Quantitative Modeling of the Neural Representation of Nouns and Phrases

Kai-min Kevin Chang

CMU-LTI-11-006

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:

Marcel A. Just, Carnegie Mellon University
Tom M. Mitchell, Carnegie Mellon University
Charles Kemp, Carnegie Mellon University
Brian Murphy, University of Trento

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
In Language and Information Technologies

QUANTITATIVE MODELING OF THE NEURAL REPRESENTATION OF NOUNS AND PHRASES

QUANTITATIVE MODELING OF THE NEURAL REPRESENTATION OF NOUNS AND PHRASES

A thesis submitted in partial
fulfillment of the requirements for the degree of
Doctor of Philosophy

By

Kai-min Kevin Chang
Carnegie Mellon University, 2006
Master of Science in Computer Science

August, 2011
Carnegie Mellon University

ABSTRACT

Recent advances in brain imaging and machine learning technologies offer a significant new approach to studying language processing in humans. For the first time, theories regarding how linguistic concepts are processed can be directly validated and grounded by the patterns of brain activity while people comprehend words and phrases. In this dissertation, we used functional magnetic resonance imaging (fMRI) to study the cortical systems that underpin semantic processing of various linguistic concepts, including nouns of concrete object *(e.g. dog),* adjective-noun phrases *(e.g. strong dog)*, and noun-noun concept combinations *(e.g. corn coat).*

The thesis of this research is that the distributed pattern of brain activity encodes the meanings of linguistic concepts and an intermediate semantic representation can be used to model how brain represents and processes conceptual knowledge in terms of more primitive semantic features. Our effort in multivariate analysis shifts the focus of fMRI analysis from characterizing the location of brain activity (traditional univariate approaches) toward understanding how patterns of brain activity differentially encode information in a way that distinguishes among different stimuli. By postulating that the brain activity is based on an intermediate semantic level of representation and subsequently learning the correspondence between semantic features and observed brain activity, this work provides a neural account of some existing linguistic theories and furthermore enables a predictive theory that is capable of extrapolating the model of the brain activity to previously unseen words and phrases.

MAIN RESULTS

1. By postulating that the brain activity is based on an intermediate semantic level of representation (derived from word co-occurrence statistics or feature norming studies), this work enables a computational model that can help predict brain activity for a new stimulus, based on its relation to the semantic level of representation.
2. The difference in brain activity when contemplating an isolated noun *(e.g. dog)* vs. the same noun modified by an adjective *(e.g. strong dog)* can be detected by machine learning classifiers and modeled by vector-based semantic composition model to provide a neural account of how people use adjectives to modify the meaning of the noun.
3. The distributed pattern of brain activity contains sufficient signal to decode between a property-based interpretation *(e.g. a coat that is bright yellow)* and a relation-based interpretation *(e.g. a coat that is used to protect corn)* of the identical visual stimuli *(e.g. corn coat)*, and furthermore, provides a neural account of how relation-based interpretation are more accessible to humans.
4. Bayesian probabilistic analysis offers a new approach to characterize semantic representation by inferring the most likely feature structure directly from the patterns of brain activity. The neurally-inspired semantic representation is consistent with some existing conjectures regarding the role of different brain areas in processing different psycholinguistics features.

This thesis is approved for recommendation to the Graduate Council.

Thesis Directors:

Professor Marcel Adam Just
Center for Cognitive Brain Imaging, Carnegie Mellon University

Professor Tom Mitchell
Machine Learning Department, Carnegie Mellon University

Thesis Committee:

Professor Charles Kemp
Department of Psychology, Carnegie Mellon University

Dr. Brian Murphy
Centre for Mind/Brain Sciences, University of Trento

THESIS DUPLICATION RELEASE

I hereby authorize the Carnegie Mellon University Libraries to duplicate this thesis when needed for research and/or scholarship.

Agreed ____

Refused ____

ACKNOWLEDGEMENTS

I thank my thesis advisors, Dr. Marcel Just and Dr. Tom Mitchell, for their supervision. I have learned a great deal from their remarkable vision and determination for great research. I thank my thesis committee, Dr. Charles Kemp and Dr. Brian Murphy, for their guidance in helping me complete this thesis work. I also like to thank my M.S. advisors, Dr. Jack Mostow and Dr. Joseph Beck, for their wholeheartedness in the early development of my academic career.

I also thank the wonderful support from my wife, Yi-Chia Wang, who is also pursuing her Ph.D. We have grown together along the journey. Finally, I like to express my most sincere gratitude to my parents. Their decision to immigrate the family to Canada and dedication to children's education fundamentally shaped my life and character.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

We are at an especially opportune time in the history of the study of human cognition. Brain imaging technology allows us to directly observe and model brain activity associated with cognitive processes. Techniques from statistics and machine learning allow us to construct quantitative computational models that describe these cognitive brain processes. Furthermore, they allow us to construct mental state decoders that accurately predict certain aspects of thought from measured brain activity. In addition to the scientific impact of better understanding the representation and processing of human cognition, this research will lead to many applications and broad impacts. For example, a brain-computer interface (BCI) device that could decode internal speech may enable locked-in patients to communicate. Pattern classifiers can be used to identify processing abnormalities in autism or detect neuropsychiatric illnesses such as schizophrenia.

Computational neurolinguistics is an emerging research area which integrates recent advances in computational linguistics and cognitive neuroscience, with the objective of developing cognitively plausible models of language and gaining a better understanding of the human language system. It builds on research in decoding cognitive states from recordings of neural activity, and computational models of lexical representations and sentence processing. Advances in computational neurolinguistics require close collaboration between neuroscience, language technology, cognitive psychology, and machine learning. To this end, my thesis work helps advance existing work and initiates new research.

How humans represent meanings of individual words and how lexical semantic knowledge is combined to form concepts, phrases, or sentences are issues fundamental to the study of human language. Recent advances in functional magnetic resonance imaging (fMRI) provide a significant new approach to studying semantic representations in humans by making it possible to directly observe brain activity while people comprehend words and sentences. fMRI measures the hemodynamic response (changes in blood flow and blood oxygenation) related to brain activity in the human brain. Images can be acquired at good spatial resolution and reasonable temporal resolution – the activity level of 15,000 - 20,000 brain volume elements (voxels) of about 50 mm^3 each can be measured every second. Studies have shown that the fMRI signal is proportional to the local average neuronal activity, although the relationship is modulated by many factors (Heeger & Ress, 2002; Logothetis et al., 2001; Bandettini & Ungerleider, 2001).

Traditional analysis of fMRI data uses the General Linear Model (GLM) (Friston, 2005) to find voxels whose activation time-series reflect alternations between experimental conditions of interest (e.g. periods when a task is being performed versus a rest period). The GLM analysis is often referred to as mass univariate analysis since significance tests are typically performed at every voxel in the brain. Mass univariate analysis has identified the cortical activation associated with language processing to be strongly lateralized in the left cerebral hemisphere and involving a network of regions in the frontal, temporal, and parietal lobe. Bookheimer (2002), summarizing contemporary fMRI analyses, identified the role of the left inferior frontal lobe (Broca's area) in semantic processing, the role of the temporal robe (Wernicke's area) in organization of categories of objects and concepts, and the role of the right hemisphere in comprehending contextual and figurative meaning, although left temporoparietal activation outside the classical "Wernicke area" and left prefrontal activation outside the classical "Broca's area" were also reported (Binder et al., 1996). Most of the fMRI studies in language processing utilize mass-univariate approaches to contrast regions of brain areas in processing conditions like ambiguity

(Rodd, Davis, & Johnsrude, 2005), novelty (Saykin et al., 1999), syntax (Dapretto & Bookheimer, 1999), metaphor (Mashal et al., 2007), and sentence comprehension (Just et al., 1996).

Haxby et al. (2001) was one of the first studies to apply multivariate analysis to study the distributed patterns of fMRI activity. They showed a distinct pattern of response in ventral temporal cortex could be found while participants viewed faces and objects, and furthermore, that machine learning classifiers could be used to decode what stimuli the participants were viewing. Their results supported a distributed and overlapping representation of faces and objects. Since then, multivariate analyses of fMRI activity have shown that classifiers can be trained to decode which of several visually presented objects or object categories a person is contemplating, given the person’s fMRI-measured brain activity (Cox & Savoy, 2003; O'Toole et al., 2005; Haynes & Rees, 2006; Mitchell et al., 2004). Moreover, multivariate analyses of fMRI activity have shown that classifiers can be trained to decode the visual and subjective contents of the human brain (Kamitani & Tong, 2005), the orientation of invisible stimuli (Haynes & Rees, 2005), lie detection (Davatzikos, 2005), stream of consciousness (Haynes & Rees, 2005), speech content and speaker identity (Formisano et al., 2008).

1.1 Problem Statement

Given these succesess in multivariate analysis of fMRI activity, it is interesting to ask whether a similar approach can be used to study the representation of linguistic concepts like nouns and phrases. Thus, the first question we ask is:

1. Does the distribution of neural activity encode sufficient signal to decode linguistic concepts like nouns and phrases?

One research direction is to investigate the granularity that the distributed patterns of activity encodes. For instance, how many different semantic categories *(e.g. tools, dwellings)* or objects *(e.g. hammer, house)* can machine learning classifiers decode? Can classifiers discriminate the subtle difference of a noun *(e.g. dog)* from an adjective-noun phrase *(e.g. strong dog),* where the adjectives are expected to emphasize certain semantic properties of the nouns *(e.g. the physical attribute of the dog)*? Moreover, most of the current mental state decoding research focuses on stimuli that are rich in visual input or brain activities in early visual areas (Ishai et al., 2000; Cox & Savoy, 2003; Kay et al., 2008; Thirion et al., 2006; Harrison & Tong, 2009; Haynes & Rees, 2005). Do classifiers obtain their discriminative power from distinguishing the brain activity of low-level visual perceptions, or are they capable of decoding the higher-level characterization of semantic differences? Does the distributed pattern of brain activity contain sufficient signal to decode the differences in different interpretations *(e.g. a coat that is bright yellow* vs. *a coat that is used to protect corn*) of the same visual stimuli *(e.g. corn coat)*? Finally, what is the effect of different stimuli *(e.g. pictures vs. text-labels)* on classifier performance?

Despite the early success of mental state decoding research, discriminative classification provides a characterization of only a particular dataset, and does not reveal the underlying principles that would allow for generalization to other stimuli. One way to obtain this extensibility is to construct a model which postulates that the brain activity is based on an intermediate semantic level of representation. Then the model can predict the activation for a new stimulus, based on its relation to the semantic level of representation. In effect, this is how regression models typically generate predicted values. A regression model that successfully models the intermediate semantic factors underpinning object knowledge would have this generative capability. Thus, the second question we ask is:

2. Can intermediate semantic representation be used to model how the brain composes the meaning of words or phrases in terms of more primitive semantic features?

