University of Pennsylvania
ScholarlyCommons

Publicly Accessible Penn Dissertations

2017

Shades Of Meaning: Capturing Meaningful Context-Based Variations In Neural Patterns

Elizabeth Musz
University of Pennsylvania, lisa.musz@gmail.com

Follow this and additional works at: https://repository.upenn.edu/edissertations

 Part of the Cognitive Psychology Commons, and the Neuroscience and Neurobiology Commons

Recommended Citation
Musz, Elizabeth, "Shades Of Meaning: Capturing Meaningful Context-Based Variations In Neural Patterns" (2017). *Publicly Accessible Penn Dissertations*. 2492.
https://repository.upenn.edu/edissertations/2492

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/edissertations/2492
For more information, please contact repository@pobox.upenn.edu.

Shades Of Meaning: Capturing Meaningful Context-Based Variations In Neural Patterns

Abstract
When cognitive psychologists and psycholinguists consider the variability that arises during the retrieval of conceptual information, this variability it is often understood to arise from the dynamic interactions between concepts and contexts. When cognitive neuroscientists and neurolinguists think about this variability, it is typically treated as noise and discarded from the analyses. In this dissertation, we bridge these two traditions by asking: can the variability in neural patterns evoked by word meanings reflect the contextual variation that occurs during conceptual processing? We employ functional magnetic resonance imaging (fMRI) to measure, quantify, and predict brain activity during context-dependent retrieval of word meanings. Across three experiments, we test the ways in which word-evoked neural variability is influenced by the sentence context in which the word appears (Chapter 2); the current set of task demands (Chapter 3); or even undirected thoughts about other concepts (Chapter 4). Our findings indicate that not only do the neural patterns evoked by the same stimulus word vary over time, but we can predict the degree to which these patterns vary using meaningful, theoretically motivated variables. These results demonstrate that cross-context, within-concept variations in neural responses are not exclusively due to statistical noise or measurement error. Rather, the degree of a concept's neural variability varies in a manner that accords with a context-dependent view of semantic representation. In addition, we present preliminary evidence that prefrontally-mediated cognitive control processes are involved in expression of context-appropriate neural patterns. In sum, these studies provide a novel perspective on the flexibility of word meanings and the variable brain activity patterns associated with them.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Psychology

First Advisor
Sharon L. Thompson-Schill

Keywords
concepts, fmri, language, psycholinguistics, semantic memory

Subject Categories
Cognitive Psychology | Neuroscience and Neurobiology

This dissertation is available at ScholarlyCommons: https://repository.upenn.edu/edissertations/2492

SHADES OF MEANING: CAPTURING MEANINGFUL CONTEXT-BASED VARIATIONS IN NEURAL PATTERNS

Elizabeth Musz

A DISSERTATION

in

Psychology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

_____

Sharon L. Thompson-Schill, Ph.D.

Professor of Psychology

Graduate Group Chairperson

_____

Sara R. Jaffee, Ph.D.

Professor of Psychology

Dissertation Committee

John C. Trueswell, Professor of Psychology

Geoffrey K. Aguirre, Associate Professor of Neurology

ACKNOWLEDGEMENT

This work was made possible thanks to generous financial and institutional support, including a graduate research fellowship from the National Science Foundation and a Penn Behavioral and Cognitive Neuroscience Training Grant from the National Institutes of Health. Additionally, this work has greatly benefited from constructive feedback from my graduate committee members, John Trueswell and Geoff Aguirre. Their helpful comments, questions, and ideas throughout my dissertation research and qualifying exams have concretely increased the quality of my research and my writing.

I am also grateful for my peers at the Center for Cognitive Neuroscience, and all current and previous members of the Thompson-Schill lab since 2010. I am especially thankful for the helpful research-related discussions and life-related conversations that I have had with Christine Boylan, Marc Coutanche, Jen DeSantis, Nina Hsu, Heath Matheson, Nathan Tardiff, and Matt Weber. My biggest thanks go to Eiling Yee, who has been an immensely positive and formative role model throughout my research career. When I first started working in the lab, Eiling made science seem fun and do-able. Her patience, persistence, and kindness motivated me to do my best. In working with Eiling, I have learned to embrace the problem-solving process and relish in the excitement that comes with designing and conducting novel scientific research. I'm very grateful to have her as a mentor and collaborator.

Of course, none of these experiences would have ever been possible without the guidance and support of Sharon Thompson-Schill. I have learned and grown so much in the seven years since she first welcomed me into her lab as a research assistant. Sharon is a major source of inspiration for me in every aspect of her career: she is a productive and impactful scientist, an effective public communicator and energetic educator, and an inclusive leader in the field. On top of all that, she is a very thoughtful and caring person. Sharon is an ideal PhD advisor, because she approaches every research question with equal parts scrutiny and enthusiasm. She has granted me a lot of independence and ownership of my work throughout my graduate studies, and my research skills are

stronger because of it. I am so glad and lucky that her mentorship has shaped my academic development.

I also appreciate the moral support that my friends and family have provided throughout this process. I am especially grateful to my partner Ryan Bradley, who has been by my side at every step along the way. He knows more about MVPA and the left prefrontal cortex than anyone in his (very unrelated) research field ever ought to. Thank you to Ryan for reading countless paper drafts, for talking through my analysis codes, and for intently listening to me muse about my experiments on our walks to the dog park and Rittenhouse Square. Ryan makes all of my pursuits more enjoyable and worthwhile, and my personal and my academic life are both happier and more rewarding thanks to his company.

The findings from Chapter 2 have been published as: Musz, E., & Thompson-Schill, S. L. (2017). Tracking competition and cognitive control during language comprehension with multi-voxel pattern analysis. *Brain and language, 165*, 21-32. Reprinted with permission from Elsevier.

The findings from Chapter 4 have been published as: Musz, E., & Thompson-Schill, S. L. (2015). Semantic variability predicts neural variability of object concepts. *Neuropsychologia, 76*, 41-51. Reprinted with permission from Elsevier.

ABSTRACT

SHADES OF MEANING: CAPTURING MEANINGFUL CONTEXT-BASED VARIATIONS IN NEURAL PATTERNS

Elizabeth Musz

Sharon L. Thompson-Schill

When cognitive psychologists and psycholinguists consider the variability that arises during the retrieval of conceptual information, this variability it is often understood to arise from the dynamic interactions between concepts and contexts. When cognitive neuroscientists and neurolinguists think about this variability, it is typically treated as noise and discarded from the analyses. In this dissertation, we bridge these two traditions by asking: can the variability in neural patterns evoked by word meanings reflect the contextual variation that occurs during conceptual processing? We employ functional magnetic resonance imaging (fMRI) to measure, quantify, and predict brain activity during context-dependent retrieval of word meanings. Across three experiments, we test the ways in which word-evoked neural variability is influenced by the sentence context in which the word appears (Chapter 2); the current set of task demands (Chapter 3); or even undirected thoughts about other concepts (Chapter 4). Our findings indicate that not only do the neural patterns evoked by the same stimulus word vary over time, but we can predict the degree to which these patterns vary using meaningful, theoretically motivated variables. These results demonstrate that cross-context, within-concept variations in neural responses are not exclusively due to statistical noise or measurement error. Rather, the degree of a concept's neural variability varies in a manner that accords with a

context-dependent view of semantic representation. In addition, we present preliminary evidence that prefrontally-mediated cognitive control processes are involved in expression of context-appropriate neural patterns. In sum, these studies provide a novel perspective on the flexibility of word meanings and the variable brain activity patterns associated with them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

"δὶς ἐς τὸν αὐτὸν ποταμὸν οὐκ ἂν ἐμβαίης" –Heraclitus, b. 535 B.C.
("You could not step twice into the same river.")

In a physical world that is continually in flux, the ways that we think of the world may likewise be subject to change. If we never step into the same river twice, then our understanding of the concept *river* may also vary from one moment to the next. That is, our concept of an object or a thing in the world—which is abstracted away from any one instance of that object—must be flexible enough to accommodate change. This includes within-object variation (a river during your first step into it, versus your second), and between-object variation among unique instances of that concept (e.g., the Danube and the Nile are both instances of the concept *river*, despite their numerous differences).

One major source of this variation is the context in which a concept is encountered. Here, "context" is a general term that refers to "everything else" that is present or ongoing while the concept is accessed. Context includes the information that co-occurs with the object, including an individual's current task and goals, spatiotemporal details, the other things in the surrounding scene or linguistic phrase, etc. Due to the presence of other information that occurs along with the thought of a concept, no concept is ever accessed in a "context-free" fashion (Casasanto & Lupyan, 2015). Meanings change because the instances of concepts and contexts in which they are embedded exist in an ever-changing world.

The context-dependence of meanings and how they are expressed in brain activity is the central topic of this thesis. In the following chapters, we employ functional magnetic resonance imaging (fMRI) to measure, quantify, and predict brain activity during context-dependent retrieval of conceptual knowledge. Across three fMRI experiments, we explore how neural responses to the same stimulus item can vary from one moment to the next. Further, in each study, we predict the degree to which a stimulus item's concurrent neural response changes across various contexts. Taken together, this research program lends support to the theory that there is a wealth of meaningful

information about our thought processes carried by the context-mediated variations in our encounters with concepts. These studies illustrate the theoretical impact and methodological utility of studying and predicting item-level, cross-context variations in word meanings and their corresponding neural responses.

The central theme in this work is the examination of how the neural response patterns evoked by the same concept can vary depending on the context in which the concept is retrieved. In Chapters 3 and 4, we employ experimental paradigms in which subjects think about the same concept at different points in the experiment, while the surrounding context changes over time. We find that neural activity during semantic retrieval of object concepts is variable; two separate thoughts of the same concept yield two different brain activity patterns. Further, not only do the patterns evoked by the same concept vary over time, but we can predict the degree to which these patterns vary using meaningful, theoretically motivated variables. The findings from these studies indicate that cross-context, within-concept variations in neural responses are not exclusively due to statistical noise or measurement error. Rather, the degree of a concept's neural variability varies in a manner that accords with specific hypotheses about semantic representation.

In addition to describing semantic variables that contribute to measurable neural variation in conceptual processing, we have also explored possible mechanisms that enable the context-dependent retrieval of word meanings. In Chapters 2 and 3, we examine the control processes that are deployed when subjects use contextual information to guide attention toward task-relevant and context-appropriate aspects of a stimulus word's meaning, amidst competition from task-irrelevant information. In these experiments, the behavioral task context explicitly directs subjects to retrieve specific information about each stimulus item. In some cases, these contexts bias retrieval toward information that is relatively infrequent or weakly activated, such that stronger, alternative information might compete for activation. In these chapters, we test the hypothesis that prefrontally-mediated cognitive control processes bias semantic retrieval toward task-relevant and context-appropriate information. Under this framework, increases in prefrontal response should predict increases in the context-appropriateness of a stimulus item's resulting neural pattern.

Below, we briefly review the applications of neuroimaging techniques to study object representations, and how the fMRI studies in this thesis differ from the most common approaches to this topic. Then, we motivate the proposal that left-lateralized regions of prefrontal cortex are involved in biasing retrieval toward context-appropriate information. Finally, we summarize the interconnections between these studies.

1. Traditional fMRI Approaches to studying object representation

To gain insight into the brain areas that support the retrieval of object information, neuroscientists have used fMRI to measure changes in blood oxygenation level-dependent (BOLD) signal while human subjects view experimental stimuli that name or depict real-world objects. In traditional univariate analyses, researchers measure contrasts in the average response amplitude elicited by various stimulus conditions; these comparisons are performed either within a single voxel or averaged across a larger region of interest (ROI). Such experiments have revealed a number of brain regions throughout cortex where the overall magnitude of BOLD response varies for different object categories, such as animals versus tools (Mahon & Caramazza, 2009), or for different object attributes, such as color-related or action-related information (Chao & Martin, 1999).

In contrast to testing for average changes in a region's overall response magnitude, more recent fMRI studies have examined unique response patterns that are spread across small subsets of voxels. Unlike traditional, univariate-based analysis of mean activation and spatial averaging, multi-voxel pattern analysis (MVPA) samples from the signals contained in the patterns of activity among multiple, spatially distributed voxels. Applications of MVPA to fMRI data have revealed that the information contained across these spatially distributed activity patterns encode fine-grained distinctions between different object stimuli. For example, spatially distributed voxels in ventral temporal (VT) cortex exhibit distinct response patterns when subjects view pictures of objects from one stimulus category, compared to stimuli from another category.

The logic of the MVPA approach is based on the notion of similarity: similar stimuli should exhibit relatively similar response patterns, and conversely, stimuli that are dissimilar from one another should elicit response patterns that are relatively dissimilar.

For instance, in a seminal study by Haxby and colleagues (2001), the multi-voxel patterns (MVPs) in VT cortex exhibited similar responses to different pictures of chairs, and these chair-evoked patterns were relatively more similar to one another than they were to patterns evoked by pictures of shoes. The neural similarity between two stimulus-evoked MVPs can be computed using measures of vector proximity (e.g., Pearson or Spearman correlation, cosine similarity, Euclidean distance) or linear separability (Weber et al., 2009). Recent applications of MVPA have demonstrated that multi-voxel activity patterns distributed throughout object-selective regions of ventral temporal cortex encode distinctions between both broad-level categories, such as animate versus inanimate objects (Kriegeskorte et al., 2008; Clarke & Tyler, 2014), and more fine-grained differences among within-category exemplars, such as beetles versus moths (Connolly et al., 2012; Weber et al., 2009).

These studies perform the impressive technical feat of detecting subtle relationships between object stimuli that are measured from admittedly noisy signals. However, to detect these multivariate, object-evoked signals, most of these studies employ highly constrained experimental paradigms, and perform analysis decisions that filter the data toward certain kinds of signals. Indeed, it is almost trivially true that any scientific inquiry is limited by the lens with which it studies the phenomenon of interest, and the questions it chooses to ask of the collected data. In this case, the methodological decisions of traditional MVPA studies carry with them some consequential assumptions about the nature of the underlying representations of the experimental stimuli. Namely, researchers have primarily studied neural representations under conditions in which variations in object-evoked thoughts, and variations in the resulting BOLD signals, were minimized.

For example, most MVPA studies present each stimulus item several separate times throughout the experiment, and then average across the MVPs evoked upon each separate presentation of the same stimulus. This averaging method yields a single composite neural pattern for each stimulus, thereby discarding the aspects of the stimulus' patterns that varied across instances. Averaging across stimulus presentations is a generally useful tool for fMRI analyses, as it boosts the ratio of signal to noise. This sort of within-stimulus averaging is most powerful when it is applied to brief

presentations of short and isolated events, where one can assume a canonical response profile (Ben-Yakov et al., 2012). However, the assumption of a canonical, stable response is violated if individuals conceive of a concept in different ways at different times. By averaging across presentations, these studies limit the neural characterization of each concept to its common activation across presentations, and discard any variability in activated object properties that might have occurred over time and contextual shifts. It is precisely this intrinsic variation which we wish to assess.

Furthermore, MVPA studies often measure the neural patterns evoked under task conditions that encourage subjects to consistently recruit the same information about an object upon each repeated stimulus presentation. In a seminal MVPA study by Mitchell and colleagues (2008), subjects were shown all of the experimental stimuli (labeled line drawings of objects) prior to scanning, and subjects were instructed to list the specific object properties that they would think of when each stimulus was presented during the fMRI session. However, these contrived conditions bear little resemblance to the ways that we typically regard and interact with objects in our daily life. Our thoughts of the same object will vary from one moment to the next, shaped by whatever else we are thinking of at the time. When researchers constrain subjects' thoughts, the neural patterns evoked by these thoughts will be likewise constrained.

2. Leveraging within-item MVPA to study neural variability

The suite of fMRI experiments in this thesis serve as a foil to the canonical applications of MVPA to object representation. Most MVPA studies compute the neural similarity between the average MVP for one stimulus item versus the average MVP for another stimulus item. In the present studies, we measure multiple MVPs for the same stimulus item at different times in an experiment, and then compute the similarity between them. This "within-item" neural similarity quantifies the extent to which a stimulus item's evoked pattern changes across presentations. Given the physiological artifacts and statistical noise that are present in fMRI signals, one would never expect the same stimulus item to evoke two identical response patterns (i.e., the two MVPs will never be perfectly correlated with one another). Here, we contend that at least some of this observed variability is not merely due to statistical noise or measurement error, but

rather that it reflects variations in the way that the subject regards the stimuli upon the separate occasions. That is, the aspects of a stimulus item's meaning that a subject retrieves or focuses on will vary along with changes in the task demands and surrounding context of the item's presentation. To test this hypothesis, we manipulated the experimental contexts in ways that encouraged subjects to have variable and changing thoughts about the stimulus items. Then, across items within an experiment, we attempted to predict the degree of within-item neural similarity, using item-level variables that quantify the degree to which the stimulus item's meaning is expected to vary.

3. Biasing Neural Patterns: Cognitive control during semantic retrieval

Concepts and words have multiple potential interpretations, and therefore ambiguity abounds. To avoid misinterpretation or miscommunication of a word or concept's meaning, it is often necessary to select the aspects of the given stimuli that are most suited to one's current task or goals. This has meaningful behavioral consequences—we must filter out distracting information that will hinder our cognitive and behavioral performance, and focus on the aspects of the meaning that are most pertinent for the given moment. We theorize that these behavioral pressures transform the retrieved information about a concept along with its pattern of neural activation.

The ability to select among candidate information in a goal-directed manner is enacted by several well-studied cognitive control processes. Critically, these abilities allow us to select weak yet task-relevant information over strongly activated yet irrelevant information. Cognitive control is mediated by responses in the prefrontal cortex (Miller & Cohen, 2001; Fuster, 2008); in particular, left-lateralized regions of the ventrolateral prefrontal cortex are implicated in resolving competition between incompatible representations (Badre & Wagner, 2007; Thompson-Schill et al., 2005). Several univariate fMRI studies have observed increased recruitment of left ventrolateral prefrontal cortex (left vlPFC) during a variety of cognitive tasks which require selection, including verb generation, object classification, and semantic comparison (Thompson-Schill, 2003) as well as semantic fluency (Hirshorn & Thompson-Schill, 2006). This area is also involved during the co-activation of competing syntactic representations caused by

syntactic garden-path sentences (Stowe et al., 2004; Novak et al., 2005; Rodd et al. 2013). Taken together, the response profile of left vlPFC suggests that it is involved in selecting among competing information to boost activation toward the most task-relevant information.

In more recent years, fMRI studies have investigated how response fluctuations in this region are linked to the expression of task-relevant information, manifested in the robustness and clarity of multivariate representations that are encoded in distributed, posterior brain regions. Across cognitive domains, researchers have studied how the MVPs evoked by experimental stimuli are modified by changes in task demands, and furthermore, how prefrontal cortex (PFC) is critically involved in this process. The putative link between prefrontal activity and the robustness of task-relevant multivariate response patterns in posterior cortex has received empirical support from several cognitive domains.

For instance, in fMRI studies of retrieval interference in the episodic memory domain, researchers have found that increased PFC response is associated with improved memory retrieval. The neural signature of this prefrontally-mediated improvement is indexed by the clarity and distinctiveness of the stimulus-evoked multivariate patterns. In these studies, the stimulus materials presented during memory encoding are typically pictures of real-world objects (e.g., faces and scenes), and so the multivariate patterns are typically measured in object-selective regions of ventral temporal cortex. Recent fMRI studies have shown that when response increases in left prefrontal cortex, the neural patterns evoked in VT cortex during successful recognition exhibit decreased similarity to patterns evoked by distractor stimuli that were present during encoding (Kuhl et al., 2012; Wimber et al., 2015). These results can be interpreted as evidence that left PFC plays a role in selecting appropriate memories and suppressing distracting information. Furthermore, the relative presence of such information (i.e., task-relevant versus irrelevant memorial details) can be detected in the multi-voxel patterns elicited during encoding and retrieval.

Moreover, in the domain of object imagery, a recent study by Hindy and colleagues (2013) observed a similar correspondence between left PFC response and the expression of task-relevant patterns in VT cortex. When subjects were instructed to

imagine two mutually exclusive states of the same object—for example, first an intact egg and then later a cracked egg—response magnitude in left PFC predicted the degree to which the two egg-evoked MVPs exhibited distinct responses. These findings suggest that left PFC is involved in selecting the expression of each distinct object state in accordance with the current task demands.

Taken together, these findings are consistent with the proposal that left prefrontal cortex exerts top-down modulatory signals that bias the stimulus-evoked patterns in object-selective regions of VT cortex. In this thesis, we extend this proposal to the domain of lexico-semantic representations by studying word-evoked neural patterns. If left PFC is critically involved the domain-general recruitment of context-appropriate neural signals encoded in posterior brain regions, then left PFC should also modulate the retrieval of variable and context-dependent word meanings. Chapters 2 and 3 describe our investigations of trial-level correlations between left PFC response fluctuations and the expression of context-appropriate word meanings in multi-voxel activity patterns. To foreshadow our results, we observe preliminary evidence for the prediction that left PFC is involved in biasing word-evoked neural patterns toward task-relevant semantic representations. Important qualifications to our findings, and the methodological challenges of this enterprise, are addressed in the Discussion chapter.

4. Current Studies

In Chapter 2, we examine the influence of sentential context on the representations of word meanings during lexical ambiguity resolution. In this study, we exploit historical accidents in language, whereby the same word form is associated with two distinct referents (e.g., river-bank and money-bank). We measure the neural patterns evoked by each distinct word meaning, and the extent to which the stronger, more dominant meaning interferes with retrieval of each word's weaker meaning. Further, we link item-level measurements of (1) the degree of word meaning competition and (2) left vlPFC response to the expression of context-appropriate neural patterns in left anterior temporal lobe.

In Chapter 3, we measure the neural patterns evoked by the same concept under two distinct task conditions: once while subjects think about the general meaning of

concept (i.e., the stimulus word), and then later while subjects explicitly retrieve information about the taxonomic domain that the concept belongs to (i.e., living versus nonliving). We predicted that each task will bias subjects to focus on different aspects of a given concept, which will result in variable neural patterns across the two tasks. Task-dependent neural responses emerged in several brain regions, including univariate responses in left inferior frontal cortex and multivariate responses in right temporal pole and ventral temporal cortex. In these brain areas, univariate and multivariate responses to different object categories were modulated by task demands. Furthermore, at the individual item level, we tested a specific set of predictions about the relationships among the concepts and their corresponding neural patterns, and how these relationships would be altered by the task demands and by left vlPFC response. We observed weak to moderate support for these item-level hypothesis, and discuss potential avenues for future research.

In Chapter 4, we again examine the neural patterns evoked by the same concept at different moments. However, rather than explicitly biasing subjects to retrieve a specific interpretation of each stimulus word, we embed the words in equally random contexts (i.e., word lists) and measure the neural response patterns while subjects retrieve the word meanings from these random contexts in an undirected manner. We observe that a concept's degree of neural variability scales with its degree of meaning variability, such that concepts with stable meanings exhibit relatively stable patterns, and concepts with more flexible meanings will yield greater flexibility in their neural patterns across instances. This study demonstrates a direct link between meaning variability and neural variability, and has important theoretical implications for neuroscientific approaches to studying conceptual representation.

In sum, this thesis addresses the hypothesis that neural patterns change when meanings change, and that the degree of pattern change can be predicted by meaningful, theoretically motivated variables, including left vlPFC response during goal-directed semantic retrieval.

II. TRACKING COMPETITION AND COGNITIVE CONTROL DURING LANGUAGE COMPREHENSION WITH MULTI-VOXEL PATTERN ANALYSIS

1. Introduction

The field of psycholinguistics explains the resolution of lexical ambiguity as the consequence of selection between co-activated and competing interpretations of an ambiguous word. This view is akin to how researchers in the fields of perception, attention, and memory conceive of selection; namely, that it is a consequence of both bottom-up and top-down signals that drive competitive interactions between incompatible representations. In the present study, we take advantage of newly developed fMRI analysis techniques that have been usefully deployed to study the factors that influence selection and conflict resolution in domains of attention (e.g., Kamitani & Tong, 2005; Reddy et al., 2009) and memory (e.g., Kuhl et al., 2011), and apply them for the first time to track competitive interactions during language comprehension. For instance, when readers must select a weaker, subordinate meaning of an ambiguous word (e.g., a river "bank") over a stronger, dominant interpretation (e.g., a money "bank"), how (and where) does the resolution of this competition manifest in neural signals?

One useful approach for identifying interference from a task-irrelevant, competing response is to look for lingering "traces" of it in spatially distributed neural response patterns using multi-voxel pattern analyses (MVPA) of fMRI data. To accomplish this, researchers first measure the multi-voxel pattern (MVP) of activity evoked by a stimulus item, and then render this item irrelevant through a task manipulation. They then measure the MVPs elicited by another stimulus item that is somehow associated with the now-irrelevant stimulus, and determine the extent to which the MVPs evoked during the updated item resemble the responses that were evoked during the now-irrelevant, original item. In the episodic memory domain, researchers have used this technique to quantify competition during targeted memory retrieval, where the same cue simultaneously elicits two associated memories, although one of the associates is task-irrelevant (e.g., Kuhl et al., 2012; Wimber et al., 2015). Similarly, in a study of event comprehension, Hindy and

colleagues (2015) examined whether MVPs reflected the co-activation of two mutually exclusive states of the same object.

These studies have revealed that the degree of interference from the inappropriate representations, as manifested by their presence in MVPs in posterior cortical regions, was inversely predicted by increased recruitment of prefrontal cortex (PFC). We propose that PFC serves a domain-general role in biasing selection of task-relevant representations over competing alternatives. In the present study, we extend this proposal to the domain of lexical ambiguity resolution, and predict that PFC will similarly support the selection of MVPs evoked by subordinate, context-appropriate homonym meanings over dominant, context-inappropriate meanings.

1.1. Role of Left Ventrolateral Prefrontal Cortex in Lexical Ambiguity Resolution

When comprehending everyday text and speech, the vast majority of words that we encounter have some degree of fluidity in their meaning, such that a single word might refer to one of several different meanings each time it is invoked. The sentence context in which a word is embedded serves as a critical cue to the word's intended meaning. Although context serves an irrefutable role in resolving this ambiguity, the relative scope and timing of its influence is largely unresolved. How (and when) do contextual factors influence word comprehension? In order to gain traction on these questions, numerous psycholinguistic experiments have investigated the online comprehension of lexically ambiguous words, such as homographic homophones. For these words (hereafter called homonyms), the same phonemic and orthographic markers refer to two or more distinct and unrelated meanings.

Because several meanings are associated with a single word form, even context-inappropriate, alternative meanings can be inadvertently activated upon encountering a homonym. Readers and listeners must rapidly select the appropriate referent at the expense of all other possible meanings, which may require resolving competition between co-activated referents. One candidate brain region for enabling a top-down bias toward context-appropriate representations is the left vlPFC (ventrolateral prefrontal cortex). In previous fMRI investigations, left vlPFC is consistently recruited during the presentation of sentences that contain homonym words, relative to unambiguous single-

sense words (e.g., Rodd et al., 2005; 2012; Hoenig & Scheef, 2009; Vitello et al., 2014). In addition, VLPFC activity (in particular, the left-lateralized inferior frontal gyrus and inferior frontal sulcus) increases when sentences bias interpretation toward (i.e., invoke) a homonym's subordinate meaning, relative to its dominant meaning (Zempleni et al., 2007). Left vlPFC response is greatest for subordinate-biased "polarized" homonyms, whose subordinate meanings exhibit the weakest associations to the word form (Mason et al., 2007). This response profile is consistent with the role of a modulatory mechanism that biases the interpretation of ambiguous words, either by boosting selection of the context-appropriate meaning, dampening selection of the inappropriate meaning, or some combination of the two.

1.2. Role of Left Ventrolateral Prefrontal Cortex in Domain-General Conflict Resolution

More generally, beyond the domain of lexical ambiguity, this same region is consistently recruited during the resolution of competition amongst conflicting, co-activated representations (e.g., Thompson-Schill et al., 2005; January et al., 2009; Hindy et al., 2012). In fact, the act of selecting a weaker word meaning amidst interference from a competing, stronger meaning has much in common with the processes involved in the Stroop task (MacLeod, 1991; Stroop, 1935). During incongruent trials of Stroop color-word interference task, subjects must respond according to one stimulus dimension (i.e., the word's display color) and ignore a stronger yet task-irrelevant dimension of that same stimulus that would yield an incorrect response (i.e., the color referred to by the stimulus word).Whether selecting a weak, subordinate meaning of a homonym word during lexical ambiguity resolution or reporting a stimulus words' display color instead of its name, in both cases, subjects must select between two simultaneous and mutually exclusive representations. To examine the functional and anatomical correspondences between lexical ambiguity resolution and domain-general cognitive control processes, we functionally localized subject-specific, conflict-sensitive regions of left VLPFC using a Stroop interference paradigm.

1.3. The Current Study

An extensive body of psycholinguistic research indicates that the competition between potential homonym meanings is greatest when the supporting context biases readers toward the selection of a subordinate referent that is only weakly associated with the word form (e.g., river-bank) (Duffy, Morris, and Rayner, 1988; Swaab, Brown, & Hagoort, 2003; Swinney, 1979). In order to resolve this conflict between co-activated alternatives, the reader must select the subordinate yet context-appropriate meaning over the dominant yet context-inappropriate meaning. What are the neural systems that support this process? Further, what neural and psychological factors influence the degree to which a dominant, inappropriate meaning is activated? To address these questions, we tracked the competition between homonym meanings as it unfolds in the brain.

We reason that dominant and subordinate meanings should evoke distinct neural responses in regions of the brain that are sensitive to variations in lexical-semantic information. To index competition between the two meanings, we computed the similarity between their corresponding neural patterns of activation. In particular, we measured the MVPs elicited while subjects first thought about a homonym's dominant meaning, and later on, its subordinate meaning. We then examined how the degree of competition between these neural responses (i.e., their neural similarity) varied across changes in meaning frequency; sentence context; and fluctuations in left VLPFC BOLD response.

We predicted that meaning frequency would positively predict the degree of competition. That is, the association strength between a homonym word form and its dominant meaning (i.e., its meaning frequency) should predict the similarity between the dominant-biased and subordinate-biased neural patterns, such that polarized homonyms should exhibit greater within-word neural similarity than more balanced homonyms, where the meaning frequencies of the dominant and subordinate meaning are relatively more equal. Secondly, we predicted that activity in left VLPFC would be associated with the top-down selection of the context-appropriate, subordinate meaning over the inappropriate, dominant meaning, and that this would manifest as decreased competition (i.e., less within-word neural similarity) during increases in left VLPFC response. As a secondary aim, we also investigated magnitude of BOLD response during sentence

comprehension, and in particular, whether left VLPFC activity is modulated by the relative location of disambiguating sentence context.

2. Methods

2.1. Subjects

Thirteen right-handed, native English speakers (6 males), aged 20-29 years, participated in this study. Subjects were not currently taking any psychoactive medications and had no history of neurological disorders. All subjects had normal or corrected to normal vision. One additional subject was removed from analysis and replaced due to an unusually low response rate during the sentence-reading task (responded to 11% of trials, 4.4 standard deviations below the mean of all other subjects). Subjects were recruited from the University of Pennsylvania community. All subjects were paid $20/hr and gave informed consent as approved by the University of Pennsylvania Institutional Review Board.

