; 23:513–522. [PubMed: 10433263]

Valdois S, Carbonnel S, Juphard A, Baciu M, Ans B, Peyrin C, Segebarth C. Polysyllabic pseudo-word processing in reading and lexical decision: Converging evidence from behavioral data, connectionist simulations and functional MRI. Brain Research. 2006; 1085(1):149–162. [PubMed: 16574082]

Highlights

- We measured the neural variability of an object concept in distinct contexts
- We used corpus statistics to measure each concept's diversity of semantic contexts
- Concepts with low semantic variability had stable neural patterns across contexts
- Concepts with high semantic variability had more variable neural patterns
- Homonyms and polysemes showed more neural variability than single-sense nouns

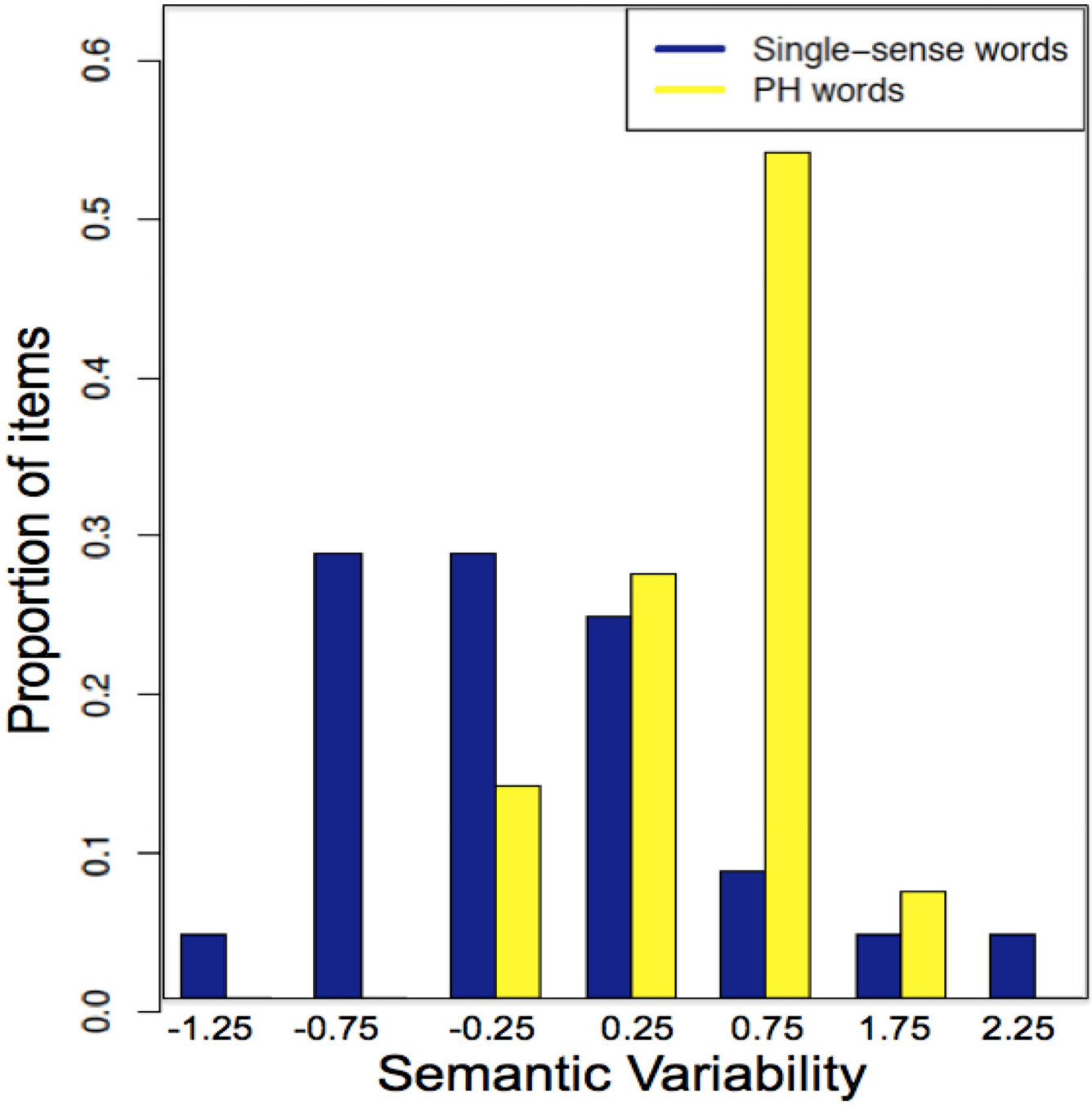

Figure 1.
Percentage of Semantic Variability (SV) scores across single-sense target words and multi-sense polysemous/homonymous (PH) words.