There have been a variety of approaches from different scientific communities trying to capture the intermediate semantic attributes and organization underlying object- and word-representation. Linguists have tried to characterize the meaning of a word with feature-based approaches, such as semantic roles (Kipper, Dang, & Palmer, 2000), as well as word-relation approaches, such as WordNet (Miller, 1995). Computational linguists have demonstrated that a word's meaning is captured to some extent by the distribution of words and phrases with which it commonly co-occurs (Church & Hanks, 1990). Psychologists have studied word meaning in many ways, one of which is through norming studies (Cree & McRae, 2003) in which human participants are asked to list the features they associate with various words. There are also approaches that treat the intermediate semantic representation as hidden (or latent) variables and use techniques like the traditional principle component analysis (PCA) and factor analysis, or the more recent hyperspace analogue to language model (HAL; Lund & Burgess, 1996), latent semantic analysis (LSA: Landauer & Dumais, 1997) and topic models (Blei, Ng, & Jordan, 2003) to recover these latent structures from text corpora. Kemp et al. (2007) have also presented a Bayesian model of inductive reasoning that incorporates both knowledge about relationships between objects, and knowledge about relationships between object properties. The model is useful to infer some properties of previously unseen stimuli, based on the learned relationships between objects. Finally, connectionists have long employed *hidden layers* in their neural networks to mediate non-linear correspondences between input and output. Hanson, Matsuka, & Haxby (2004) proposed a neural network classifier with hidden units to account for brain activation patterns, but the learned hidden units are difficult to interpret in terms of an intermediate semantic representation. For the cognitive scientists and linguists, the primary question is “how" and "why" the patterns of neural activity encode the meaning of words or concepts. In this study, we first use semantic representation derived from norming studies or corpus statistics that enable cognitive interpretation. We then use an infinite latent feature model (ILFM) with an Indian Buffet Process (IBP) (Griffiths & Ghahramani, 2005) to derive a semantic representation directly from brain activity and further show that such data-driven semantic representation is consistent with human ratings of the words.

Notice that in this work we distinguish between *semantic representation* and *semantic processing*. Whereas the former describes how the meaning of a concept is represented (e.g. is the semantic content of a word atomic or compounded?), the latter describes how processing of a concept is distributed spatially (e.g. does processing of a concept involve brain activity localized in a few voxels or distributed across a number of brain regions?). Both representation and processing can either be localized or distributed – the two need not be mutually exclusive. To foreshadow our results, we have shown distributed accounts for both semantic representation and semantic processing.

1.2 Thesis Statement

The thesis of this research is that the distributed pattern of brain activity encodes the meanings of linguistic concepts and an intermediate semantic representation can be used to model how brain represents and processes conceptual knowledge in terms of more primitive semantic features. Our goal is to build a computational model of the brain activity when people contemplate nouns and phrases. More specifically, machine learning classifiers can be trained to decode which linguistic concepts a person is contemplating. By postulating that the brain activity is based on an intermediate semantic level of representation, and subsequently learning the correspondence between semantic features and observed brain activity, this work provides a neural grounding of some existing linguistic theories and furthermore enables a predictive theory that is capable of extrapolating the model of the brain activity to previously unseen words and phrases.

1.3 Approach

To answer the questions proposed in this thesis, we designed a series of brain imaging experiments. In an object-contemplation task, participants were presented with line drawings and/or text labels of objects and were instructed to think of the same properties of the stimulus object consistently during multiple presentations of each item. fMRI recorded brain activation while people contemplated various linguistic concepts, including concrete objects *(e.g. dog)*, adjective-noun phrases *(e.g. strong dog)*, and noun-noun concept combinations *(e.g. corn coat)*. In this section, we will discuss the general experimental paradigm and modeling methodology used throughout our brain imaging studies.

1.3.1 Brain Imaging Experiments

In an object-contemplation task, participants were presented with line drawings and/or text labels of objects and were instructed to think about the same properties of the stimulus object consistently during six presentations of each item. To ensure that participants had a consistent set of properties to think about, they were each asked to generate and write a set of properties for each exemplar in a session prior to the scanning session (such as “4 legs, house pet, fed by me” for dog), however, nothing was done to elicit consistency across participants. Each item was presented six times during the scanning session, in a different random order each time. Participants silently viewed the stimuli and were asked to think about the same item properties consistently across the six presentations of the items.

Each stimulus was presented for 3s, followed by a 7s rest period, during which the participants were instructed to fixate on an X displayed in the center of the screen. There were two additional presentations of fixation, 30s each, at the beginning and end of each session, to provide a baseline measure of activity. A schematic representation of the design used in the 60 concrete objects experiment is shown in Figure 1.1.

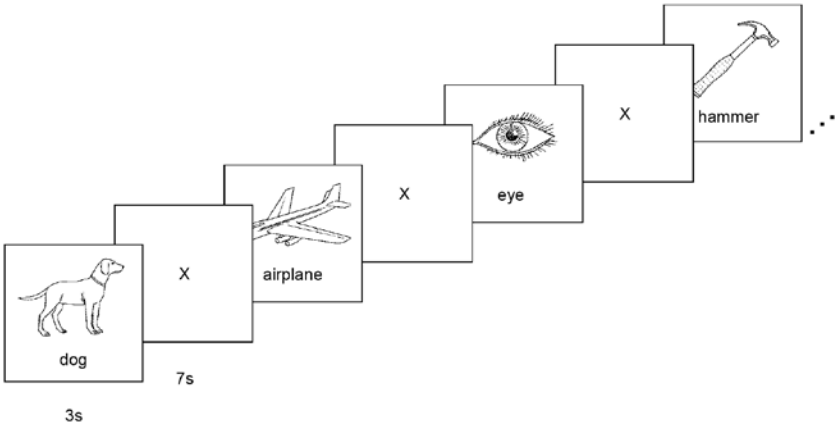


Figure 1.1 Schematic representation of the experimental design for the object-contemplating brain imaging experiment.

1.3.2 Data Acquisition and Processing

Functional images were acquired on a Siemens Allegra 3.0T scanner (Siemens, Erlangen, Germany) at the Brain Imaging Research Center of Carnegie Mellon University and the University of Pittsburgh using a gradient echo EPI pulse sequence with TR = 1000 ms, TE = 30 ms, and a 60° flip angle. Seventeen 5-mm thick oblique-axial slices were imaged with a gap of 1-mm between slices. The acquisition matrix was 64 x 64 with 3.125 x 3.125 x 5-mm voxels. The parameters have been chosen to accentuate the spatial (as opposed to temporal) distribution of neural activity. Data processing were performed with Statistical Parametric Mapping software (SPM2, Wellcome Department of Cognitive Neurology, London, UK; Friston, 2005). The data were corrected for slice timing, motion, and linear trend, and were temporally smoothed with a high-pass filter using a 190s cutoff. The data were normalized to the MNI template brain image using a 12-parameter affine transformation and resampled to 3 x 3 x 6-mm3 voxels.

The percent signal change (PSC) relative to the fixation condition was computed for each item presentation at each voxel. The mean of the four images (mean PSC) acquired within a 4s window, offset 4s from the stimulus onset (to account for the delay in hemodynamic response), provided the main input measure for subsequent analysis. The mean PSC data for each word presentation were further normalized to have mean zero and variance one to equate the variation between participants over exemplars. Due to the inherent limitations in the temporal properties of fMRI data, we consider only the spatial distribution of the neural activity after the stimuli are comprehended, and do not attempt to model the cognitive process of comprehension.

1.3.3 Decoding mental states

To find out if the distribution of neural activity encode sufficient signal to decode linguistic concepts like nouns and phrases, classifiers were trained to identify cognitive states associated with viewing stimuli from the evoked pattern of functional activity (mean PSC). Classifiers were functions f of the form: f: mean_PSC → Yi, i=1,…n, where Yi were the sixty exemplars, and mean_PSC was a vector of mean PSC voxel activation level, as described above.

Since fMRI acquires the neural activity at 15,000 – 20,000 distinct voxel locations, many of which might not exhibit neural activity that encodes word or phrase meaning, the classifier analysis selected the voxels whose responses to the different items were most stable across presentations. Voxel stability was computed as the average pair-wise correlation between all stimuli across presentations, using only the training set within each fold in the cross-validation paradigm. The focus on the most stable voxels effectively increased the signal-to-noise ratio in the data and facilitated further analysis by classifiers.

To evaluate classification performance, data were divided into training and test sets. A classifier was built from the training set and evaluated on the left-out test set. Classification results were evaluated using six-fold cross validation, where one of the six repetitions was left out for each fold. The voxel selection procedure was performed separately inside each fold, using only the training data. Since multiple classes were involved, rank accuracy was used (Mitchell et al., 2004) to evaluate the classifier. Given a new fMRI image to classify, the classifier outputs a rank-ordered list of possible class labels from most to least likely. The rank accuracy is defined as the percentile rank of the correct class in this ordered output list. Rank accuracy ranges from 0 to 1. Classification analysis was performed separately for each participant, and the mean rank accuracy was computed over the participants.

1.3.4 Modeling intermediate semantics

To find out if models of semantic representation can be used to model how the brain composes the meaning of words or phrases in terms of more primitive semantic features, regression analysis was performed to explain the systematic variances in neural activity with semantic features. There are two steps in this modeling framework. First, we represent word meaning with a vector of primitive features. Then, by learning the mapping between feature and neural activation, the generative model is capable of predicting neural activity for previously unseen words. For multi-words phrases, there is an additional step that models the semantic composition rule that governs how words are combined to form phrases. Figure 1.2 depicts the modeling framework for multi-words phrases. In this work, we use semantic feature representations derived from norming studies and corpus statistics. The details of semantic representations and semantic composition rules used in the three experiments are reported in the Method sections of the respective chapters. In the following section, we will discuss the regression model that is used in all three experiments to learn the mapping between feature and neural activation.

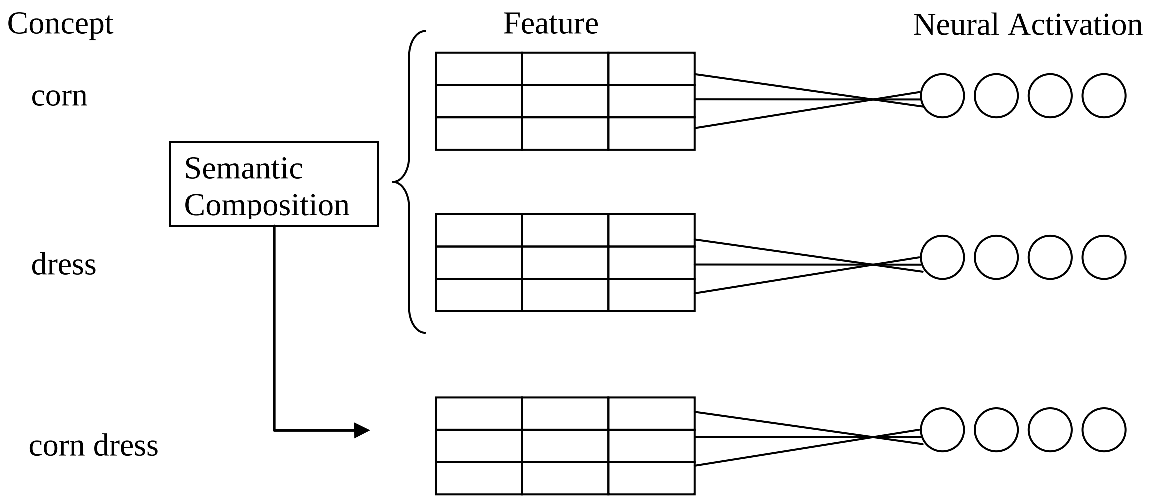


Figure 1.2 Modeling framework of the intermediate semantic representation. Each concept *(e.g. corn, dress)* is represented with a vector of features. By learning the mapping between feature and neural activation, the generative model is capable of predicting neural activity for previously unseen words. For multi-words phrases *(e.g. corn dress)* , there is an additional step that models the semantic composition rule that governs how words are combined to form phrases.

1.3.5 Learn feature-voxel mapping with regression models

In order for the generative model to make predictions for neural activity, we learn the feature-voxel mapping by training a regression model to fit the activation profile for the stimuli. The regression model examined to what extent the semantic feature vectors (explanatory variables) can account for the variation in neural activity (response variable) across the different stimuli. All explanatory variables were entered into the regression model simultaneously. More precisely, the predicted activity a_v at voxel v in the brain for word w is given by

$$a_v = \sum_{i=1}^{n} \beta_{vi} f_i(w) + \varepsilon_v$$

where $f_i(w)$ is the value of the i^{th} intermediate semantic feature for word w, β_{vi} is the regression coefficient that specifies the degree to which the i^{th} intermediate semantic feature activates voxel v, and ε_v is the model's error term that represents the unexplained variation in the response variable. Least squares estimates of β_{vi} were obtained to minimize the sum of squared errors in reconstructing the training fMRI images. An L2 regularization with lambda = 1.0 was added to prevent overfitting given the high parameter-to-data-points ratios. A regression model was trained for each of the 120 voxels and the reported R^2 is the average across the 120 voxels. R^2 measures the amount of systematic variance explained by the model. Regression results were evaluated using six-fold cross validation, where one of the six repetitions was left out for each fold.