2.2. Stimuli

2.2.1. Main Homonyms and Meaning Frequency (M1) Scores

The main testing materials consisted of 30 ambiguous words in which the two most common meanings both refer to nouns (i.e., "ball"). These noun-noun homonyms were selected from a previous norming study that had tabulated the frequency counts of various meanings of several ambiguous words (Twilley et al., 1994). In these norms, frequency scores for the most dominant word meaning (hereafter, M1) were computed by instructing behavioral subjects to generate a semantic associate for each ambiguous word. For each homonym, the authors determined the proportion of responses related to each possible meaning. In the present study, 30 of these homonyms were chosen to allow for a range of M1 scores across items (M= 0.75, SD= 0.14, Figure 1). M1 scores are weakly correlated with log word frequency, r= .29, $t(28)$= 1.66, p= .10 (Brysbaert & New, 2009).

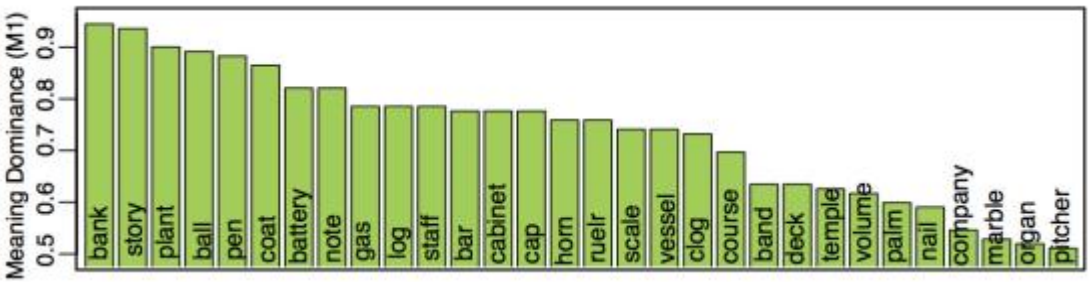

Figure 2.1

The M1 (meaning dominance) score for the dominant meaning of the 30 main homonyms (Twilley et al., 1994).

2.2.2. Filler words

In addition to the main homonym stimuli, we included a set of single-sense nouns and additional noun-noun homonyms. First, 30 single-sense nouns ("synonyms") were selected to match the dominant meaning of each main homonym. These synonyms were originally included to localize brain areas that exhibit similar MVPs in response to the dominant homonym meanings and their intended single-sense synonyms; however, this analysis failed to identify any reliable group-level effects. We will return to this null finding in the Discussion section. Second, to reduce the likelihood that subjects could predict the to-be-invoked meaning of a given homonym prior to sentence reading, we selected another 16 noun-noun homonyms and 12 single-sense nouns. Additional details about these filler word conditions are provided below.

2.2.3. Sentence Stimuli

Each of the 30 main homonyms appeared in two different sentence conditions: once in a dominant-biasing context, and once in a subordinate-biasing sentence context. There were two types of subordinate-biasing sentences: prior context (hereafter, sub-PC) and delayed context (hereafter, sub-DC). In sub-PC sentences, the homonym appeared near the end of the sentence, after the earlier words provide support for the subordinate homonym meaning. In sub-DC sentences, the homonym appeared early on in the sentence, such that the disambiguating contextual information was delayed until the end of the sentence (see Table 1). For the dominant-biasing sentences (hereafter, dom-PC),

the homonym always appeared near the end of the sentence, preceded by words that supported the dominant meaning. Additionally, each dom-PC sentence was transformed into a single-sense sentence (hereafter single-syn) by replacing the homonym with its corresponding single-sense, synonymous noun. These four sentence conditions did not differ in letter length (M= 37.2, SD= 3.6), $F(3,116)= 1.21$, $p>.3$ or number of words (M= 6.9, SD= .89), $F(3,116)= 1.31$, $p>.2$. While all subjects received the same dom-PC and single-syn sentences, assignment of main homonym to either a sub-PC or sub-DC sentence was counterbalanced across subjects.

To ensure that all sentences could be read and adequately comprehended within the 3000ms presentation duration employed during fMRI scanning, we first conducted a pilot study in which a separate group of behavioral subjects (n=6) performed a self-paced reading task with these sentence stimuli. The sentence conditions were randomly interleaved, and each sentence was presented in isolation in the center of the display screen. Subjects were instructed to press a key once they were finished reading the sentence. To confirm that subjects semantically engaged with the sentences, 40% of the sentences were followed by comprehension questions that required subjects to make "yes" or "no" responses based on content from the immediately preceding sentence. Across stimulus conditions, subjects completed reading the sentences in less than 3000ms (M= 1871ms, SD= 105), and responded to the comprehension questions with well above chance performance (M= 94.1, SD= .10). To ensure that each individual sentence would be appropriate for the 3000ms presentation timeframe, we applied conservative exclusion criteria: a sentence was removed or replaced if (1) it elicited a group average response time (RT) greater than 2500ms or (2) the RT of any one subject exceeded 2800ms.

Table 2.1

Condition	Example sentence presentation
dom-PC	The fortune teller gazed into the crystal **ball**.
sub-PC	The queen danced at her birthday **ball**.
sub-DC	The **ball** was on the queen's birthday.
single-syn	The fortune teller gazed into the crystal **orb**.
dom-DC	The **trunk** was filled with groceries.

Example sentence conditions. Each sentence's respective homonym or single-sense synonym word is highlighted in bold above, but appeared in normal font during the experimental procedure. Dom-PC = dominant meaning, prior context; Sub-PC = subordinate meaning, prior context; Sub-DC = subordinate meaning, delayed context; Sing-Syn = single-sense word, synonym to dominant meaning; dom-DC = dominant meaning, delayed context.

2.3. Design Overview

The primary goal of this procedure was to create conflict between two potential representations that might be retrieved upon the presentation of a homonym word. Findings from eye-movement studies, in which participants read sentences that contain an ambiguous word, indicate that readers require additional time to read disambiguating information that biases interpretation toward a homonym's subordinate meaning (Rayner, 1998). We created a scenario to maximize the likelihood that subjects would retrieve the dominant, previously selected meaning of a homonym during the subsequent presentation of a subordinate-biasing context.

In the first half the experiment (runs 1-4), subjects read sentences that biased the interpretation of a main homonym toward its dominant meaning. After reading the sentence, the homonym was presented in isolation, and subjects were instructed to retrieve the word meaning which had been invoked in the immediately preceding sentence (i.e., the dominant meaning). In the second half of the experiment (runs 5-6), each main homonym then reappeared in a sentence that biased interpretation toward its subordinate meaning (either sub-PC or sub-DC, see Figure 2). Subjects then again read each homonym word in isolation, this time retrieving the weaker meaning. Here, the

question was whether the retrieval of the subordinate meaning would receive interference from the dominant, previously invoked meaning.

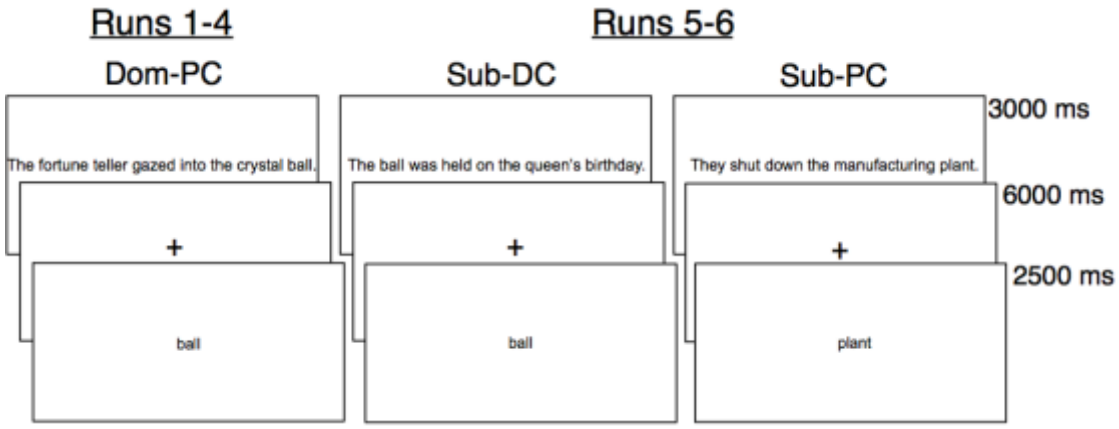

Figure 2.2

Trial structure and condition sequences. Word stimuli first appeared in a sentence, followed by an isolated presentation of the targeted homonym or synonym. The main homonyms appeared in one Dom-PC sentence in runs 1-4, and in one sub-DC or one sub-PC sentence in runs 5-6. Subjects performed the sentence-reading task during the sentence presentations and the semantic retrieval task during the word presentations. Each semantic retrieval trial was followed by a jittered inter-trial interval for 500-12,500ms during which a fixation cross was displayed.

2.4. Trial Sequences

We collected fMRI data during six acquisition runs comprising 134 trials. Each trial consisted of a 3000ms sentence presentation, followed by 6000ms fixation cross, and then the presentation of a single word from the preceding sentence (e.g., the main homonym) for 2500ms (Figure 2). Following the word presentation, a fixation cross was presented during a jittered ITI (500-12500ms). Within runs, trial orderings were randomized using Optseq2, an optimization program for sequencing trials in event-related experiments (http://surfer.nmr.mgh.harvard.edu/optseq).

Across runs 1-4, the 30 main homonyms each appeared in one dom-PC trial. In addition, a single-sense version of each dom-PC trial, in which the homonym was replaced with an unambiguous synonym (single-syn), also appeared in runs 1-4. The trial orders were pseudorandomized, such that a dom-PC trial never appeared in the same run as its single-syn counterpart. In runs 5-6, half of the main homonyms reappeared in a sub-

PC trial, and the other half appeared in a sub-DC trial. To balance the temporal distance between each homonym's dominant and subordinate presentations, subject trial sequences were yoked, such that the ordering for one subject was matched to another subject, but their sub-DC sentences were switched to sub-PC sentences, or vice-versa.

With these trial sequences, a homonym's invoked meaning could be predicted by the experiment half or a homonym's relative location in a sentence. To minimize these cues, we included sixteen filler homonyms that appeared in two different sentences, once in each experiment half. Both of its sentences biased interpretation toward the dominant meaning, and the homonym appeared early the sentence, such that the disambiguating context was delayed (i.e., dom-DC). In addition, six single-sense filler trials appeared in runs 5 and 6, such that half of the single-sense words appeared early on in their sentences, and the other half appeared later in the sentence. Runs 1-4 each consisted of 19 trials (5 minutes/run), and runs 5-6 each had 27 trials (7 minutes/run).

3. Procedure

3.1. Sentence-reading Task

Stimuli were presented using E-Prime (Psychology Software Tools). Sentences appeared in the center of the screen in Arial font subtending approximately 0.5 degrees visual angle per letter. Subjects were instructed to respond via button press once they finished reading the sentence. After 3000ms elapsed, the inter-stimulus interval (ISI) consisted of a centrally located fixation cross displayed for 6000ms. Subjects responded to the majority of trials (M = 85.1%, SD = 16.7%), and each subject indicated during a post-scan debriefing session that they had adequate time to read each sentence. Across the main homonym sentence conditions (i.e., dom-PC, sub-DC, sub-PC) there were no significant differences in response times, $F(2,24)= 1.62$, $p=.22$. Mean response times (1817ms) were consistent with the self-paced reading times from the pilot study (1871ms), $t(10.13)= -.27$, $p= .80$.

3.2. Semantic Retrieval Task

Following the sentence presentation and intervening fixation cross, a single word from the preceding sentence appeared on the screen for 2500ms. Participants were instructed to think about the meaning of this word that was supported by the sentence context that they had just read. No behavioral measures were collected during this task.

3.3. Stroop Interference Task

After completing runs 1-6, subjects completed a single run of a Stroop color identification task (cf. Hindy et al., 2012; Hindy et al., 2015). On each trial, subjects were presented with a single word and were instructed to press one of three response buttons that corresponded to the typeface color (i.e., blue, yellow, or green). The single word referred to either a color name (e.g., yellow, red) or a non-color, neutral noun (e.g., stage, tax, and farmer). Each word appeared for 1800ms followed by a 1200ms ITI. The conflict condition consisted of trials where the color name did not match the color of the typeface. In the neutral condition, the color name and typeface color matched, or a non-color, neutral noun was presented. Subjects responded correctly to 98.4% of Stroop trials. Response latencies for conflict trials (M= 721ms, (SD= 186ms) were slower than responses to neutral trials (M= 671ms, SD= 191ms), $t(12)= 7.90$, $p < .001$). In a group-level, univariate contrast of conflict versus neutral trials, left VLPFC was reliably more responsive to Stroop conflict than adjacent brain regions. The anatomical location of the top 100 conflict-responsive voxels in left VLPFC was heterogeneous across subjects (Figure 3).

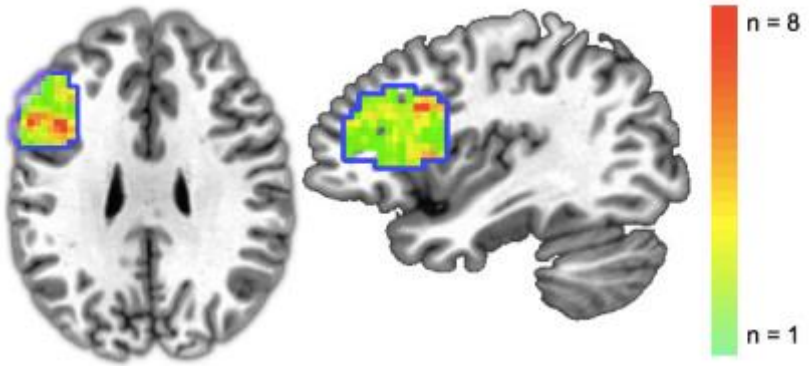

Figure 2.3

Probabilistic overlap map of the subject-specific Stroop-conflict ROIs in left pVLPFC. Anatomical constraints of left VLPFC are outlined in blue. This anatomical ROI was transformed into each subject's native brain space. In each subject, we selected the 100 voxels which yielded the highest *t*-statistics in the contrast of conflict versus neutral trials during the Stroop task. For display purposes, these subject-level masks were transformed to standardized Talaraich space and overlaid to create a group mask.

3.4. fMRI Data Acquisition

Anatomical and functional data were collected on a 3T Siemens Trio system and a 32 channel array head coil. Anatomical data consisted of 160 slices of axial T1-weighted images with 1 mm isotropic voxels (TR= 1620 ms, TE= 3.87 ms, TI=950 ms). Functional data included echo-planar fMRI collected in 44 axial slices and 3 mm isotropic voxels (TR= 3000 ms, TE= 30 ms). To approach steady state magnetization, twelve seconds preceded data acquisition in each functional run.

3.5. fMRI Preprocessing

Data preprocessing and statistical analyses were performed with AFNI (Cox, 1996) and MATLAB scripts implemented in the Princeton MVPA Toolbox (Detre et al., 2006). Functional data were sinc interpolated for slice timing correction, aligned to the mean of all function images using a seventh-order polynomial interpolation, and co-registered to the structural data. Data were then smoothed with a 4mm FWHM Gaussian kernel and z-normalized within each run.

3.6. Whole-brain Regression Analyses

We preformed two whole-brain analyses: a condition-level, univariate analysis, and an item-level, multi-voxel pattern (MVP) analysis. In both cases, a modified general linear model (Worsley & Friston, 1995) was fit to each subject's preprocessed data. Each trial segment was modeled with a canonical hemodynamic response function convolved with a boxcar that matched the duration of the trial segment (i.e., 3000ms for each sentence, 6000ms for each fixation ISI, and 2500ms for each word). For the condition-level, univariate analysis, a binary regressor was included for each sentence and word condition (i.e., dom-DC; sub-DC; sub-PC; single-syn; and dom-DC). For the item-level

MVP analysis, a unique regressor was included for each individual sentence and word presentation. For both models, scanning run and six motion parameters were modeled as covariates of no interest. For group-level, random-effects analyses, subject-level statistical maps were normalized to Talaraich space. In order to correct for multiple comparisons, minimum cluster extent was determined using AFNI's 3dClustSim (version built December 9, 2015). For this correction, we first estimated the smoothness of the data using the residual time series data using AFNI's 3dFWHMx spatial autocorrelation function. Based on a voxel-level uncorrected alpha of 0.001 (t= 4.29), Monte Carlo simulations (n=50,000) indicated a minimum cluster extent of 10 voxels for a cluster-level corrected alpha of .05.

3.7 ROI Analysis: Left VLPFC

Each Stroop-conflict ROI was anatomically constrained according to probabilistic anatomical atlases that were transformed into Talaraich space (Eickhoff et al., 2005). Left VLPFC was defined as the combination of pars opercularis (BA 44), pars triangularis (BA 45), and the anterior half of the inferior frontal sulcus. Because the Stroop task entails multiple, distinct forms of conflict (e.g., motor response, task set, and color representation), this anatomical constraint allows for the selection of cortical areas that are most likely to be involved in the cognitive process of interest. The anatomical constraint to left VLPFC ensured that this ROI reflected conflict-related processing at the level of semantic representation (cf. Hindy et al., 2012). Across subjects, this left VLPFC anatomical ROI consisted of an average of 1024 voxels (SD = 99). Within these anatomical boundaries, the Stroop-conflict ROI was further limited according to each individual subject's functional data from the Stroop color-word interference task. Specifically, the ROI was confined to the 100 voxels that exhibited the highest t-statistics for the contrast of conflict versus neutral trials. This functional constraint ensures that the voxels included in this ROI were most sensitive to conflict on a subject-specific basis.

For the ROI-based regression analyses, voxel-wise activation values were averaged across the entire Stroop-conflict left VLPFC ROI in each subject. For the condition-level analysis, we tested the same contrasts described in the whole-brain analysis. For the item-level analysis, we measured the mean BOLD signal evoked during

each main homonym's two word presentations (i.e., following its dominant- and subordinate-biasing sentences), and subtracted the average "dominant" response from the average "subordinate" response. This item-level measure serves as an index of the change in left VLPFC recruitment during the presentation of the dominant versus subordinate meaning of each main homonym.

3.8. Whole-brain Multi-Voxel Pattern Searchlight Analysis

To assess the similarity of multi-voxel, item-specific responses evoked during each word presentation, we passed a spherical searchlight with a 3-voxel radius over each voxel in the brain (Kriegeskorte et al., 2006). (The main searchlight results were also confirmed when the searchlight size was increased to a 4-voxel radius). In each searchlight volume, MVP similarity was measured as the Pearson correlation between the multi-voxel responses evoked by the dominant versus subordinate word presentations of the same main homonym. In a subject-level, parametric analysis, we used M1 scores to predict the similarity between the MVPs evoked during each homonym's dominant and subordinate presentations. Here, we estimated a separate linear regression coefficient for each subject that predicted the MVP similarity of each homonym based on its M1 score. The resulting beta value was then assigned to each searchlight center. We then used 1-sample *t*-tests to determine the cross-subject reliability of the regression coefficients. This analysis is akin to entering Pearson correlation coefficients in a second-level analysis, instead of linear regression coefficients.

4. Results

4.1. Univariate Results

4.1.1. Whole-brain Analysis

In an exploratory, whole-brain analysis, we first contrasted the responses for the various sentence conditions. The contrast between sub-DC sentences versus dom-PC sentences yielded a large area of activation in left VLPFC, extending anterior and dorsal to the Stroop-conflict functional ROI. This cluster overlapped with a cluster resulting from the contrast of sub-DC sentences versus sub-PC sentences (Figure 4). The

coordinates and peak voxel values are listed in Table 2. The contrast of sub-PC versus dom-PC did not yield any reliable above-threshold activation, nor did the contrast of single-syn sentences versus any of the three homonym sentence conditions.

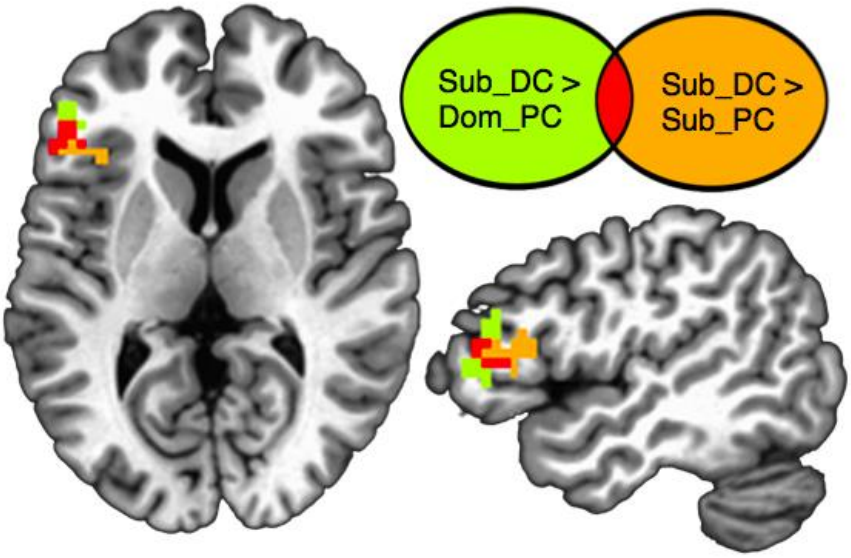

Figure 2.4

Univariate whole-brain results for BOLD responses during the sentence-reading task. Subordinate-delayed context (Sub-DC) sentences elicited a greater response than both dominant-prior context (Dom-PC) and subordinate-prior context (Sub-PC) sentences in an overlapping area of left inferior frontal gyrus. Colored voxels depict areas with above-threshold activity in a cluster-corrected group-level analysis.

4.1.2. Stroop-Conflict Selective Voxels in Left VLPFC ROI

In addition to the whole-brain analysis, we compared the mean BOLD response for each sentence condition in each subject's top 100 Stroop-selective voxels in an anatomically constrained region of left VLPFC (Figure 5). This analysis recapitulated the results that emerged at the whole brain level: mean left VLPFC response in the Stroop-conflict selective voxels was greater during the presentation of sub-DC sentences than sub-PC sentences, $t(12)= 4.20$, $p= .001$, and for sub-DC sentences versus dom-PC sentences $t(12)= 3.50$, $p= .004$. In addition, mean response was greater for sub-DC sentences versus single-syn sentences, $t(12)= 2.46$, $p= .03$.

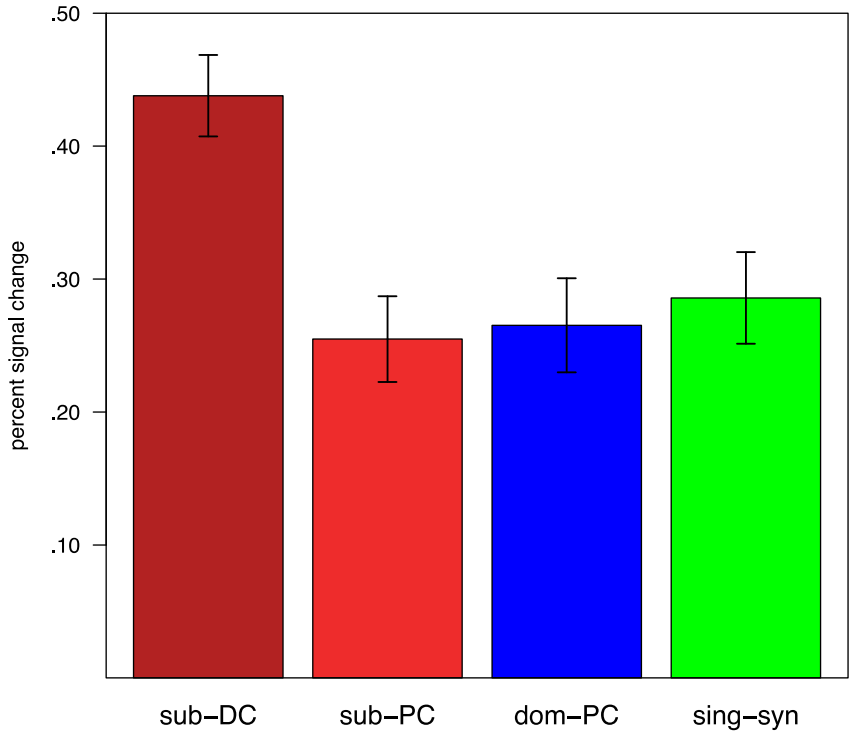

Figure 2.5

Group-average responses during sentence comprehension in left posterior ventrolateral prefrontal cortex, limited to subject-specific, Stroop-conflict selective voxels. Comparisons between sentence conditions were performed within each subject. Sub-DC = subordinate meaning, delayed context; Sub-PC = subordinate meaning, prior context; Dom-PC = dominant meaning, prior context; Sing-Syn = single-sense word, synonym to dominant meaning

4.2. Multi-voxel Searchlight Results

4.2.1. Role of Meaning Frequency

We used a whole-brain, multi-voxel searchlight analysis to examine the similarity between the MVPs evoked during the dominant-biased versus subordinate-biased version of the same homonym. In a group-level analysis, we performed a random-effects analysis using the statistical maps yielded by each subject's searchlight results, in which the linear regression coefficient for M1 was assigned to the searchlight centers. Across subjects, we identified a cluster of 21 searchlight volumes in left anterior temporal lobe (ATL) in which M1 scores reliably predicted the similarity between the MVPs evoked by the dominant- and subordinate-biased presentations of a main homonym (see Table 2 and

Figure 6a), $t(12)= 5.45$, $p= .0001$ (mean $r= .22$, $SD= .13$). In the MVPs sampled in these searchlight volumes, the greater the homonym's M1 score, the greater the similarity in the responses evoked by the two versions of the same ambiguous word. This relationship was positive in all 13 subjects (Figure 6b). Follow-up analyses at the peak left ATL searchlight, in which homonyms were separated based on the relative location of the subordinate-biasing sentence context (i.e., sub-DC or sub-PC) indicate that this result holds when the analysis is limited to the main homonyms that had appeared in sub-DC sentences, $t(12)= 2.88$, $p= .01$, and marginally holds for the sub-PC homonyms alone as well $t(12)= 2.04$, $p= .06$.

Table 2.2

Task	**Effect**	**Location**	**Peak coordinates (*x*,*y*,*z*)**	**Cluster extent**	**Peak *t*-statistic**
Sentence Reading	BOLD response: sub-DC > sub-PC	left inferior frontal gyrus	-46, 26, 8	17	$t(12)= 6.35$, $p< .001$
Sentence Reading	BOLD response: sub-DC > dom-PC	left inferior frontal gyrus	-46,35,11	10	$t(12)= 6.53$, $p< .0001$
Semantic Retrieval	With-word neural similarity: positively correlated with meaning dominance	left anterior temporal lobe	-37,-7,-31	21	$t(12)= 5.45$, $p< .0001$
Semantic Retrieval	Within-word neural similarity: negatively correlated with Stroop-conflict selective lVLPFC response	left anterior temporal lobe			$t(12)= -3.14$, $p< .01$

Whole-brain, group level results. Responses in left inferior frontal gyrus increased during the presentation of subordinate-delayed context (sub-DC) sentences, relative to subordinate-prior context (sub-PC) and dominant-prior context (dom-PC) sentences. During the subsequent presentation of each sentence's homonym word, within-word multi-voxel pattern similarity positively correlated with meaning dominance in left anterior temporal lobe (left ATL). In addition, within-word pattern similarity in the peak left ATL searchlight sphere negatively predicted BOLD response in Stroop conflict-sensitive regions of left ventrolateral prefrontal cortex (lVLPFC).

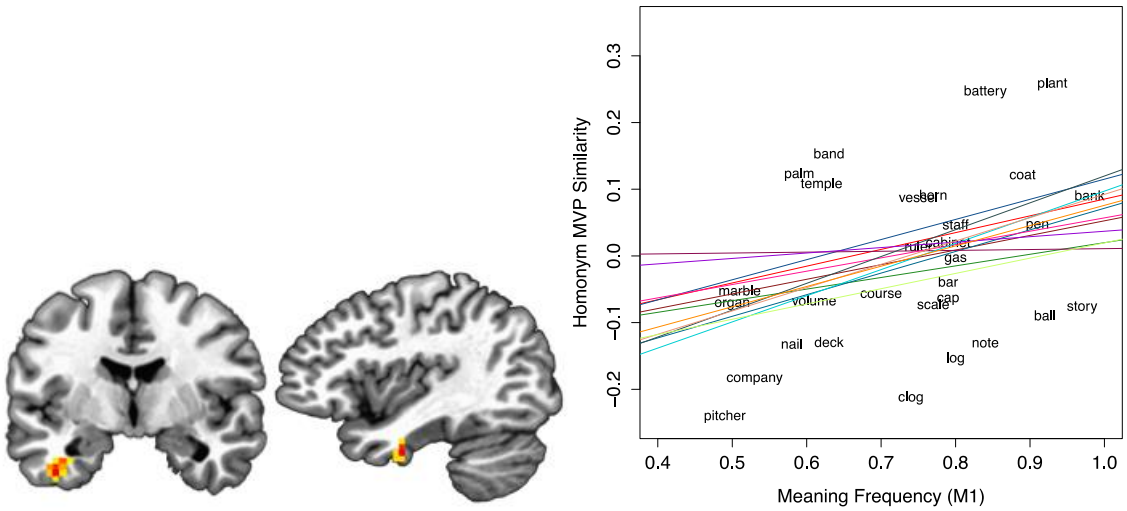

Figure 2.6

(a) In left anterior temporal lobe, meaning frequency (M1) predicted the similarity between the neural patterns evoked during the semantic retrieval of a homonym's dominant and subordinate meanings. (b) The positive relationship between multi-voxel pattern (MVP) similarity and meaning frequency was present in all 13 subjects. The linear trend for each subject is depicted in a different color. Item-level results in a single subject are depicted in the background.

Neural similarity was computed using Pearson's *r*, a similarity measure that is assumed to be largely independent of the absolute magnitude of univariate response. To confirm that the MVP similarity effects we observed in left ATL reveal information that is not redundant to univariate effects, we submitted the neural similarity values to a confirmatory, within-subject regression at the peak left ATL searchlight. For this regression analysis, we used four independent variables to predict M1 scores: neural similarity between the dominant and subordinate MVPs; mean univariate activity during the dominant retrieval; mean univariate activity during the subordinate retrieval; and the interaction between the mean univariate activity during each retrieval period (cf. Ritchey et al., 2012; Wing et al., 2015). Across subjects, the beta coefficient for MVP similarity continued to reliably predicted M1 scores, even with mean univariate response included in the model, $t(12)= 6.4$, $p< .001$ ($M= .21$, $SD= .12$). This confirmatory analysis minimizes the possibility that the searchlight results in left ATL are driven by mean activation differences.

In a follow-up analysis, we extracted the Pearson correlation coefficients for each main homonym at this peak left ATL searchlight center, and then used these values as a seed in a whole-brain analysis to predict changes in univariate response between the dominant versus subordinate word presentations. No reliable above-threshold activity emerged from this analysis.

4.2.2. Role of left VLPFC Response

We also examined role the relationship between left VLPFC activity and dominant and subordinate MVP similarity. For this analysis, we selected the neural similarity values from the peak left ATL searchlight center where within-word neural similarity had exhibited the positive correlation with M1 scores. We then correlated changes in Stroop-selective left VLPFC response during the homonym presentations with the MVP similarities in this peak left ATL searchlight. Across subjects, increases in left VLPFC response from the dominant to subordinate word presentation reliably predicted decreases in the neural similarity between the subordinate and dominant word presentations of the same homonym in left ATL, $t(12)= -3.14$, $p= .01$ ($M= -.07$, $SD= .08$). This relationship was negative in 10 out of 13 subjects.

To further investigate the effects of left VLPFC response on within-word neural similarity, we also performed an exploratory whole-brain searchlight analysis. Here, the change in left VLPFC response in subject-specific Stroop-conflict voxels between the subordinate versus dominant word presentation were used as predictors of MVP similarities in searchlights passed over the entire brain volume. This analysis failed to yield any reliable results at the whole-brain, group level.