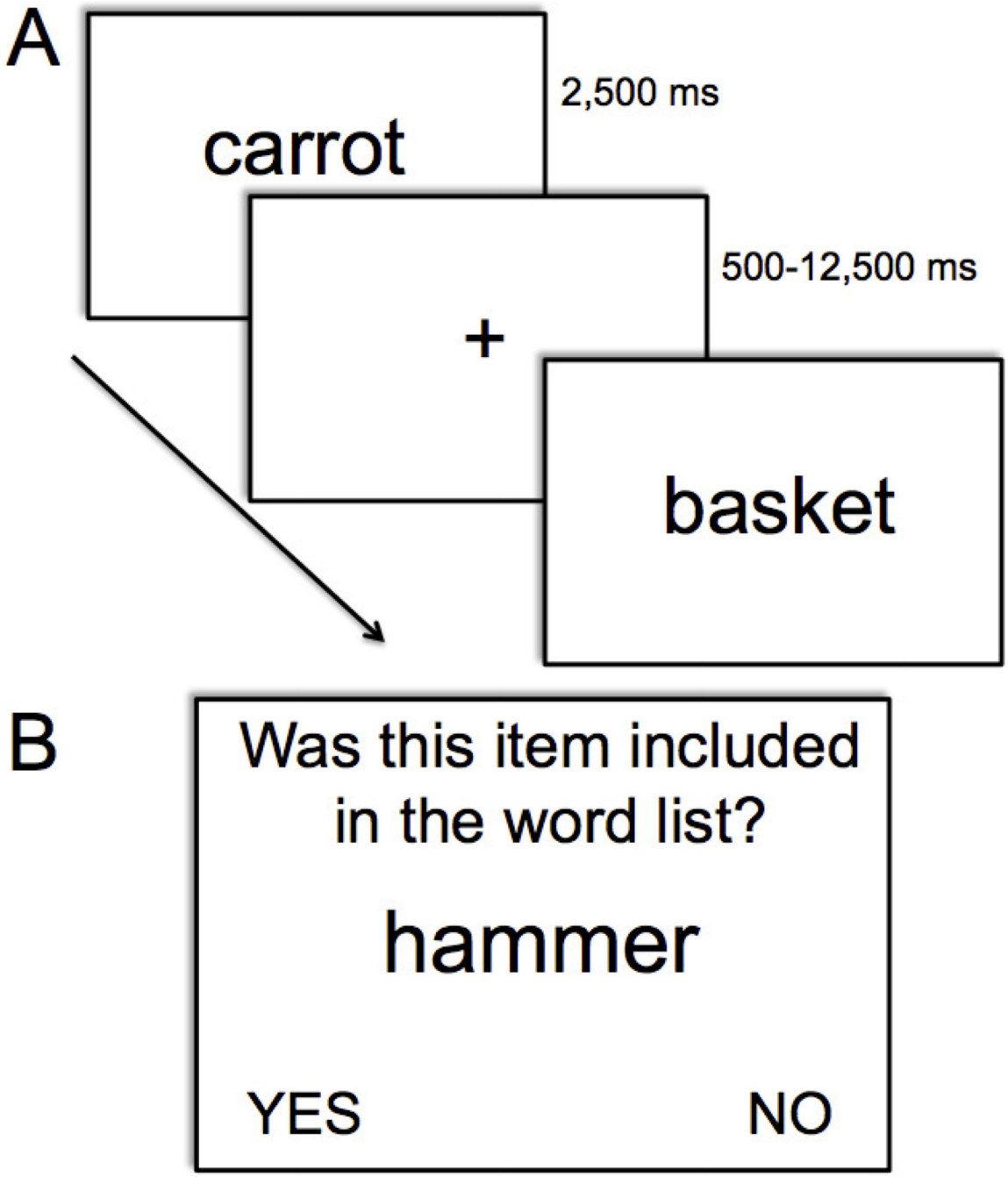

Figure 2.
Stimulus presentation and experimental task. (A) Words appeared for 2.5s, followed by a fixation cross of variable duration. (B) After each word list presentation, subjects performed old/new judgments, where half of the words were context items from the list. Responses were self-paced and made via button press.