Linear regression assumes a linear dependency among the variables and compares the variance due to the independent variables against the variance due to the residual errors. While the linearity assumption may be overly simplistic, it reflects the assumption that fMRI activity often reflects a superposition of contributions from different sources, and has provided a useful first order approximation in the field (Mitchell et al., 2008). Neural networks may be used to learn non-linear correspondences between semantic features and neural activity. However, the high parameter-to-data-points ratios will make non-linear methods more prone to overfitting. Thus, the choice of linear methods over non-linear methods is prompted by the amount of data points and not that one is more realistic than the other.

2 QUANTITATIVE MODELING OF THE NEURAL REPRESENTATIONS OF OBJECTS: HOW SEMANTIC FEATURE NORMS CAN ACCOUNT FOR FMRI ACTIVATION

2.1 Introduction

Recent multivariate analyses of fMRI activities have shown that discriminative classifiers, such as Support Vector Machines (SVM), are capable of decoding mental states associated with the visual presentation of categories of various objects, given the corresponding neural activity signature (Cox & Savoy, 2003; O'Toole et al., 2005; Norman et al., 2006; Haynes & Rees, 2006; Mitchell et al., 2004; Shinkareva et al., 2008). This shifts the focus of brain activation analysis from characterizing the location of neural activity (traditional univariate approaches) toward understanding how patterns of neural activity differentially encode information in a way that distinguishes among different stimuli. However, discriminative classification provides a characterization of only a particular set of training stimuli, and does not reveal the underlying principles that would allow for extensibility to other stimuli. One way to obtain this extensibility is to construct a model which postulates that the brain activity is based on a hidden intermediate semantic level of representation. Here we develop and study a model that achieves this extensibility through its ability to predict the activation for a new stimulus, based on its relation to the semantic level of representation.

In the present work, functional Magnetic Resonance Imaging (fMRI) data is used to study the hidden factors that underpin the semantic representation of object knowledge. In an object-contemplation task, participants were presented with 60 line drawings of objects with text labels and were instructed to think of the same properties of the stimulus object consistently during each presentation. Given the neural activity signatures evoked by this visual presentation, a multivariate multiple linear regression model is estimated, which explains a significant portion of systematic variance in the observed neural activities. In terms of semantic attributes of the stimulus objects, our previous work (Mitchell et al., 2008) showed that semantic features computed from the occurrences of stimulus words within a trillion-token text corpus that captures the typical use of words in English text can predict brain activity associated with the meaning of these words. The advantage of using word co-occurrence data is that semantic features can be computed for any word in the corpus – effectively any word in existence. Nonetheless, these semantic features were assessed implicitly through word usage and may not capture what people retrieve when explicitly recalling features of a word. Moreover, despite the success of this model, which uses co-occurrences with 25 sensorimotor verbs as the feature set, it is hard to determine the optimal set of features. In this paper, we draw our attention to the intermediate semantic knowledge representation and experiment with semantic features motivated by other scientific communities.

Here we model the intermediate semantic knowledge with features from an independently performed feature norming study (Cree & McRae, 2003), where participants were explicitly asked to list features of 541 words. Our results suggest that 1) object features derived from a behavioral feature norming study can explain a significant portion of the systematic variance in the neural activity observed in our object-contemplation task. Moreover, we demonstrate how a generative classifier[a] that includes an intermediate semantic representation 2)

[a] We use the term *generative classifier* to refer to a classifier that bases its prediction on a generative theory through some intermediate semantic representation. It is not the same as the typical usage of a generative model in Bayesian community, although one can adopt a fully Bayesian approach that models the intermediate semantic representation as latent variables.

generalizes better across participants, compared to a discriminative classifier that does not utilize such an intermediate semantic representation, and 3) enables a predictive theory that is capable of predicting fMRI neural activity well enough that it can successfully match words it has not yet encountered to their previously unseen fMRI images with accuracies far above chance levels, which simply cannot be done with a discriminative classifier.

2.2 Material and Methods

Participants. Nine right-handed adults (5 female, age between 18 and 32) from the Carnegie Mellon community participated and gave informed consent approved by the University of Pittsburgh and Carnegie Mellon Institutional Review Boards. Two additional participants were excluded from the analysis due to head motion greater than 2.5 mm.

Experimental paradigm. The stimuli were line drawings and noun labels of 60 concrete objects from 12 semantic categories with 5 exemplars per category. Most of the line drawings were taken or adapted from the Snodgrass and Vanderwart (1980) set and others were added using a similar drawing style. Table 2.1 lists the 60 stimuli. Stimuli that were not part of Cree and McRae’s (2003) feature norming study (discussed later) are marked with asterisks.

Table 2.1 List of 60 Words. Stimuli that are not part of Cree and McRae's (2003) feature norming study are marked with asterisks.

Categories	**Exemplars**
Animal	Bear, cat, cow, dog, horse
Body part	Arm*, eye*, foot*, hand*, leg*
Building	Apartment, barn, church, house, igloo*
Building part	Arch*, chimney*, closet, door, window*
Clothing	Coat, dress, pants, shirt, skirt
Furniture	Bed, chair, desk, dresser, table
Insect	Ant, bee*, beetle, butterfly, fly*
Kitchen	Bottle, cup, glass*, knife, spoon
Man-made objects	Bell*, key, refrigerator*, telephone, watch*
Tool	Chisel, hammer, pliers, saw*, screwdriver
Vegetable	Carrot, celery, corn, lettuce, tomato
Vehicle	Airplane, bicycle*, car, train, truck

To ensure that each participant had a consistent set of properties to think about, they were asked to generate and write a set of properties for each exemplar in a separate session prior to the scanning session (such as *cold*, *knights*, *stone* for *castle*). However, nothing was done to elicit consistency across participants.

The entire set of 60 stimuli was presented 6 times during the scanning session, in a different random order each time. Participants silently viewed the stimuli and were asked to think of the same item properties consistently across the 6 presentations. Each stimulus was presented for 3s, followed by a 7s rest period, during which the participants were instructed to fixate on an X displayed in the center of the screen. There were two additional presentations of the fixation, 31s each, at the beginning and at the end of each session, to provide a baseline measure of activity. A schematic representation of the design is shown in Figure 2.1.

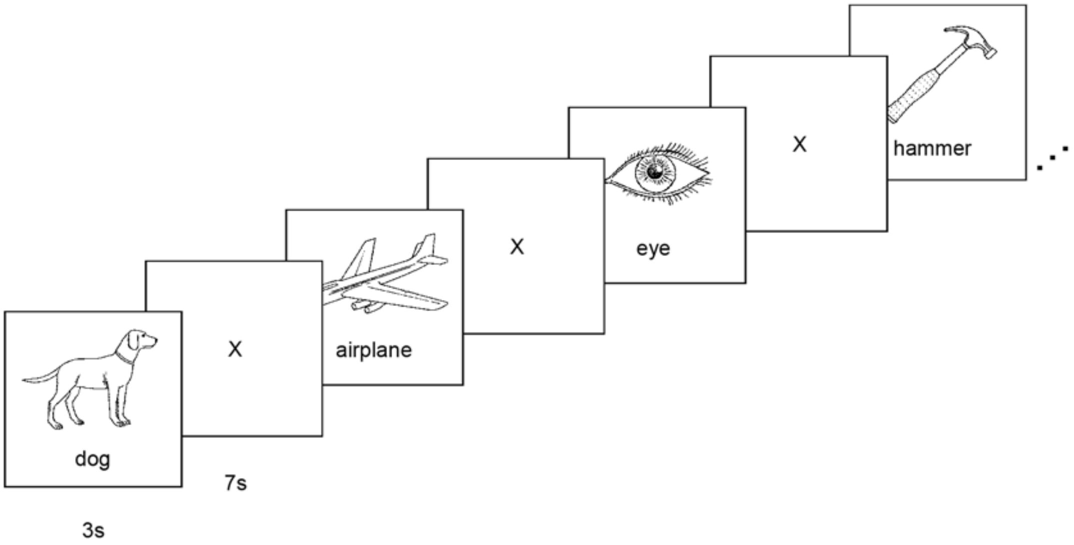


Figure 2.1 Schematic representation of experimental design for the 60 word experiment.

Data acquisition. Functional images were acquired on a Siemens Allegra 3.0T scanner (Siemens, Erlangen, Germany) at the Brain Imaging Research Center of Carnegie Mellon University and the University of Pittsburgh using a gradient echo EPI pulse sequence with TR = 1000 ms, TE = 30 ms and a 60° flip angle. Seventeen 5-mm thick oblique-axial slices were imaged with a gap of 1-mm between slices. The acquisition matrix was 64 x 64 with 3.125 x 3.125 x 5-mm voxels.

Data processing and analysis. Data processing and statistical analysis were performed with Statistical Parametric Mapping software (SPM2, Wellcome Department of Cognitive Neurology, London, UK; Friston, 2005). The data were corrected for slice timing, motion, linear trend, and were temporally smoothed with a high-pass filter using 190s cutoff. The data were normalized to the MNI template brain image using 12-parameter affine transformation.

The data were prepared for regression and classification analysis by being spatially normalized into MNI space and resampled to 3x3x6 mm^3 voxels. We try to keep approximately the same acquisition voxel size which has been used in many of our previous studies and is adequate for a list of different cognitive tasks. Voxels outside the brain or absent from at least one participant were excluded from further analysis. The percent signal change (PSC) relative to the fixation condition was computed for each object presentation at each voxel. The mean of the four images (mean PSC) acquired within a 4s window, offset 4s from the stimulus onset (to account for the delay in hemodynamic response) provided the main input measure for subsequent analysis. The mean PSC data for each word or picture presentation were further normalized to have mean zero and variance one to equate the variation between participants over exemplars.

Furthermore, our theoretical framework does not take a position on whether the neural activation encoding meaning is localized in particular cortical regions. Shinkareva et al. (2007) identified single brain regions that consistently contained voxels used in identification of object categories across participants. The brain locations that were important for category identification were similar across participants and were distributed throughout the cortex where various object properties might be neurally represented. Thus, we consider all cortical voxels and allow the training data to determine which locations are systematically modulated by which aspects of word meanings. The main analysis selected the 120 voxels whose responses to the 60 different items were most stable across presentations (many previous analyses had indicated that 120 was a useful set size for our purposes). Voxel stability was computed as the average pairwise correlation between 60-item vectors across presentations.

The stable voxels were located in multiple areas of the brain. Figure 2.2 shows voxel clusters from the union of stable voxels from all nine participants. As shown, many of these locations are in occipital, occipital-temporal, and occipital-parietal areas, with more voxels in the left hemisphere. Table 2.2 lists the distribution of the 120 voxels selected by the stability measure for each participant, sorted by major brain structures and size of clusters.