4.2.3. Left Anterior Temporal Lobe Results: Role of Left VLPFC and Meaning Frequency

In a subject-level linear regression analysis, we predicted the neural similarity values observed in the peak left ATL searchlight by modeling separate covariates for M1 scores and change in left VLPFC response. Across subjects, the covariates for M1 and change in left VLPFC both reliably predicted neural similarity in left ATL, even when both covariates were simultaneously included in the model, $t(12)= 7.97$, $p= .0001$ for the M1 covariate, and $t(12)= -2.76$, $p= .02$ for the left VLPFC covariate. Moreover, M1

scores and changes in left VLPFC response were not reliably correlated across subjects, $t(12) = -.23$, $p = .81$ ($M = -.01$, $SD = .17$).

5. Discussion

Several neural and behavioral factors have been implicated in semantic ambiguity resolution, including left VLPFC response, homonym-level properties (i.e., meaning frequency), and sentence-level characteristics (i.e., the relative location of disambiguating context). We examined the role of these factors while tracking the outcome of ambiguity resolution using online, item-level neural measures. Our analyses revealed that these three factors each impact the neural correlates of lexical ambiguity resolution. In turn, we discuss each finding and the implications for psycholinguistic models of ambiguity resolution.

5.1. Univariate Findings During Sentence Reading

We first examined changes in BOLD response while subjects read ambiguous noun-noun homonyms within sentence contexts. A whole-brain analysis revealed that BOLD response in left VLPFC was modulated by meaning frequency, such that activity here was greater for subordinate-biasing versus dominant-biasing sentences. However, this effect was limited to subordinate-biasing sentences in which the disambiguating context was delayed (sub-DC). Additionally, in an overlapping set of voxels in left VLPFC, an effect of context position emerged for subordinate-biasing sentences, such that responses were greater when the disambiguating context followed the homonym (sub-DC) compared to when the context preceded it (sub-PC). This pattern of results was recapitulated in an fROI-based analysis, in which we selected subject-specific voxels in left VLPFC that were most responsive to conflict during a Stroop color-word interference task. This approach is important, because there have been suggestions that left VLPFC is a highly heterogeneous region, and subject-specific analyses are necessary to localize activity associated with the distinct process of interest (Fedorenko et al., 2010).

Taken together, these findings confirm the role of left VLPFC in sentence reinterpretation and resolving competition between co-activated representations. The increased recruitment that we observed here is consistent with a scenario in which a

frequency-based probabilistic choice is made between the alternative meanings, and then the meaning is updated if the selected nterpretation does not fit with the subsequent disambiguating context (Zempleni et al., 2007).

5.2. Multivariate Findings During Semantic Retrieval

In addition to examining neural activity during sentence reading, we also measured the neural activity that followed this disambiguation process, once the context had biased interpretation toward a particular homonym meaning. In previous work (Musz & Thompson-Schill, 2015), we have demonstrated the utility of within-item, cross-context neural similarity analyses by showing that the MVP similarity elicited by the same word across different presentations can be predicted by item-level semantic properties. In the present experiment, we employed sentence contexts to bias semantic retrieval toward one of two specific and distinct homonym meanings. We predicted that the neural representation evoked by the same word in the two different contexts would vary, such that these two different meanings would evoke variable neural patterns. Further, we examined the effects of switching the context (and hence the meaning) while holding the word form constant, such that a previously invoked meaning is rendered inappropriate and potentially distracting. Thus, retrieval of the subordinate meaning would require the subject to disregard a salient yet contextually inappropriate word meaning in favor of the weaker representation of the same word.

5.2.1. Meaning Frequency Predicts Within-Word Neural Similarity in Left Anterior Temporal Lobe

We first tested whether meaning frequency correlated with the extent to which subordinate-biased activity patterns resemble dominant-biased MVPs during retrieval of a subordinate meaning. A whole-brain searchlight analysis revealed that, in left ATL, the association strength of the dominant meaning (i.e., M1) predicted the degree of neural similarity between the dominant and subordinate-biased MVPs. Crucially, this effect emerged during the time period that followed the homonym's appearance in a sentence that biased interpretation toward its subordinate meaning. That is, even *after* the subordinate meaning had been supported via linguistic context, the neural patterns in left

ATL still resembled those evoked by the dominant meaning. This finding adds to a growing literature on the role of meaning dominance during lexical ambiguity resolution. These investigations have largely found that the dominant meaning of a homonym interferes with the selection of a subordinate homonym, and this competition between co-activated meanings leads to processing costs (Pacht & Rayner, 1993; Rayner et al., 1994) and increased recruitment of left VLPFC (which we also observed during the presentation of sub-DC sentences). However, the majority of previous studies focused on the time interval during which a subject first encounters the subordinate-biased homonym (cf. Gorfein et al., 2001). In the present analysis, the neural pattern evoked by subordinate meaning was measured six seconds after this meaning had already been invoked in the preceding sentence. Thus, in addition to the competition that arises when a homonym word meaning is first accessed or reinterpreted, we found evidence of competition even after the word meaning has been resolved.

This finding indicates that even when the dominant meaning is rendered irrelevant by an earlier, subordinate-biasing sentence context, it nevertheless competes for activation. A host of previous behavioral research corroborates this finding. Several studies on reading times have revealed that subjects experience processing delays (manifested in increased reading times and regressive eye movements) while selecting in the subordinate meaning of a homonym, even when the supporting linguistic context has supported its interpretation (cf. Duffy et al., 1998; Pacht & Reyner, 1993; Sereno et al., 2006). This performance decrement, termed the "Subordinate Bias Effect" (SBE) has been demonstrated under several experimental conditions in which a previous context is provided to bias interpretation toward the subordinate meaning (e.g., paragraph titles, immediately preceding uses of the subordinate meaning, etc.).

Behavioral studies have found, however, that the SBE can in fact be eliminated by a strong subordinate-biasing preceding context, but only for ambiguous words that are only moderately biased (8-30% strength of the subordinate meaning). For polarized homonyms, in which the strength of the subordinate meaning was very weak (8% or less), the interference from the dominant meaning could not be fully eliminated (Wiley & Rayner, 2002). In a related study, Rodd and colleagues (2012) investigated the extent to which lexical-semantic re-turning can rapidly occur. Subjects performed a free

association task, in which they were presented with a homonym word and were instructed to generate a semantic associate. Twenty minutes beforehand, subjects in the primed condition listened to sentences that invoked the homonyms' subordinate meanings. Relative to unprimed subjects, the primed group was more likely to subsequently generate words related to the subordinate meanings. However, the priming effect was relatively modest: although the proportion of subordinate associates of polarized homonyms increased fivefold (e.g., from 2% to 10%), subjects were still far more likely to produce an associate of the dominant meaning. Taken together, these results suggest that even strong subordinate-biasing contexts cannot override the unintended dominant meaning if it has a very high frequency.

An eyetracking study by Huettig and Altmann (2007) provides a particularly striking demonstration of the interference from context-inappropriate, dominant homonym meanings. In a visual word paradigm, subjects viewed an array of four objects, where some of these objects depicted a homonym's subordinate meaning (e.g., a pig pen) and either its dominant meaning (a writing pen) or an object related in shape to the dominant meaning (e.g., a sewing needle). During the auditory presentation of a subordinate-biasing sentence context, fixations increased for the dominant competitor, and even for an object related in shape to the dominant referent, relative to unrelated control objects. Looks to these competitor objects can be interpreted as evidence that the dominant meaning was activated, despite the contextual support for the subordinate meaning.

Whereas those authors found evidence of transient, online activation of dominant meanings via eye fixations, we tracked the activation of homonym meanings as manifested in the similarity of their evoked neural signals. The neural similarity effects emerged in a left-lateralized subregion of the anterior temporal lobe. This area has been previously associated with increased recruitment during the retrieval of multiple ambiguous word meanings. In a recent study on homonym comprehension, Whitney et al. (2011) found that BOLD activity in this same region was sensitive to the number of homonym meanings that were retrieved. Additionally, Snijders et al. (2009) reported increased activity in an overlapping region of left mid-inferior temporal gyrus (BA 20) while subjects read homonyms that were embedded in equibiasing sentence contexts,

such that two alternative interpretations of the ambiguous word were equally plausible. In conjunction with our effect, these findings suggest that responses in this subregion of left ATL track the activation of several co-activated interpretations of ambiguous words.

These findings are also consistent with a host of previous research that points to a critical role for left ATL in semantic memory. This area's role in semantic processing has been established by convergent findings from patient studies, neuroimaging studies, and brain stimulation research (e.g., Patterson et al. 2007; Visser et al., 2010; 2011; Rogers et al., 2006; Pobric et al., 2010). In fact, our identified searchlight cluster directly overlaps with a site recently identified as critical for semantic processing: Binney et al. (2010) found that BOLD response in this same subregion of left ATL increases while healthy subjects perform a synonym judgment task, and that Semantic Dementia patients with damage to this region exhibit impaired performance on the same task.

In light of the extant findings that implicate left ATL in conceptual processing, we suggest that the MVP similarities that we have identified here reflect the co-activation of the meanings associated with two alternative interpretations of the same homonymous word. However, we cannot conclusively attribute our effects to the activation of semantic information. In a preliminary, whole-brain analysis, we attempted to localize brain areas in which neural similarity tracked semantic relatedness. We compared the similarity between MVPs evoked during the semantic retrieval of dominant-biased homonyms and their intended unambiguous synonym (e.g., "ball"; "orb"). This analysis did not yield any reliable neural similarity effects in response to semantically related versus unrelated homonym-synonym word pairs. Further, we did not find any areas in which neural similarity continuously scaled with subjective, numerical ratings of semantic relatedness.

To further characterize the M1- and left VLPFC-predicted MVPs that we identified in left ATL, we performed follow-up analyses in the peak searchlight volume. In particular, we compared the relative similarities between the MVPs evoked during retrieval of each item's dominant-biased (e.g., sphere-ball); subordinate-biased (e.g., dance-ball); and dominant-synonym (e.g., "orb") presentations. This analysis revealed that the synonym MVPs were more similar to the dominant-biased patterns (mean r=.02) than they were to the subordinate-biased patterns (mean r= .001), $t(12)$= 2.07, p= .06. We also checked whether M1 or left VLPFC activity could predict a synonym's relative

MVP similarity match to the dominant-biased homonym presentation, versus its similarity to the subordinate-biased presentation. We observed a positive relationship between left VLPFC response and meaning match, such that the synonym pattern's relative similarity to the dominant-biased versus subordinate-biased presentation is predicted by increases in left VLPFC response, $t(12)= 2.63$, $p= .02$. That is, when left VLFPC response increases during the subordinate-biased word presentation, its resemblance to the synonym pattern decreases, relative to the similarity between the dominant and synonym MVPs. In contrast, M1 did not reliably predict increases in a synonym's match to the dominant versus subordinate-biased word presentation, $t(12)= .68$, $p= .51$.

These post-hoc findings in left ATL suggest that the neural patterns observed here might encode abstract, conceptual information about word meanings. Alternatively, it is possible that our neural similarity effects in left ATL could reflect the activation of lexical representations that serve as an interface between word form and meaning. With the current data and paradigm, we are unable to determine whether the MVPs that we identified in left ATL represent lexical versus conceptual information (or some combination of the two). Our interpretations of the effects in this region are limited, because although we can predict within-word neural similarity using two parametric, item-level measures (i.e., M1 and left VLPFC response) which have strong theoretical and empirical support for predicting lexical-semantic competition (cf. Twilley et al.; Rodd et al., 2005), we are nevertheless unable to describe the dimensions that govern the observed similarities. Future research will benefit from more extensively characterizing the nature of the representational similarity space evoked by lexical stimuli in left ATL.

To more conclusively determine whether left ATL activity reflects the co-activation of competing word meanings, future analyses should interrogate neural patterns evoked by additional noun-noun homonyms, and several synonyms for both dominant and subordinate homonym meanings. Additionally, more elaborate and in-depth behavioral measures of stimulus processing during sentence comprehension and semantic are necessary to make any strong claims about the extent to which disambiguating linguistic contexts might influence the resulting neural patterns. The present study is the first step in applying a combination of behavioral and fMRI

multivariate analysis techniques to advance our understanding of how people interpret ambiguous linguistic input (see also Danelli et al., 2015). The current work demonstrates the promise and utility of this approach.

5.2.2. Left VLPFC Activity Negatively Predicts Within-Word Neural Similarity in Left Anterior Temporal Lobe

The meaning frequency effects in left ATL suggest that the dominant meaning of polarized homonym words might always be retrieved, regardless of context. But does biasing context have *any* effect on the activation of the dominant meaning? To address this question, we tested whether BOLD response in left VLPFC tracks decreases in neural similarity between the activation patterns evoked by context-appropriate and context-inappropriate homonym meanings. This analysis revealed that when left VLPFC response increases during the subordinate meaning retrieval, within-word neural similarity decreases in left ATL. We suggest that the reductions in neural similarity reflect the task-driven expression of the subordinate, contextually appropriate word meaning, and its distinction from the initial, contextually inappropriate dominant meaning, thereby increasing the dissimilarity between their corresponding neural patterns. When a comprehender must resolve the interference caused by alternative meanings of a single word form, left VLPFC may act as a top-down modulatory signal to bias neural patterns toward the contextually appropriate representation.

Empirical support for this proposal comes from both our own data in the same set of subjects, and from numerous other studies. In the present study, we demonstrated that left VLPFC response is associated with the reinterpretation of homonym meanings, in which a subordinate meaning must be selected over an initially activated dominant meaning. Moreover, during the Stroop conflict task, responses here increased during conflict trials, during which distracting information (i.e., incongruent color names) must be ignored. Further, evidence from converging methods, including patient lesion data, TMS, and fMRI demonstrate that this region is activated during, or is necessary for, selecting contextually-appropriate meanings of ambiguous words (Thompson-Schill et al., 2005; Bedny et al., 2007, 2008; Rodd et al. 2005, 2012; Ihara et al., 2014); completing sentences with multiple alternative responses (Robinson et al., 2005);

generating verbs with many semantic competitors (Thompson-Schill et al., 1997); and overriding misinterpretations of syntactically ambiguous sentences (January et al. 2009; Rodd et al., 2010).

The linear effect of left VLPFC response magnitude on neural similarity suggests that multiple homonym meanings compete for activation during the semantic retrieval of a single meaning, and that left VLPFC tracks the resolution of this conflict. This result is compatible with a handful of other studies that have reported a relationship between left VLPFC activity and dissimilarity between MVPs evoked by competing stimuli elsewhere in the brain. In a recent study by Hindy and colleagues (2015), in early visual cortex, the neural dissimilarity between MVPs evoked by two incompatible states of the same object (e.g., a cracked versus intact egg) was predicted by increased left VLPFC response during the presentation of the object in its second state.

Likewise, MVPA studies in the domain of episodic memory, recent studies have found that recruitment of frontal cortex during the encoding (Kuhl et al., 2012) and the retrieval (Wimber et al., 2015) of updated memories predicts decreased competition from earlier memories. One interesting possibility is that episodic interference from older memories may have played a role in the present study as well. In our paradigm, dominant meanings were presented in the first half of the experiment, followed by the subordinate meanings in the second half. Perhaps subjects experienced episodic interference from the memory event of comprehending and retrieving the dominant meaning earlier in the experiment. However, it is unclear how various sources of potential interference (e.g., episodic or semantic) might interact and influence lexical ambiguity resolution. This open and interesting question warrants further study.

Taken together with our findings, we propose that left VLPFC serves as a domain-general, top-down control signal that suppresses competition between co-activated neural representations, and that the outcome of this modulatory role can be identified in the dissimilarity between neural patterns evoked in posterior cortical areas. However, although the pattern-predicted increase in left VLPFC response was reliable across subjects, it was not robust at the whole-brain level. Rather, the relationship between left VLPFC response and left ATL neural similarity was identified through the fROI-based analyses, in which we limited our analyses to the fluctuations in BOLD

response in subject-specific, Stroop-conflict sensitive regions of left VLPFC. Why did this relationship fail to emerge at the whole-brain level? One possibility is that there are individual differences in the extent to which left VLPFC is recruited while subjects retrieve a context-appropriate homonym meaning. In fact, previous fMRI studies on lexical ambiguity resolution have found that prefrontal recruitment during the retrieval of subordinate meanings can be predicted by individual differences in reading span (Mason & Just, 2007) and behavioral performance during a semantic interference task (Hoenig & Scheef, 2009). Additional research is necessary to determine the subject-specific variables associated with pattern-predicted activity in left VLPFC.

Although the current study focused on the role of left VLPFC, other studies indicate that additional brain regions also participate in cognitive control processes (e.g., right prefrontal cortex and anterior cingulate cortex (ACC)), particularly when an overt response is required. For instance, along with left VLPFC, responses in ACC and right PFC increase during judgements of homonym words (Bedny et al., 2008; Chan et al., 2004; Hoenig & Scheef, 2009) and during incongruent trials of the Stroop task (Macleod & MacDonald, 2000). In contrast, BOLD response in ACC and right PFC was not modulated by sentence condition in our whole-brain analysis. The lack of reliable activity in these regions has also been observed in other fMRI studies that, similar to our experimental paradigm, measured BOLD response during passive comprehension of homonyms embedded in sentence contexts (e.g., Rodd et al., 2009; Vitello et al., 2014; Zempleni et al., 2007). This differential response profile suggests that the recruitment of brain regions implicated in cognitive control processes depends on the specific task demands (Milham et al., 2001).

5.3. Conclusions

The representation of multiple lexical-semantic representations of the same homonym word across contexts, and how these representations might compete for activation, has not been extensively studied. The data reported here suggest that not only do ambiguous word meanings compete for selection in left ATL, but also that the extent of their competition is driven by both bottom-up features (frequency-based form-to-meaning associations) and top-down neural signals (left VLPFC response magnitude).

We present the first step in identifying the representational mechanisms that given rise to successful resolution of semantic ambiguity.

III. CATEGORY TYPICALITY MODULATES GOAL-DIRECTED RETRIEVAL OF LIVING AND NONLIVING THINGS

1. Introduction

Humans possess the important and impressive ability to represent the same object as an instance of several different meaningful categories. For example, a pine tree can be represented as a member of both the "things that are living" category and the "things that are immobile" category. Moreover, we can dynamically select the representation of the object that is most appropriate to the task, or category, at hand.

If we can have different thoughts about the same object, then one might predict that the neural responses evoked by these different thoughts would also vary. In recent years, neuroscientists have leveraged multi-voxel pattern analysis (MVPA) to study how neural representations, manifested in the spatially-distributed activity patterns that are evoked by pictures of objects, are altered by changes in attention, experience, and task demands. This line of research has revealed that large swaths of the brain including the ventral temporal cortex (Harel et al., 2014; Senoussi et al., 2016), fronto-parietal regions (Erez & Duncan, 2015; Bracci & Op de Beeck, 2017), and prefrontal cortex (Hanson & Chrysikou, 2017; Bugatus et al., 2017) exhibit flexible and task-dependent neural response profiles, such that the distinctions, associations, and commonalities amongst stimuli are enhanced once such boundaries and groupings become behaviorally relevant (Carlson et al., 2014). These studies serve as elegant demonstrations of how an observer's behavioral goals can exert influences on object perception, and how these effects are manifested in changes to the tuning properties of multivariate activity patterns throughout cortex.

One outstanding question, currently unaddressed by the extant neuroimaging literature, is how task-dependent neural changes are transformed during competition from conflicting representations of the same object. During the dynamic activation of task-relevant object information, stimulus features that are salient yet task-irrelevant might compete for selection. This ensuing conflict could potentially attenuate task-relevant neural responses and hinder behavioral performance. How is this competition resolved?

The example described in the first paragraph provides a particularly striking demonstration of one such competitive scenario. Behavioral research has demonstrated that living/nonliving judgments of real-world stimuli (e.g., "fire"; "daisy") reflect persistent interference from information about an object's degree of perceived animacy (Babai et al., 2010; Goldberg & Thompson-Schill, 2009; Zaitchik et al., 2008a; 2008b). That is, judgments are delayed and less accurate for "atypical" living and nonliving things—which are objects whose living status and ostensible animacy are incongruent—relative to "typical" items, for which these dimensions align. Among living things (LTs), performance is worse for apparently inert entities like plants, relative to active entities, like animals. For nonliving things (NLTs), performance suffers for things that appear to self-generate movement, like vehicles and celestial bodies, relative to their stationary counterparts.

During living/nonliving judgments of atypical LTs, which appear stationary, and atypical NLTs, which appear active, how is semantic conflict resolved? Findings from behavioral research across the lifespan indicate that this pattern of impairments is strongest when cognitive control abilities are limited or compromised. In young children, executive functioning skills uniquely predict the accuracy of LT vs. NLT judgments after controlling for age and vocabulary (Zaitchik et al., 2014). Moreover, in healthy adults, when cognitive control processes are given insufficient time to operate (i.e., during speeded judgments), university students' and biology professors' responses exhibit the motion-focused bias (Goldberg & Thompson-Schill, 2009). The bias also co-occurs with declines in executive functioning in both patients with Alzheimer's Disease and in healthy elderly adults (Zaitchik et al., 2008a; 2008b).

This convergent evidence suggests that cognitive control processes are involved in recruiting task-relevant representations of atypical living and nonliving things amidst competition from prepotent information which would yield an incorrect judgment. These goal-directed biases toward task-relevant information are thought to occur via top-down modulatory signals from the prefrontal cortex (e.g., Frith, 2000; Mechelli et al., 2004; Miller & Cohen, 2001; Noppeney et al., 2006). In particular, the ventrolateral regions of left inferior frontal cortex (LIFC) are proposed to serve as a domain-general, dynamic filtering mechanism that biases neural responses toward task-relevant information while

gating task-irrelevant information (Shimamura, 2000; Kan & Thompson-Schill, 2005; Chrysikou et al., 2014).

Although no previous neuroimaging studies have directly tested for typicality effects during living/nonliving judgments, two recent studies have observed increases in univariate LIFC response amplitude for atypical object stimuli during judgments of basic-level category membership (e.g., during judgments atypical versus typical fruits, vehicles, mammals, and clothing; Liu et al., 2013; Santi et al., 2016). For example, LIFC response increases during judgments of whether an olive is a fruit, versus judgments of whether an apple is a fruit. If the LIFC is critically involved in selecting task-relevant information amidst competition from task-irrelevant information, then responses in this region should predict the degree to which category-veridical information is recruited during living/nonliving judgments of atypical living and nonliving things.

1.1 The Present Study

This experiment investigates the multivariate activity patterns elicited by the same set of object stimuli under two distinct task conditions, where the tasks differed in the extent to which they required retrieval of category-related information, and the word stimuli varied in the extent to which their conceptual referents possess category-typical object features. The present study provides two key contributions to the MVPA literature on task-dependent neural representations during object processing. One is the examination of neural response patterns during retrieval of object information that is accessed through lexical stimuli (i.e., object names) as opposed to the neural patterns elicited during object perception, via pictures or drawings of objects. The use of word stimuli instead of visual stimuli mitigates confounds that exist between visual form and category identity (Rice et al., 2014; Coggan et al., 2016). Furthermore, there is a dearth of fMRI studies that examine the experience dependence of neural responses using non-pictorial object stimuli (but cf. Malone et al., 2016; Peelen et al., 2014). Using lexical stimuli allows us to examine whether the neural effects that have been consistently observed under conditions of object viewing generalize to other routes of accessing object information. Secondly, unlike previous fMRI investigations of task-dependent MVPA effects, the present study tests specific, item-level predictions about how neural

responses should change across tasks, and which items should experience greater neural changes than others. Further, we test for links between these item-level changes in neural representations and trial-level modulations in LIFC response amplitude. Our full set of predictions are listed below.

1.2 Hypotheses

1.2.1 Goal-Directed Semantic Retrieval During Living/Nonliving Judgments

We compared behavioral and neural responses while subjects performed living/nonliving judgments on stimulus items that varied in their degree of category typicality. Category typicality characterizes the degree to which an item is semantically related to (i.e., shares features with) other members of its own category, versus members of opposing categories (Rosch & Mervis, 1975), and the degree to which an object's feature co-occurrences match those of other category members (McRae et al., 1999; Plaut, 1996). Atypical items share features with both category members and non-members, and hence their category membership is relatively more ambiguous. We predicted that judgments of atypical living and nonliving things would generate semantic conflict, due to the co-activation of information that would lead to two mutually exclusive judgments. This conflict should manifest in slower and less accurate behavioral judgments of atypical items, relative to typical category members. Additionally, judgments of atypical items should elicit increased BOLD activity in brain areas associated with cognitive control (i.e., left inferior frontal cortex), which are thought to exert top-down signals that bias activation toward task-relevant information (Miller & Cohen, 2001).

1.2.2. Cross-Context Multivariate Pattern Changes

In addition to the predictions posed above, we also tested for changes in multivariate responses patterns while subjects thought about the same set of experimental stimuli under two distinct task conditions. We first measured neural responses while subjects performed an undirected semantic encoding task, and then later while subjects made an explicit judgment about each item's domain membership in the living/nonliving

judgment task. We then measured changes in each item's neural responses from one task to the next.

We predicted that subjects would retrieve information about each item's category identity during living/nonliving judgments. In regions of ventral temporal and parietal (VTP) cortex that are sensitive to lexico-semantic information, retrieval of this information should manifest in more similar neural responses among stimuli from the same category, and more distinct neural responses between items from opposing categories (i.e., increased category selectivity). In contrast, the neural patterns should exhibit relatively weaker category-level distinctions during the undirected semantic encoding task, when the task demands do not require explicit retrieval of this information.

1.2.3. Category Typicality and Pattern Change

We predicted that item-level measures of category typicality would modulate cross-task changes to the neural response patterns. In particular, category typicality should predict the degree to which an item neurally resembles members of its own category, versus members of opposing categories. We predicted that typical items would exhibit category-related information during both tasks, and consequently, their neural response patterns should be relatively similar across tasks. In contrast, for atypical items, category-related information should be weakly activated during the undirected task and more strongly activated during the living/nonliving judgment task, leading to greater cross-task pattern change.

1.2.4. LIFC Response and Pattern Change

We also predicted that degree of cross-task pattern change would scale positively with increases in item-level LIFC response. If this region is involved in recruiting task-relevant information, then activity here should predict the extent to which an item's neural pattern changes from one task to the next. In particular, LIFC response should be positively correlated with increases the expression of category-selective neural patterns during the living/nonliving judgment task, relative to the semantic encoding task.

According to this proposal, LIFC response re-weights the information represented in the multivariate patterns, such that the activation of task-relevant features is strengthened.

2. Methods

2.1. Subjects

Participants in this study included twenty-three right-handed, native English speakers, all aged 18-28 years old (11 males). Subjects had no history of neurological disorders and were not currently taking any psychoactive medications. All subjects had normal or corrected to normal vision. Subjects were recruited from the University of Pennsylvania community and were compensated $20/hr for their participation and up to $14 in bonuses based on their behavioral task performance. All subjects provided informed consent as approved by the University of Pennsylvania Institutional Review Board. Seven subjects were removed from the analysis due poor task performance (n=2); failure to stay awake during testing (n=1); and excessive head motion (n=4), yielding a final sample size of sixteen participants (7 males).

2.2. Stimuli

The stimuli were comprised of 120 names of real-world animals and objects, including items from three basic-level taxonomic categories: 40 living things (hereafter LTs), 40 man-made artifacts (hereafter ARTs), and 40 nonliving natural kinds (hereafter NATs). A set of 20 un-pronounceable non-words were also included. These non-words were created by shuffling the letter ordering of randomly selected items from each category, including 7 LTs, 7 NATs, and 6 ARTs. The non-words were included so that we could identify brain voxels in which BOLD response is modulated by presentations of real words versus non-words (see Section 3.6.2).

2.2.1 Stimulus norming

The 40 selected stimulus items from each semantic category (i.e., LTs, ARTs, and NATs) were selected from an initial pool of 400 items. The items in each domain (i.e., LT and NLT) were then randomly sorted into subsets of 100-130 items. These item subsets were then included in two separate surveys: a “Typicality” survey, and an

"Activity" survey. Fifty unique Amazon Mechanical Turk workers participated in each survey.

2.2.2. Typicality Ratings

Each survey was comprised of pseudo-randomly selected subsets of items, such that each subset was comprised of all living things (LTs) or all nonliving things (NLTs, i.e., ARTs and NATs). Participants in the Typicality Survey were instructed to think of all the possible objects that belong to the items in their assigned category (e.g., all the LTs in the world), and the characteristics (e.g., the appearance and behaviors) that are most common among the category members. Participants were instructed to rate each individual item by the extent to which it shared features with other members of its own category. Each stimulus word appeared on the screen one at a time, along with the prompt: "How typical is this item of the category X?" where "X" was either "LIVING THING" OR "NONLIVING THING." Items were rated on a continuous scale from 0 to 100 (Figure 1).

2.2.3. Activity Ratings

Participants in the Activity Survey were instructed to rate the extent to which each item exhibits activity. The survey instructions explained that participants should consider each item's frequency of activity; the perceptual strength of the activity (e.g., can it be seen, heard, smelled); and the extent to which each item requires energy (e.g., food, fuel, electricity) to function. Item names appeared on the screen one at a time along with the prompt: "To what extent does this thing exhibit activity?" Participants selected their response on a linear sliding scale, with 0 labeled as "Completely Inactive" and 100 labeled as "Very Active". Each survey participant exclusively rated items of the living or the nonliving domain, and there was no overlap in the participants to took the Typicality Survey and those who took the Activity Survey.

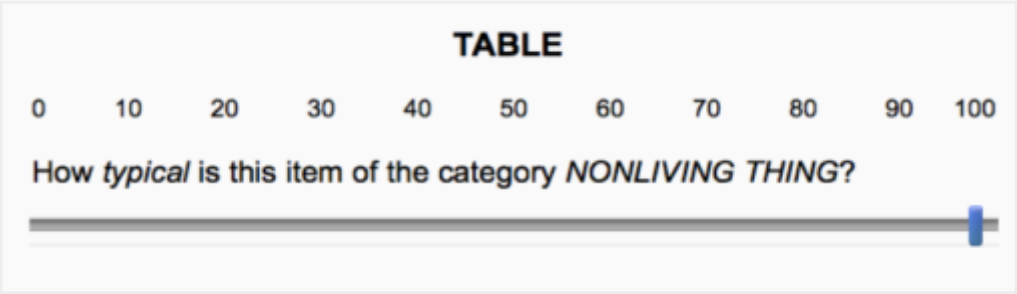

Figure 3.1

An example prompt from the Typicality Ratings survey. The typicality of each item as rated on a continuous, sliding scale from 0 to 100. The items included in each survey were exclusively living things or nonliving things.

2.3. Stimulus Selection

For each of the three basic-level taxonomic categories, we selected forty items such that they would meet two criteria. First, we selected items to create a wide, continuous range of typicality ratings across the items included in each category. Additionally, we wanted to ensure that typicality ratings were not correlated with other psycholinguistic variables (e.g., word length; word frequency; contextual diversity). The selected stimulus items meet both criteria (see Appendix A for item ratings, and Appendix B for correlations with psycholinguistic variables).

To test whether typicality ratings varied by basic-level taxonomic category, we submitted the ratings to a three-way analysis of variance (ANOVA). There was a main effect of category on typicality ratings, $F(1,117)= 51.03$, $p<.001$. Planned comparisons indicated that the selected NATs received lower typicality ratings ($M= 44.7$, $SD= 13.3$) than LTs ($M= 58.6$, $SD= 20.6$), and both LTs and NATs received lower ratings than ARTs ($M= 78.0$, $SD= 7.5$). Given that the distributions of typicality ratings greatly vary across the three categories, we focused our analyses of typicality effects at the within-category level, rather than collapsing across category distinctions. We adopted this approach for two primary reasons. First, it avoids the confound between category membership and differences in the distributions of the typicality ratings. Second, it does not assume that typicality operates the same way for all taxonomic categories, because the defining characteristics of typicality are category-dependent.