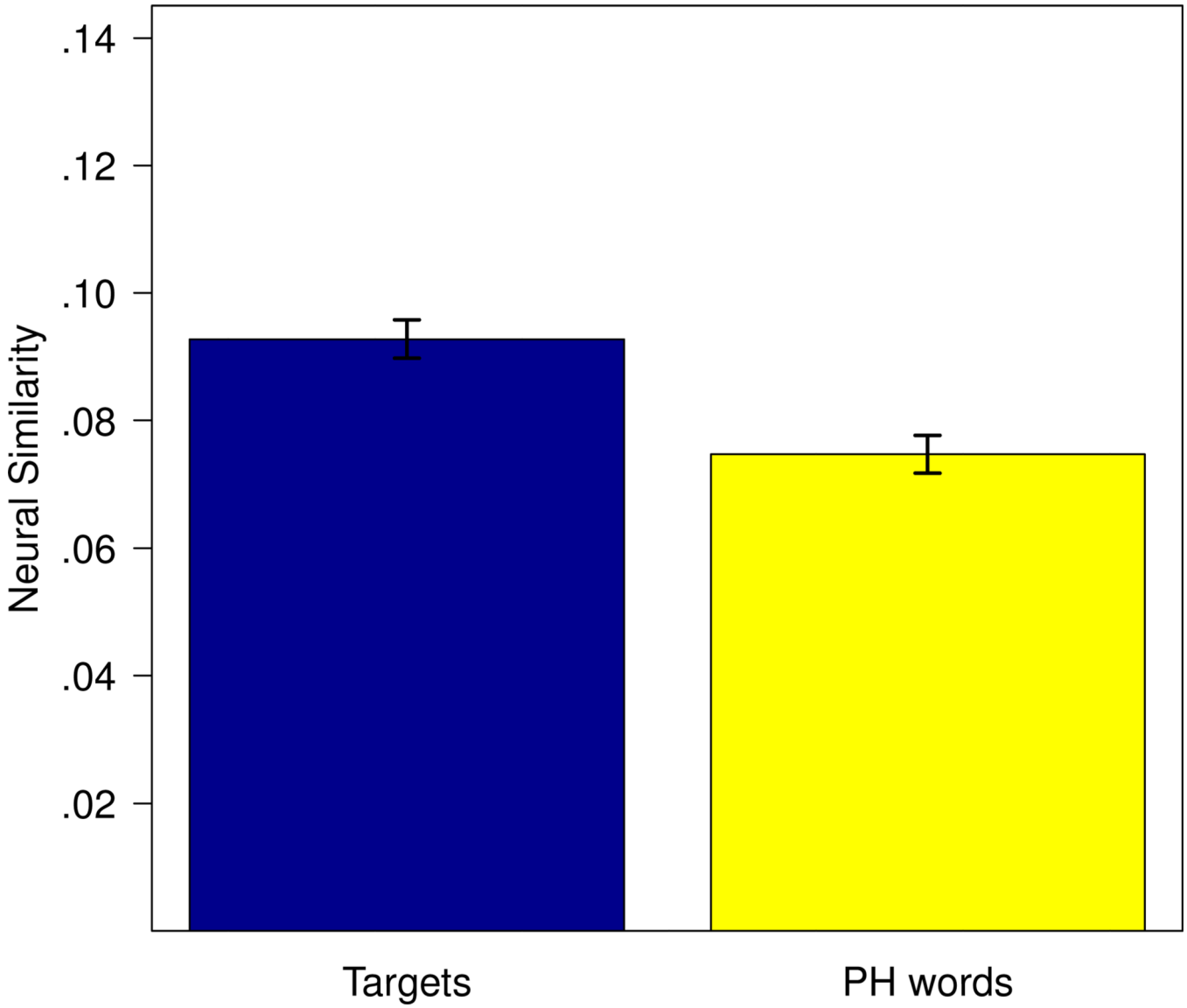

Figure 3.
Average neural similarity by word type in subjects' selected 500 voxels chosen from distributed grey matter voxels. Error bars reflect within-subject standard error.