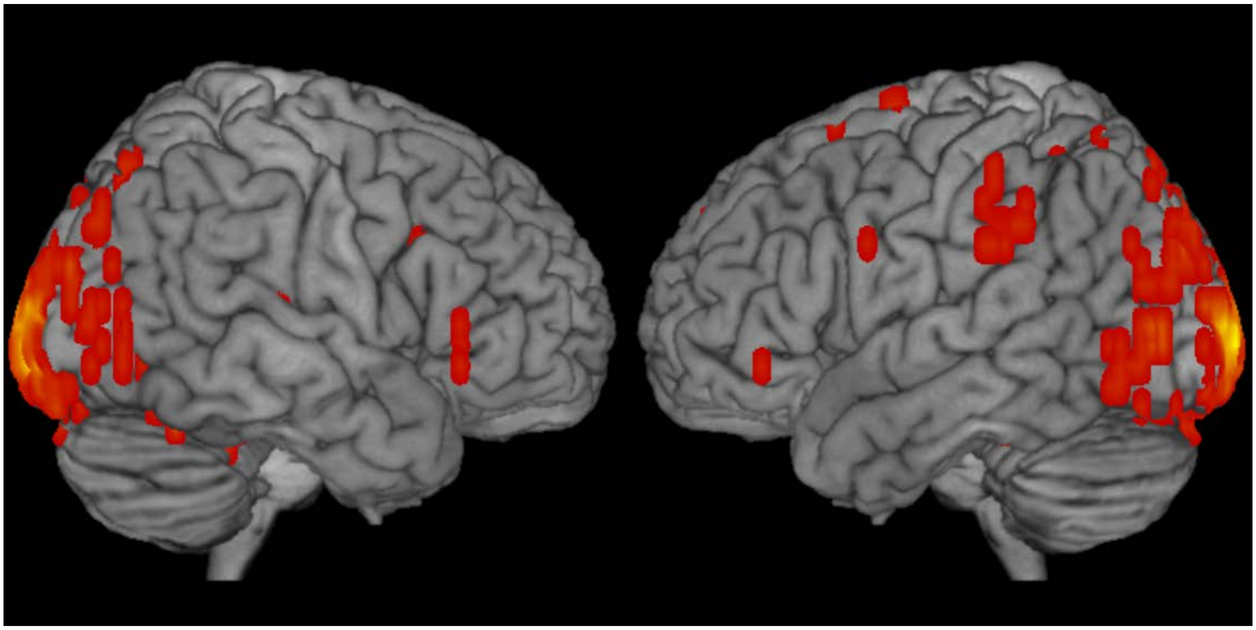


Figure 2.2 Voxel clusters from the union of stable voxels from all nine participants. Many of these locations are in occipital, occipital-temporal, and occipital-parietal areas, with more voxels in the left hemisphere.

Table 2.2 Locations (MNI centroid coordinates) and sizes of the voxel clusters selected by the stability measure.

Participant	Label	X	Y	Z	Voxels[b]	Radius
P1	*Occipital*					
	R Fusiform Gyrus	31.5	-50.4	-10	24	7.02
	L Fusiform Gyrus	-26.9	-50.9	-11.7	21	6.13
	L Occipital Middle	-20.1	-98.7	6	21	6.03
	L Occipital Inferior	-15.1	-91.1	-10.2	13	5.22
	R Occipital Middle	34.9	-76	13	6	4.72
	R Calcarine	6.2	-91.1	4	6	4.17
P2	*Medial Temporal*					
	L Parahippocampal Gyrus	-25	-42.2	-15	6	3.79
	Occipital					
	R Calcarine	15.5	-96	-0.9	70	9.73
	L Calcarine	-16.6	-98.6	-4.1	22	7.1
	L Cuneus	-20.6	-60	15.6	5	3.51

[b] The number of voxels per participant is less than 120 because of a cluster size threshold of 5 voxels used in this table.

Participant	Label	X	Y	Z	Voxels	Radius
P3	*Parietal*					
	L Precuneus	-5.6	-57.5	24	5	2.65
	Occipital					
	R Calcarine	18.2	-93.5	2.8	75	11.26
	L Occipital Middle	-17.1	-98.3	-1.5	28	7.73
P4	*Temporal*					
	R Fusiform Gyrus	36.5	-40.1	-23	6	5.72
	Parietal					
	L Supramarginal Gyrus	-53.8	-33.1	33	10	4.56
	L Parietal Inferior	-35.4	-39.6	43	6	3.31
	R Parietal Superior	19.4	-63.7	56.4	5	3.51
	Occipital					
	L Fusiform	-28.6	-53.1	-14	12	6.59
	R Occipital Middle	32	-86.7	19.5	12	5.36
	L Occipital Superior	-13.2	-84.7	40	9	5.69
	L Occipital Middle	-31.6	-87.5	24	9	5.4
	R Lingual	13.3	-101.2	-7.5	8	4.02
P5	*Temporal*					
	L Fusiform Gyrus	-31.5	-42.9	-18.8	15	5.3
	R Fusiform Gyrus	34.4	-41.6	-16.2	13	4.51
	Occipital					
	L Lingual	-14.9	-89.7	-2.3	44	7.75
	R Calcarine	20.5	-94.6	-2.9	35	8.45
P6	*Medial Temporal*					
	R Parahippocampal Gyrus	25.9	-47.5	-13.2	10	6.47
	Occipital					
	R Calcarine	17.3	-96.6	-1.1	51	10.92
	L Occipital Middle	-19.4	-97.8	-3.1	23	8.56
	L Fusiform Gyrus	-23.8	-49.5	-9.2	13	5.68
	R Fusiform Gyrus	30	-71.9	-9.6	5	3.76
P7	*Temporal*					
	L Fusiform Gyrus	-28.8	-46.1	-16.5	20	5.96
	Occipital					
	R Calcarine	8.8	-96.1	-2.1	35	7.93
	R Fusiform Gyrus	31.2	-49.9	-14.9	21	5.65
	L Calcarine	-16	-98.8	-5.2	8	3.97
	L Lingual	-9.8	-88.8	-11.1	7	4.17

Participant	Label	X	Y	Z	Voxels	Radius
P8	*Temporal*					
	L Temporal Inferior	-45.5	-67.2	-7.7	14	5.09
	Occipital					
	R Lingual	7.7	-87.9	-6.4	43	9.64
	L Occipital Middle	-18.2	-97.4	-1.9	28	8.48
	R Calcarine	11.9	-100.3	-0.6	10	6.9
P9	*Temporal*					
	L Fusiform Gyrus	-31.8	-39.8	-21.3	11	5.04
	R Temporal Inferior	45	-64.4	-3.6	5	3.61
	Medial Temporal					
	R Parahippocampal Gyrus	23.8	-42	-15	16	5.05
	Occipital					
	R Calcarine	20.6	-98	-2.5	19	5.42
	L Occipital Middle	-16.4	-102	4.5	8	5.61
	L Occipital Middle	-26.8	-88.4	35.1	7	4.15
	L Lingual	-20.3	-44.8	-10	6	4.16
	R Occipital Middle	37.5	-78.8	38.4	5	3.68

For classifier analysis, voxel stability was computed using only the training set within each fold in the cross-validation paradigm. For within-participants analysis, where the training data consist of 5 of the 6 presentations and the testing data consist of the remaining presentation, the voxel stability was computed using only the training data for that particular participant. For between-participants analysis, where the training data consists of 8 of the 9 participants and the testing data consist of the remaining participant, the voxel stability was computed using only the training data for the 8 participants. The focus on the most stable voxels effectively increased the signal-to-noise ratio in the data and also served as a dimensionality reduction tool that facilitated further analysis by classifiers.

2.3 Approach

In this study, we model hidden factors that underpin semantic representation of object knowledge with a multivariate multiple linear regression model. We adopt a feature-based representation of semantic knowledge, in which a word's meaning is determined by a vector of features. Two competing models based on Cree and McRae (2003)'s feature norming study were developed and evaluated using three types of criteria. The three types of evaluation criteria are a regression fit to the fMRI data, the ability to decode mental states given a neural activation pattern, and the ability to distinguish between the activation of two previously unseen objects. Figure 2.3 depicts the flow chart of our approach.

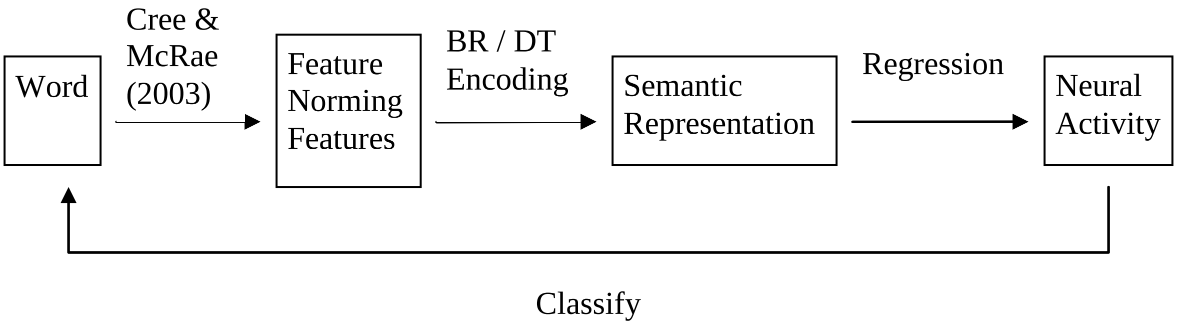


Figure 2.3 The flow chart of the generative model. First, the feature norming features associated with the word are retrieved from Cree & McRae (2003). Secondly, the feature norming features are encoded into BR or DT knowledge types, which constitute the semantic representation. Then, a linear regression model learns the mapping between the semantic representation and fMRI neural activity. Finally, a nearest neighbor classifier uses the predicted neural activity generated by the regression model to decode the mental state (word) associated with an observed neural activity.

2.3.1 Feature norming study

One way to characterize an object is to ask people what features an object brings to mind. Cree and McRae's (2003) semantic feature norming studies asked participants to list the features of 541 words. Fortunately, 43 of these words were included in our fMRI study. The words were derived from five domains that include living creatures, nonliving objects, fruits, and vegetables. The features that participants produced were a verbalization of actively recalled semantic knowledge. For example, given the stimulus word *house*, participants might report features such as *used for living, made of brick, made by humans*, etc. Such feature norming studies have proven to be useful in accounting for performance in many semantic tasks (Hampton, 1997; McRae et al., 1999; Rosch & Mervis, 1975).

Because participants in the feature norming study were free to recall any feature that came to mind, the norms had to be coded to enable further analysis. Two encoding schemes, Cree and McRae's (2003) brain region (BR) scheme and Wu and Barsalou's (2009) detailed taxonomic (DT) encodings, were compared. BR encoding was based on a knowledge taxonomy that adopts a modality-specific view of semantic knowledge. That is, the semantic representation of an object is assumed to be distributed across several cortical processing regions known to process related sensory input and motor output. BR encoding therefore groups features into knowledge types according to their relations to some sensory/perceptual or functional processing regions of the brain. For example, features for *cow* like *eats grass* would be encoded as visual-motion, *is eaten as beef* as function, and *is animal* as taxonomic in this scheme. By contrast, DT encoding captures features from four major perspectives: entity, situation, introspective, and taxonomic, which are further categorized into 37 hierarchically-nested specific categories. For example, features for *cow* like *eats grass* would be encoded as entity-behavior, *is eaten as beef* as function, and *is an animal* as superordinate. Adapted from Cree and McRae (2003), Table 2.3 lists the features and the corresponding BR and DT encodings for the words *house* and *cow*. Also, Table 2.4 and Table 2.5 list all the classes and knowledge types in BR and DT encodings that are relevant to our stimulus set.

The analyses below are applied only to those 43 of the 60 words in our study that also occurred in Cree and McRae's study. The missing stimuli are marked with asterisks in Table 2.1. A matrix was thus constructed for each of the two types of encodings of the feature norms, of size 43 exemplars by the number of knowledge types (10 for BR encoding and 27 for DT encoding, which have non-zero entries). A row in the matrix corresponds to the semantic representation for an exemplar, where elements in the row correspond to the number of features (for that exemplar) categorized as particular knowledge types. Normalization consists of scaling the row vector of feature values to unit length. Consequently, these matrix representations encoded the meaning of each exemplar in terms of the pattern distributed across different knowledge types. For example, the word *house* would have a higher value in the *visual form and surface properties* knowledge type, as opposed to *sound* or *smell*, because people tended to recall more features that described the appearance of a house rather than its sound or smell.