For binary, condition-level comparisons between typical and atypical category members, the items in each category were sorted by typicality score, and the 20 items with the highest typicality scores were labeled as "typ" (i.e., typical) items, while the items with the 20 lowest typicality scores were labeled as "atyp" (i.e., atypical) items. In the following analyses, we will investigate how behavioral performance and changes in neural activity are modulated by binary, condition-level differences in category identity and category typicality, as well as by item-level, continuous variation in typicality scores within each category.

2.4. Relationship between Typicality Scores and Activity Scores

Within each category, we z-scored the raw typicality and activity ratings, such that each score quantifies the number of standard deviations by which an item's rating was above or below the mean score of its respective category. We then measured the correlation between typicality and activity scores for each category. These variables were strongly positively correlated for LTs, such that higher activity scores were associated with greater typicality, r= 0.91; and moderately negatively related for ARTs (r= -.79) and NATs, (r= -.72), such that higher category typicality scores were associated with lower activity scores. These relationships indicate that degree of activity is strongly related to category membership for each of the three categories, which is consistent with previous behavioral observations of links between category identity and motion-related information (Zaitchik et al., 2014). However, the relationship between category typicality and strength of motion information was strongest among LTs (R^2 = .83), while the activity scores predicted relatively smaller proportion of the variance in typicality scores for other two categories (R^2 = .51 for NATs and R^2 = .62 for ARTs).

2.5. Design Overview

The fMRI experiment was divided into two parts: Part A was composed of runs 1-10, and Part B constituted runs 11-14. Subjects read task instructions and completed practice trials for Part A and Part B immediately before runs 1 and 11, respectively. All stimulus items were presented twice in Part A and once in Part B. Each experiment part involved a distinct experimental task and instructions that encouraged subjects to process

the stimulus items in a particular manner (Figure 2). The Part A task, semantic encoding, was designed to promote elaborative thoughts about each stimulus item in an unbiased, undirected manner. In contrast, the Part B task, living/nonliving judgments, required subjects to explicitly retrieve the item's living/nonliving status. Although subjects were aware that the task and instructions would change at some point in the experiment, they did not know the exact task they would be performing for Part B until right before run 11 began.

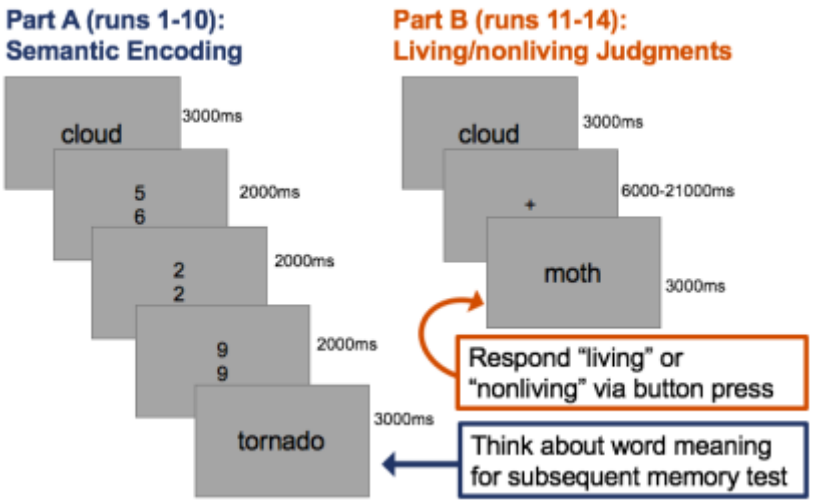

Figure 3.2

Schematic of experiment design. BOLD response was measured while subjects processed each stimulus word during two distinct tasks. After each scanning run in Part A, subjects performed a self-paced, yes/no recognition memory task. Each stimulus word appeared twice in Part A and once in Part B.

3. Procedure

3.1. Part A: General Semantic Encoding

In Part A, each stimulus item appeared twice, in two separate and randomly assigned scanning runs. Each scanning run included 28 items: eight items each from the NAT, ART, and LT categories, and 4 scrambled words. Subjects performed a semantic encoding task, in which they were instructed to think about the meaning of each individual item during its word presentation, and to remember this item in preparation for a recognition memory task that would immediately follow each scanning run (cf. Musz &

Thompson-Schill, 2015). Each trial consisted of a single stimulus item centrally presented on the screen for 2500ms. Subjects were told to ignore any trials that featured non-words, as these items would not be included in the subsequent recognition memory tests. At the end of each scanning run, after the scanner turned off, and before the next scanning run began, subjects performed a self-paced, yes/no recognition memory test via button press. Each memory test included six "hit" items that appeared in the immediately preceding scanning run, and six "lure" items which did not appear at any other point in the experiment. Each set of hit and lure items consisted of 2 LTs, 2 ARTs, and 2 NATs. Data from any fMRI subject who scored below 50% on any of the ten memory tests was removed and replaced in subsequent analysis (n=1).

To obtain single-trial estimations of BOLD response for each individual word presentation, it was necessary to space out the trials over time, because increasing the inter-trial interval (ITI) between two stimulus presentations minimizes the overlap between their hemodynamic response functions. However, we did not want subjects to use the time during the ITIs to rehearse the stimulus items, as this would reduce our ability to measure the contrast between BOLD responses during stimulus presentations versus during baseline measures. In an earlier, preliminary pilot experiment, we separated the word presentations with either (1) a fixation ITI, in which subjects are instructed to clear their mind and patiently wait for the next trial, or (2) a number parity task (described below). Preliminary data and debriefing with pilot fMRI subjects indicated that the number parity ITI increased subjects' alertness and level of engagement during the scanning session, and did not impair their ability to engage in elaborative encoding of the word meanings for subsequent recognition memory performance.

During the number parity task, two digits between 0-9 appeared on the screen for 2000ms, one above the other. Subjects were instructed to add the two digits together, and then to response via button press according to whether the sum of the two digits yielded an even or an odd number (Hulbert & Norman, 2015). During the scanning runs, the number parity trials were interleaved with the semantic encoding trials, such that subjects performed one trial of the semantic encoding task, followed by three trials of the number parity task (Figure 2). Each scanning run was approximately four minutes long. To further incentivize subjects to perform with high accuracy on both tasks— but to

emphasize that subjects should prioritize good performance on the semantic encoding task— we awarded subjects an extra $1 for each run in Part A in which their accuracy was above 95% on the recognition memory task, and above 50% on the number parity task. Average accuracy on the recognition memory task was 92% (*SD*= 5%) and average accuracy on the number parity task was 89% (*SD*= 6%).

3.2. Part B: Living/Nonliving Judgments

Right before the 11th scanning run, subjects were told that they would see the same stimulus items, but now their task is to judge whether each item referred to a living thing (LT) or nonliving thing (NLT). The written instructions reminded subjects that living things are biological organisms that can grow, reproduce and die, and that living things require a food source to survive, while nonliving things do not have these characteristics. The instructions provided some example items from each category, and then subjects performed four practice trials of the living/nonliving judgment task. The example items and practice trials were comprised of items that received high typicality ratings (e.g., "bear"; "book") and low typicality ratings (e.g., "petunia"; "rain") from the initial large pool of stimulus items but were ultimately not selected for the final stimulus set.

During runs 11-14, each stimulus item re-appeared once, randomly assigned to one of the four final runs. Each run consisted of 30 stimulus items, including ten randomly selected items from each category. No non-word stimuli were included. Each item appeared in the center of the screen for 3000ms. Subjects were instructed to make their LT vs. NLT judgement at any point while the stimulus item appeared on the screen. The word presentation remained on the screen until the full 3000ms elapsed. Subjects were discouraged from rushing their response, as they would have the entire 3000ms duration to make their judgment via button press (Figure 1). Subjects did not receive feedback on their task performance.

Stimulus presentations were separated by a fixation ITI, during which a centrally located fixation cross was presented for 6,000 to 21,000ms. During this time, subjects were instructed to clear their mind and wait for the next trial to appear. We chose to employ a fixation ITI in Part B instead of the number parity task ITI from Part A because

of the differences in response demand characteristics between the semantic encoding task and the living/nonliving judgment task. The latter task requires subjects to make an overt and explicit judgment of the stimulus items via button press during the stimulus presentation, while the former task does not. We were concerned that requiring subjects to alternate between living/nonliving judgments and number parity odd/even judgments would potentially impair performance on the main task of interest (i.e., the living/nonliving judgments). Stimulus sequences and timing schedules were developed using optseq2 (http://surfer.nmr.mgh.harvard.edu/optseq). Stimulus timings and visual presentations were controlled by E-Prime 2.0 software (Psychology Software Tools).

Each scanning run in Part B lasted approximately five minutes, and subjects were awarded an extra $1 for each scanning run in which their average accuracy at the living/nonliving judgment task exceeded 95%. Trials in which subjects responded incorrectly or failed to respond during the item presentation were modeled as covariates of no interest in subsequent fMRI analyses. Subjects correctly responded to an average of 97% trials per run (*SD*= 4%). Data from subject who did not perform above chance on the living/nonliving judgment task during each scanning run was removed and replaced in subsequent analysis (n=1).

3.3. fMRI data acquisition

Functional and anatomical data were collected with a 64-channel array head coil on a 3T Siemens Prisma system. The structural data included axial T1-weighted localizer images with 160 slices and 1 mm isotropic voxels (TR = 1850 ms, TE = 3.91 ms, TI = 1100 ms). For each run, we collected 81 axial slices (2mm isotropic voxels) of echoplanar fMRI data (TR = 2000 ms, TE = 30 ms). Twelve seconds preceded data acquisition in each functional run to approach steady-state magnetization.

3.4. fMRI Pre-processing and Statistical Analyses

Image preprocessing and statistical analyses were performed using the AFNI software package (Cox, 1996). The time series data were initially preprocessed to remove the influence of various sources of noise, and to yield better estimates of BOLD signal. First, images were sinc interpolated to correct for differences in slice acquisition time due

to the interleaved slice order within each 2000ms TR. Then, each individual volume was spatially registered to the first volume of the first scanning run, because this volume was acquired closest in time to the high-resolution anatomical scan. Next, the data were de-spiked, such that any large values not attributive to the physiological processes were removed from the data.

Additional pre-processing was applied to the data depending on the dimensionality and spatial scale of the signal that was targeted by each distinct analysis. For the univariate analyses, the subject-level data were normalized to a common template and smoothed with an 8mm FWHM Gaussian kernel prior to statistical analyses, and the signal was scaled to percentage signal change. For the multivariate analyses, the data were smoothed with a 4mm FWHM Gaussian kernel and z-normalized within each run, and the data remained in each subject's native brain space during the subsequent statistical analyses.

For both analyses, a modified general linear model (GLM) was fit to each subject's preprocessed data. Each stimulus item presentation was modeled with a canonical hemodynamic response function convolved with a boxcar function that matched the duration of the trial. Data from Part A and Part B were analyzed separately, because they involved different tasks during the un-modeled baseline ITI periods (i.e., the number parity task and fixation cross presentations, respectively). For both types of analyses, scanning run and six motion parameters were modeled as covariates of no interest, along with error and omission trials in the Part B living/nonliving judgment task.

The GLMs in the univariate analyses targeted differences in average BOLD response magnitude across different categories and levels of typicality, while the GLMs in the multivariate analyses estimated BOLD response in spatially distributed activity patterns evoked during the individual presentations of each stimulus item. For the univariate analysis, the condition-level GLMs yielded a unique beta estimate for each condition of interest at each individual voxel in a subject's brain map. For the multivariate analyses, small subsets of spatially distributed voxels were first selected from each subject's brain map, and then the pattern of activity across these voxels were submitted to further statistical tests. The voxel selection criteria for the multivariate analyses are described in Section 3.6 below.

3.5. Univariate Whole-Brain Analyses

3.5.1. Condition-Level Effects of Category and Typicality

Each stimulus presentation was modeled according to its category membership and typicality status, yielding six covariates of interest: LT_typ, LT_atyp, NAT_typ, NAT_atyp, ART_typ, and ART_atyp. In a group-level analysis, a three-way repeated measures ANOVA was performed at every voxel (category = fixed factor with three levels: LT, ART, or NAT; typicality status = fixed factor with two levels: typical or atypical; and subject = random factor with 16 levels). We tested for main effects of category; typicality; and interactions between these factors. Planned a priori statistical comparisons were performed to test the effect of typicality within each of the three categories using paired *t*-tests. This analysis was performed separately on Part A and Part B data.

3.5.2. Item-Level Effects of Typicality

We examined the parametric effect of typicality score on BOLD response in each category. Each stimulus presentation was modeled according to its category membership, along with a continuous value that was specific to each item presentation. This parametric regressor modeled each item's typicality score, relative to its other category members. We then performed group-level, single-sample *t*-tests versus zero to test for voxels that exhibited a linear relationship between the item-level continuous scores and the trial-level fluctuations in BOLD response. This analysis was performed separately on Part A and Part B data.

3.5.3. Region of Interest Analysis: Left inferior frontal cortex

An anatomical region of interest (ROI) mask of ventrolateral regions of left inferior frontal cortex (LIFC) was created using a probabilistic anatomical atlas included in the AFNI software package (Eickhoff et al., 2005). This mask included pars opercularis (BA 44), pars triangularis (BA 45), and the anterior half the inferior frontal sulcus (cf. Musz et al., 2017; Hindy et al., 2012). For the group-level univariate analyses,

the voxel-wise beta coefficients for each condition of interest were averaged across this entire LIFC ROI mask (Figure 3a).

For the item-level multivariate analyses, the LIFC mask was translated to each subject's native brain space, and the beta weights for each stimulus presentation were averaged across the entire ROI mask. For each item, we subtracted its average response during the semantic encoding task from the average response during the living/nonliving judgments. This measurement indexes the item-level change in average LIFC response between the two tasks. Across items within a category, we then z-scored these values, such that each value reflects an item's average change in LIFC response from Part A to Part B, relative to all other items in its category. These values were computed at the individual-subject level, and they were used to predict degree of pattern change in the multivariate analyses (see Section 3.6.3. and 3.6.5 below).

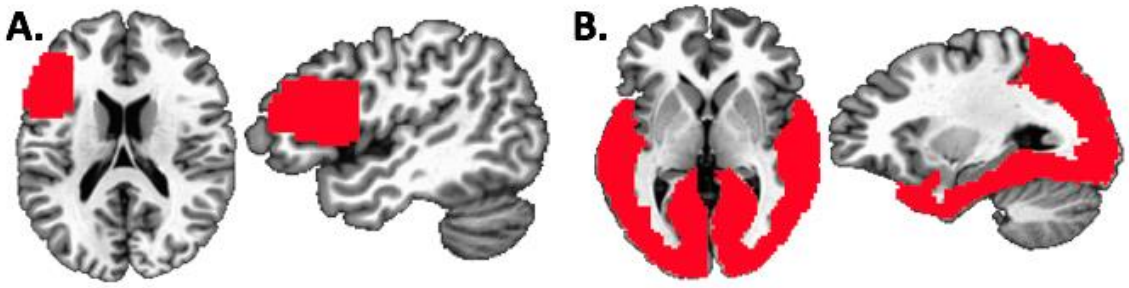

Figure 3.3

Anatomical region of interest (ROI) masks. The ROI mask in left inferior frontal cortex (3a) and the ROI mask covering bilateral gyri in the temporal, parietal, and occipital lobes (3b).

3.6. Multivariate Analyses

Small subsets of spatially distributed voxels were selected from each subject's brain map. The voxel selection criteria are described below. After selecting subsets of voxels, we then extracted the beta estimates for each item presentation in each selected voxel. The set of beta estimates for a given stimulus presentation constituted a multi-voxel pattern (MVP) evoked by that item. We extracted the MVPs for each item presentation in each experiment part (e.g., two MVPs per item in Part A, and one MVP per item in Part B, excluding error and omission trials). We then performed Pearson

correlations to compute the neural similarity between the MVPs evoked by each item in each part, and the neural similarities between MVPs evoked by different stimulus items.

3.6.1. Multivariate Feature Selection: Searchlight Analysis

In this analysis, we extracted the MVPs evoked during each item presentation by sampling small subsets of spatially contiguous voxels. We passed a spherical searchlight with a 4-voxel (8mm) radius over each voxel in each subject's brain map in native space (Kriegeskorte et al., 2006). In each searchlight volume, we extracted the MVPs evoked by each item presentation, and measured the neural similarities between the MVPs in order to test specific hypotheses (see Section 3.6.3 below). The statistical values yielded by these comparisons were then assigned to the center voxel of each searchlight volume. Each subject's searchlight map was then normalized to a standard template and submitted to group-level, random effects analyses. This exploratory voxel selection approach allows for the examination of regionally specific effects that reliably occur in the same spatial location across the subject sample.

3.6.2. Multivariate Feature Selection: ROI Analysis

In this analysis, we extracted MVPs for each item presentation from subsets of voxels that were both anatomically and functionally constrained, such that we could identify the brain voxels that are most likely to be sensitive to the effects of interest. Here, we aim to functionally localize voxels that encode semantic and lexical information, and to anatomically localize brain regions that consistently show such effects across a range of diverse tasks and subject populations in previous fMRI investigations. Previous neuroimaging studies indicate that large swaths of fusiform gyri, angular gyri, and the temporal lobes are sensitive to semantic content, including distinctions between taxonomic categories (Binder et al., 2009; Binder & Desai, 2011; Fairhall & Caramazza, 2013; Martin, 2007) and object identity (Clarke & Tyler, 2014). Thus, in this analysis, we only sampled voxels from inferior parietal, lateral temporal, and ventral temporal cortex (Figure 3b). We created this ventral-temporal-parietal (hereafter "VTP") anatomical ROI mask by combining bilateral temporal, parietal, and occipital regions labeled in the MNIA probabilistic anatomical atlases in the AFNI

software package (Eickhoff et al., 2005). This anatomical ROI mask was then transformed and applied to the native brain space of each individual subject.

To functionally select voxels within each subject's anatomical VTP ROI mask, we computed three separate statistical contrasts at each masked voxel. We then ranked the masked voxels according their statistical values (i.e., their *t*-statistics for a given functional contrast). For each contrast, the VTP ROI voxels with the highest X statistical values were included in the ultimate ROI mask for that subject. To examine whether our effects of interest were robust across a range of ROI mask sizes, the value of X ranged from the top 500 to 5,000 voxels in increments of 500, yielding 10 unique masks per contrast for each individual subject. Each functional contrast involved comparisons between stimulus conditions in the Part A data (i.e., runs 1-10) that were orthogonal to the main comparisons of interest (i.e., category typicality and category identity). One functional contrast quantified the extent to which the two repeated presentations of the stimulus items in Part A elicited similar a voxel-wise BOLD timecourse, averaged across all items (hereafter "stable" VTP voxels; cf. Mitchell et al., 2008). A second functional contrast targeted voxels where responses increased during the semantic encoding task versus the number parity task (hereafter "W>#" VTP voxels). The third functional contrast targeted voxels that responded more to presentations of the critical stimulus words versus the scrambled non-words during the semantic encoding task (hereafter "W>NW" VTP voxels). The following analyses were performed by extracting MVPs from the beta coefficients for each item in each of these ROI masks, or from the set of voxels included in each searchlight volume as described in Section 3.6.1. above. We report the statistical values for mask sizes from the middle of this range (2,500-voxel masks), although the graphical figures will indicate the reliability of each effect across the whole span of ROI mask sizes.

3.6.3. Within-Item, Neural Similarity Analysis: Predicting Cross-Context Pattern Changes

To quantify the degree to which an item pattern changed from the Part A task to the Part B task, we measured the difference between an item's within-task neural similarity and its between-task neural similarity. We extracted each item's three MVPs:

the two elicited during the semantic encoding task in Part A, and the one during the item's living/nonliving judgment in Part B. We computed the pairwise similarities between each of these three patterns, and averaged across the two between-task correlations to obtain a single estimate of between-task neural similarity (Figure 4). We quantified cross-context pattern change as an item's within-task neural similarity minus its average between-task similarity. This metric quantifies the extent to which an item's evoked neural pattern has changed during the Part B living/nonliving judgment, relative to the Part A semantic encoding task. In an alternative version of this analysis, we first averaged each item's two Part A patterns together, and then computed between-task similarity between this average Part A MVP and the Part B MVP. The group-level results reported in Section 4.4 were unchanged when cross-context pattern change was computed by first averaging the two Part A patterns together, or averaging together the two separate between-task neural similarity values (Figure 4).

We predicted that degree of pattern change would negatively scale with typicality scores, such that typical category members would exhibit greater neural similarity (i.e., less pattern change) across the two tasks. In contrast, the MVPs of atypical category members would have to undergo greater changes across tasks in order to explicitly think of these items as category members during the Part B living/nonliving task. Additionally, we predicted that LIFC activity would positively predict cross-context pattern change. If LIFC response is associated with increases in the selection and expression of task-relevant information, then activity here should predict the degree of cross-context pattern change.

In separate analyses, we tested whether each of these two item-level variables (i.e., typicality scores and LIFC response) exhibited a linear relationship to cross-context pattern change. For each ROI or searchlight volume, we computed the correlation between pattern change values and each of these variables. These correlations were performed separately for each variable and for each taxonomic category (i.e., LT, ART, or NAT). We then employed single-sample *t*-tests to determine whether these cross-item correlations were reliably different from zero across the group of subjects.

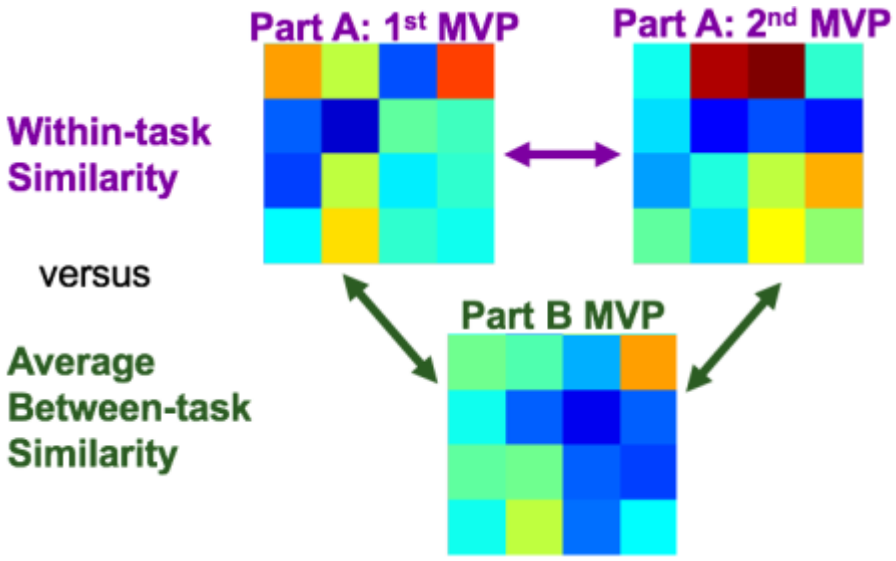

Figure 3.4

Diagram depicting the measure of cross-context pattern change that was computed for each item. Each colored grid represents a multi-voxel pattern (MVP) of beta estimates for the presentation of a stimulus item during three separate times: twice during the semantic encoding task in Part A (the MVPs labeled in purple) and once during the living/nonliving judgment task in Part B (the MVP labeled in green). For each item, we computed the pairwise similarity between each MVP pair using Pearson correlations. To obtain a measure of cross-context pattern change for each item, we subtracted its average between-task pairwise similarity from its within-task similarity.

3.6.4. Between-Category, Cross-context Neural Similarity Analysis

In this analysis, instead of directly comparing the MVPs from Part A and Part B to one another, we compared the data from each part to a category-level model of semantic similarity (Kriegeskorte et al., 2008). This model poses specific predictions regarding the relative similarity between the neural activity patterns evoked by the various experimental stimuli; namely, that the neural similarities between MVPs from the same semantic category (e.g., LTs versus other LTs) should be relatively high, and the neural similarity observed between MVPs of items from different categories (e.g., LTs versus NATs, and LTs versus ARTs) should be relatively low (Figure 5a). In each ROI or searchlight volume, we measured the strength of the correspondence (i.e., the Pearson correlation) between (1) the predicted category-level similarity model and (2) the observed neural similarities between every pairwise comparison of MVPs (Figure 5b). These correlations were performed separately for Part A and Part B data. For the Part A

data, the two MVPs that corresponded to an item's two stimulus presentations were averaged together prior to computing the neural similarities between every item pairing.

The group-level analyses were performed in each ROI or searchlight volume, and consisted of three random-effects analyses. First, we tested for the reliability of nonzero correlations between the category-level model and the observed neural similarities in Part A, using single-sample *t*-tests. We then repeated this analysis using the Part B data instead. Finally, we tested whether the neural data in Part A and Part B reliably differed in the extent to which they matched the category-level model, via paired *t*-tests. We predicted that the Part B neural data would exhibit the category-level similarity structure, and that the correspondence to this similarity structure would greater in Part B data than the Part A data.

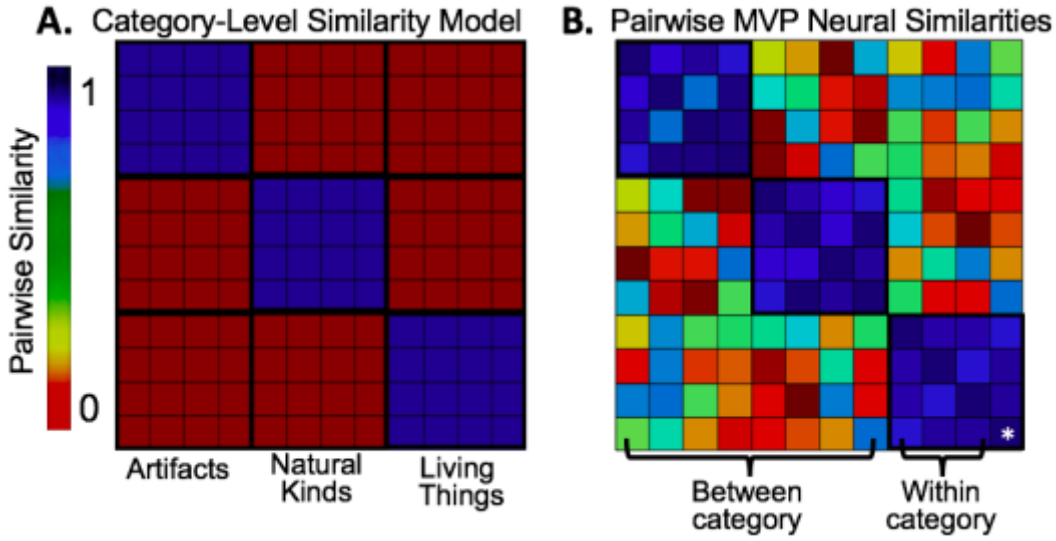

Figure 3.5

Example pairwise similarity matrices, constructed with four stimulus items per category. There were 40 stimuli per category in the actual experiment. The color of each cell indicates the pairwise similarity (the Pearson correlation coefficient) between two stimulus items. Figure 5a depicts the similarity structure predicted by the category-level model, in which within-category neural similarity is greater than between-category similarity. Figure 5b depicts an example neural similarity matrix derived from simulated data. To compute item-level measures of category selectivity, each item's average between-category neural similarity is subtracted from its average within-category neural similarity. For the example stimulus item from the "living things" category that is marked with a white asterisk, Figure 5b indicates the within- and between-category neural similarity values that would be extracted from this item's row of the matrix.

3.6.5. Item-level Measures of Neural Category Selectivity

In addition to testing whether the neural data from each experiment part conformed to the predicted category-level similarity structure, we also computed item-level measures of category selectivity. Here, category selectivity is defined as the extent to which an item is more similar to members of its own category, versus members of the opposing categories (Kriegeskorte et al., 2008; Iordan et al., 2015). For each item MVP, we computed (1) its average pairwise similarity to all other members of its own category (i.e., average within-category similarity) and (2) its average pairwise similarity to all other items from the other two categories (i.e., average between-category similarity). To obtain a measure of each item's category selectivity, we subtracted its average between-category similarity value from its average within-category similarity value (Figure 4b). This measurement was computed twice for each item: once using the data in Part A (i.e., each Part A item MVP to all other Part A MVPs) and once in Part B (i.e., each Part B item MVP to all other Part B MVPs).

After obtaining each item's neural category selectivity in Part A and Part B, we then tested whether category selectivity scaled with typicality scores. During Part A, we predicted that typicality scores would predict the degree of item-level category selectivity, because relatively more typical items should share more features with their own category and less features with the opposing category, manifesting in relatively greater within- versus between-category distinctions in their neural patterns. Additionally, we predicted that typicality scores would negatively scale with *increases* in category selectivity for Part A to Part B. That is, not only will the patterns of atypical

items experience greater changes from Part A to Part B (as predicted in Section 3.6.3), but they will change in a particular way. Namely, their patterns should exhibit relatively greater similarity to other category members, and less similarity to non-members, in Part B versus Part A. In contrast, typical item patterns should exhibit neural category selectivity in both parts, and hence experience smaller changes in the degree of category selectivity across the two experiment parts.

In addition to the predicted relationships between typicality scores and category selectivity, we also tested a key prediction about the role of LIFC response and the recruitment of task-relevant (i.e., category-selective) neural patterns. We predicted that LIFC response should predict *increases* neural category selectivity from Part A to Part B. If LIFC is critically involved in recruiting task-relevant information manifested in neural activity patterns in VTP cortex, then increases in LIFC response should be associated with increases in neural category selectivity. The full set of predictions for relationships between LIFC response, typicality scores, and item-level multivariate patterns are listed in Table 1.

Table 3.1

Predicted Relationships between Item-Level Multivariate Measures and Item-Level Variables.

Item-level Measure	**Experiment Part**	**Item-Level Variable**	**Predicted Relationship**
Cross-context pattern change	n/a	typicality	negative
Cross-context pattern change	n/a	LIFC response	positive
Category Selectivity	Part A	typicality	positive
Category Selectivity	Part B - Part A	typicality	negative
Category Selectivity	Part B - Part A	LIFC response	positive

3.7. Multiple Comparison Corrections

After performing the univariate and multivariate searchlight analyses for each individual subject's functional data, the resulting statistical brain maps were submitted to

group-level, random-effects analyses. Procedures for multiple comparison correction differed for univariate versus multivariate analyses. For the univariate analyses, minimum cluster extent was determined using AFNI's 3dClustSim (version built May 21, 2017). For this correction, we first estimated the smoothness of the residual time series data using AFNI's 3dFWHMx spatial autocorrelation function. Based on a voxel-level uncorrected alpha of 0.001, Monte Carlo simulations (n= 50,000) indicated a minimum cluster extent of 191 voxels for a cluster-corrected alpha of .05. For the multivariate searchlight analyses, we utilized a non-parametric permutation version of 3dClustSim for cluster-size thresholding, as this method does not make any assumptions about the spatial correlation structure of the functional data (cf. Cox et al., 2017; Eklund et al., 2016). This is approach is particularly well-suited for searchlight analyses, because the voxel-level estimates of spatial smoothness from the univariate data are not the only source of smoothness in the statistical maps that are yielded by the multivariate searchlight analysis.