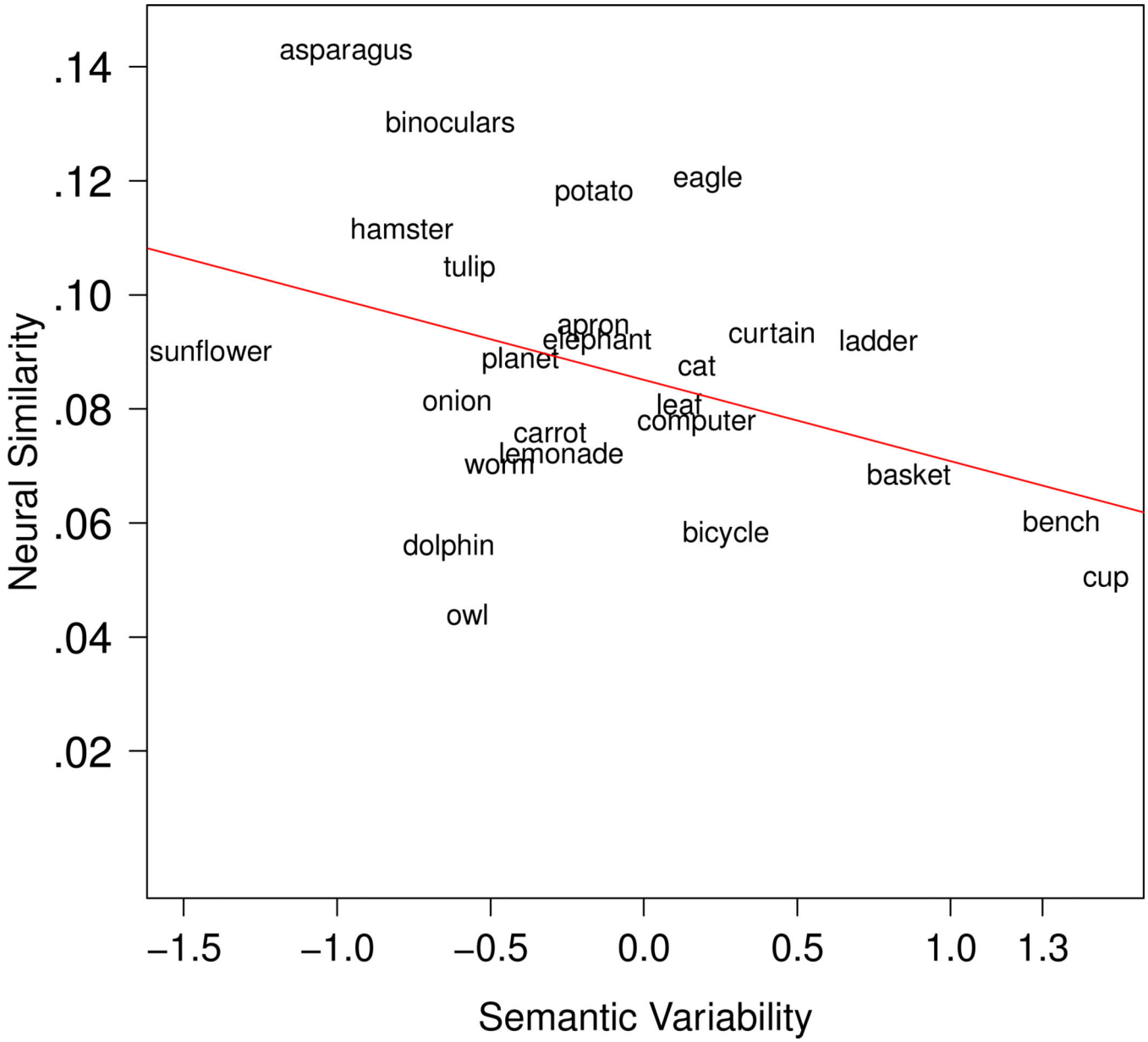

Figure 4.

Relationship between target words' semantic variability (SV) scores and within-item neural similarity, averaged across subjects. Correlations were calculated in each individual subject's 500 selected voxels. Depicted results are averaged across subjects.