Table 2.3 Example of Concepts from Feature Norms

Concept	Feature	BR Encoding	DT Encoding
House	Made by humans	Encyclopedic	Origin
	Is expensive	Encyclopedic	Systemic property
	Used for living in	Function	Function
	Used for shelter	Function	Function
	Is warm	Tactile	Internal surface property
	A house	Taxonomic	Superordinate
	Is large	Visual-form and surface properties	External surface property
	Made of brick	Visual-form and surface properties	Made of
	Has rooms	Visual-form and surface properties	Internal component
	Has bedrooms	Visual-form and surface properties	Internal component
	Has bathrooms	Visual-form and surface properties	Internal component
	Is small	Visual-form and surface properties	External surface property
	Has doors	Visual-form and surface properties	External component
	Has windows	Visual-form and surface properties	External component
	Made of wood	Visual-form and surface properties	Made of
	Has a roof	Visual-form and surface properties	External component
Cow	Lives on farms	Encyclopedic	Location
	Is stupid	Encyclopedic	Evaluation
	Is domestic	Encyclopedic	Systemic property
	Eaten as meat	Function	Function
	Eaten as beef	Function	Function
	Used for producing milk	Function	Function
	Is smelly	Smell	External surface property
	Moos	Sound	Entity behavior
	An animal	Taxonomic	Superordinate
	An mammal	Taxonomic	Superordinate
	Is white	Visual-color	External surface property
	Is black	Visual-color	External surface property
	Is brown	Visual-color	External surface property
	Has 4 legs	Visual-form and surface properties	External component
	Has an udder	Visual-form and surface properties	External component
	Is large	Visual-form and surface properties	External surface property
	Has legs	Visual-form and surface properties	External component
	Has eyes	Visual-form and surface properties	External component
	Produces milk	Visual-motion	Entity behavior
	Eats grass	Visual-motion	Entity behavior
	Produces manure	Visual-motion	Entity behavior
	Eats	Visual-motion	Entity behavior

Table 2.4 Cree and McRae (2003)'s Brain Region (BR) Encoding Scheme

Class	Knowledge Type	Frequency	Example
Visual	Visual color	32	Celery <is green>
	Visual form and surface properties	252	House <is made of bricks>
	Visual motion	22	Cow <eat grass>
Other primary sensory-processing	Smell	2	Barn <is smelly>
	Sound	7	Cat <behavior – meows>
	Tactile	20	Bed <is soft>
	Taste	3	Corn <tastes sweet>
Functional	Function	142	Hammer <used for pounding>
Miscellaneous	Taxonomic	62	Skirt <clothing>
	Encyclopedic	132	Car <requires gasoline>

Table 2.5 Wu and Barsalou (2009)'s Detailed Taxonomic (DT) Encoding Scheme

Class	Knowledge type	Frequency	Example
Entity	Associated abstract entity	1	Church <associated with religion>
	Entity behavior	26	Cat <behavior – meows>
	External component	139	Chair <has 4 legs>
	External surface property	85	Celery <is green>
	Internal Component	24	Airplane <has engines>
	Internal surface property	12	Corn <tastes sweet>
	Larger whole	3	Spoon <part of table setting>
	Made-of	47	House <is made of bricks>
	Quantity	3	Butterfly <different types>
	Systemic property	36	Knife <is dangerous>
Situation	Action	9	Screwdriver <is hand held>
	Associated entity	24	Shirt <worn with ties>
	Function	116	Hammer <used for pounding>
	Location	38	Keys <found on chains>
	Origin	5	Tomato <grows on vines>
	Participant	17	Desk <used by students>
	Time	5	Coat <worn for winter>
Taxonomic	Coordinate	1	Cup <a mug>
	Individual	0	N/A
	Subordinate	9	Pants <e.g. jeans>
	Superordinate	52	Skirt <clothing>
	Synonym	0	N/A
Introspective	Affect emotion	0	N/A
	Cognitive operation	0	N/A
	Contingency	12	Car <requires gasoline>
	Evaluation	10	Dog <is friendly>
	Negation	0	N/A

2.3.2 Regression model

Our generative model attempts to predict the neural activity (mean PSC), by learning the correspondence between neural activation and object features. Given a stimulus word, *w*, the first step (deterministically) encoded the meaning of *w* as a vector of intermediate semantic features, using BR or DT. The second step predicted the neural activity level of the 120 most stable voxels in the brain with a multivariate multiple linear regression model. The regression model examined to what extent the semantic feature vectors (explanatory variables) can account for the variation in neural activity (response variable) across the 43 words. R^2 measures the amount systematic variances explained in the neural activation data. All explanatory variables were entered into the regression model simultaneously. More precisely, the predicted activity a_v at voxel *v* in the brain for word *w* is given by

$$a_v = \sum_{i=1}^{n} \beta_{vi} f_i(w) + \varepsilon_v$$

where $f_i(w)$ is the value of the i^{th} intermediate semantic feature for word *w*, β_{vi} is the regression coefficient that specifies the degree to which the i^{th} intermediate semantic feature activates voxel *v*, and ε_v is the model's error term that represents the unexplained variation in the response variable. Least squares estimates of β_{vi} were obtained to minimize the sum of squared errors in reconstructing the training fMRI images. This least squares estimate of the β_{vi} yields the maximum likelihood estimate under the assumption that ε_v follows a Normal distribution with zero mean. A small L2 regularization with lambda = 0.5 was added to avoid rank deficiency.

The use of a linear regression model to model the hidden factors is not new to analysis of neural activity. Indeed, both linear regression analysis and Statistical Parametric Mapping (SPM)

- the most commonly used technique for fMRI data analysis - belong to the more general mathematical paradigm called Generalized Linearized Models (GLM). GLM is a statistical inference procedure that models the data to partition the observed neural response into components of interest, confounds, and error (Friston, 2005). Specifically, GLM assumes a linear dependency among the variables and compares the variance due to the independent variables against the variance due to the residual errors. While the linearity assumption underlying the general linearized model may be overly simplistic, it reflects the assumption that fMRI activity often reflects a superposition of contributions from different sources, and has provided a useful first order approximation in the field.

The intermediate semantic features associated with each word are therefore regarded as the hidden factors or sources contributing to the object knowledge. The trained regression model then weights the influence of each source and linearly combines the contribution of each factor to produce an estimate of the resulting neural activity. For instance, the neural activity image of the word *house* may be different from that of *cow* in that the contribution from the factor corresponding to the item's *function* (what it is used for) plays a more significant part for *house* and that the contribution from the *sensory* factor plays a more significant part for *cow*, as depicted in the sensory/functional theory.

2.3.3 Classifier model

Classifiers were trained to identify cognitive states associated with viewing stimuli from the evoked pattern of functional activity (mean PSC). Classifiers were functions *f* of the form: *f: mean_PSC* → *Y*i, *i*=1,…n, where *Y*i were the sixty exemplars, and *mean_PSC* was a vector of mean PSC voxel activation level, as described above. To evaluate classification performance, data were divided into training and test sets. A classifier was built from the training set and evaluated on the left-out test set.

In this study, two classifiers were compared: a Support Vector Machine (SVM) classifier that does not utilize a hidden layer representation and a nearest neighbor classifier that utilizes a hidden layer representation learned in the regression analysis. The SVM classifier (Guyon, Boser, & Vapnik, 1993) is a widely-used discriminative classifier that maximizes the margin between exemplar classes. The SVM classifier is implemented in a software package called SVM-light, which is an efficient implementation of SVM by Thorsten Joachims and can be obtained from http://svmlight.joachims.org. On the other hand, the nearest neighbor classifier proposed here uses the estimated regression weights to generate predicted activity for each word. The regression model first estimates a predicted activation vector for each of the 60 objects. Then, a previously unseen observed neural activation vector is identified with the class of the predicted activation that had the highest correlation with the given observed neural activation vector.

Our approach is analogous in some ways to research that focuses on lower-level visual features of picture stimuli to analyze fMRI activation associated with viewing the picture (O'Toole et al., 2005; Hardoon et al., 2007; Kay et al., 2008). A similar generative classifier is used by Kay et al. (2008) where they estimate a receptive-field model for each voxel and classify an activation pattern in terms of its similarity to the predicted brain activity. Our work differs from these efforts, in that we focus on encodings of more abstract semantic features signified by words and predict brain activity based on these semantic features, rather than on visual features that encode visual properties.

2.4 Results

Using feature norms to explain the variance in neural activity. The regression models were assessed in terms of their ability to explain the variance in neural activity patterns. A multivariate multiple linear regression was run for each participant, using either BR or DT encoding as explanatory variables, and average neural activity (mean PSC) across 120 most stable voxels as response variables. Specifically, DT encoding (with its 27 independent variables) accounted for an average of 58% of the variance in neural activity, whereas BR encoding (with its 10 independent variables) accounted for an average of 35% of the variance. R^2 is higher for DT than for BR for all 9 of the participants, as shown in Table 2.6. Notice that DT encoding outperforms BR encoding in explaining the variance in neural activity pattern, even though Cree and McRae (2003) found that the two encodings produce similar results in their hierarchical clustering analysis of behavioral data and that they both can be used to explain the tripartite impairment pattern in category-specific deficit studies. This difference may, however, simply be due to the different number of parameters (explanatory variables) that the two regression models use. Akaike information criterion (AIC) is a measure of the goodness of fit that accounts for the tradeoff between the accuracy and complexity of different models and is invariant to the number of parameters. The relative values of AIC scores are used for *model selection* among a class of parametric models with different numbers of parameters, with the model with lowest AIC being preferred. The BR decoding yields an average AIC score of -37.18, whereas the DT encoding yields an average AIC score of -23.93. Thus, it appears that the difference in regression fit may be due to the different number of parameters that the two regression models use. We further explore this issue in the discussion section.

Table 2.6 Regression Analysis R^2

Model	Mean	SD	Participants								
			P1	P2	P3	P4	P5	P6	P7	P8	P9
BR	0.35	0.07	0.47	0.30	0.29	0.43	0.31	0.29	0.36	0.29	0.39
DT	0.58	0.04	0.61	0.56	0.53	0.62	0.59	0.59	0.64	0.52	0.58

The regression models produce a predicted neural activity pattern for each word, which can be compared to the observed pattern. For example, Figure 2.4 shows one slice of both the observed and the predicted neural activity pattern for the words *house* and *cow*. In each case, the predicted activity is more similar to the observed activity of the target word than to the other word.