4. Results

4.1. Behavioral results: Living/Nonliving Judgments

Behavioral performance across the different stimulus categories and typicality levels were compared using two-factor (category × typicality) repeated measures ANOVA. Comparisons of task accuracy across the different conditions revealed a main effect of category, $F(1,15) = 4.13$, $p = .03$. Planned paired comparisons indicate that accuracy for ARTs (M= 99%, SD= 2%) was greater accuracy on LT trials (M= 96%, SD= 6%), $t(15) = -2.71$, $p = .02$. Accuracy between ART and NAT trials did not reliably differ across subjects, $t(15) = -1.73$, p= .10 (NAT M= 98%, SD= 3%), nor did accuracy for LT versus NAT trials, $t(15) = -1.42$, $p = .18$. Within each category, we performed follow-up paired *t*-tests comparing accuracy for typical versus atypical category members. Task accuracy did not reliably differ by typicality in any category.

To minimize the influence of outlier values, we compared response times using subjects' median response times for each condition. A two-factor repeated measures ANOVA revealed a main effect of category, $F(2,30) = 4.83$, $p = .02$. Response latencies to NAT trials were delayed (M= 1330ms, SD= 221ms), relative to both ART trials (M=

1232ms, *SD*= 159ms), *t*(15)= -2.80, *p*= .01, and to LT trials (*M*= 1240ms, *SD*= 155ms), *t*(15) = -2.70, *p*= .02. There was also a main effect of typicality, *F*(1,15)= 10.6, *p*=.005, although the interaction between category and typicality was not reliable, *F*(2,30)= 1.54, *p*= .21. We tested the effect of typicality in each category using paired sampled *t*-tests. Responses to LT_atyp trials were slower than LT_typ trials, *t*(15)= -3.36, *p*= .004. However, responses did not vary by typicality for NAT trials, *t*(15)= 0.44, *p*= .7, or ART trials, t(15)= -1.21, *p*= .24 (Figure 6).

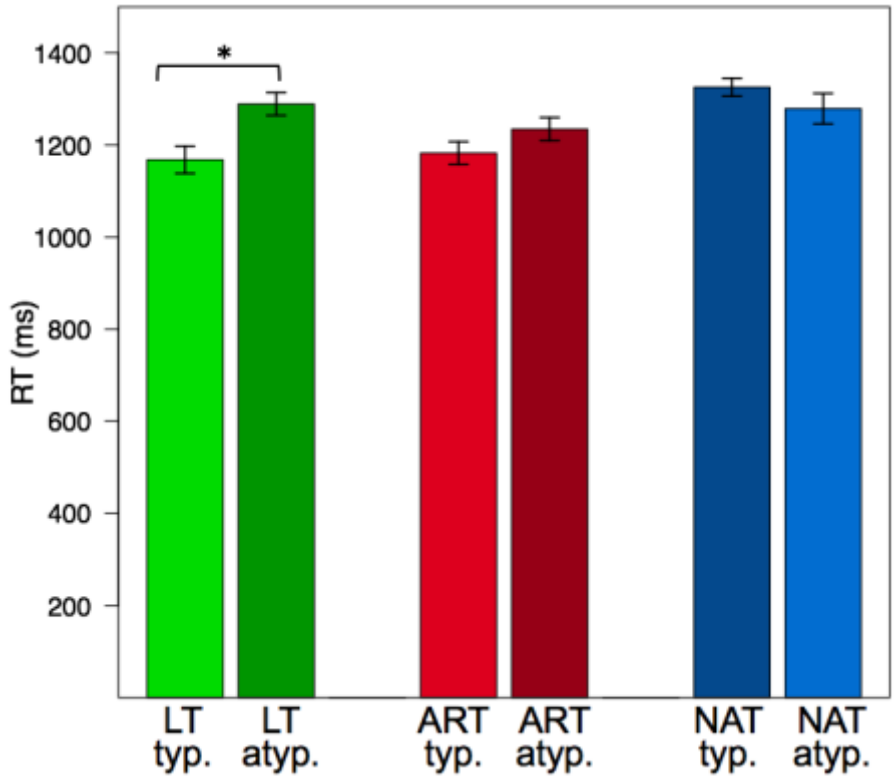

Figure 3.6

Mean response latencies for each category condition during the living/nonliving judgment task in Part B. Error bars indicate within-subject standard error (Cousineu, 2005). LT = "living things"; ART = "Artifacts"; NAT = "Natural Kinds"; typ. = "typical"; atyp. = "atypical." Asterisk indicates *p*<.005.

In addition to testing for binary differences between RTs for typical and atypical trials, we also tested whether we could predict continuous differences in RTs using the typicality and activity scores. In this analysis, we correlated the activity scores and the typicality scores with each subject's trial-level response latencies. Across subjects,

typicality scores predicted faster response latencies for both LTs, $t(15)= -4.6$, $p= .0003$ (mean $r= -.19$, $SD= .17$) and ARTs, $t(15)= -2.28$, $p= .04$ (mean $r= -.10$, $SD= .17$) but not for NATs, $t(15)= -.09$, $p= .93$ (Figure 7). In contrast, activity scores predicted faster responses only for LTs $t(15)= -3.76$, $p= .002$ (mean $r= -.16$, SD= .18). There were no reliable relationships between activity scores and RTs for NATs (mean $r= .03$, $SD= .14$) or for ARTs (mean $r= -.03$, $SD= .15$). In subsequent fMRI analyses, we used the item-level typicality scores to predict neural responses, because these scores predicted the behavioral signatures of semantic conflict (i.e., response latencies) for both LTs and ARTs.

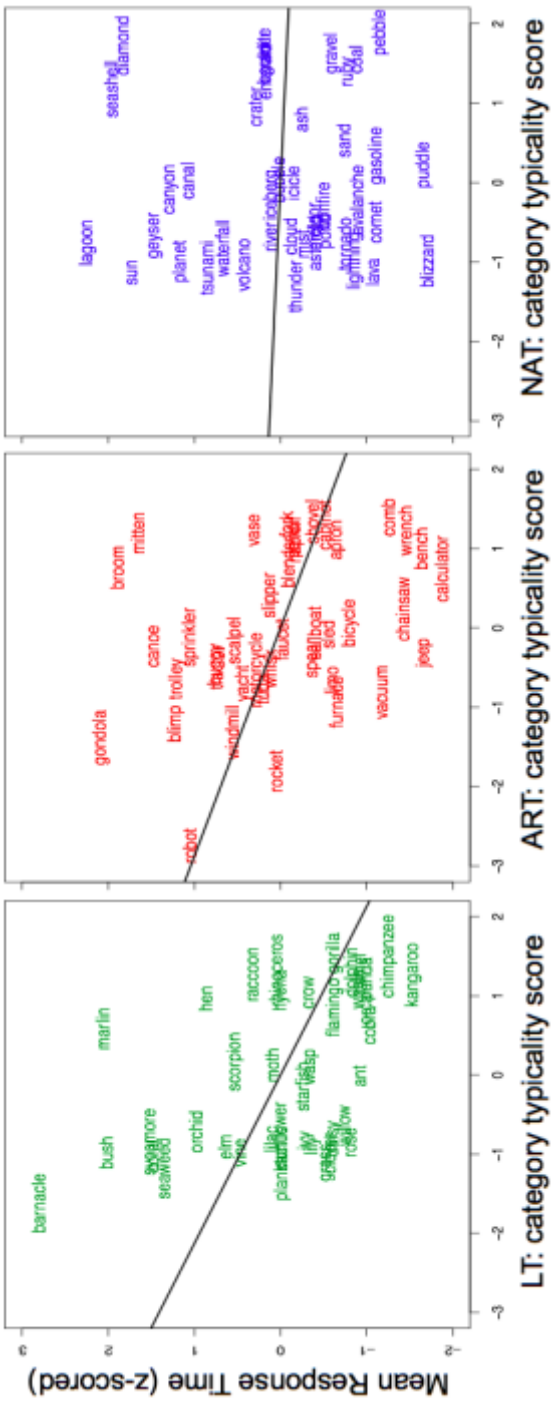

Figure 3.7

Relationship between within-category typicality scores and average response time for each category. The three plots share the same y-axis. Black trend lines indicate the slope of the linear relationship between the two variables. The location of each item on the y-axis (the z-scored RT value) depicts the central tendency across subjects.

4.2. Univariate Results

4.2.1. Part B Category and Typicality Effects: Whole-Brain Results

The statistical brain maps were submitted to a 3×2 (category by typicality) ANOVA with subjects designated as a random factor. A main effect of category emerged in four clusters, including left inferior frontal gyrus, medial frontal gyrus, left inferior parietal lobule, and left inferior temporal gyrus (Figure 8). In the peak voxel of each cluster, follow-up paired comparisons revealed that BOLD responses in these regions were greatest during living/nonliving judgments of NATs, relative to both LTs and ARTs, and that response was greater for LTs than ARTs (Table 2). There were no above-threshold effects of typicality, nor an interaction between category and typicality.

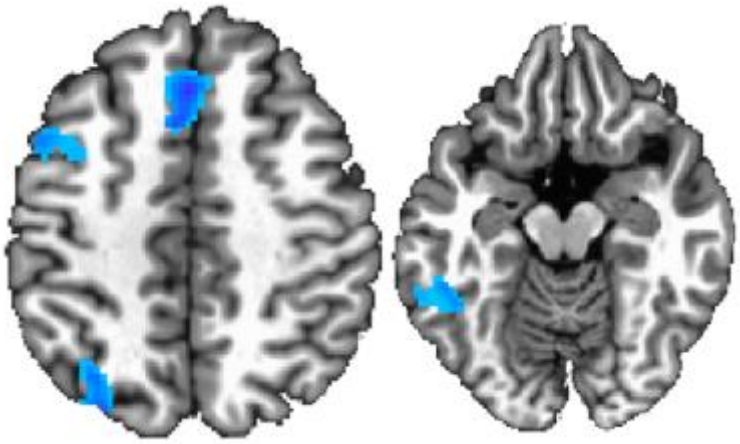

Figure 3.8

BOLD response during living/nonliving judgments varied by category in four clusters. Follow-up comparisons in the peak voxel of each cluster indicate that responses in these regions increase during judgments of natural kinds, relative to artifacts and living things, and during judgments of living things, relative to artifacts. Clusters include medial frontal gyrus, left inferior frontal cortex, and left inferior parietal lobule (shown in the axial brain image on the left), and left inferior temporal gyrus (shown in the axial brain image on the right).

Table 3.2

Peak voxel locations for category-level univariate effects in Part B.

Brain Region	Cluster Extent	x	y	z	Peak F-statistic (Category)	Peak T-statistic (ART > NAT)	Peak T-statistic (LT > NAT)	Peak T-statistic (LT > ART)

Left inferior frontal cortex	950	-39	-29	16	17.84	-5.9	-2.53	3.78
Left inferior parietal lobule	642	-29	-75	38	19.98	-6.08	-2.97	3.41
Left inferior temporal gyrus	346	-55	-45	-16	15.82	-5.64	-3.21	2.2
Left medial frontal gyrus	210	-1	-25	44	14.99	-5.85	-1.4	4.5

For each category, we also tested whether BOLD response amplitude was modulated by typicality scores. In this analysis, the item-level typicality scores were entered as parametric regressor to predict trial-level changes in BOLD response for each category. Two clusters exhibited changes in BOLD response that were linearly related to typicality scores for LTs. In a cluster of 1187 voxels in LIFC (peak voxel coordinates: x= -45, y= 23, z= 20), greater typicality scores predicted decreases in BOLD response during nonliving/living judgments of LTs. A 248-voxel cluster in right supramarginal gyrus (peak voxel coordinates: x= 57, y= -49, z= 30) showed the reverse pattern: here, greater typicality scores predicted increases in BOLD response. Neither NAT nor ART typicality scores showed any relation to trial-level changes in BOLD response at the whole-brain level.

4.2.2. Part B Category and Typicality Effects: ROI Analysis

The values of average percent signal change in each subject's LIFC ROI were submitted to a 3×2 (category by typicality) ANOVA with subjects designated as random factor. We observed a main effect of category, $F(2,30)= 3.61$, $p= .001$. Follow-up paired comparisons between each category indicated that response increased for NATs, relative to ARTs, $t(15)= 4.29$, $p= .001$. Additionally, trending results suggest that BOLD response increased for LTs versus ARTs, $t(15)= 2.07$, $p= .06$, and for NATs versus LTs, $t(15)= 2.07$, $p= .06$ (Figure 9a).

This analysis also revealed a main effect of typicality in LIFC, $F(1,15)= 5.40$, $p= .03$, and an interaction between category and typicality, $F(1,15)= 3.43$, $p= .05$. Planned follow-up comparisons between typical and atypical trials within each category indicate

that judgments of atypical LTs recruited increased LIFC response, relative to typical LTs, *t*(15)= 4.58, *p*= .0003 (Figure 9b). These binary typicality effects did not occur for ARTs, *t*(15)= 1.48, *p*= .16 or NATs, *t*(15)= -.59, *p*= .56.

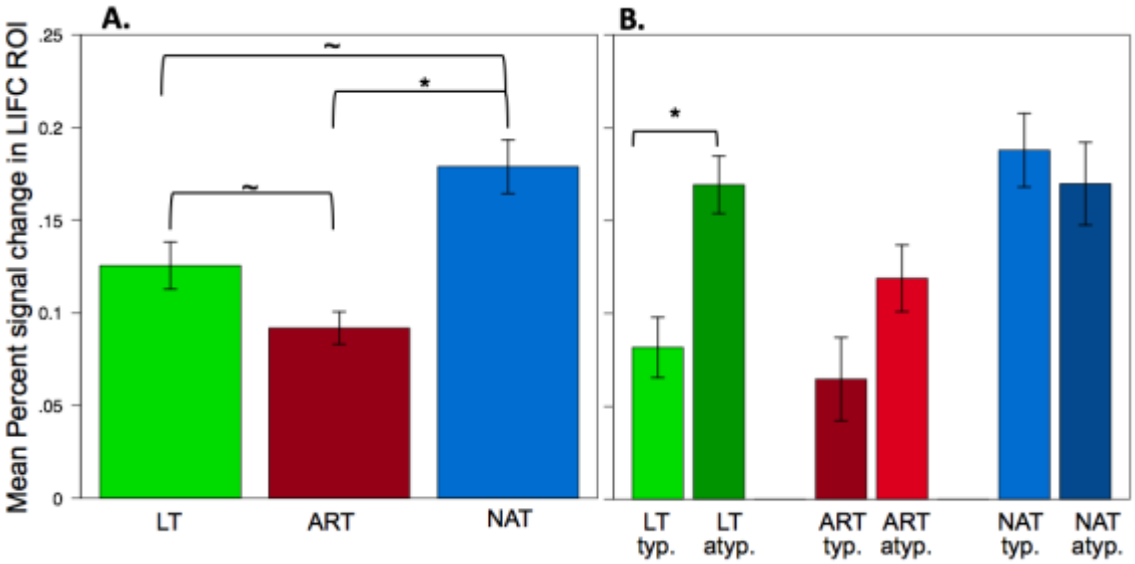

Figure 3.9

Average BOLD response in LIFC ROI masks during Part B living/nonliving judgments. Figure 9a shows differences in response by category and Figure 9b shows how responses within each category vary by typicality status. Asterisks indicate $p < .05$ and tildes indicate $p < .07$. Error bars indicate within-subject standard error.

For each subcategory, we submitted subjects' average LIFC beta coefficient for the typicality parametric regressor to single-sample *t*-tests versus zero. For LTs, this analysis recapitulated the results that were observed in the whole-brain analysis: mean LIFC response negative scaled with typicality scores for LTs, *t*(15)= -8.22, *p*= .0001. This negative relationship between continuous typicality scores and average trial-level LIFC response was also present for ARTs, *t*(15)= -2.20, *p*= .05, but not for NATs, *t*(15)= -1.3, *p*= .21.

4.3. Part A: Category and Typicality Effects

Although the task demands of the semantic encoding task in Part A did not require subjects to explicitly access information about each item's category identity and

its category typicality, we tested whether BOLD response during Part A was nevertheless modulated by these factors. We repeated the 3×2 ANOVA described for Part B above to tests for contrasts between category identity and typicality status during the semantic encoding task in Part A. We failed to find any reliable group-level effects of either factor, or an interaction between them. These null results also persisted in the LIFC ROI analysis.

4.4. Multivariate Results

4.4.1. Searchlight Analysis: Predictors of Cross-context, Within-Item Pattern Change

In each searchlight volume, we tested whether (1) typicality scores or (2) average LIFC response increase would predict the degree to which neural activity patterns changed from Part A to Part B. For each category and in each searchlight volume, we separately computed the correlation between typicality scores and degree of pattern change, and between LIFC response and degree of pattern change. The correlation across category members was assigned to the searchlight center. Each subject's three category searchlight maps were then normalized to standard space to test for reliable effects in the group-level analyses in single-sample *t*-tests versus zero.

For ARTs, there was a negative relationship between typicality and pattern change in two clusters of searchlight centers, one centered in left angular gyrus (105 searchlight centers, peak searchlight center x= -48, y= -56, z= 28) and left inferior temporal gyrus (81 searchlight centers, peak searchlight center x= -50, y= -50, z= -19) (Figure 10a). In these voxels, typicality scores negatively predicted degree of pattern change, such the greater the typicality score, the less the item MVPs changed from the baseline semantic encoding task in Part A to the living/nonliving judgment in Part B. We failed to detect any searchlight clusters which showed above-threshold relationships between typicality scores and degree of pattern change for LTs or NATs.

For NATs, item-level increases in mean LIFC response positively predicted MVP pattern change in left fusiform gyrus (79 searchlight centers, peak searchlight x= -39, y= -35, z= -14) (Figure 10b). In these searchlights, increases in LIFC response predicted increases in degree of pattern change from Part A to Part B. No above-threshold searchlight clusters for emerged for LT or NAT items.

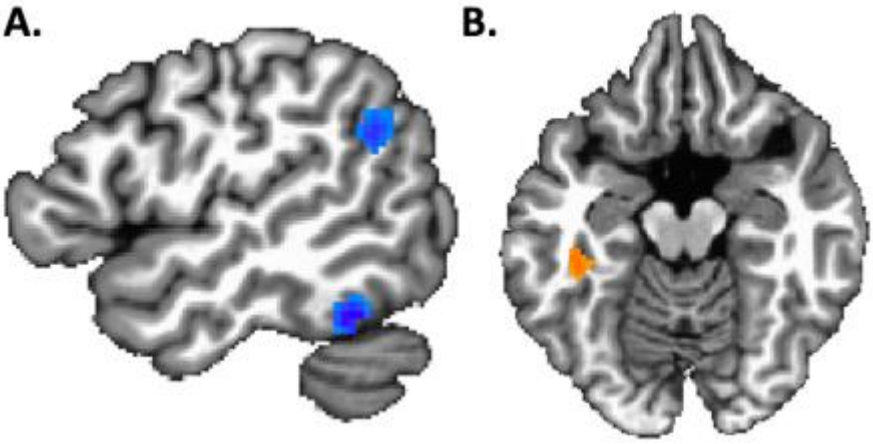

Figure 3.10

Brain regions in which a continuous variable exhibited a linear relationship with cross-context pattern change for one stimulus category. Figure 10a depicts two clusters, one in left angular gyrus and one in left inferior temporal gyrus, in which degree of typicality negatively predicted cross-context pattern change for artifact items. Figure 10b depicts one cluster in left fusiform gyrus, cross-context pattern changes were reliably predicted by increases in mean LIFC response for the natural kind stimuli.

4.4.2. ROI Analysis: Predictors of Cross-context, Within-Item Pattern Change

In each VTP ROI mask, we tested whether degree of pattern change could be predicted by typicality scores or by LIFC response. For ARTs, typicality scores negatively predicted degree of pattern change in all three ROI masks, $t(15)= -2.28$, $p= .04$ (mean $r= -.08$, SD= .13) for the stable masks; $t(15)= -2.96$, $p=.01$ (mean $r= -.08$, $SD=$ .11) for the W>NW ROI masks, and $t(15)= -2.24$, $p= .04$ (mean= -.07, $SD= .13$) for W># ROI masks. The negative relationship between item typicality and degree of pattern change was robust across almost all masks sizes in each ROI (Figure 11). There was no reliable linear relationship between typicality scores and degree of pattern change for LTs or NATs in any VTP ROI.

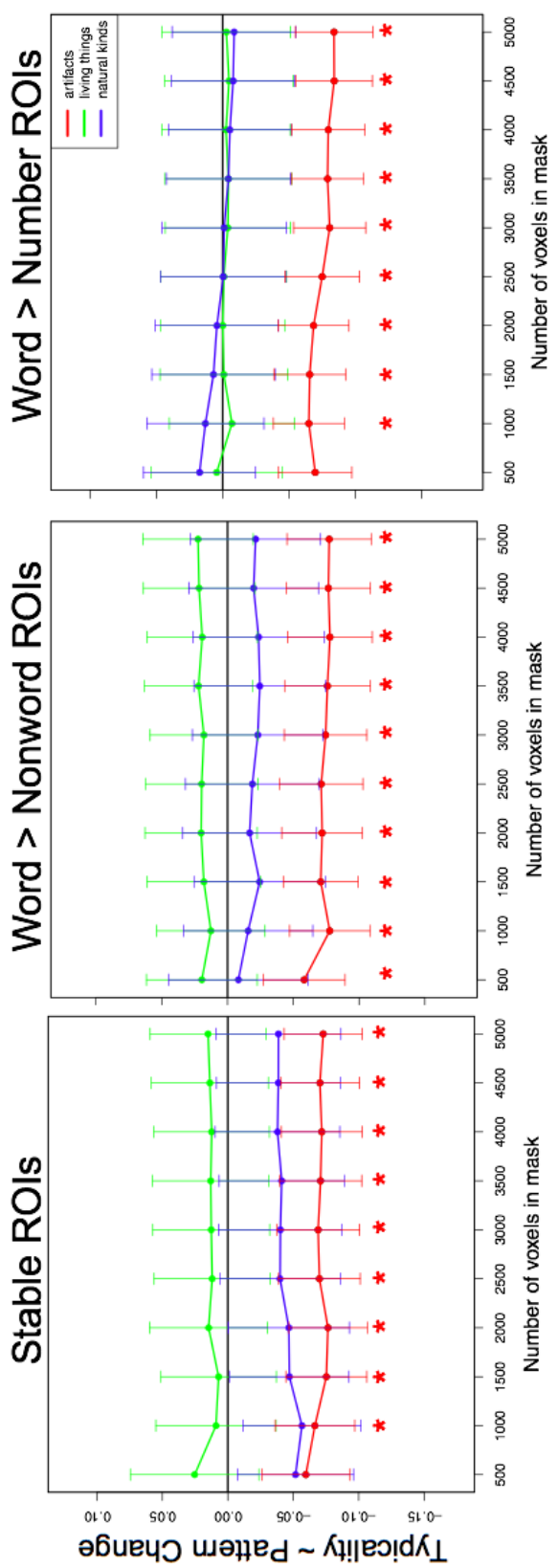
Stable ROIs
Word > Nonword ROIs
Word > Number ROIs
artifacts
living things
natural kinds
Typicality ~ Pattern Change
Number of voxels in mask
Number of voxels in mask
Number of voxels in mask

Figure 3.11

The average correlations between typicality scores and cross-context, within-item pattern change in each ROI mask, for each stimulus category. The three plots share the same y-axis and figure legend. For artifacts, typicality scores negatively predicted degree of pattern change for the multi-voxel patterns extracted from each ROI mask. This relationship was absent for both the living thing stimuli and the natural kind stimuli. Error bars depict standard error of the mean for each category. Asterisks indicate the reliability of the correlations versus zero at the $p<.05$ level. Asterisk colors correspond to category labels in the plot legend.

For NATs, LIFC response positively scaled with degree of pattern change in both the stable ROI mask, $t(15)= 2.32$, $p= .04$ (mean r= .10, $SD= .17$), and the W>\# ROI mask, $t(15)= 2.60$, $p= .02$ (mean $r= .10$, $SD= .15$), but not in the W>NW mask, $t(15)= 1.69$, $p= .11$ (mean $r= .07$, $SD= .18$). This positive relationship was robust across all mask sizes in the stable and W>\# ROIs masks, but was absent for either the LT or ART data (Figure 12). For NATs, the MVPs in these masks exhibited greater cross-context pattern changes when LIFC response increased during Part B.

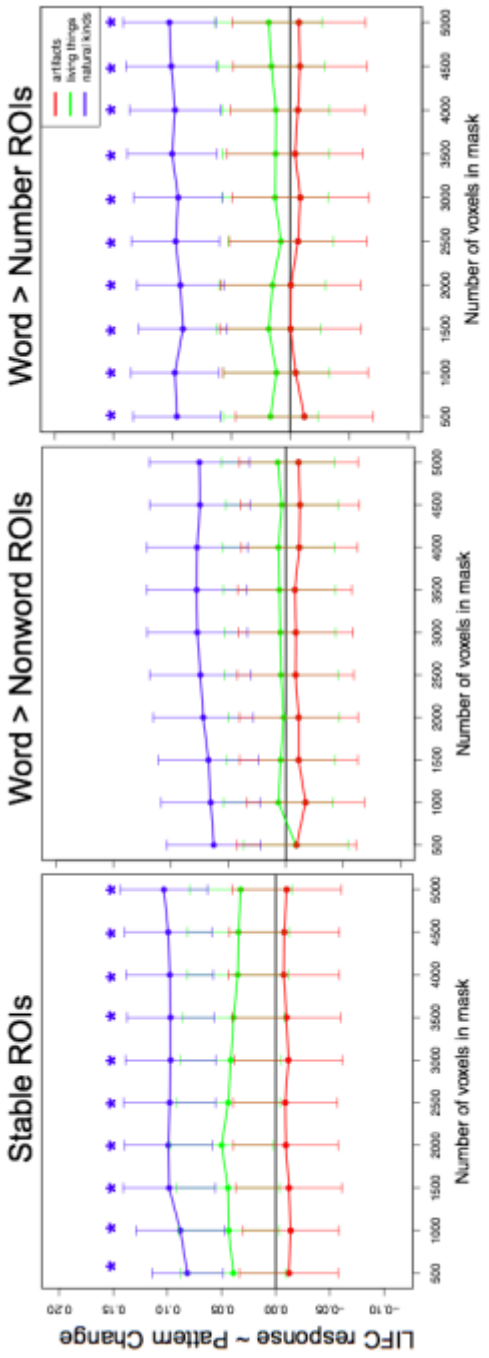
Stable ROIs
Word > Nonword ROIs
Word > Number ROIs
LIFC response ~ Pattern Change
Number of voxels in mask
artifacts
living things
natural kinds

Figure 3.12

The average correlations between average response in the LIFC ROI and cross-context, within-item pattern change in each ROI mask, for each category. The three plots share the same y-axis and figure legend. For natural kinds, the relationship between mean LIFC response and pattern change was reliably positive for the multi-voxel patterns extracted from the stable ROI mask (left) and the word vs. number ROI mask (right). This relationship was absent for both the living thing stimuli and the artifact stimuli. Error bars depict standard error of the mean for each category. Asterisks indicate $p<.05$.

4.4.3. Searchlight Analysis: Changes in Category-Level Similarity Structure

In each searchlight volume, we computed the correlation between the category-level similarity model (Figure 4a) and the observed pairwise neural similarities between each MVP, separately for each experiment part. The Pearson correlation coefficient, which quantifies the degree to which the neural data matches the category-level similarity model, was assigned to each searchlight's center voxel. After warping subjects' searchlight maps to a common template, we tested for reliable category-level distinctions in the Part A and Part B data separately, and then tested whether the strength of the category-level distinctions change from Part A to Part B.

The whole-brain analyses failed to reveal any above-threshold searchlights with reliable matches between the category-level similarity model and the neural data from either Part A or from Part B. However, paired comparisons between each part's match to the category model revealed a cluster in right temporal pole (Figure 13). In this 75-voxel cluster of searchlight centers (peak searchlight coordinates: x= 51, y= 17, z= -14), the correspondence between the category-level model and the neural data increased from Part A to Part B. However, the effect in this searchlight cluster is just below threshold, corrected alpha = .07.

In follow-up analyses, we examined effects at this peak searchlight volume in each individual subject. The MVPs at this peak searchlight volume show an increase in the match to the category-level model, $t(15)= 3.0$, $p= .01$, (Part A mean $r= -.12$, Part B mean $r= .14$). However, neither the Part A nor Part B neural data showed a reliable correspondence to the category level model, $t(15)= 1.8$, $p= .09$ for Part B, and $t(15)= -1.8$, $p= .09$ for Part A. We also tested the effects in this peak searchlight volume when limiting analyses to only two categories at a time. This analysis revealed that, from Part A to Part B, distinctions between LTs and NATs increased, $t(15)= 2.56$, $p= .02$, as well as

distinctions between LTs and ARTs, $t(15)= 3.45$, $p= .01$, but not between ARTs and NATs, $t(15)= .22$, $p= .83$.

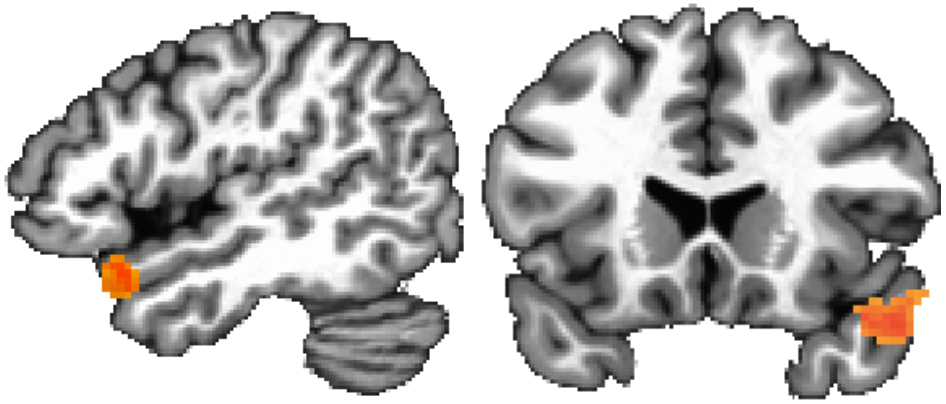

Figure 3.13

In 75 searchlight volumes centered in the right anterior temporal pole, multi-voxel activity patterns exhibited increased category-level distinctions during the living/nonliving judgment task in Part B, relative to the semantic encoding task in Part A. This effect was not above threshold at the whole-brain level, corrected $p<.07$.

4.4.4. ROI Analysis: Changes in Category-Level Similarity Structure

We also tested the correspondence between the category-level similarity model and the observed neural similarities in the VTP ROIs. In the "stable" masks, the neural data from Part A was negatively correlated with the category-level model, $t(15)= -2.36$, $p= .03$ (mean $r= -.16$, $SD= .27$). This finding suggests that the MVPs evoked during the semantic encoding task exhibited relatively greater pairwise similarities to members of other categories, relative to members of their own category. However, the direction of this effect reversed in Part B, such that the MVPs derived from the stable mask trended toward a positive correspondence to the category-level model, $t(15)= 1.94$, $p= .07$ (mean $r= .15$, $SD= .31$). This directional change in the relationship between the category model and the neural data from Part A to Part B was robust across all mask sizes (Figure 14). Within-subjects paired *t*-tests in each voxel mask confirmed that the effects reliably differed by experiment part, $t(15)= 3.09$, $p= .01$. Match to the category-level model also increased from Part A to B for almost all voxel mask sizes in the W>NW ROIs, $t(15)= 2.47$, $p= .03$, such that the data in Part B showed a reliable correspondence to the

category-level model, $t(15)= 2.42$, $p= .03$ (mean $r= .20$, $SD= .32$), while the data in Part A showed neither positive nor negative relationship to the data, $t(15)= -.28$, $p= .78$ (mean $r= -.02$, $SD= .27$) (Figure 14). In the W>## ROIs, larger mask sizes (3500-5000) voxels also positively matched category-level model, $t(15)= 2.13$, $p= .05$ (mean $r= .17$, $SD= .32$) for the 4000-voxel mask, and the increase in match from Part A to Part B also trends in the predicted direction.