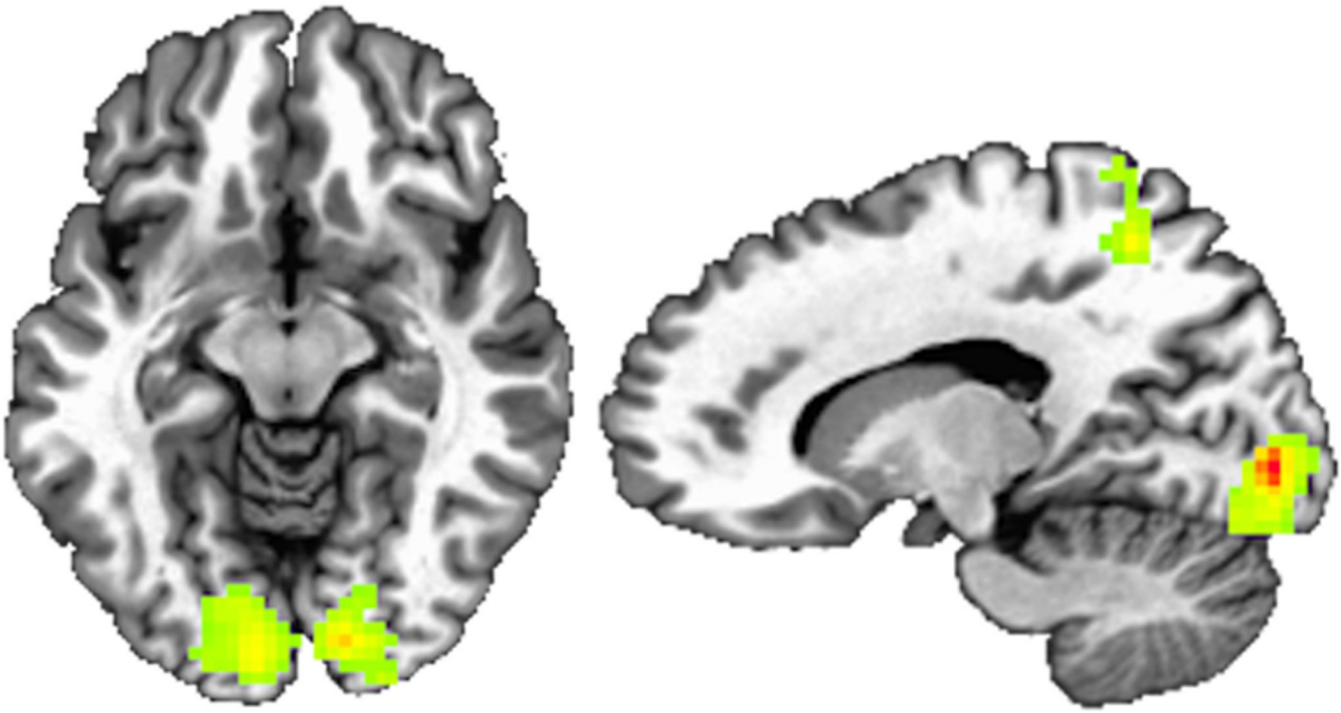

Figure 5.
Searchlight centers that exhibited more neural similarity for single-sense target words than PH words. Peak voxels are centered in the right lingual gyrus, extending into the left lingual gyrus. Sagittal view depicts this result in the right lingual gyrus and in the right superior parietal gyrus.

Figure 6.
Searchlight centers that exhibited more average neural similarity for single-sense words than PH words. Clusters are centered in the superior parietal lobule bilaterally.