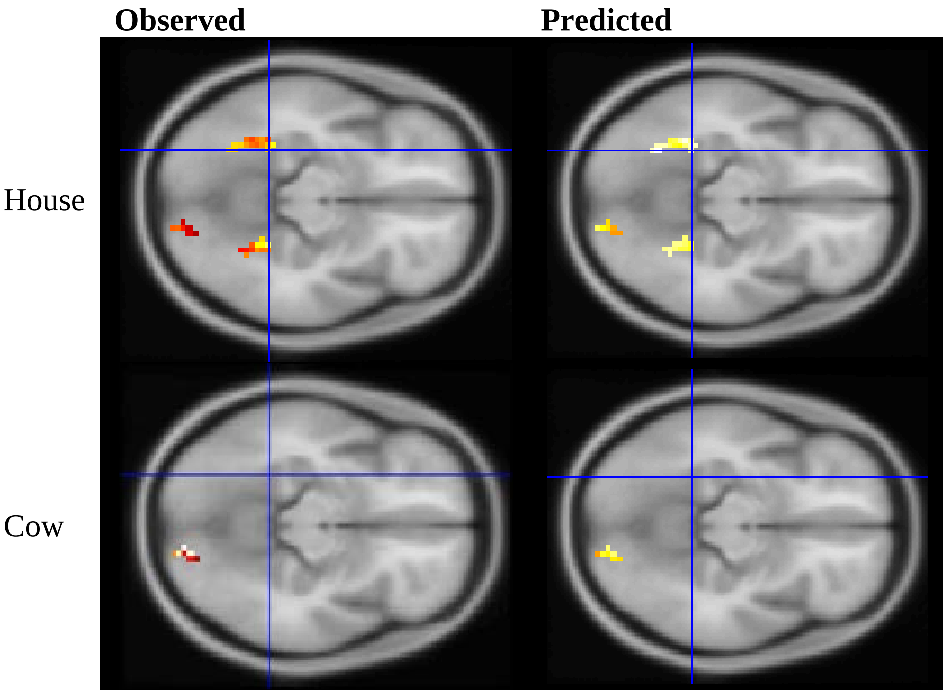
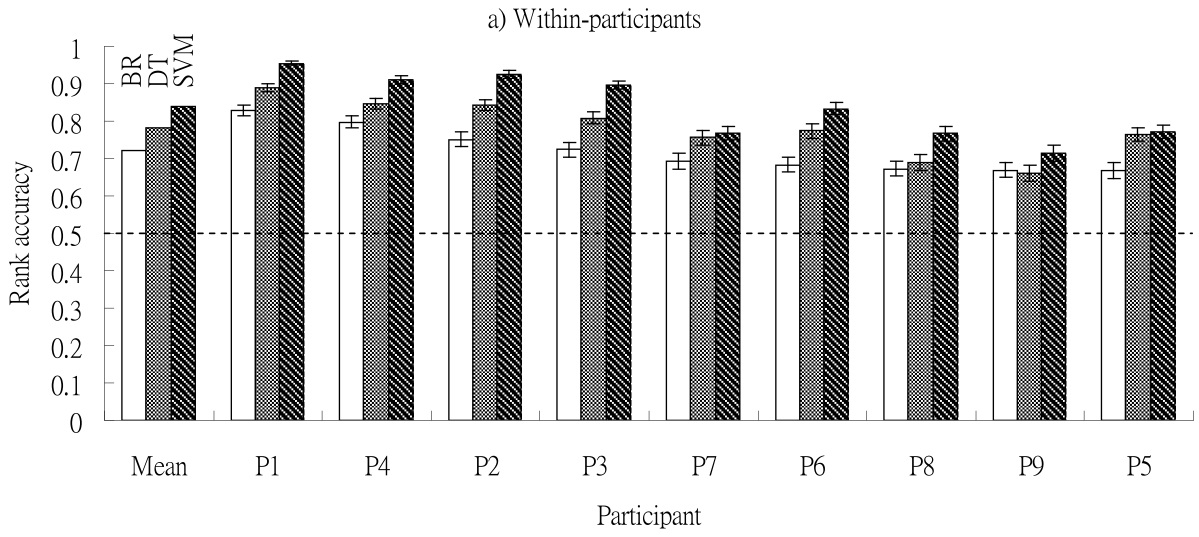


Figure 2.4 Observed vs. predicted neural activities at left parahippocampal gyrus (Brodmann area 37, coordinates -28.125, -43.75, -12) for the stimulus words *house* and *cow*. The observed neural activity vector is taken from participant P1, whereas the predicted neural activity vector is estimated by the regression model with BR encoding as explanatory variables and 120 most stable voxels as response variables. In each case, the predicted activity is more similar to the observed activity of the target word than to the other word, suggesting that the predicted activity may be useful to classify words.

Classifying mental states. Given that the semantic feature vectors can account for a significant portion of the variation in neural activity, the predictions from the regression model can be used to decode mental states of individual participants. This was effectively a 43-way word classification task, where the attributes were neural activity vectors and the classes were 43 stimulus items. This analysis can be performed both within participants (by training the classifier on a subset of the participant's own data and then testing on an independent, held-out subset) and between-participants (training on all-but-one participants' data and testing on the left-out one).

For the within-participants analysis, a regression model was developed from the data from 4 out of 6 presentations of a participant and applied to the average activation of the two remaining presentations of the same participant, using a nearest neighbor classifier to classify the neural activity pattern. A regression model using BR or DT encoding classified the items from the held-out presentations with an average of 72% and 78% rank accuracy, respectively. Since multiple classes were involved, *rank accuracies* are reported, which measure the percentile rank of the correct word within a list of predictions made by the classifier (Mitchell et al., 2004). The rank accuracy for each participant, along with the 95% confidence interval, is reported in Figure 2.5. The confidence interval is estimated by random sampling of the dataset 10,000 times with replacement and subsequently computes the classifier performance. All classification accuracies were significantly ($p < 0.05$) different from a chance level of 50% determined by permutation testing of class labels. DT encoding performed significantly better ($p < 0.05$) than BR encoding for 7 out of 9 participants. Furthermore, the generative classifiers were compared with the SVM classifier which does not utilize a hidden layer representation. The SVM classifier, which achieved an average of 84% rank accuracy, performed significantly ($p < 0.05$) better than the two generative classifiers for 7 out of 9 participants.

For the between-participants analysis, a regression model was developed from the data from 8 out of 9 participants and applied to the average activation of all possible pairs of presentations in the remaining participant, using a nearest neighbor classifier to classify the neural activity pattern. A regression model using BR or DT encoding classified the items from the held-out subject with an average of 68% and 70% rank accuracy, respectively. The rank accuracy for each participant, along with the 95% confidence interval, is reported in Figure 2.5. The confidence interval is estimated by random sampling of the dataset 10,000 times with replacement and subsequently computes the classifier performance. All classification accuracies were significantly ($p < 0.05$) different from a chance level of 50% determined by permutation testing of class labels. For 7 out of 9 participants, the difference between BR and DT encoding was not significantly ($p < 0.05$) different. Furthermore, the generative classifiers were compared with the SVM classifier which does not utilize a hidden layer representation. Unlike in the within-participants classification, the SVM here performed poorly, achieving a mean rank accuracy of only 63%, and obtaining a significantly ($p < 0.05$) lower rank accuracy than the two generative classifiers for 5 out of 9 participants.



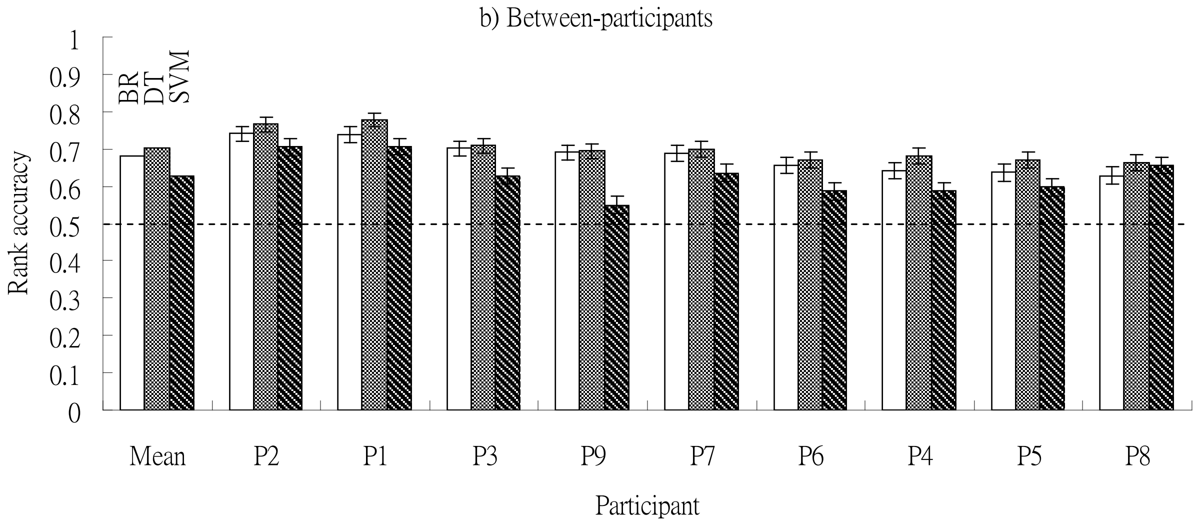


Figure 2.5 Decoding mental states given neural activation pattern. A discriminative SVM classifier, which utilizes no hidden layer representation, is compared to two generative nearest neighbor classifiers which extend the regression model, with BR or DT as the explanatory variables. The dashed line indicates chance level at 50%. Participants are sorted according to rank accuracy of the BR model. a) Within-participants analysis, b) Between-participants analysis. Whereas the discriminative SVM classifier performs the best in the within-participants classification, the generative classifiers generalize better in the between-participants classification.

Distinguishing between the activation of two unseen stimuli. Can the predictions from the regression model be used to classify the mental states of participants on words that were never seen before by the model? In other words, can the regression model generalize to make predictions for a previously unseen word, given the values of the independent variables (the semantic features) for that word? To test this possibility, all possible pairs of the 43 words were held out (one pair at a time) from the analysis, and a multivariate multiple linear regression model was developed from the data of the remaining 41 words, with semantic feature vectors (either the BR or DT encoding) as the explanatory variables, and observed neural activity vectors (mean PSC across 120 most stable voxels) as the response variables. The estimated regression weights were then used to generate the predicted activation vector for the two unseen words, based on the feature encodings of those two words. Then, the observed neural activation vector for the two unseen words was identified with the class of the predicted activation vector with which it had the higher correlation.

A regression model using BR or DT encoding correctly classified an average of 65% and 68% of the unseen words, respectively. The classification accuracy for each participant, along with the 95% confidence interval estimated by 10,000 bootstrapped samples, is reported in Figure 2.6. All classification accuracies were significantly ($p < 0.05$) higher than a chance level of 50% determined by permutation testing of class labels. Unlike the case in the regression analysis and word classification, there is no clear difference in the ability of the two encoding schemes to distinguish between two unseen words. For 1 participant, the BR encoding performed significantly better than the DT encoding, but for 2 other participants, the DT performed significantly better. There are no significant differences between BR and DT encoding for the remaining 6 participants.

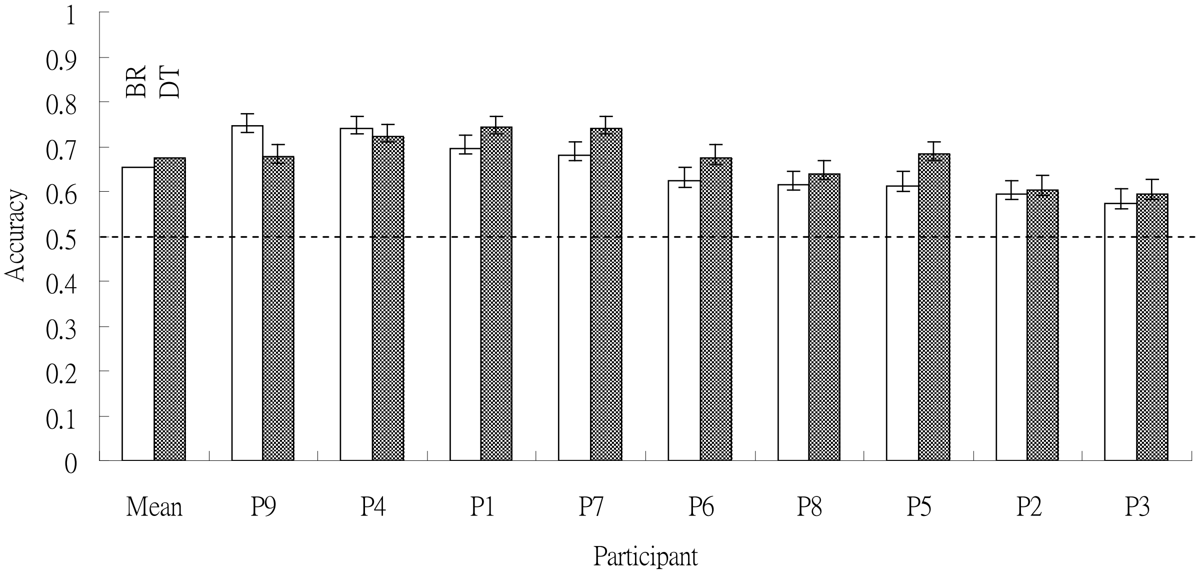


Figure 2.6 Distinguishing between two unseen words. Two generative nearest neighbor classifiers which extend the regression model, with BR or DT encoding as explanatory variables, are shown. The dashed line indicates chance level at 50%. Participants are sorted according to accuracy of BR encoding.