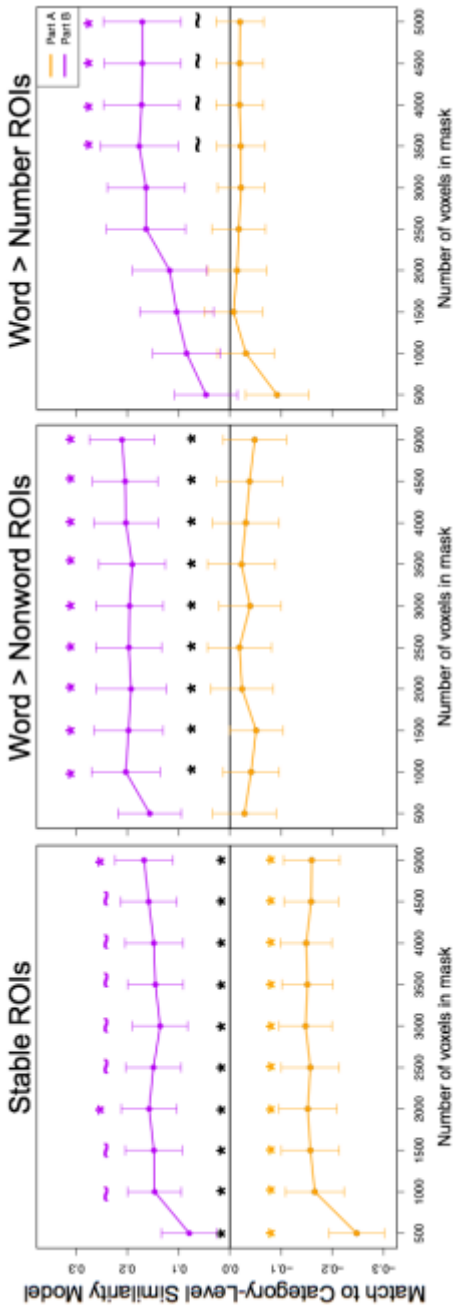

Figure 3.14

Group-level results depicting the correlations between the category-level model of similarity structure (see Figure 4a) and the observed neural data in each ROI mask. The three plots share the same y-axis and figure legend. Asterisks indicate significance at the $p<.05$ level and tildes indicate statistical trends, $p<.10$. Orange and purple symbols indicate reliable non-zero correlations that correspond to the dataset labels in the plot legend. Black symbols indicate reliable pairwise differences in category-level similarity between the two experiment parts. Error bars indicate standard error of the mean.

4.4.5. Links between category typicality, LIFC response, and item-level category selectivity

The above results indicate that degree of cross-context pattern change can be negatively predicted by typicality scores for ARTs, and positively predicted by LIFC response for NATs. Furthermore, in right temporal pole and in each ROI mask, category-level distinctions increased in Part B relative to Part A. Given these two sets of findings, we tested for links between them. That is, for artifacts, do the observed typicality-predicted pattern changes result in more category-selective patterns in Part B? Similarly, for natural kind stimuli, do the LIFC-predicted pattern changes exhibit increases category selectivity during Part B? We performed follow-up analyses for each of these stimulus categories to test whether these cross-context, within-item pattern changes are associated with increases in category selectivity.

4.4.6. ROI Results: Artifact atypicality predicts increases in category selectivity

In this analysis, we computed the change in each item's MVP category selectivity across the two tasks. This was accomplished by measuring each item's average within-category neural versus between-category neural similarity in Part B versus in Part A (Figure 4b). We predicted that item typicality would predict the degree an item's category selectivity would change from Part A to Part B, such that atypical items would exhibit increasingly category-selective responses (see Table 1). In the stable and W>$\#$ ROI masks, we observed a negative relationship between typicality scores and increases in item-level category selectivity for ARTs in the stable masks, $t(15)= -2.54$, $p= .02$ (mean $r= -.09$, $SD= .13$), and the W>$\#$ masks, $t(15)= -2.40$ $p= -.03$ (mean r= -.09, $SD= .14$), and a trending negative relationship in the W>NW masks, $t(15)= -1.85$, $p= .08$ (mean $r= -.07$, $SD= .15$). That is, ARTs with lower typicality scores experienced greater increases in their category selectivity from Part A to Part B than items with higher typicality scores.

In addition to the VTP ROIs, we also tested for relationships between typicality and changes in category selectivity in the peak searchlight centers in left angular gyrus and left inferior temporal gyrus which had exhibited cross-task pattern changes that negatively scaled with typicality for ARTs (Figure 10a). However, typicality did not correlate with changes in category selectivity in either of these searchlight volumes.

Additionally, there were no reliable relationships between typicality scores and changes in category selectivity in the peak right temporal pole searchlight which showed cross-task increases category selectivity (Figure 13).

4.4.7. ROI Results: No relationships between LIFC response and increased category selectivity

We also tested whether the trial-level changes in LIFC response, which predicted cross-context pattern changes for NATs, also predicted cross-context increases in category selectivity (Figure 12). We had predicted that, across tasks, increases LIFC response would be associated with increases in category-selective neural patterns. None of the tested VTP ROI masks exhibited reliable linear relationships between these two variables (Figure 15). Neural responses in the left fusiform gyrus peak searchlight, which had shown LIFC-related increases in cross-context pattern change, also did not show LIFC-predicted increases in category selectivity (Figure 10b). In addition, no reliable effects emerged for LTs or ARTs in any VTP ROI, or in the peak right temporal pole searchlight.

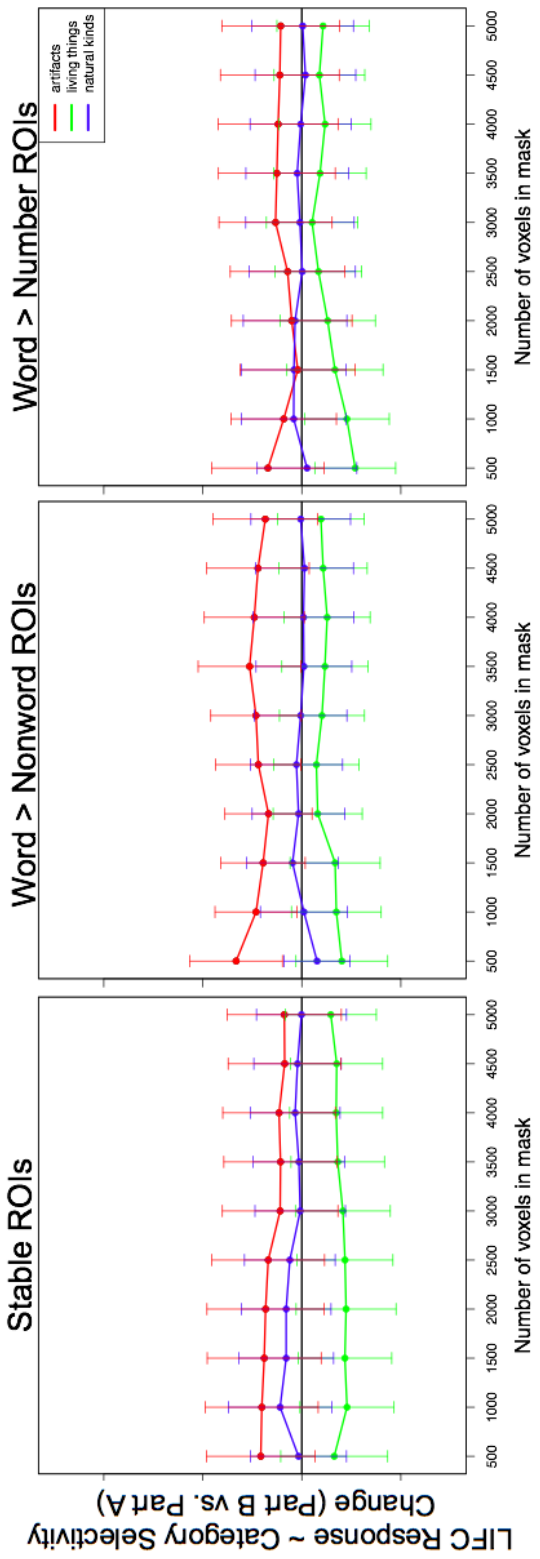
LIFC Response ~ Category Selectivity
Change (Part B vs. Part A)
Stable ROIs
Word > Nonword ROIs
Word > Number ROIs
Number of voxels in mask
artifacts
living things
natural kinds
500
1000
1500
2000
2500
3000
3500
4000
4500
5000

Figure 3.15

Relationship between item-level average response in left inferior frontal cortex (LIFC) and item-level changes in category selectivity from Part A to Part B.

4.4.8. ROI Results: Relationship between category typicality scores and category selectivity during semantic encoding

In addition to the cross-task predictions for changes in category selectivity, we had also predicted that, during the semantic encoding task in Part A, item-level typicality scores would predict the degree of category-relevant information in the multivariate patterns. Our reasoning was that, because typical category members possess more category-related information, their corresponding patterns would inherently exhibit high category selectivity, even when this information is not explicitly task-relevant. For each category, we computed the correlation between typicality scores and item-level neural measures of category selectivity in Part A (Figure 4b). For the ART stimuli, there were no reliable relationships between these two variables. For LT stimuli, typicality scores positively predicted degree of category selectivity in some of the stable and W>NW masks. For NATs, typicality was negatively related to category selectivity for the MVPs in some of the stable ROI masks, such that more typical NATs elicited less category-selective patterns during the semantic encoding task (Figure 16).

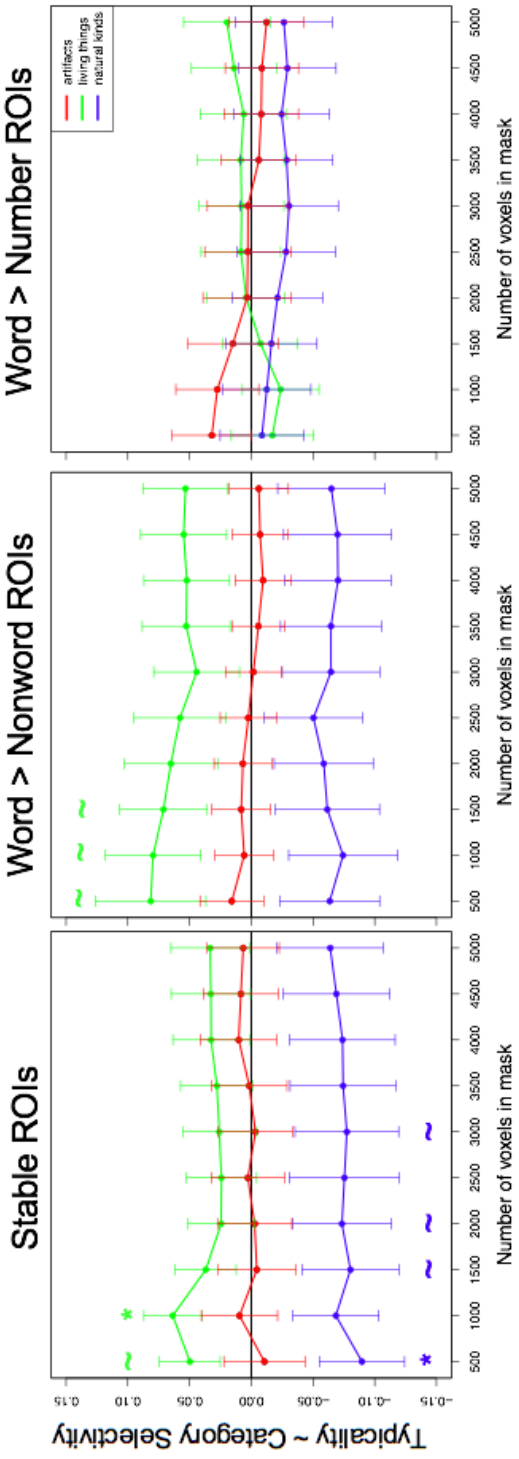
Stable ROIs
Word > Nonword ROIs
Word > Number ROIs
Typicality ~ Category Selectivity
Number of voxels in mask
artifacts
living things
natural kinds

Figure 3.16

Relationship between category typicality scores and item-level neural category selectivity in Part A. For NATs, increased typicality was associated with less category-selective neural patterns during the semantic encoding task, while some ROIs showed the reverse pattern for LTs. The three plots share the same y-axis and figure legend. Asterisks denote $p<.05$ and tildes denote $p<.10$ for the category of the same color. Error bars indicate standard error of the mean.

5. Discussion

We examined how neural and behavioral responses are modulated by category typicality and changes in LIFC activity during semantic retrieval of manmade artifacts, living things, and nonliving natural kinds. We obtained univariate and multivariate measures of BOLD response while subjects thought about the same set of stimuli under two distinct task conditions: once while thinking about each item's meaning in an undirected manner, and once while explicitly judging each item's living/nonliving status. Overall, we found that neural responses were modulated by task demands. Univariate responses varied by both category and typicality during living/nonliving judgments, but not during general semantic encoding. Moreover, voxels in right temporal pole and bilateral ventral temporal and parietal cortex exhibited reliable increases in category-selective multivariate responses once these distinctions were task-relevant. We interpret these findings as evidence of context-dependent retrieval of semantic information.

Additionally, item-level multivariate analyses revealed that, for some object categories, cross-task changes in neural patterns correlated with either typicality scores or changes in LIFC response in ways that were consistent with our hypotheses. However, several of our predictions regarding item-level measures of category selectivity were not borne out by the data, and most of these analyses yielded inconclusive results. For instance, we failed to observe any relationships between increases multivariate measures item-level category selectivity and univariate LIFC response for any object category. Additionally, the correspondences that we observed between item typicality ratings and neural category selectivity were elusive. Below, we summarize and discuss the most striking results from these data. We then consider theoretical and methodological explanations for the divergences between our hypotheses and the observed results. We

conclude with recommendations for refining these methods in a way that resolve some of the new questions raised by these experiments.

5.1. Typicality Effects on Behavioral Responses and Univariate Activity

We observed binary effects of category typicality for LTs, such that living/nonliving judgments of atypical LTs were delayed and also recruited an increased LIFC response, relative to typical LTs. Moreover, across items in the LT category and the ART category, continuous ratings of typicality were negatively correlated with RT and LIFC response. In contrast, no binary or continuous effects of category typicality were observed for NAT stimuli. Although the observed judgement delays for atypical items is consistent with previous studies, our typicality effects were limited to LT stimuli, and only emerged for the ART stimuli when typicality was treated as a continuous rather than binary variable. In contrast, a previous study reported binary typicality effects for all three stimulus categories (Goldberg & Thompson-Schill, 2009). One possible explanation for difference in typicality effects between the present study and the previous one is the difference in response demands. Subjects in the present study had three-fold increase in the time allotted for their judgment (3000ms maximum limit in our study, versus 1000ms in Goldberg & Thompson-Schill, 2009), and subjects here spent nearly twice as long considering their judgment prior to selecting their ultimate response (approximately 1300ms on average in the present study, versus 650ms previously).

Perhaps the processing disadvantage for atypical NAT and ART stimuli is weak enough that it can be eliminated with additional processing time. This could explain the lack of effects for these items, both in response delays and univariate LIFC response. In contrast, the typicality advantage for LTs might be so robust that it emerges even under relatively unspeeded conditions. In fact, the LT typicality effect is often observed using tasks that require self-paced responses (Zaitchik et al., 2014; Opfer & Siegler, 2004). Additionally, our stimulus selection and the segregation of our analyses by basic-level category might have concealed underlying typicality effects that are present when regarding all NLTs together (i.e., collapsing across the ART/NAT distinction). In fact, behavioral data indicates that NATs received lower typicality ratings than ARTs, and that NAT versus ART judgments are delayed (Figure 6) and elicit greater responses from

LIFC (Figure 9). In addition to category typicality effects at the subordinate category level (i.e., NAT vs. ART), it will be worthwhile, in future analyses, to consider typicality effects for the entire domain of NLTs.

In contrast to the typicality effects that emerged during category judgments, response in LIFC—or elsewhere in the brain—was not modulated by category or typicality when this information was not explicitly task relevant (i.e., during the semantic encoding task). One potential avenue for future research is to investigate interactions between task demands and typicality effects. It is certainly possible that behavioral and/or neural typicality effects could emerge even when category-related information is not relevant for the task at hand. However, this was not the case in the present study, under the conditions of the general semantic encoding task.

5.2. Category Effects on Behavioral Responses and Univariate Activity

Neural and behavioral responses during the living/nonliving judgments were also modulated by stimulus category. Judgments of NAT items were delayed and elicited increased BOLD response in several brain areas, including two frontal regions (i.e., LIFC and MFG), left fusiform gyrus, and left inferior parietal lobule. These distinctions are consistent with some category-level effects that have been previously reported in the literature (Devlin et al., 2002). However, the category-level distinctions which are most commonly reported in fMRI studies of object processing, such as the medial/lateral fusiform dissociation between living things (e.g., animals) and artifacts (e.g., tools) were not observed in the present data (cf. Chouinard & Goodale, 2010).

In contrast to the category-level differences observed during Part B, BOLD response during the semantic encoding task in Part A did not vary by category. One possible explanation for this result is that subjects were not considering an item's category membership while performing the general semantic encoding task, and the information that subjects retrieved about each item during that task did not systematically vary by category. One additional factor to consider is the task that subjects performed in-between stimulus presentations (i.e., the number parity judgments). Perhaps a more neutral baseline is required to detect BOLD contrasts that are modulated by category. In fact, several previous studies have identified reliable category-level effects in BOLD

response during similar encoding tasks (e.g., picture naming) when BOLD response was contrasted to a neutral, fixation baseline as in Part B (Zannino et al., 2010; Garn et al., 2009).

5.3. Cross-Context Changes in Multivariate Activity Patterns

5.3.1. Increases in Category-Level Distinctions in VTP ROIs and right temporal pole

We tested for changes in multivariate activity patterns while subjects thought about the same set of items in two different ways. We predicted that, once category-level distinctions become task relevant (i.e., during the living/nonliving judgment task), the observed neural activity patterns would exhibit greater category selectivity, relative to when these distinctions are unrelated to the task demands (i.e., during the semantic encoding task). Across the three stimulus categories, we observed increases in category-level semantic similarity structure in each VTP ROI mask and in a cluster of searchlight volumes in the right temporal pole. In these voxels, the relative similarities amongst the neural activity patterns shifted from Part A to Part B, such that the neural responses evoked by members of the same category became relatively more similar to one another, and less similar to patterns evoked by non-category members. The localization of these effects is broadly consistent with previous findings. Responses in right anterior temporal lobe are consistently associated semantic processing (Binney et al., 2010; Lambon Ralph et al., 2009), and previous MVPA studies have identified multivariate patterns throughout regions of ventral temporal cortex (Clarke & Tyler, 2014; Kriegeskorte et al., 2008) and parietal cortex (Fairhall & Caramazza, 2013) where responses vary by object category. Furthermore, the observed context-dependent changes in multivariate activity patterns is congruent with several previous reports in the fMRI literature, in which category-level distinctions increase when such information is task relevant (cf. Carlson et al., 2014, Ritchie et al., 2014).

These findings can be interpreted as a change in the type of information that the neural responses express. For instance, the information in the activity pattern that pertains to category identity might be more strongly activated when this information is recruited by the task demands. In such a scenario, activity patterns evoked by items with similar category identities will exhibit increased similarity when the activation of category

information is increased. This interpretation is consistent with a view in which the relative activation strengths of the various aspects of a concept's identity will vary across contexts, depending on which information is relevant for the current task or context at hand.

5.3.2. Category Typicality and LIFC Response Predict Within-Item, Cross-Context Changes

In addition to testing for cross-context changes in category-level distinctions amongst the MVPs, we also tested whether, across items, we could predict the degree of item-level change from Part A to Part B. On an item level, we tested whether our two measures of semantic conflict (i.e., category typicality and increases in LIFC response) could predict the degree of change in the multivariate patterns evoked by each item across the two experimental tasks. We predicted that atypical items would undergo greater changes to their MVPs, and that their patterns would exhibit increases in category selectivity from Part A to Part B. We found partial support for these hypotheses in the neural patterns evoked by the artifacts (ARTs) and natural kind stimuli (NATs).

5.3.3. Item-level Changes to Artifacts Relate to Category Typicality

For artifact stimuli, typicality scores negatively predicted degree of cross-task pattern change for artifacts in the VTP ROIs as well as left-lateralized fusiform gyrus and angular gyrus. In these regions, an item's relative category typicality predicted the degree to which its evoked activity pattern exhibited a similar response in both Part A to Part B. This result suggests that the neural representations of atypical artifacts undergo greater changes once these items are explicitly thought of as nonliving things. However, although this finding indicates that the neural patterns are indeed changing from Part A to Part B, but it does not address *how* the patterns are changing. We hypothesized that, due to changes in the task demands, the activation strength of category-relevant information would increase in Part B relative to Part A. To test this hypothesis more directly, we measured each item's degree of category selectivity in each experiment part. In each VTP ROI mask, artifact typicality scores negatively predicted increases in category selectivity from Part A to Part B. That is, the neural patterns evoked by atypical artifacts became

increasingly similar to other ARTs and increasingly different from NATs and LTs once category-level information was relevant to perform the task at hand. Conversely, typical ART items experience smaller changes in category selectivity from Part A to B.

One possible explanation for this pattern of results is that perhaps typical ARTs always exhibit more category-selective patterns than atypical ARTs, even under unbiased conditions (e.g., during the semantic encoding task), and hence their Part A patterns already resemble their Part B patterns. In a follow-up analysis, we tested whether item-level typicality scores predict degree of category selectivity during Part A exclusively. This analysis failed to reveal any reliable relationship between category typicality and category selectivity during the semantic encoding task for ARTs (Figure 15). Even though the activity patterns of atypical ARTs exhibited increased changes and increased category selectivity from Part A to Part B, it is not necessarily because the activity patterns of typical items already exhibit category selectivity in Part A. To summarize: although we can predict the degree of change in ART patterns across contexts, and the direction of the changes these patterns undergo, we are unable to further describe or characterize their evoked patterns during the semantic encoding task. Next, we turn to the findings for cross-context changes in the neural patterns evoked by the natural kind stimuli.

5.3.4. Item-Level Changes to Natural Kinds

For the natural kind stimuli, degree of cross-context pattern change was predicted by item-level increases in average LIFC response. This positive relationship was limited to searchlight volumes in left fusiform gyrus, and the word-selective VTP ROIs (i.e., W > NW and W > #). In these voxels, increases in LIFC response from Part A to Part B were associated with increases in neural pattern changes across the two tasks. This result indicates that, for natural kind stimuli, LIFC activity is linearly related to cross-task changes the multivariate activity patterns that are expressed in posterior regions of ventral temporal and parietal cortex.

After observing the relationship between LIFC response and cross-context pattern change, we tested whether LIFC also predict increased category selectivity to Part A to Part B. Although changes in LIFC response predicted degree of cross-context pattern

change in general, LIFC activity did not predict increases in category selectivity in particular. We observed no reliable relationship between trial-level increases in LIFC response and increases in category selectivity in any tested ROI, either for NATs or for the other two stimulus categories (Figure 15).

5.3.5. Findings for Living Things Stimuli

Although both the behavioral and univariate LIFC typicality effects were largest amongst the LT stimuli, we were unable to predict degree of cross-context MVP pattern change or increases in category selectivity for these items. Along with NATs and ARTs, the LTs as a category exhibited increased category-level distinctions between their neural patterns in Part B relative to Part A (Figure 14). However, on an individual item level, we were unable to predict these increases. Although we performed feature selection in both an exploratory (i.e., searchlight analysis) and in a principled manner (i.e., functional and anatomical ROIs), it is possible that we nevertheless failed to identify voxels in which the predicted effects emerge for LTs. It is also possible that the spatial scale at which we sampled the multivariate patterns (2mm cubed voxels) or the searchlight volume size (8mm radius) were not the optimal spatial scales for detecting these effects.

In general, it is unclear why the effects that we have observed did not consistently emerge for all three tested stimulus categories; we did not have any a priori predictions that were specific to any one category. Why did only the ART stimuli exhibit the predicted relationships between typicality and increases in category-selectivity, and why were LIFC-predicted cross-task pattern changes limited to the NAT stimuli? The heterogeneity in the pattern of results across LTs, NATs, and ARTs might indicate that the changes in the response patterns across our two tasks are qualitatively different for different categories. Perhaps the semantic content of one category versus another might influence the proclivity of neural pattern changes, or our sensitivity to detect these changes at the level of spatial resolution and in the voxels which we selected. Future analyses might benefit from separately performing multivariate feature selection for each individual stimulus category.

5.6. Relationship between typicality scores and neural category selectivity

In our original hypotheses, we had predicted that an item's degree of neural category selectivity would depend on its degree of category typicality, such that more typical category members would evoke neural patterns that are more similar to items of the same category, and less similar to items from opposing categories. In particular, we had hypothesized that this positive relationship would occur in the Part A data, when the item-level neural patterns were evoked during the general semantic encoding task. Our reasoning here was that, even under these "unbiased" conditions (i.e., when subjects are not explicitly probed to think about category membership), that category information would nevertheless be encoded and activated during semantic retrieval of the most representative category members. That is, a typical living thing like *tiger* would evoke a neural pattern that is inherently more "LT"-like, and more distinct from items like *broom* and *sand*. We found weak support for this hypothesis for LTs in some VTP ROI masks, and opposing evidence for NATs in the stable ROI masks (Figure 16).

Although it is challenging to speculate about this collection of weakly significant observations, because several factors may contribute to the absence of an effect, it is nevertheless worth considering some possible explanations for our findings. One possibility is that the instructions and behavioral demands of the semantic encoding task inadvertently encouraged item-level individuation of the stimuli, such that the retrieval of category-level information was avoided or minimized. During the Part A task, subjects were encouraged to think about each word meaning in preparation for a subsequent recognition memory test. One potential strategy for this task is to focus on the most memorable and distinctive aspects of the word meaning, which might lead to item-specific neural patterns that are highly individuated. Furthermore, during the subsequent memory test, subjects were required to distinguish between items that had been presented in the scanning run (i.e., "hit" items), versus "lure" items. However, the lure items were drawn from the same categories as the hit items, and thus it might have been challenging to distinguish between hits and lures without focusing more on item-specific details and less on category-general details during semantic encoding.

If subjects were retrieving highly specific, idiosyncratic, and memorable thoughts about items during Part A, then perhaps it is less surprising that typicality scores did not

predict the category selectivity of the neural patterns evoked during those thoughts. It remains an open and compelling question: what kind of behavioral task would be better suited to encourage subjects to semantically encode word meanings in a way that elicits inadvertent but not explicit retrieval of category membership? In the behavioral priming literature, there are examples of facilitated performance on an orthogonal task (e.g., lexical decision tasks; a pronunciation task) when target stimuli are preceded by same-category primes (cf. Lucas, 2000; Thompson-Schill et al., 1998). Perhaps a relationship between category typicality and category-selective neural responses would be more likely to manifest under task conditions that require less focus on distinctive item-level characteristics.

5.7. Study Limitations and Future Directions

Future investigations of context-dependent changes in multivariate patterns might benefit from adopting an experimental paradigm similar to the one employed here, but with several methodological adjustments. First, the power and sensitivity to detect multivariate pattern changes might be stronger if the experimental tasks are better matched. In the present study, the numerous differences between the Part A and Part B might have resulted in neural patterns that are less directly comparable. The two tasks had different response demands: unlike the living/nonliving judgment task in Part B, the semantic encoding task in Part A did not require an overt and explicit judgment and behavioral response during the stimulus word presentations. Subjects also performed different tasks during the stimulus inter-trial intervals: in Part A, subjects performed number parity judgments between word presentations, and in Part B, they merely viewed a fixation cross and waited for the next stimulus to appear.

Moreover, future investigations might want to predict a specific similarity structure in Part A and then examine the extent to which this similarity structure persists or fades once those distinctions are no longer task-relevant in Part B. In the present study, we had no specific predictions about the similarity structure that should manifest in the Part A neural patterns, other than that perhaps the typical items would exhibit increased category selectivity, relative to the atypical items. This hypothesis was unsupported.

Additionally, the multivariate feature selection would benefit from more principled voxel selection methods. An independent localizer, with a different set of stimulus items or with a pilot fMRI subject group, could be used to localize areas that are sensitive to the distinctions of interest. Then, a main experiment that is informed by the independent preliminary data could test how the strength of the studied distinctions change under different task demands.

5.8. Summary

This study investigated how category typicality modulates behavioral judgments; univariate activity; and changes in multivariate activity patterns in response to several object concepts, including living things, artifacts, and natural kinds. We found that degree of category typicality predicts delays in domain-level judgments and increases in LIFC response for both living things and man-made artifacts. Additionally, we identified voxels in which the multivariate activity patterns undergo changes, depending on the current context and task demands. Taken together, these results demonstrate that thoughts about living and nonliving things evoke variable and context-dependent multivariate activity patterns, and that these pattern changes reflect the enhancement of the within-stimulus aspects and between-stimulus relationships that are most relevant to the task at hand.

IV: SEMANTIC VARIABLITY PREDICTS NEURAL VARIABLITY OF OBJECT CONCEPTS

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 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.

Table 4.1

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

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.

Table 4.2

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, $p<.05$.

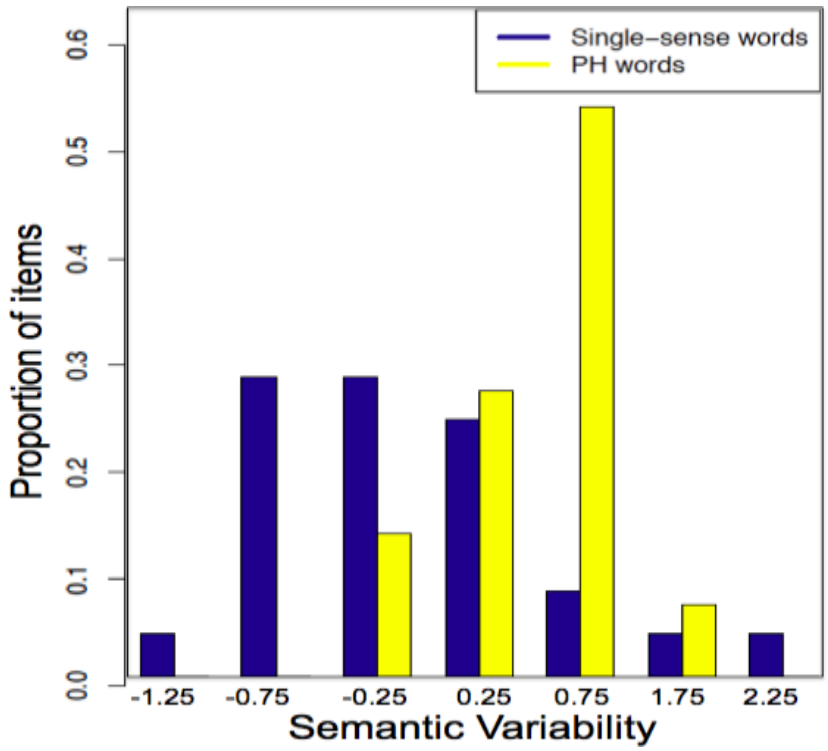

Figure 4.1

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

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.