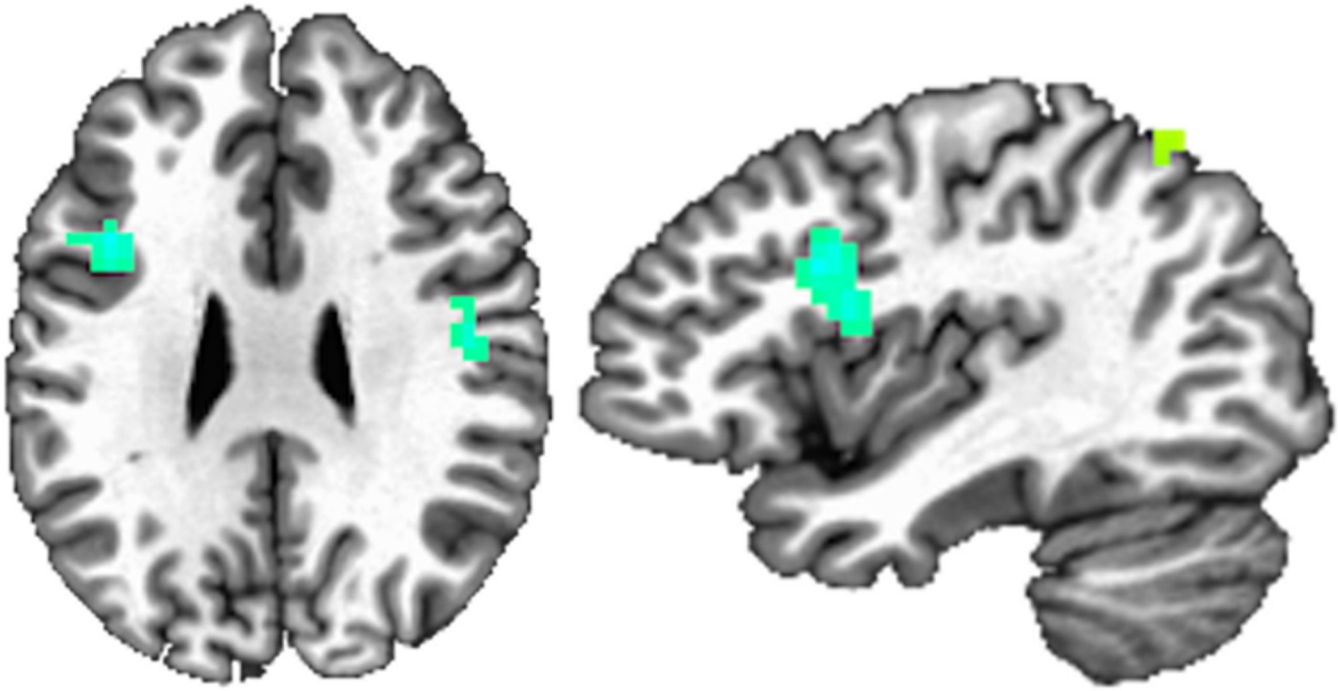

Figure 7.
Searchlight centers that exhibited more average neural similarity for PH words than single-sense target words. Clusters are centered in the left inferior frontal gyrus (pars Triangularus) and right postcentral gyrus.

Figure 8.

In searchlights centered in the left lingual gyrus and left fusiform gyrus, semantic variability scores were inversely correlated with average neural similarity across single-sense target words.