2.5 Discussion

The results indicate that the features from an independent feature norming study can be used in a regression model to explain a significant portion of the variance in neural activity in this 43-item word-picture stimulus set. Moreover, the resulting regression model is useful for both decoding mental states associated with the visual presentation of 43 items and distinguishing between two unseen items. Although the proposed generative nearest neighbor classifier that utilizes a hidden layer does not outperform a discriminative SVM classifier in the within-participants classification, it does outperform the SVM classifier in between-participants classification, suggesting that the hidden, semantic features do provide a mediating representation that generalizes better across participants. Furthermore, the hidden factors allow us to extrapolate the neural activity for unseen words, which simply cannot be done in a discriminative classifier.

Comparing the generative classifier and discriminative classifier. There appears to be a double dissociation between the two classifier approaches and within- versus between-participants generalization. Whereas an SVM-based discriminative classifier achieves the best classification accuracy in within-participants analysis, the generative classifier outperforms an SVM-based model which does not utilize such intermediate representations in a between-participants analysis. In fact, there is a strong negative correlation (p = -0.79) between the within-participants difference and the between-participants difference between the models. That is, the better SVM is, relative to DT, at decoding brain activity within participants, the worse SVM is, again relative to DT, at decoding brain activity across participants. This pattern of results suggests the SVM-based classifier may be picking up some idiosyncratic patterns that do not generalize well across participants and that good generalization across participants may require broad, large-scale patterns that are used in our set of intermediate semantic features.

A discriminative SVM classifier attempts to learn the function that maximizes the margin between exemplar classes across all presentations/subjects. While this strategy is the current state-of-the-art classification technique and indeed yields the best performance in within-participants classification, it works less well in between-participants classification when there is not sufficient data to learn complex functions that would capture individual differences (or when that the function is too complicated to learn). On the contrary, the regression model does not attempt to model the differences in neural activity across presentations/subjects. Instead, the regression model averages out the differences across presentation/subjects and learns to estimate the average of the neural activity that is available in the training data. Specifically, the regression model learns the correspondence between neural activation and object features that accounts for the most systematic variance in neural activity across the 43 words. The advantage is two-fold. First, sample mean is the uniformly minimum-variance unbiased estimator of population mean of neural activity. Thus, to predict the neural activity of a previously unseen presentation or individual, one of the best unbiased estimators is the average of the neural activity of the same word available in the training data. But simply taking the sample mean does not allow prediction of a previously unseen word – there is no data for it. Thus, by learning the correspondence between neural activation and object features, the regression model has the second advantage that it can extrapolate to predict the neural activity for unseen words, as long as there is access to the object features of the unseen words, which can be assumed given access to the large scale feature-norming studies and the various linguistic corpora.

Encoding feature norming features into knowledge types. In our analysis, we encode the feature norming features into knowledge types. The generative models work with knowledge types, not with knowledge content. For instance, it would matter for the models whether a *house* is associated more often with surface property, but not the exact property like *is large* or *is small*. As another example, it matters that a *cow* is associated more often with entity behavior, but it does not matter what type of behavior the cow executes (e.g. *eat grass* or *produce milk*). The model discriminates between a *house* and a *cow* by the pattern distributed across different knowledge types (e.g. a *house* is described with more surface properties and a *cow* is described with more entity behaviors), but not the actual features listed (e.g. a *house is large* and a *cow eats grass*). Thus, our intermediate semantic representation encodes word meaning at the level of knowledge types. From this viewpoint it is less surprising that this type of intermediate representation generalizes well across participants. Good generalization across participants may require broad, large-scale patterns, while idiosyncratic patterns may be related to more fine-scale patterns of activity that do not survive the inter-participants differences in anatomy.

Comparing BR and DT encoding. Different encodings (e.g. BR or DT) on the same feature norming set, however, led to different regression fits and classification accuracies. The DT encoding outperformed BR encoding in the regression analysis and in within-participants mental state classification, but the phenomenon diminishes in between-participants mental state classification and when distinguishing between two unseen stimuli. The former finding is surprising at first, since Cree and McRae (2003) reported that the two encodings performed similarly in their hierarchical clustering analysis in explaining seven behavioral trends in category deficits. The difference obtained between the two types of feature norm encodings in their account of brain activation data could have arisen because one encoding is truly superior to the other, but there are also technical differences between the models that merit consideration. Specifically, the phenomenon called *overfitting* refers to a regression model with more predictor variables being able to better tune to the data and as a result overfit. Consequently, the DT regression model with its encoding of 27 knowledge types (independent variables) would overfit more easily to data than a BR regression model that utilizes 10 knowledge types.

The overfitting phenomenon can be considered more precisely by examining each model’s performance under the three evaluation criteria, which, though correlated, measure different constructs and have different profiles. First, the regression fit measures the amount of systematic variance explained by the regressor variables, and their ability to re-construct the neural images. Second, the word classification accuracy measures the degree to which the predicted neural image is useful for discriminating among stimuli. Third, classification on novel stimuli measures how well the model generalizes to previously unseen words. Whereas regression analysis is performed on all available data, classification analysis (especially classification of novel stimuli, in our case distinguishing between two unseen words) is *cross validated* (train and test on different data set) and is less prone to overfitting.

To compare the two encoding schemes while equating the number of independent variables, a step-wise analysis was performed to gradually enter additional variables in the regression model, instead of entering all of them simultaneously. As the number of knowledge types included in the DT encoding increases, the regression fit keeps increasing, as shown in Figure 2.7a, but the classification accuracy on novel stimuli, shown in Figure 2.7b, increases at first but peaks and gradually decreases – clear evidence of overfitting. With fewer knowledge types, the BR encoding overfits less to the data and generalizes better to unseen words. Moreover, the performance of the BR encoding peaks when about 6 knowledge types are entered into the regression model, reaching an average accuracy of 68%, whereas the performance of the DT encoding peaks when about 8 knowledge types are used, reaching an average accuracy of 77%. Notice that, although the BR and DT encodings are constructed subject to different criteria, the features of the two encoding schemes that are found to be the most important in the step-wise analysis are similar. The underlying semantic features that provide the best account of the neural activation data consist of taxonomic and visual features (e.g. *visual color, visual motion,* and *function* for the BR encoding and *internal component*, *entity behavior,* and *associated entity* for the DT encoding). Table 2.7 and Table 2.8 show the ranked order list of each of the BR knowledge type and each of the DT knowledge type's ability to classify mental state (within-participants analysis, averaged over participants), respectively. To test whether a classifier was significantly better than chance, we first computed its overall accuracy for each subject, yielding a distribution of *N* accuracies, where *N* is the number of subjects. Treating this distribution as a random value, we performed a two-tailed T-test of whether its mean exceeds chance performance of 0.5. Our significance criterion was $p < 0.01$, without correction for multiple comparisons. Thus the superficial differences between BR and DT feature encoding schemes lessen or disappear in the light of more sensitive assessments, and the modeling converges on some core encoding features that provide a good converging account of the data.

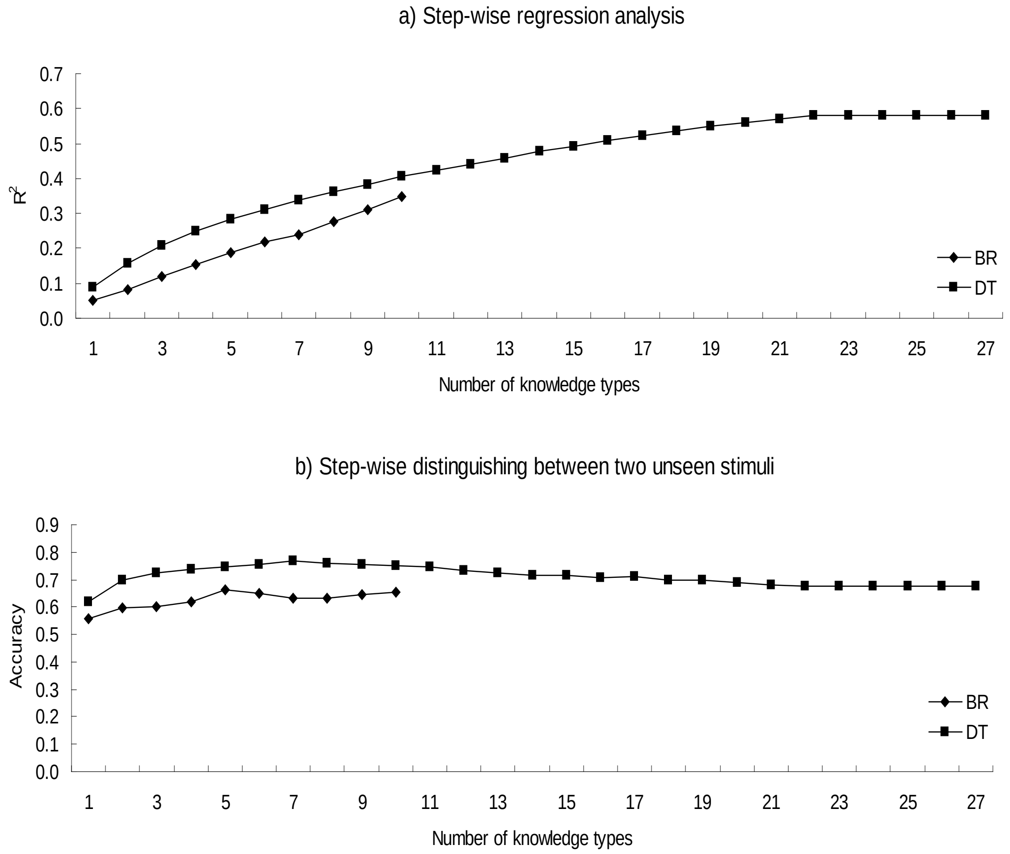


Figure 2.7 Step-wise analysis. a) Step-wise regression analysis, b) step-wise distinguishing between two unseen stimuli. With finer distinction of knowledge types, DT encoding is more prone to overfitting than BR encoding. As the number of knowledge types in DT encoding is increased, the regression fit keeps increasing, but classification accuracy on unseen stimuli increases at first but peaks and gradually decreases – clear evidence of overfitting. With fewer knowledge types, BR overfits to a lesser extent.

Table 2.7 Each BR knowledge type's ability to classify mental states. Asterisks mark classifiers whose performance is significant better than a chance level of 0.5 ($p < 0.01$).

Knowledge Type	Accuracy
Visual-color	0.58*
Visual-motion	0.58*
Function	0.53*
Sound	0.53*
Taxonomic	0.52*
Tactile	0.52*
Encyclopedic	0.51*
Smell	0.51
Taste	0.51
Visual-form and surface properties	0.50

Table 2.8 Each DT knowledge type's ability to classify mental state. Asterisks mark classifiers whose performance is significant better than a chance level of 0.5 ($p < 0.01$).