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).

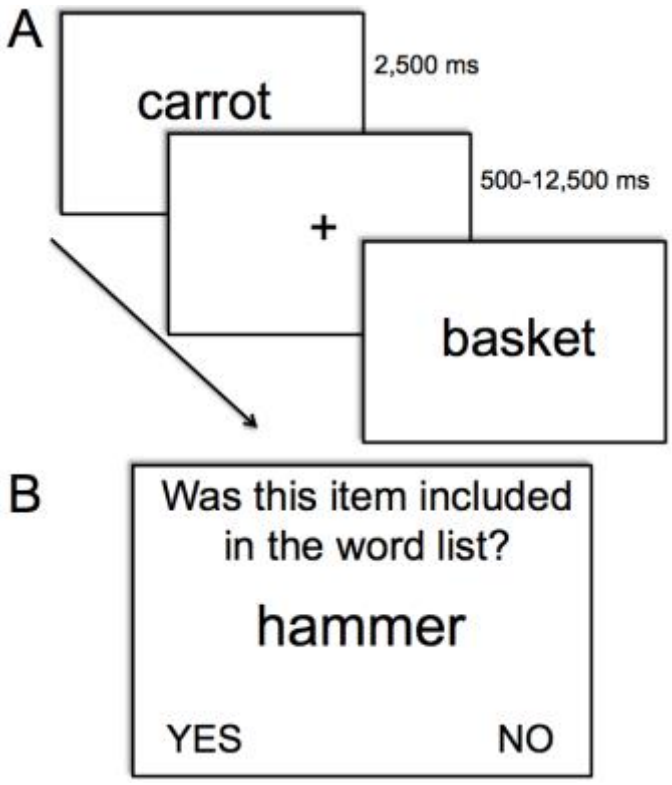

Figure 4.2

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.

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.

3.1 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.

3.2. 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.

3.3. 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.

3.3.1 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 C.

3.3.2 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.

4. Results

4.1 Whole-brain distributed patterns

4.1.1 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).

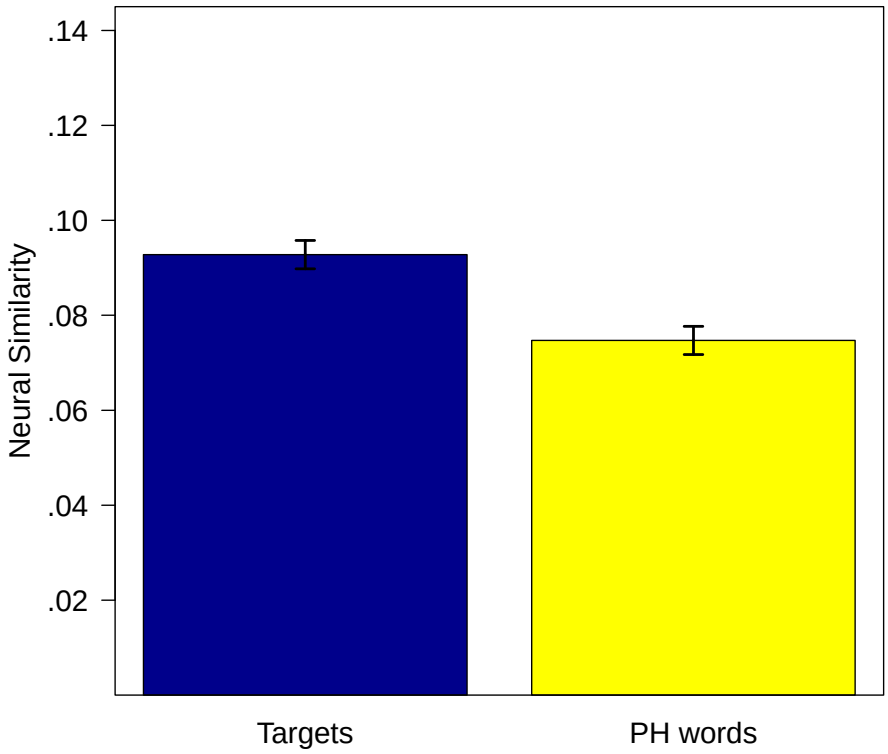

Figure 4.3

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

4.1.2. 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).

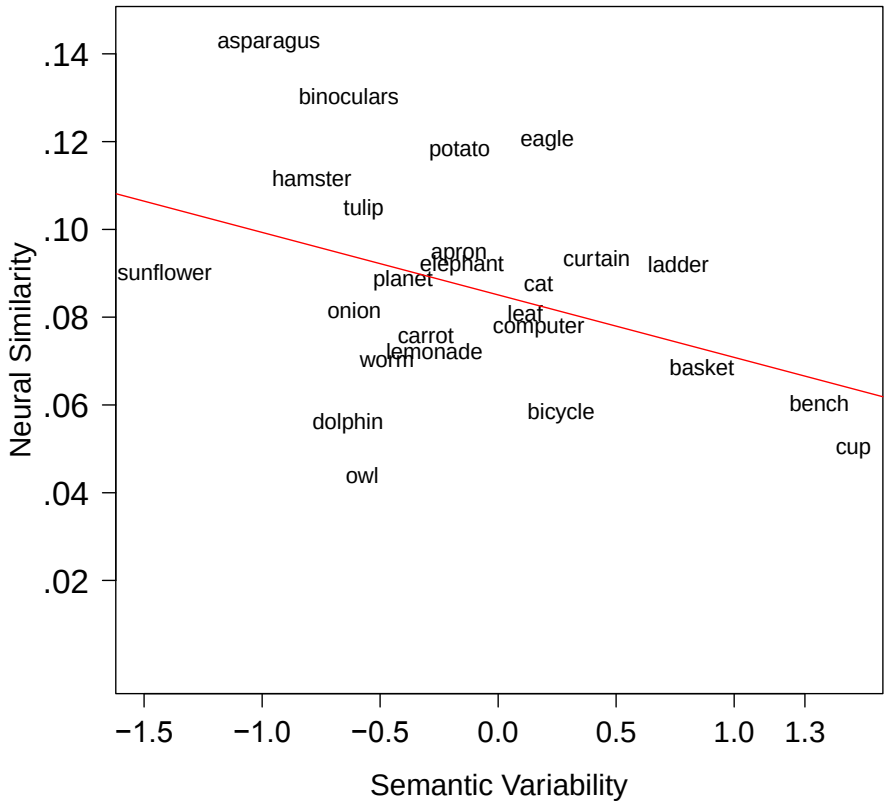

Figure 4.4

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.

4.2 Searchlight Localized patterns

4.2.1. 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).

Table 4.3

Peak searchlight centers from whole-brain analysis

	Cluster Extent	x	y	z	peak *t*-value	Brain region	Similarity result
Comparison 1: Neural similarity by word type	685	14	-85	-1	13.5	R. lingual gyrus	targets > PH
		-8	-83	-13	-7.5	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	3.8	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

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.

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.

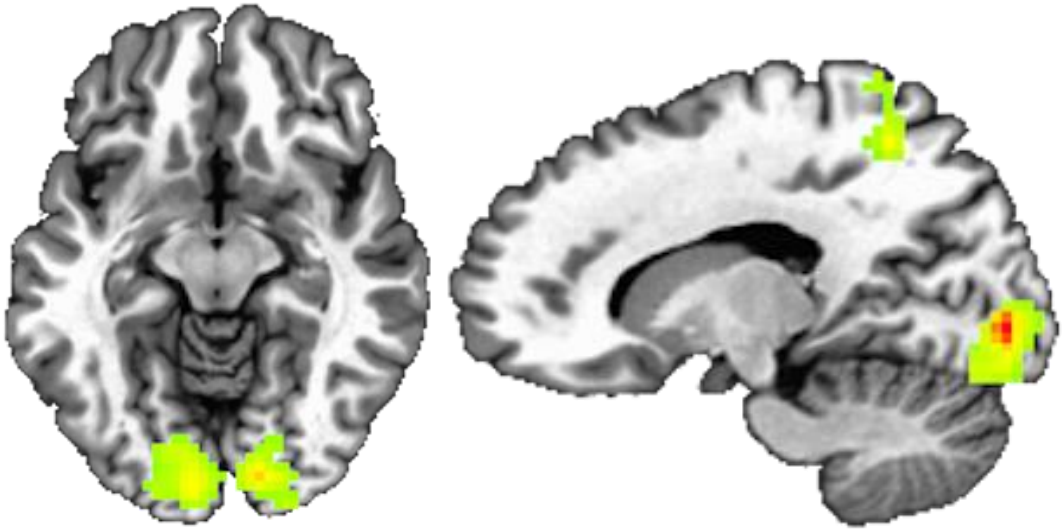

Figure 4.5

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 4.6

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

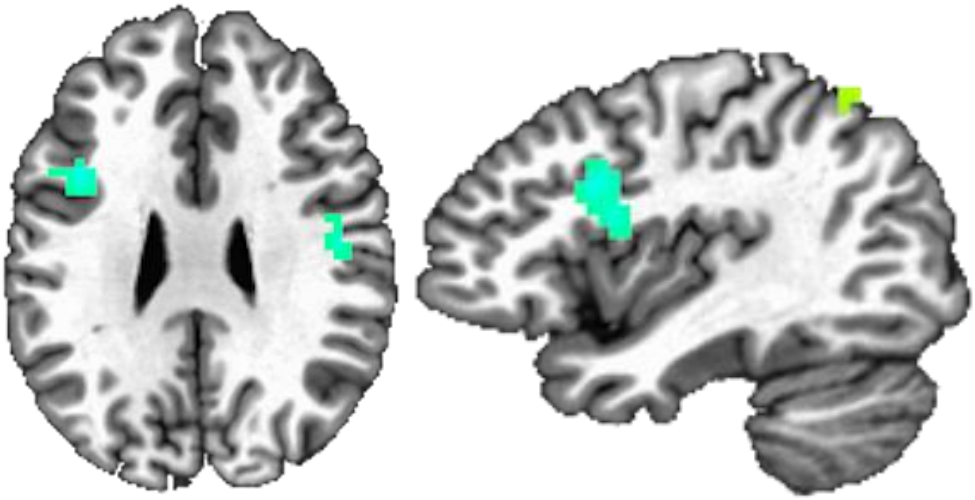

Figure 4.7

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.

4.2.2. 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.

Figure 4.8

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.

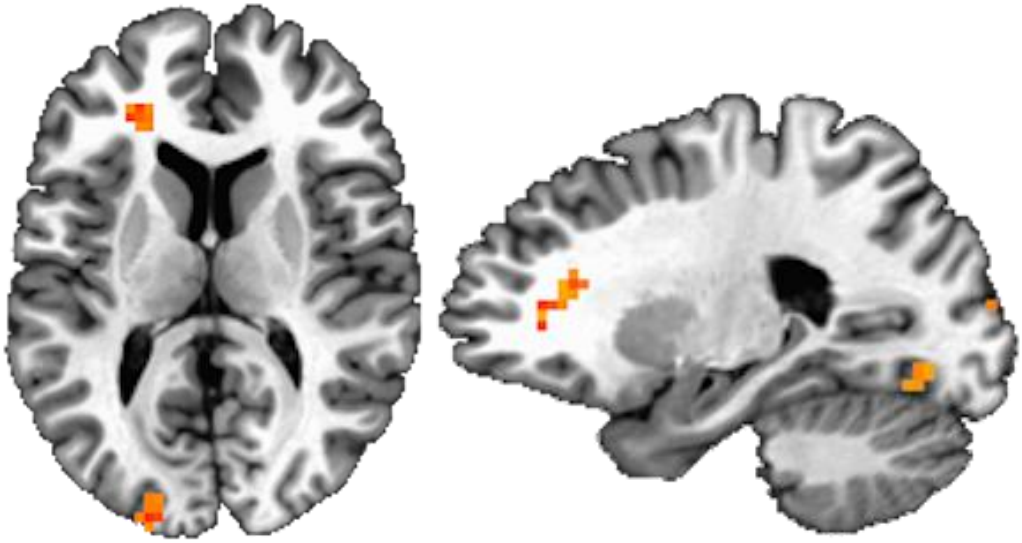

Figure 4.9

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 4.10

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.

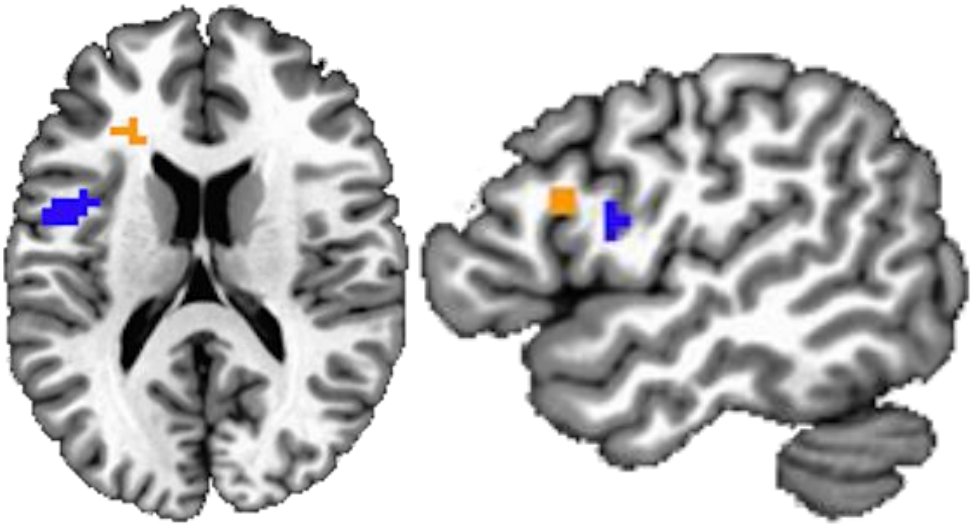

Figure 4.11

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.

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. 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.

5. 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.

5.1 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.

5.2 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.

5.3 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 multi-voxel 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) 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.

5.4 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.

5.5 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.

5.6 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 C). 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.

V. DISCUSSION

Many cognitive neuroscientists seek to purposefully isolate or distill thoughts about concepts down to common features that can be observed through stable neural patterns. While this mode of inquiry is consistent with one of the key features of a thought—namely, namely that it coheres— it necessarily discards information which may be just as important to the precise shapes of these thoughts. That is, the variation and flexibility of a concept embedded in a context. Without this variability, we could imagine that thoughts would be far too rigid to accomodate the transformation of our thoughts of, for instance, the swiftly changing river bed from our introductory remarks. We have not only found that words elicit variable brain patterns (the same variation which other experimenters seek to minimize), but that these variations encode meaningful information about the concept. Rather than theorizing that thinking about a concept (or reading a word) invokes a stable neural pattern and a stable meaning regardless of what you are doing with that word, the fMRI studies described in this dissertation show that words elicit variable brain patterns. Not only are these variations meaningful, but they are meaningful in different ways. The degree of a stimulus item's neural variability can be influenced by the sentence context in which the word appeared (Chapter 2); the task you are performing with the word (Chapter 3); or even other concepts that you were thinking about at that moment, or just beforehand (Chapter 4). We also present preliminary evidence that prefrontally-mediated cognitive control processes are involved in expression of context-appropriate neural patterns. In sum, these studies provide a novel perspective on the flexibility of word meanings and the variable brain activity patterns associated with them.

In Chapter 2, we showed that a single stimulus word with two different meanings can evoke two different patterns. In left anterior temporal lobe, this within-word, cross-context pattern dissimilarity is predicted by measures of homonym meaning frequency and left vlPFC response. In Chapter 3, we then examined responses when the conceptual referent of the word does not change, but rather the task-relevant features of a given concept are manipulated by task demands. Once certain stimulus features (i.e., category-related information) were made task-relevant, these distinctions increased in the neural

patterns evoked in distributed regions of ventral temporal and parietal cortex. In Chapter 4, we again tested neural patterns that evoked by same conceptual referent at different times in the experiment, but we studied word-evoked neural pattern variability under spontaneous, undirected task conditions. We observed a correspondence in the diversity of a word's invoked meanings, indexed by text co-occurrence statistics, and the variability in the neural patterns that it evokes in gray matter voxels distributed throughout cortex. Taken together, these studies demonstrate the utility of measuring and predicting the neural variability among separate presentations of the same stimulus item, rather than discarding this variability by averaging over it.

5.1. Characterizing the Information Contained in Variable Neural Patterns: Assumptions and Future Directions

In the future, it will be important to further characterize the information that is represented in the variable neural patterns that we observed. In the present studies, we claim that the neural patterns that we measured were sensitive to semantic information that is retrieved upon reading a stimulus word. Based upon this premise, we further contend that the variability in the neural patterns that we observed across contexts was due to variable retrieval of word meanings. In Chapters 2 and 3, we attempted to find support for the claim that the observed neural patterns encoded conceptual information. To accomplish this, in addition to measuring and predicting within-word neural similarity, we also tested for between-word neural similarities. Following the similarity logic of MVPA studies, if the observed neural patterns contain information about word meanings, then words with similar meanings should evoke similar patterns, and word with less similar meaning should evoke less similar patterns.

In Chapter 2, we found a cluster of searchlight volumes in left ATL where the degree of within-homonym, cross-meaning neural similarity was predicted by item-level measures of meaning frequency and trial-level fluctuations in left vlPFC response. To further address whether the variable neural patterns in left ATL reflected the expression of variable meanings, we correlated each homonym's two distinct meaning-evoked neural patterns with a third pattern: that evoked by a single-sense synonym of the dominant meaning. In support of our claim that these voxels reflect information about

word meanings, the synonym-evoked patterns exhibited greater neural similarity to the dominant-biased homonym patterns than they did to the subordinate-biased homonym patterns. Moreover, the degree of left vlPFC response scaled positively with the degree of the same-meaning neural similarity advantage: when left vlPFC response increased during the retrieval of a subordinate homonym meaning, the word's left ATL pattern exhibited relatively less similarity to the synonym, versus the similarity between the dominant meaning and the synonym. These results provide preliminary evidence that the neural patterns observed in left ATL are indeed sensitive to word meanings.

However, a more powerful and conclusive way to identify meaning-sensitive voxels would be to compare all stimulus-evoked neural patterns to one another, and to predict the neural similarities amongst all stimulus items, rather than specific stimulus pairings or triads. In Chapter 3, we tested for correspondences between the predicted neural similarities within and between all members and non-members of three basic taxonomic object categories (i.e., living things, artifacts, and natural kinds). At several spatial scales in ventral temporal and parietal cortex, we identified voxels in which the observed neural similarities corresponded to a category-level model of semantic similarity. However, this neural-semantic similarity correspondence was task-dependent: it was only observed when subjects were explicitly required to retrieve category-related information about each stimulus word. When subject's thoughts about each word meaning were unconstrained (i.e., during the semantic encoding task), we failed to identify any voxels where neural pattern similarity correlated with the predicted category-level similarity relationships. Further, we were unable to predict the degree to which an item would exhibit category-selective patterns, or the degree to which this category selectivity would increase from the semantic encoding task to the category judgment task.

This null finding could be interpreted as evidence that subjects did not retrieve category-related information during the semantic encoding task, or at least not in a way that conformed with the predictions in the model. To further characterize the information represented in the observed neural patterns, we recommend exploratory, data-driven analysis techniques, such as hierarchical clustering or multi-dimensional scaling (MDS). These methods project high-dimension similarity spaces onto a simpler, lower-dimension space. It may be possible to use MDS to observe the similarity relationships among the

word-evoked neural patterns and then interpret the aspects of the stimulus dimensions that are represented in the selected voxels (Harel et al., 2012). With these exploratory approaches, one might better characterize and address the underlying representations that are encoded in variable response patterns that we observe.

Additionally, to further ensure that the observed neural patterns reflect word meanings, future studies on this topic could employ additional feature selection and voxel localization criteria. In the present experiments, we targeted voxels in an exploratory manner (i.e., whole-brain searchlight analyses) and with more targeted functional contrasts (e.g., voxels that responded more to words versus numbers) and anatomical constraints (e.g., regions in ventral temporal cortex). However, the gold standard for feature selection would be able to first independently verify that the selected voxels reflect semantic content, either in an additional set of fMRI subjects, or with a separate set of similar experimental stimuli. These independent test samples could also help carefully delineate the boundaries of the hypothesis space, which is otherwise subject to selection among numerous free parameters in the data analysis (e.g., the smoothing kernel for the functional data; the number of voxels to sample; etc.) that are challenging to approach in a principled way.

5.2. Item-level predictors of neural pattern variability: reliable but weak correlations

The previous section outlines the ways in which we were limited in our ability to characterize the neural similarities that we observed. However, although we were not able to fully describe the similarities among the observed neural patterns, we were nevertheless able to predict the variability in their signals over time, using theoretically motivated, item-level variables. For instance, in Chapter 4, although all word-evoked neural patterns varied across contexts, the degree of variation across words was not random. Rather, it systematically conformed to our hypotheses regarding which word patterns should vary more than others. We must nevertheless acknowledge a caveat here. While all three studies found reliable group-level relationship between hypothesized item-level variables and the degree of cross-context neural variability, the observed correlations were relatively weak. For example, across subjects, the average correlation between semantic variability and item-level neural similarity in distributed gray matter

voxels in Chapter 4 was r= -.12. These small correlations indicate that there are several other factors contributing to the variability in the observed neural signals; these additional sources of variances are currently unaccounted for. This presents an exciting opportunity for future research to further explore the other potential psychological factors that might contribute to the observed neural variability. While the effects that we observed may be subtle, we nevertheless contend that measuring and predicting neural variability, rather than discarding it as noise, is a promising way to use neuroimaging to study the dynamic and flexible nature of cognition. In future work, additional cognitive neuroscience methods, especially those with high temporal resolution (such as magnetoencephalography and intracranial electroencephalography), will be particularly well-suited to study within-stimulus, cross-context variations in neural signals.

5.3. Relationship between LIFC response and context-appropriate, word-evoked multivariate signals

One main theme of the present work is the role of cognitive control processes in the recruitment of weak yet context-appropriate word interpretations when stronger yet inappropriate interpretations compete for selection. We proposed that left vlPFC would be critically involved in resolving this competition, and that this resolution would result in expression of task-relevant neural activity patterns. We found partial support for this hypothesis in Chapter 2 and Chapter 3. In Chapter 2, trial-level fluctuations in left vlPFC predicted the degree to which neural patterns in left ATL exhibited distinct activity patterns for distinct meanings of homonym words. However, this relationship was only observed in a post-hoc analyses, after identifying patterns in left ATL that scaled with a different item-level predictor of within-word similarity (i.e., meaning frequency scores). In an exploratory whole-brain searchlight analysis, we failed to identify any voxels that exhibited prefrontally-mediated neural similarity. In Chapter 4, we observed relationships between left vlPFC response and within-word neural similarity that were limited to a single stimulus category (i.e., natural kind stimuli), but the left-vlPFC changes in the neural patterns did not result in increased category selectivity, as our hypotheses predicted. Taken together, these findings provide limited support for the proposal that left

vlPFC biases the multivariate neural patterns evoked by variable word meanings in ventral temporal cortex.

In contrast, evidence of links between univariate left vlPFC response and multivariate VT patterns have been observed in other cognitive domains (e.g., object imagery and episodic memory). It is possible that more reliable relationships between word-evoked pattern variability and left vlPFC responses would emerge with other experimental paradigms. Another possibility is that the neural patterns evoked by word meanings are far more spatially distributed and spatially dynamic, relative to the picture-evoked patterns observed in relatively circumscribed brain areas during object perception and imagery. Indeed, identifying the neuroanatomical loci of the conceptual system continues to be a central pursuit in the field of cognitive neuroscience.

In addition, to directly test the role of left vlPFC response in the recruitment of context-appropriate neural patterns, future studies could use noninvasive brain stimulation (e.g., transmagnetic stimulation or transcranial direct current stimulation) to disrupt activity to this region, and observe degrees of subsequent expression of task-relevant multivariate patterns (for an example of one such paradigm in the domain of visual attention, cf. Lee & D'Esposito, 2012). But before testing the putative causal role of left lvPFC in expressing task-relevant neural patterns that reflect word meanings, one would first have to identify a paradigm in which the links between these two neural signatures of conflict resolution are much more robust and reliable than they are in the present studies.

One potential avenue for future investigations would be to study competition among context-dependent meanings that are associated with pictorial stimuli, rather than lexical stimuli. Several recent fMRI experiments have employed behavioral training paradigms to imbue visual stimuli with meaning over repeated experiences, and then observed experience-dependent changes in the object-evoked neural representations in visual regions (e.g., Hsu et al., 2014; Persichetti et al., 2015; Clarke et al., 2016). It might be fruitful to adopt a similar paradigm to study semantic conflict using visual object stimuli, and the putative role of left vlPFC in sculpting their corresponding neural signals.

5.4. Conclusions

In referring to instances of concepts that we observe, we use a common label. For example, the word "river" is applied to several different objects, none of which necessarily appear or behave in the exact same way. Further, the qualities of a single river can change over time, even in the instance of a footstep. The instances of concepts that we encounter in the real world are shaped by their surroundings, and so the meanings intended by their lexical referents are context-dependent as well. In the present set of studies, we have marshaled evidence to debunk the "same word, same meaning" theory, which is frequently contradicted by our daily life experiences. Adopting such a theory is certainly necessary if one wishes to isolate the essence of a concept at high precision in neural signals, but we can never fully characterize the neural correlates of semantic representation without expanding this theory to carefully account for the natural and necessary variation in these concepts as they are shaped by the dynamic variations in stimuli that invoke them. Variation exists in the world, in our thoughts, and in our brains. The findings from the present set of experiments illustrate that this variation and can be both measured and predicted in neural signals.

BIBLIOGRAPHY

Anderson, A.J., Murphy B., & Poesio M. (2014). Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. *Journal of Cognitive Neuroscience, 26*, 658-681.

Badre, B., & D'Esposito, M. (2007). Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *Journal of Cognitive Neuroscience, 19,* 2082-2099.

Baron, S. G., & Osherson, D. (2011). Evidence for conceptual combination in the left anterior temporal lobe. *NeuroImage, 55*(4), 1847–1852.

Bedny, M., McGill, M., & Thompson-Schill, S. L. (2008). Semantic Adaptation and Competition during Word Comprehension. *Cerebral Cortex, 18*(11), 2574–2585.

Bedny, M., Hulbert J.C., & Thompson-Schill S.L. (2007). Understanding words in context: the role of Broca's area in word comprehension. *Brain Research, 1146*, 101–114.

Ben-Yakov, A., Honey, C. J., Lerner, Y., & Hasson, U. (2012). Loss of reliable temporal structure in event-related averaging of naturalistic stimuli. *Neuroimage, 63*(1), 501-506.

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

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in cognitive sciences, 15*(11), 527-536.

Binney, R. J., Embleton, K. V., Jefferies, E., Parker, G. J., & Lambon Ralph, M. A. (2010). The ventral and inferolateral aspects of the anterior temporal lobe are crucial in semantic memory: evidence from a novel direct comparison of distortion-corrected fMRI, rTMS, and semantic dementia. *Cerebral Cortex, 20*(11), 2728-2738.

Brysbaert, M. & New, B. (2009). 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. 41*, 977–990.

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., & Ma, J. (2014). Reaction time for object categorization is predicted by representational distance. *Journal of Cognitive Neuroscience, 26*(1), 132-142.

Casasanto, D. & Lupyan, G. (2015). All Concepts are Ad Hoc Concepts. In Concepts: New Directions. E. Margolis & S. Laurence (Eds.) Cambridge: MIT Press.

Chan, A.H., Liu, H.L., Yip, V., Fox, P.T., Gao, J.H., Tan, L.H., 2004. Neural systems for word meaning modulated by semantic ambiguity. *NeuroImage, 22*, 1128–1133.

Chouinard, P. A., & Goodale, M. A. (2010). Category-specific neural processing for naming pictures of animals and naming pictures of tools: an ALE meta-analysis. *Neuropsychologia*, *48*(2), 409-418.

Church, J.A., Balota, D.A., Petersen, S.E., & Schlaggar, B.L. (2011). Manipulation of length and lexicality localizes the functional neuroanatomy of phonological processing in adult readers. *Journal of Cognitive Neuroscience, 23,* 1475-1493.

Clarke, A., & Tyler, L.K. (2014). Object-Specific Semantic Coding in Human Perirhinal Cortex. *Journal of Neuroscience, 34(14*), 4766-4775.

Clarke, A., Pell, P. J., Ranganath, C., & Tyler, L. K. (2016). Learning warps object representations in the ventral temporal cortex. *Journal of cognitive neuroscience*.

Chrysikou, E. G., Weber, M. J., & Thompson-Schill, S. L. (2014). A matched filter hypothesis for cognitive control. *Neuropsychologia*, 62, 341–355.

Coggan, D. D., Liu, W., Baker, D. H., & Andrews, T. J. (2016). Category-selective patterns of neural response in the ventral visual pathway in the absence of categorical information. *Neuroimage*, *135*, 107-114.

Coutanche, M. N., & Thompson-Schill, S. L. (2015). Creating Concepts from Converging Features in Human Cortex. *Cerebral Cortex*, *25*(9), 2584–2593.

Cox, R.W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computational Biomedical Research, 29*, 162–173.

Danelli, L., Marelli, M., Berlingeri, M., Tettamanti, M., Sberna, M., Paulesu, E., & Luzzatti, C. (2015). Framing effects reveal discrete lexical-semantic and sublexical procedures in reading: an fMRI study. *Frontiers in Psychology*, *6*, 1328.

Detre, G., Polyn, S. M., Moore, C., Natu, V., Singer, B., Cohen, J., et al. (2006). The multi-voxel pattern analysis (MVPA) toolbox. Poster presented at the Annual Meeting of the Organization for Human Brain Mapping (Florence, Italy).

Devlin, J. T., Russell, R. P., Davis, M. H., Price, C. J., Moss, H. E., Fadili, M. J., & Tyler, L. K. (2002). Is there an anatomical basis for category-specificity? Semantic memory studies in PET and fMRI. *Neuropsychologia*, *40*(1), 54-75.

Duffy, S.A., Morris, R.K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, *27*, 429-446.

Eickhoff, S.B., Stephan, K.E., Mohlberg, H., Grefkes C., Fink, G.R., Amunts, K, Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage*, *25*, 1325–1335.

Fairhall, S. L., & Caramazza, A. (2013). Brain regions that represent amodal conceptual knowledge. *Journal of Neuroscience*, *33*(25), 10552-10558.

Fedorenko, E., Hsieh, P.-J., Nieto-Castanon, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194.

Frith, C.D. (2000). The role of the dorsolateral prefrontal cortex in the selection of action. In: Monsell S, Driver (eds). *Control of cognitive processes.* Attention and Performance XVIII. MIT Press, Cambridge.

Garn, C. L., Allen, M. D., & Larsen, J. D. (2009). An fMRI study of sex differences in brain activation during object naming. *Cortex*, *45*(5), 610-618.

Gorfein DS, editor. (2001). On the consequences of meaning selection: perspectives on resolving lexical ambiguity. Washington (DC): American Psychological Association.

Graves, W. W., Desai, R., Humphries, C., Seidenberg, M. S., & Binder, J. R. (2010). Neural Systems for Reading Aloud: A Multiparametric Approach. *Cerebral Cortex, 20(8),* 1799–1815.

Hanson, G. K., & Chrysikou, E. G. (2017). Attention to Distinct Goal-relevant Features Differentially Guides Semantic Knowledge Retrieval. *Journal of Cognitive Neuroscience*.

Hargreaves, I. S., White M., Pexman P. M., Pittman D., Goodyear B. G. (2012). The question shapes the answer: the neural correlates of task differences reveal dynamic semantic processing. *Brain and Language, 120*, 73–78.