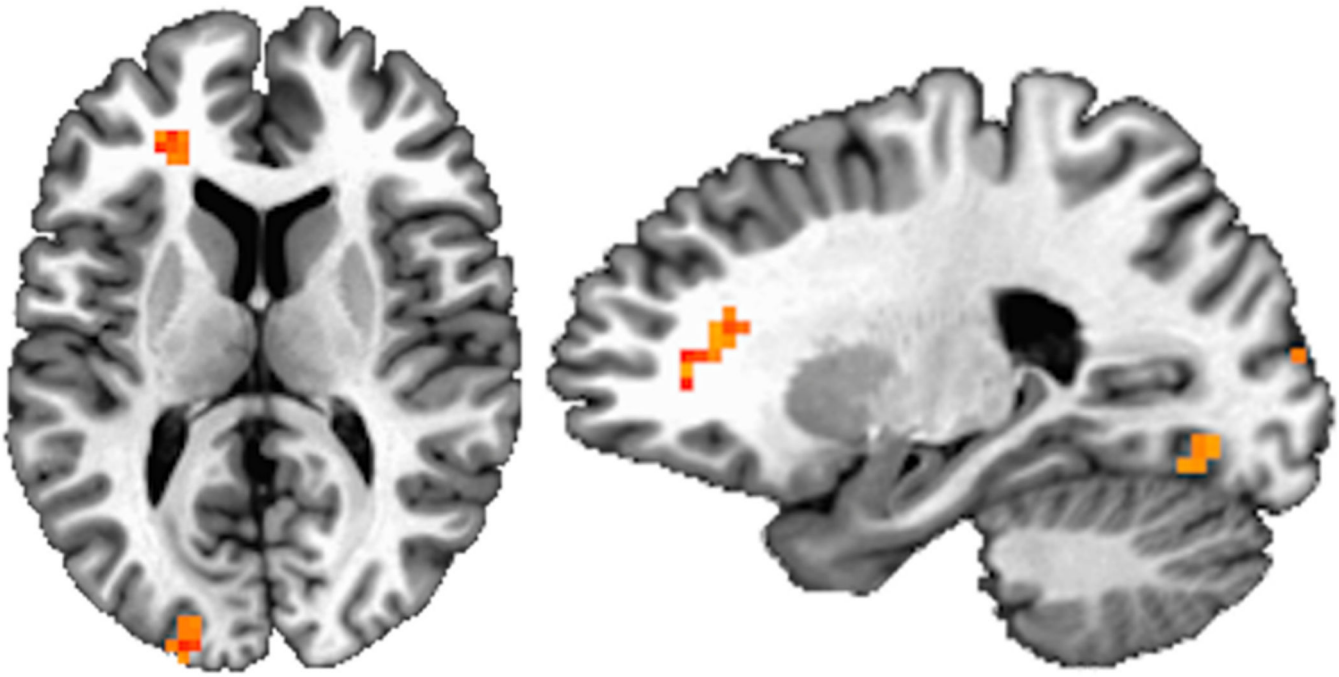

Figure 9.

In peak searchlight centers in the left inferior frontal gyrus and surrounding left anterior cingulate, semantic variability scores were inversely correlated with average neural similarity across single-sense target words. Effects in left lingual gyrus are depicted as well.

Figure 10.

Whole-brain searchlight results in the left lingual gyrus. In 31 contiguous searchlight centers, (1) target words exhibited more neural similarity than PH words and (2) SV scores inversely correlated with neural similarity across single-sense target words.

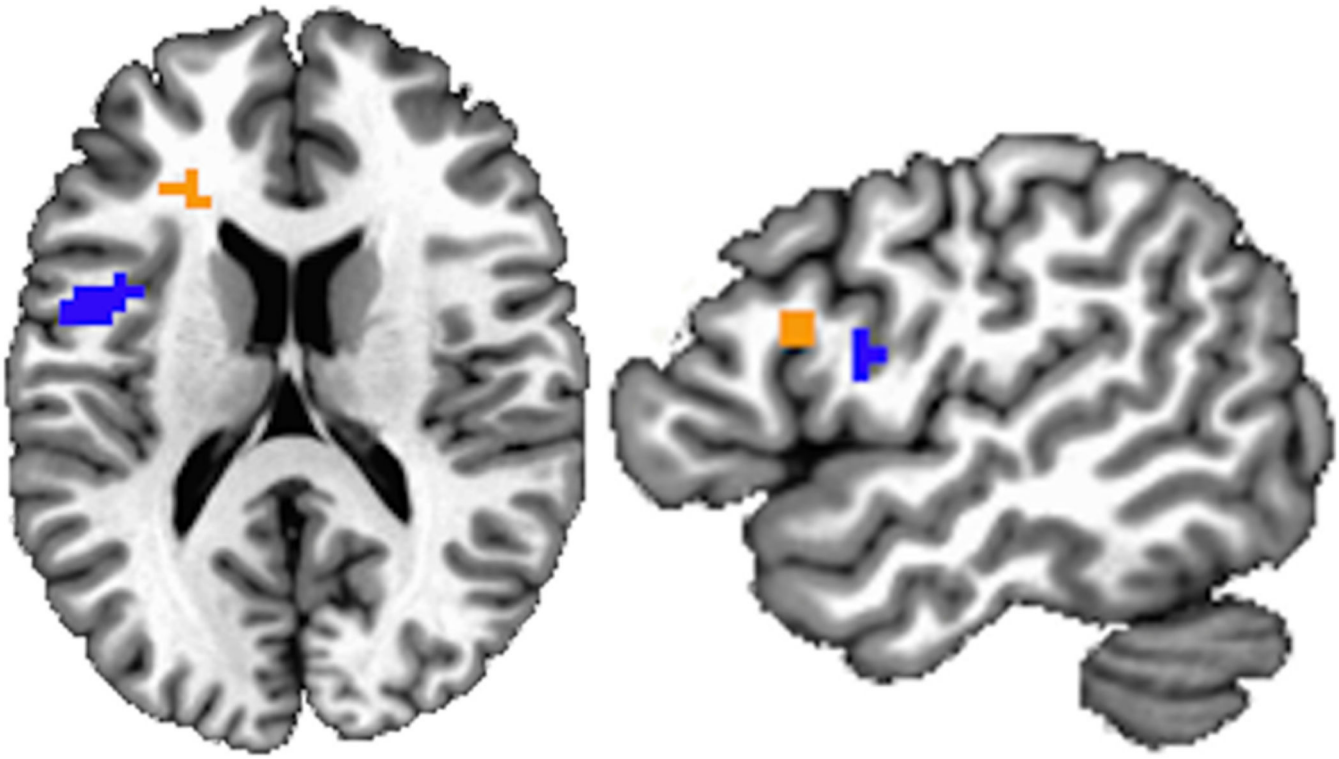

Figure 11.