Harrison, S.A., & Tong F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature, 458*, 632–635.

Hindy, N.C., Solomon, S.H., Altmann G.T.M., & Thompson-Schill, SL. (2013). A Cortical Network for the Encoding of Object Change. *Cerebral Cortex.*

Hindy, N.C., Altmann G.T.M., Kalenik E., & Thompson-Schill S.L. (2012). The effect of object state-changes on event processing: do objects compete with themselves? *Journal of Neuroscience, 32,* 5795–5803.

Hirshorn, E. A., & Thompson-Schill, S. L. (2006). Role of the left inferior frontal gyrus in covert word retrieval: Neural correlates of switching during verbal fluency. *Neuropsychologia*, *44*(12), 2547–2557.

Hoenig, K., & Scheef, L. (2009). Neural correlates of semantic ambiguity processing during context verification. *NeuroImage*, *45*(3), 1009–1019.

Hoffman, P., Rogers, T.T., & Ralph, M.A.L. (2011). 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, 23(9),* 2432–2446.

Hoffman, P., Lambon Ralph, M.A., & Rogers, T.T. (2012). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods. 45*, 718-730.

Huettig, F., & Altmann, G. T. M. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Visual Cognition, 15*(8), 985–1018.

Hulbert, J.C. & Norman, K. A. (2015). Neural Differentiation Tracks Improved Recall of Competing Memories Following Interleaved Study and Retrieval Practice. *Cerebral Cortex, 25(10),* 3994-4008.

Hsu, N. S., Schlichting, M. L., & Thompson-Schill, S. L. (2014). Feature diagnosticity affects representations of novel and familiar objects. *Journal of cognitive neuroscience*.

Hsu, N. S., Kraemer, D. J., Oliver, R. T., Schlichting, M. L., & Thompson-Schill, S. L. (2011). Color, context, and cognitive style: Variations in color knowledge retrieval as a function of task and subject variables. *Journal of Cognitive Neuroscience, 23(9),* 2544–2557.

Ihara, A. S., Mimura, T., Soshi, T., Yorifuji, S., Hirata, M., Goto, T., … Fujimaki, N. (2015). Facilitated Lexical Ambiguity Processing by Transcranial Direct Current Stimulation over the Left Inferior Frontal Cortex. *Journal of Cognitive Neuroscience*, *27*(1), 26–34.

January, D., Trueswell, J. C., & Thompson-Schill, S. L. (2009). Co-localization of Stroop and syntactic ambiguity resolution in Broca's area: Implications for the neural basis of sentence processing. *Journal of Cognitive Neuroscience*, *21*(12), 2434–2444.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685.

Kan, I.P., & Thompson-Schill, S. L. (2004). Selection from perceptual and conceptual representations. *Cognitive, Affective, & Behavioral Neuroscience*, 4(4), 466–482.

Klein, D.E., & Murphy, G.L. (2001). The Representation of Polysemous Words. *Journal of Memory and Language. 45*, 259–282.

Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences, 11*, 229–235.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126-1141.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National academy of Sciences of the United States of America*, *103*(10), 3863-3868.

Kuhl, B. A., Bainbridge, W. A., & Chun, M. M. (2012). Neural Reactivation Reveals Mechanisms for Updating Memory. *Journal of Neuroscience*, *32*(10), 3453–3461.

Kuhl, B. A., Rissman, J., Chun, M. M., & Wagner, A. D. (2011). Fidelity of neural reactivation reveals competition between memories. *Proceedings of the National Academy of Sciences*, *108*(14), 5903–5908.

Lee, S-H., Kravitz, D.J., & Baker, C.I. (2012). Disentangling visual imagery and perception of real-world objects. *NeuroImage*, *59*, 4064–4073.

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review*, *7*(4), 618-630.

MacLeod, C. M., & MacDonald, P. A. (2000). Interdimensional interference in the Stroop effect: Uncovering the cognitive and neural anatomy of attention. *Trends in cognitive sciences*, *4*(10), 383-391.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*, 163.

Malone, P.S., Glezer, L.S., Kim, J., Jiang, X., & Riesenhuber, M. (2016). Multivariate Pattern Analysis Reveals Category-Related Organization of Semantic Representations in Anterior Temporal Cortex. *Journal of Neuroscience, 36(39)*, 10089-10096.

Martin, A. (2007). The representation of object concepts in the brain. *Annu. Rev. Psychol.*, *58*, 25-45.

Mason, R. A., & Just, M. A. (2007). Lexical ambiguity in sentence comprehension. *Brain Research*, *1146*, 115–127.

McRae, K., Cree, G. S., Westmacott, R., & De Sa, V. R. (1999). Further evidence for feature correlations in semantic memory. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *53*(4), 360.

McRae, K. & Jones, M. (2012). Semantic Memory. In D. Reisberg (Ed.), *Oxford Handbook of Cognitive Psychology*. New York, NY: Oxford University Press, Inc.

Mechelli, A., Price, C. J., Friston, K. J., & Ishai, A. (2004). Where bottom–up meets top–down: Neuronal interactions during perception and imagery. *Cerebral Cortex, 14*, 1256–1265.

Milham, M. P., Banich, M. T., Webb, A., Barad, V., Cohen, N. J., Wszalek, T., & Kramer, A. F. (2001). The relative involvement of anterior cingulate and prefrontal cortex in attentional control depends on nature of conflict. *Cognitive Brain Research*, *12*(3), 467–473.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*, 167–202.

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

Murphy, K., Bodurka, J., & Bandettini, P.A. (2007). How long to scan? The relationship between fMRI temporal signal to noise ratio and necessary scan duration. *NeuroImage*, *34,* 565-574.

Musz, E., & Thompson-Schill, S. L. (2015). Semantic Variability Predicts Neural Variability of Object Concepts. *Neuropsychologia, 76*, 41–51.

Musz, E., & Thompson-Schill, S. L. (2017). Tracking competition and cognitive control during language comprehension with multi-voxel pattern analysis. *Brain and Language, 165*, 21-32.

Noppeney U, Price CJ, Penny WD, Friston KJ. (2006). Two distinct neural mechanisms for category-selective responses, *Cerebral Cortex, 16*, 437–45.

Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective, & Behavioral Neuroscience, 5*(3), 263–281.

Pacht, J. M., & Rayner, K. (1993). The processing of homophonic homographs during reading: Evidence from eye movement studies. *Journal of Psycholinguistic Research, 22,* 252-271.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976–987.

Peelen, M. V., & Caramazza, A. (2012). Conceptual Object Representations in Human Anterior Temporal Cortex. *Journal of Neuroscience*, *32*(45), 15728–15736.

Peelen, M. V., He, C., Han, Z., Caramazza, A., & Bi, Y. (2014). Nonvisual and visual object shape representations in occipitotemporal cortex: evidence from congenitally blind and sighted adults. *Journal of Neuroscience*, *34*(1), 163-170.

Pereira, F., Botvinick, M., & Detre, G. (2011). Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial Intelligence, 194*, 240–252.

Persichetti, A. S., Aguirre, G. K., & Thompson-Schill, S. L. (2015). Value is in the eye of the beholder: early visual cortex codes monetary value of objects during a diverted attention task. *Journal of cognitive neuroscience*.

Plaut, D. C. (1996). Relearning after damage in connectionist networks: Toward a theory of rehabilitation. *Brain and language*, *52*(1), 25-82.

Pobric, G., Jefferies, E., & Lambon Ralph, M. A. (2010). Category-Specific versus Category-General Semantic Impairment Induced by Transcranial Magnetic Stimulation. *Current Biology*, *20*(10), 964–968.

Polyn, S., Norman, K.A., & Kahana, M.J. (2008). A context maintenance and retrieval model of organization processes in free recall. *Psychological Review, 116*, 129-156.

Ralph, L.M.A, Hoffman, P. G., & Jefferies, E. (2009). Conceptual knowledge is underpinned by the temporal pole bilaterally: Convergent evidence from rTMS. *Cerebral Cortex, 19*(4), 832-838.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372–422

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*(3), 191–201.

Rayner, K., & Frazier, L. (1989). Selection mechanisms in reading lexically ambiguous words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(5), 779.

Rayner, K., Pacht, J.M., & Duffy, S.A. (1994). Effects of prior encounter and global discourse bias on the processing of lexically ambiguous words: Evidence from eye fixations. *Journal of Memory and Language, 33*, 527-544.

Reddy, L., Kanwisher, N. G., & VanRullen, R. (2009). Attention and biased competition in multi-voxel object representations. *Proceedings of the National Academy of Sciences*, *106*(50), 21447–21452.

Rice, G.E., Watson, D.M., Hartley, T., Andrews, T.J. (2014). Low-level image properties of visual objects predict patterns of neural response across category-selective regions of the ventral visual pathway. *Journal of Neuroscience, 26*, 8837–8844.

Ritchey, M., Wing, E. A., LaBar, K. S., & Cabeza, R. (2013). Neural Similarity Between Encoding and Retrieval is Related to Memory Via Hippocampal Interactions. *Cerebral Cortex*, *23*(12), 2818–2828.

Ritchie, J. B., Tovar, D. A., & Carlson, T. A. (2015). Emerging object representations in the visual system predict reaction times for categorization. *PLoS Comput Biol*, *11*(6), e1004316.

Robinson, G., Shallice, T., & Cipolotti, L. (2005). A failure of high level verbal response selection in progressive dynamic aphasia. *Cognitive Neuropsychology*, *22*(6), 661–694.

Rodd, J. M. (2004). The Neural Mechanisms of Speech Comprehension: fMRI studies of Semantic Ambiguity. *Cerebral Cortex, 15*(8), 1261–1269.

Rodd, J. M., Johnsrude, I. S., & Davis, M. H. (2012). Dissociating Frontotemporal Contributions to Semantic Ambiguity Resolution in Spoken Sentences. *Cerebral Cortex, 22*(8), 1761–1773.

Rodd, Jennifer M., Vitello, S., Woollams, A. M., & Adank, P. (2015). Localising semantic and syntactic processing in spoken and written language comprehension: An Activation Likelihood Estimation meta-analysis. *Brain and Language, 141*, 89–102.

Schurz, M., Sturm, D., Richlan, F., Kronbichler, M., Ladurner, G., & Wimmer, H. (2010). 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, 49(3),* 2649–2661.

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-Specific Delay Activity in Human Primary Visual Cortex. *Psychological Science, 20(2)*, 207–214.

Shimamura, A.P. (2000). The role of the prefrontal cortex in dynamic filtering. *Psychobiology, 28,* 207–218.

Shinkareva, S.V., Malave, V.L., Mason, R.A., Mitchell, T.M., & Just, M.A. (2011). Commonality of neural representations of words and pictures. *NeuroImage, 54,* 2418–2425.

Simmons, W.K., Reddish, M., Bellgowan, P.S.F., Martin, A. (2010). The selectivity and functional connectivity of the anterior temporal lobes. *Cerebral Cortex, 20,* 813-825.

Simpson, G.B. & Krueger, M.A. (1991). Selective access of homograph meanings in sentence context. *Journal of Memory and Language, 30,* 627-643.

Snijders, T. M., Vosse, T., Kempen, G., Van Berkum, J. J. A., Petersson, K. M., & Hagoort, P. (2009). Retrieval and Unification of Syntactic Structure in Sentence Comprehension: an fMRI Study Using Word-Category Ambiguity. *Cerebral Cortex, 19*(7), 1493–1503.

Steyvers, M., & Malmberg, K.J. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 760–766.

Stowe, L. A., Paans, A. M. ., Wijers, A. A., & Zwarts, F. (2004). Activations of "motor" and other non-language structures during sentence comprehension. *Brain and Language, 89*(2), 290–299.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643.

Swaab, T., Brown, C., & Hagoort, P. (2003). Understanding words in sentence contexts: The time course of ambiguity resolution. *Brain and Language*, *86*(2), 326–343.

Swine, D.A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18*, 645-659.

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

Thompson-Schill, S.L., D'Esposito, M., Aguirre, G.K., & Farah, M.J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. *Proceedings of the National Academy of Science, 94*, 14792–14797.

Thompson-Schill, S. L., Kurtz, K. J., & Gabrieli, J. D. (1998). Effects of semantic and associative relatedness on automatic priming. *Journal of Memory and Language*, *38*(4), 440-458.

Thompson-Schill, S.L., D'Esposito, M., & Kan, I.P. (1999). Effects of repetition and competition on activity in left prefrontal cortex during word generation. *Neuron, 23*, 513–522.

Thompson-Schill, S. L., Bedny, M., & Goldberg, R. F. (2005). The frontal lobes and the regulation of mental activity. *Current Opinion in Neurobiology*, *15*(2), 219–224.

Twilley, L. C., & Dixon, P. (2000). Meaning resolution processes for words: A parallel independent model. *Psychonomic Bulletin & Review*, *7*(1), 49–82.

Twilley, L. C., Dixon, P., Taylor, D., & Clark, K. (1994). University of Alberta norms of relative meaning frequency for 566 homographs. *Memory & Cognition*, *22*(1), 111–126.

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

Visser, M., Jefferies, E., & Ralph, M. L. (2010). Semantic processing in the anterior temporal lobes: a meta-analysis of the functional neuroimaging literature. *Journal of Cognitive Neuroscience*, *22*(6), 1083–1094.

Visser, M., & Ralph, M. L. (2011). Differential contributions of bilateral ventral anterior temporal lobe and left anterior superior temporal gyrus to semantic processes. *Journal of Cognitive Neuroscience*, *23*(10), 3121–3131.

Vitello, S., Warren, J. E., Devlin, J. T., & Rodd, J. M. (2014). Roles of frontal and temporal regions in reinterpreting semantically ambiguous sentences. *Frontiers in Human Neuroscience*, *8*.

Whitney, C., Jefferies, E., & Kircher, T. (2011). Heterogeneity of the Left Temporal Lobe in Semantic Representation and Control: Priming Multiple versus Single Meanings of Ambiguous Words. *Cerebral Cortex*, *21*(4), 831–844.

Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature Neuroscience*, *18*(4), 582–589.

Wing, E. A., Ritchey, M., & Cabeza, R. (2015). Reinstatement of Individual Past Events Revealed by the Similarity of Distributed Activation Patterns during Encoding and Retrieval. *Journal of Cognitive Neuroscience*, *27*(4), 679–691.

Worsley, K.J. & Friston, K.J. (1995). Analysis of fMRI Time-Series Revisited— Again. *Neuroimage, 2,* 173-181.

Zaitchik, D., & Solomon, G. (2008a). Animist thinking in the elderly and in patients with Alzheimer's disease. *Cognitive Neuropsychology*, *25*, 27–37.

Zaitchik, D., & Solomon, G. (2008b). Inhibitory mechanisms and impairment in domain-specific reasoning: Studies of healthy elderly adults and patients with Alzheimer's disease. *Proceedings of the Cognitive Science Society, 30*, 141.

Zaitchik, D., Iqbal, Y., & Carey, S. (2014). The Effect of Executive Function on Biological Reasoning in Young Children: An Individual Differences Study. *Child Development*, *85*(1), 160–175.

Zannino, G. D., Buccione, I., Perri, R., Macaluso, E., Gerfo, E. L., Caltagirone, C., & Carlesimo, G. A. (2010). Visual and semantic processing of living things and artifacts: an fMRI study. *Journal of cognitive neuroscience*, *22*(3), 554-570.

Zempleni, M.-Z., Renken, R., Hoeks, J. C. J., Hoogduin, J. M., & Stowe, L. A. (2007). Semantic ambiguity processing in sentence context: Evidence from event-related fMRI. *NeuroImage*, *34*(3), 1270–1279.

APPENDIX A

Chapter 3 stimulus words, sorted by domain, category, and typicality scores

Item Name	Domain	Category	Typicality	Raw Typicality	Z-scored Typicality	Raw Activity	Z-scored Activity
robot	nonliving	artifact	atypical	57.4	-2.74	42.2	-1.11
rocket	nonliving	artifact	atypical	64.4	-1.80	24.5	-1.97
gondola	nonliving	artifact	atypical	67.6	-1.38	61.6	-0.17
windmill	nonliving	artifact	atypical	68.2	-1.31	41.4	-1.15
blimp	nonliving	artifact	atypical	69.0	-1.20	53.2	-0.58
furnace	nonliving	artifact	atypical	71.0	-0.93	70.2	0.25
vacuum	nonliving	artifact	atypical	71.9	-0.81	55.6	-0.46
tuba	nonliving	artifact	atypical	72.2	-0.77	73.1	0.39
yacht	nonliving	artifact	atypical	72.8	-0.69	49.8	-0.74
limo	nonliving	artifact	atypical	73.0	-0.66	48.8	-0.79
trolley	nonliving	artifact	atypical	73.2	-0.64	40.1	-1.21
motorcycle	nonliving	artifact	atypical	74.0	-0.53	33.0	-1.56
whisk	nonliving	artifact	atypical	74.1	-0.51	69.2	0.20
tractor	nonliving	artifact	atypical	74.3	-0.49	47.6	-0.85
buggy	nonliving	artifact	atypical	74.6	-0.46	54.9	-0.49
spear	nonliving	artifact	atypical	75.2	-0.38	78.5	0.65
jeep	nonliving	artifact	atypical	75.7	-0.31	38.4	-1.29
canoe	nonliving	artifact	atypical	76.3	-0.23	60.9	-0.21
scalpel	nonliving	artifact	atypical	76.9	-0.16	77.5	0.60
faucet	nonliving	artifact	atypical	77.0	-0.14	66.6	0.07
sprinkler	nonliving	artifact	typical	77.2	-0.12	45.4	-0.95
sled	nonliving	artifact	typical	77.4	-0.08	63.8	-0.06
sailboat	nonliving	artifact	typical	77.6	-0.06	44.2	-1.02
bicycle	nonliving	artifact	typical	78.5	0.07	44.8	-0.98
chainsaw	nonliving	artifact	typical	79.9	0.24	45.1	-0.97
slipper	nonliving	artifact	typical	81.1	0.41	89.4	1.18
calculator	nonliving	artifact	typical	83.6	0.74	75.9	0.52
broom	nonliving	artifact	typical	83.8	0.76	76.4	0.55
blender	nonliving	artifact	typical	84.3	0.83	45.8	-0.94
bench	nonliving	artifact	typical	85.6	1.00	95.0	1.45

napkin	nonliving	artifact	typical	86.4	1.11	89.4	1.18
apron	nonliving	artifact	typical	86.4	1.11	88.7	1.14
pencil	nonliving	artifact	typical	86.6	1.14	89.3	1.17
mitten	nonliving	artifact	typical	87.1	1.20	95.1	1.45
wrench	nonliving	artifact	typical	87.3	1.23	76.8	0.57
vase	nonliving	artifact	typical	87.3	1.23	95.7	1.48
cabinet	nonliving	artifact	typical	87.8	1.30	96.1	1.50
fork	nonliving	artifact	typical	87.9	1.31	90.8	1.24
shovel	nonliving	artifact	typical	88.1	1.34	85.2	0.97
comb	nonliving	artifact	typical	88.5	1.39	85.4	0.98
barnacle	living	living things	atypical	25.4	-1.62	15.4	-1.05
plankton	living	living things	atypical	33.6	-1.22	41.9	-0.19
seaweed	living	living things	atypical	34.5	-1.18	18.9	-0.94
grass	living	living things	atypical	36.0	-1.10	14.4	-1.08
coral	living	living things	atypical	37.7	-1.02	18.5	-0.95
clover	living	living things	atypical	38.3	-0.99	15.4	-1.05
bush	living	living things	atypical	38.5	-0.98	14.1	-1.09
vine	living	living things	atypical	38.5	-0.98	15.6	-1.04
cactus	living	living things	atypical	39.6	-0.93	9.8	-1.23
elm	living	living things	atypical	40.1	-0.90	12.4	-1.15
lily	living	living things	atypical	40.3	-0.89	20.4	-0.89
rose	living	living things	atypical	41.2	-0.85	17.0	-1.00
sycamore	living	living things	atypical	41.3	-0.84	11.6	-1.18
ivy	living	living things	atypical	41.7	-0.82	18.9	-0.94

lilac	living	living things	atypical	42.0	-0.81	10.9	-1.20
daisy	living	living things	atypical	42.2	-0.80	17.1	-1.00
sunflower	living	living things	atypical	43.2	-0.75	22.2	-0.83
orchid	living	living things	atypical	44.0	-0.71	18.2	-0.96
willow	living	living things	atypical	45.6	-0.63	15.0	-1.07
starfish	living	living things	atypical	55.7	-0.14	35.2	-0.41
ant	living	living things	typical	58.4	-0.01	78.4	0.99
wasp	living	living things	typical	60.8	0.10	86.7	1.26
moth	living	living things	typical	61.2	0.12	76.8	0.94
scorpion	living	living things	typical	62.1	0.17	72.3	0.80
marlin	living	living things	typical	70.6	0.58	82.1	1.11
cobra	living	living things	typical	71.6	0.63	72.0	0.79
orca	living	living things	typical	75.8	0.83	78.1	0.98
flamingo	living	living things	typical	76.3	0.86	68.8	0.68
hen	living	living things	typical	78.5	0.96	70.4	0.73
crow	living	living things	typical	80.2	1.05	79.0	1.01
hyena	living	living things	typical	80.7	1.07	81.0	1.08
whale	living	living things	typical	81.0	1.09	71.9	0.78

panda	living	living things	typical	83.7	1.22	69.9	0.72
kangaroo	living	living things	typical	84.7	1.26	84.8	1.20
raccoon	living	living things	typical	84.8	1.27	78.0	0.98
camel	living	living things	typical	85.0	1.28	69.1	0.69
dolphin	living	living things	typical	85.4	1.30	87.6	1.29
rhinoceros	living	living things	typical	85.9	1.32	76.1	0.92
chimpanzee	living	living things	typical	89.6	1.50	84.2	1.18
gorilla	living	living things	typical	90.4	1.54	82.0	1.11
thunder	nonliving	natural kinds	atypical	27.3	-1.30	46.8	-0.46
sun	nonliving	natural kinds	atypical	29.6	-1.13	32.3	-0.95
lava	nonliving	natural kinds	atypical	29.7	-1.13	37.9	-0.76
tsunami	nonliving	natural kinds	atypical	30.4	-1.07	11.4	-1.64
volcano	nonliving	natural kinds	atypical	30.9	-1.04	46.1	-0.49
planet	nonliving	natural kinds	atypical	31.5	-0.99	43.5	-0.57
lightning	nonliving	natural kinds	atypical	31.6	-0.99	22.1	-1.28
blizzard	nonliving	natural kinds	atypical	31.6	-0.99	25.0	-1.19
waterfal	nonliving	natural kinds	atypical	33.7	-0.82	19.0	-1.39
mist	nonliving	natural kinds	atypical	34.4	-0.77	60.3	-0.01
tornado	nonliving	natural kinds	atypical	34.5	-0.77	11.2	-1.65

asteroid	nonliving	natural kinds	atypical	34.6	-0.76	27.2	-1.12
lagoon	nonliving	natural kinds	atypical	34.6	-0.76	77.7	0.56
river	nonliving	natural kinds	atypical	35.6	-0.68	26.0	-1.16
cloud	nonliving	natural kinds	atypical	35.7	-0.67	60.3	-0.01
geyser	nonliving	natural kinds	atypical	35.9	-0.66	34.8	-0.86
pond	nonliving	natural kinds	atypical	36.5	-0.61	81.9	0.70
meteor	nonliving	natural kinds	atypical	37.4	-0.55	27.1	-1.12
comet	nonliving	natural kinds	atypical	38.1	-0.50	19.1	-1.39
bonfire	nonliving	natural kinds	atypical	40.9	-0.29	36.3	-0.81
avalanche	nonliving	natural kinds	typical	41.9	-0.21	19.5	-1.37
iceberg	nonliving	natural kinds	typical	42.7	-0.15	73.9	0.44
canyon	nonliving	natural kinds	typical	43.6	-0.08	85.0	0.81
icicle	nonliving	natural kinds	typical	44.7	0.00	91.3	1.02
canal	nonliving	natural kinds	typical	45.1	0.03	71.1	0.35
bubble	nonliving	natural kinds	typical	45.3	0.04	67.6	0.23
puddle	nonliving	natural kinds	typical	47.7	0.23	84.5	0.79
gasoline	nonliving	natural kinds	typical	49.3	0.34	87.1	0.88
sand	nonliving	natural kinds	typical	51.8	0.53	89.5	0.96

ash	nonliving	natural kinds	typical	55.3	0.80	88.2	0.91
crater	nonliving	natural kinds	typical	57.5	0.96	91.2	1.01
seashell	nonliving	natural kinds	typical	60.4	1.18	90.9	1.00
ruby	nonliving	natural kinds	typical	63.4	1.40	95.9	1.17
emerald	nonliving	natural kinds	typical	63.4	1.40	92.9	1.07
coal	nonliving	natural kinds	typical	65.4	1.55	91.9	1.04
boulder	nonliving	natural kinds	typical	65.7	1.57	93.3	1.08
gravel	nonliving	natural kinds	typical	66.5	1.63	92.0	1.04
granite	nonliving	natural kinds	typical	66.7	1.65	92.4	1.05
diamond	nonliving	natural kinds	typical	67.8	1.73	92.5	1.06
pebble	nonliving	natural kinds	typical	70.0	1.89	93.2	1.08

APPENDIX B

Correlations between category-level typicality scores and psycholinguistic variables.

Category	Word Length	Word Frequency	Contextual Diversity
LT typicality	0.17	-0.11	-0.08
ART typicality	-0.06	0.06	0.16
NAT typicality	-0.06	-0.13	-0.16

Word frequency and contextual diversity tabulated from the SUBTLEX database, a corpus composed of 50 million words from spoken language subtitles and transcripts (cf. Brysbaert & New, 2009).

APPENDIX C

Supplementary group-level results: correlations between semantic variability and neural similarity for additional voxel set sizes

The main text of the paper reports results where neural similarity is sampled in the top 500 voxels throughout each subject's brain, where "top voxels" are ones that maximally respond to each individual stimulus event, versus baseline. We also performed the neural similarity by word type comparison (i.e., Comparison 1) and the SV-neural similarity correlation (i.e., Comparison 2) by measuring item-wise neural similarity at other voxel set sizes. Specifically, we computed these comparisons in each subject by calculating neural similarity in the top X voxels, where X was 10,000; 7,000; 5,000; 2,000; 1,000; 750; 500; 250; 100; 50; and 25 voxels. Neural similarity was consistently higher among single-sense target words than PH words when the top 100-750 voxels were selected. Additionally, correlations between semantic variability and neural similarity were reliably negative across subjects when the top 250-2000 voxels were selected. In addition to the 500-voxel results reported in the paper, the other significant results are reported below.

Voxels sampled	Comparison 1: Neural similarity by word type	Comparison 2: Item-wise SV and neural similarity relationship
100	$t(20)= 2.94$, $p< .01$	
250	$t(20)= 3.48$, $p< .01$	$t(20)= -2.28$, $p= .03$
750	$t(20)= 2.30$, $p= .03$	$t(20)= -2.21$, $p= .04$
1000		$t(20)= -2.13$, $p= .05$
2000		$t(20)= -2.01$, $p= .06$

Principle Components Analysis on Semantic Variability Measures from Table 1

We assessed the shared variance cross the seven variable ratings with principal-components analysis (PCA). This technique is useful for finding latent patterns in high-dimensional data. The PCA aided us in interpreting the shared variance underlying the variables (listed in Table 4.1). The resulting component scores are listed in the table below.

Table C2. Principal components analysis on variables listed in Table 1.

Component	Eigenvalue	Percentage of variance	Cumulative percentage
1	3.04	43	43
2	1.82	26	69
3	1.10	16	85
4	0.49	7	92
5	0.30	4	96
6	0.16	2	99
7	0.09	1	100

These resulting component scores reflect weighted combinations of the seven variables from Table 1. These scores can be compared to the original variables, to determine which original variables loaded most highly on the principal component. Squared-cosine, a measure of the similarity between a principal component's vector and a variable's vector in high-dimensional space, is one way to describe the loading strength. Higher squared-cosine values, particularly those above 1, indicate that a variable contributed to the principal component.

Squared-cosine values between each variable and the first principal component

Table C3. Squared-cosine values between the first principal component and each variable listed in Table 1.

Variable	Cos2
1	0.81
2	0.86
3	0.81
4	0.34
5	0.12
6	0.07
7	0.02

It is standard practice to retain all principal components with eigenvalues above 1. When we retain the top three eigenvalues and enter these three dimensions as regressors in a multiple regression model to predict the neural data, the model did not robustly explain the variance in item-wise neural similarity across subjects. Additionally, none of the three individual regressors reliably predicted the neural data across subjects. However, when the first principal component alone was used to predict neural similarity in a single regression model, the regressor reliably predicted neural similarity across subjects at two voxel set sizes. These results are provided in the table below.

Table C4. Correlations between the first principal component and neural similarity at varying set sizes of whole-brain voxels

Voxels sampled	Group Results: first principal component and neural similarity correlation
250	$t(20)= -2.18$, $p= .04$
500	$t(20)= -2.26$, $p= .03$

Neural similarity predicted by individual SV variables from Table 4.1

The main text of the paper reports results where neural similarity is predicted by a composite measure of SV. This measure was developed by combining seven variables which measure semantic variability from a variety of methods and corpora (see Table 4.1). Many of these variables also individually predict the observed neural similarity. The table below reports the variables which individually correlated with neural similarity, at

varying set sizes of whole-brain voxels. Variables 5 and 6 did not individually predict neural similarity.

Table C5. Group results for correlation between individual SV variables and neural similarity

oxels Sampled	Variable 1	Variable 2	Variable 3	Variable 6	Variable 7
50	$t(20)= -2.26$, $p= .04$	$t(20)= -2.22$, $p= .04$		$t(20)= -2.29$, $p= .03$	
00		$t(20)= -2.11$, $p= .05$	$t(20)= 2.10$, $p= .05$	$t(20)= -3.31$, $p< .01$	$t(20)= -2.72$, $p= .01$
50				$t(20)= -3.10$, $p< .01$	$t(20)= -2.76$, $p= .01$
000				$t(20)= -3.40$, $p< .01$	$t(20)= -3.34$, $p< .01$
000				$t(20)= -3.40$, $p< .01$	$t(20)= -3.16$, $p< .01$
000				$t(20)= -2.72$, $p= .01$	$t(20)= -2.64$, $p= .02$
000				$t(20)= -2.42$, $p= .03$	$t(20)= -2.40$, $p= .03$

Correlations between Semantic Variability Variables and other Semantic Variables

While SV moderately correlates with several semantic variables, the individual SV variables do not all strongly correlate with the semantic variables listed in Table 2 of the main text. The table below lists each individual variable used to create SV, and its correlation with the semantic variables listed in Table 2 of the main text.

Table C6. Correlations between individual SV variables and various stimulus characteristics. Concreteness, Familiarity, and Imageability ratings 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. *p<.05

Stimulus Characteristic	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable 6	Variable 7
Concreteness	0.42*	0.44*	0.40*	0.10	.34	.12	-0.20
Familiarity	0.46*	0.49*	0.42*	0.16	.06	-0.02	.12
Imageability	-0.07	-0.13	-0.14	0.18	.14	-0.08	0.35*
Word length	0.49*	0.47*	0.53*	0.25	-0.12	.02	0.32*
Number of phonemes	0.44*	0.48*	0.50*	0.21	-0.10	.01	0.33*
Number of synonyms	0.46*	0.46*	0.47*	0.20	-0.08	-0.06	0.45*
Word frequency	.72*	0.80*	0.75*	0.40*	-0.15	-0.07	.26