Whole-brain searchlight results in the left inferior frontal gyrus. The categorical effects from Comparison 1 are depicted in blue, in which PH words exhibited more neural similarity than target words. The orange voxels show the parametric effects from Comparison 2, in which item-wise semantic variability scores inversely predicted neural similarity. The center of mass of the parametric effects is in the left anterior cingulate.

Table 1

Variables included in the development of Semantic Variability (SV) Scores

	Authors	Corpus	Method	Variables
1	Brysbaert & New (2009)	SUBTLEX US	movie counts	number of movies in which the word occurs in the subtitles
2,3	Hoffman et al. (2012)	British National Corpus; TASA corpus	document counts	number of paragraphs in which word occurs
4,5	Hoffman et al. (2012)	British National Corpus; TASA corpus	LSA	In high-dimensional space, the distances between all of a word's paragraphs
6	Pereira et al. (2011)	Wikipedia articles	Topic Modeling	Number of topics in which a word occurs
7	Pereira et al. (2011)	Wikipedia articles	Topic Modeling	Probability that word occurs in its most dominant topic, where a word's topic inclusion probabilities must sum to 1

Table 2

Summary of linguistic features of the word stimuli.

Stimulus characteristics	Target words	PH words	Correlation with SV
Semantic variability (SV)	−0.09 (.70)	0.51 (.46)*	--
Concreteness	604 (30)	585 (18)*	−0.25
Familiarity	519 (25)	540 (34)	0.37*
Imageability	592 (47)	578 (49)	−0.24
Word length	6.08 (1.93)	4.53 (1.19)*	−0.46*
Number of phonemes	5.21 (.83)	3.67 (.49)*	−0.44*
Number of syllables	2.08 (1.79)	1.33 (1.05)*	−0.47*
Word frequency	25.36 (27.16)	74.67 (67.35)*	0.59*

Table 2. Values are means with standard deviations. Concreteness, Familiarity, and Imageability ratings were rated on a 100–700 scale and were obtained from the MRC psycholinguistic database (Coltheart, 1981) and were available for 80%; 85%; and 83% of the items, respectively. Norms for word frequency were obtained from the WebCelex database (Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; http://celex.mpi.nl) and reflect word frequencies per million instances.

Asterisks in PH words column denote significant differences between Target and PH word groups; in Correlation column, asterisks denote significant Pearson correlations between SV and stimulus characteristic, $p<.05$.

Table 3

Peak searchlight centers from whole-brain analysis

	Extent	x	y	z	peak *t*-value	Brain region	Neural similarity result
Comparison 1: Neural similarity by word type	685	14	−85	−1	13.50	R. lingual gyrus	targets > PH
		−8	−83	−13	−7.50	L. lingual gyrus	
	83	−37	11	26	5.16	L. inferior frontal gyrus (pars Triangularis)	PH > targets
	73	17	−52	50	−5.65	R. superior parietal lobule	targets > PH
	37	−28	−58	50	−4.24	L. superior parietal lobule	targets > PH
	37	59	−1	8	3.80	R. superior temporal gyrus	PH > targets
	36	−13	−7	−16	5.05	L. parahippocampal gyrus	PH > targets
	24	47	−13	26	4.16	R. postcentral gyrus	PH > targets
Comparison 2: item-wise SV and neural similarity relationship	61	−7	−91	−1	−4.23	L. superior occipital gyrus	inversely predicts SV
	30	8	32	50	3.80	R. superior medial gyrus	predicts SV
	24	−19	−73	−10	−3.42	L. fusiform gyrus	inversely predicts SV
	22	−25	35	11	−4.23	L. inferior frontal gyrus (pars Triangularis)	inversely predicts SV

Table 3. Clusters of searchlight centers that were reliably sensitive to differences between word types (Comparison 1) or semantic variability (SV) differences (Comparison 2). In Comparison 1, three clusters exhibited greater neural similarity for single-sense target words than PH words, and four clusters showed the reverse pattern. In Comparison 2, three regions contained searchlight centers where SV negatively predicted neural similarity; the reverse relationship was found in an additional searchlight cluster. Each cluster is thresholded at $p < .05$, corrected for multiple comparisons. Talairach coordinates and anatomical labels indicate the peak searchlight center location of each cluster. L., left; R., right.