Knowledge Type	Accuracy
Internal component	0.59*
Entity behavior	0.58*
Associated entity	0.56*
Made of	0.56*
Location	0.56*
Contingency	0.55*
Function	0.55*
Subordinate	0.54*
Systemic property	0.54*
Evaluation	0.53*
Participant	0.53*
External component	0.53*
Action	0.53*
External surface property	0.53*
Superordinate	0.52*
Larger whole	0.52*
Time	0.52*
Internal surface property	0.52*
Origin	0.52*
Quantity	0.51
Associated abstract entity	0.51*
Coordinate	0.51
Affect emotion	0.50
Cognitive operation	0.50
Individual	0.50
Negation	0.50
Synonym	0.50

Comparing feature norming features and word-co-occurrence features. The various models described here were compared to a similar analysis that used features derived from word co-occurrence in a text corpus (Mitchell et al., 2008). In that model, the features of each word were its co-occurrence frequencies with each of 25 verbs of sensorimotor interaction with physical objects, such as *push* and *see*. Co-occurrence features produced an average R^2 of 0.71 when accounting for the systematic variance in neural activity, an average rank accuracy of 0.82 when classifying mental states within-participants, an average rank accuracy of 0.75 when classifying mental states across-participants, and an average accuracy of 0.79 when distinguishing between two previously unseen stimuli. While the performance in rank accuracy when classifying mental states is not statistically different ($p < 0.05$) from that of DT encoding, the advantage of the co-occurrence model in distinguishing between two unseen stimuli is statistically significant ($p < 0.05$). One explanation may be that the encoded object-by-knowledge-type matrices are sparse and heavily weighted in a handful of knowledge types (e.g. visual knowledge types). Feature norming may have fared better if the features corresponded more closely to the types of interactions with objects that are suggested by the 25 sensorimotor verbs. The shortcoming of feature norming in accounting for participants' thoughts when they think about an object is that participants may fail to retrieve a characteristic but psychologically unavailable feature of an object. For example, for an item like *celery*, the attribute of *taste* may be highly characteristic but relatively unavailable. By contrast, using a fixed set of 25 verbs ensures that all 25 will play a role in the encoding. One way to bring the two approaches together is to ask participants in a feature norming study to assess 25 features of an object that correspond to the verbs.

Regardless of whether one uses feature norms or text co-occurrences, choosing the *best* set of semantic features is a challenging problem. For example, it is not clear from the analyses above whether a different set of 25 verbs might not provide a better account. To address these issues, additional modeling was done with corpus co-occurrence features using the 485 most frequent verbs in the corpus (including the 25 sensorimotor verbs reported in Mitchell et al., 2008). A greedy algorithm was used to determine the 25 verbs among the 485 that optimize the regression fit. The greedy algorithm easily overfitted the training data and generalized less well to unseen words. Mitchell et al. (2008) hand-picked their 25 verbs according to some conjectures concerning neural representations of objects. Similarly, it might be worthwhile to consider some conjectures revealed in behavioral feature norming studies when picking the set of co-occurrence semantic features. Further study is required.

Voxel selection method. One property of this study is that it focused on only the most stable voxels, which may have biased the findings in favor of encodings of visual attributes of the items. The voxel selection procedure increases the signal-to-noise ratio and serves as an effective dimensionality reduction tool that empirically derives regions of interest by assuming that the most informative voxels are those that have activation patterns that are stable across multiple presentations of the set of stimuli. The ability of our models to perform classification across previously unseen words suggests we have, to some extent, successfully captured this intermediate semantic representation. Whether the voxels extracted by this procedure correspond to the human semantic system may be task-dependent. For instance, in our task where the stimulus presentations consist of line drawings with text labels, the voxels extracted by this procedure are mostly in the posterior and occipital regions, since our stimuli consist of easily depicted objects and the visual properties of the stimuli are the most invariant part of the stimuli.

Indeed, visual features are among the most important features that account for our neural activation data. If the stimulus presentation consists of only line drawings or text labels, different sets of voxels might be selected. Shinkareva et al. (2007) studied the exact question of the neural representation of pictures versus words. They applied similar machine learning methods on fMRI data to identify the cognitive state associated with viewings of 10 words (5 tools and 5 dwellings) and, separately, with viewings of 10 pictures (line drawings) of the objects named by the words. In addition to selecting voxels from the whole brain, they also identified single brain regions that consistently contained voxels used in identification of object categories across participants. We performed a similar analysis to restrict the analysis space to some predetermined regions of interests. That is, instead of selecting 120 voxels from the whole brain, the voxel selection is applied separately to the frontal lobe, temporal lobe, parietal lobe, occipital lobe, fusiform gyrus, and hippocampus. When only a single region of interest is considered, the highest category identification in the within-participant mental state decoding task is achieved when analysis space is restricted within the occipital lobe, as shown in Table 2.9. However, other regions of interests like the parietal lobe and the fusiform gyrus also carry important information to decode mental state between participants and to distinguish between the activation of two previously unseen words. Indeed, selecting voxels from the whole brain yields the best category identification in the classifier analysis.

Table 2.9 Restricting analysis space through ROIs

a) Regression fit to the fMRI data (R^2)

Model	All	Frontal	Temporal	Parietal	Occipital	Fusiform	Hippocampus
BR	0.35	0.27	0.27	0.32	0.30	0.38	0.24
DT	0.58	0.55	0.55	0.58	0.56	0.61	0.52

b) Ability to decode mental states, within participants (rank accuracy)

Model	All	Frontal	Temporal	Parietal	Occipital	Fusiform	Hippocampus
BR	0.72	0.57	0.60	0.64	0.70	0.67	0.52
DT	0.78	0.58	0.62	0.66	0.77	0.69	0.53

c) Ability to decode mental states, within participants (rank accuracy)

Model	All	Frontal	Temporal	Parietal	Occipital	Fusiform	Hippocampus
BR	0.68	0.47	0.47	0.57	0.59	0.61	0.50
DT	0.70	0.46	0.47	0.56	0.60	0.60	0.49

d) Ability to distinguish between the activation of two previously unseen words (accuracy)

Model	All	Frontal	Temporal	Parietal	Occipital	Fusiform	Hippocampus
BR	0.65	0.60	0.57	0.66	0.62	0.69	0.49
DT	0.68	0.61	0.60	0.69	0.64	0.70	0.51

2.6 Conclusions and Contributions

The results indicate that features from an independently performed feature norming study or word co-occurrence in web corpus can explain a significant portion of the variance in neural activity in this task, suggesting that the features transfer well across tasks, and hence appear to correspond to enduring properties of the word representations. Moreover, the resulting regression model is useful for decoding mental states from their neural activation pattern. The ability to perform this classification task is remarkable, suggesting that the distributed pattern of neural activity encodes sufficient signal to discriminate differences among stimuli.

Our major contribution is to shift the focus to the hidden factors that underpin semantic representation of object knowledge. Functional neuroimaging research has been focused on attempting to identify of the functions of cortical regions. Here we present one of the first studies to investigate some intermediate cortex-wide representations of semantic knowledge and further apply it in a classification task. Akin to the recent multivariate fMRI analysis which shifted the focus from localizing brain activity toward understanding how patterns of neural activity encode information in an intermediate semantic representation, we take one further step and ask 1) what intermediate semantic representation might be encoded to enable such discrimination and 2) what is the nature of this representation?

There are several advantages to work with an intermediate semantic representation. In this study, we have demonstrated how learning the mapping between feature and neural activation enables a predictive theory that is capable of extrapolating the model of the neural activity to previously unseen words, which cannot be done with a discriminative classifier. Another advantage of working with an intermediate semantic representation is that features in the intermediate semantic representation are more likely to be shared across experiments. For example, in one experiment, the participant may be presented the word *dog*, while the word *cat* is shown in another experiment. Even though the individual category differs, there are many features that are shared (e.g. is a pet, has 4 legs, etc.) between the two words. Learning the mapping between features and voxel activation instead of the mapping between categories and voxel activation may facilitate data to be shared across experiments. This is especially important when brain imaging data are relatively more expensive to acquire and that many classifier techniques would perform significantly better if more training data were available.

Although we propose a specific implementation of the hidden layer representation with a multivariate multiple linear regression model estimated from features of a feature norming study, we do not necessarily commit to this specific implementation. We look forward to future research to extend the intermediate representation and experiment with different modeling methodologies. For instance, the intermediate semantic representation can be derived from research done in other related scientific characterizations of meaning, such as WordNet, LSA, or topic models. Another direction is to experiment with different modeling methodologies, such as neural networks which model non-linear functions or generative models of neural activities from a fully probabilistic, Bayesian perspective.

3 QUANTITATIVE MODELING OF THE NEURAL REPRESENTATION OF ADJECTIVE-NOUN PHRASES TO ACCOUNT FOR FMRI ACTIVATION

3.1 Introduction

Given these early succesess in using fMRI to discriminate categorial information and to model lexical semantic representations of individual words, it is interesting to ask whether a similar approach can be used to study the combinatorial aspects of human language. How is lexical semantic knowledge combined to form complex concepts? Does the distributed pattern of brain activity differ when a person is thinking about a dog versus when a person is thinking about a particular dog that is strong and muscular? To address these questions, we designed an object-contemplation task, where human participants were presented with 12 text labels of unmodified objects *(e.g. dog)* and modified objects *(e.g. strong dog)*. They were instructed to think of the properties of the stimulus object, while their brain activities were recorded by fMRI.

Mitchell and Lapata (2008) presented a framework for representing the meaning of phrases and sentences in vector space. They discussed how an additive model, a multiplicative model, a weighted additive model, a Kintsch (2001) model, and a model which combines multiplication and addition can be used to model human behavior in similiarity judgments when human participants were presented with a reference containing a subject-verb phrase *(e.g., horse ran)* and two landmarks *(e.g., galloped, dissolved)* and asked to choose which landmark was most similiar to the reference (in this case, *galloped*). They compared the composition models to human similarity ratings and found that all models were statistically significantly correlated with human judgements. Moreover, the multiplicative and combined model performed signficantlly better than the non-compositional models. In this study, vector-based models of semantic composition were used to model neural activation patterns obtained while subjects comprehended adjective-noun phrases. Our approach is similar to that of Mitchell and Lapata (2008) in that we compared additive and multiplicative models to non-compositional models in terms of their ability to model human data. Our work differs from these efforts because we focus on modeling neural activity while people comprehend adjective-noun phrases.

In section 2, we describe the experiment and how functional brain images were acquired. In section 3, we perform group-level analysis to determine brain regions that are activated in different experimental conditions. In section 4, we apply classifier analysis to see if the distributed pattern of neural activity contains sufficient signal to discriminate among phrases. In section 5, we discuss a vector-based approach to modeling the lexical semantic knowledge using word occurrence measures in a text corpus. Two composition models, namely the additive and the multiplicative models, along with two non-composition models, namely the adjective and the noun models, are used to explain the systematic variance in neural activation. Section 6 distinguishes between two types of adjectives that are used in our stimuli: attribute-specifying adjectives and object-changing adjectives. Classifier analysis suggests people interpret the two types of adjectives differently. Finally, we discuss some of the implications of our work and suggest some future studies.

3.2 Methods

Participants. Nineteen right-handed adults (aged 18 - 32) from the Carnegie Mellon community participated and gave informed consent approved by the University of Pittsburgh and Carnegie Mellon Institutional Review Boards. Four additional participants were excluded from the analysis due to head motion greater than 2.5 mm.

Experimental Paradigm. The stimuli were text labels of 12 concrete nouns from 4 semantic categories with 3 exemplars per category. The 12 nouns were *bear*, *cat*, *dog* (animal);

bottle, cup, knife (utensil); *carrot, corn, tomato* (vegetable); *airplane, train,* and *truck* (vehicle; see Table 3.1). The fMRI neural signatures of these objects have been found in previous studies to elicit differentiable neural activity. The participants were also shown each of the 12 nouns paired with an adjective, where the adjectives are expected to emphasize certain semantic properties of the nouns. For instance, in the case of *strong dog,* the adjective is used to emphasize the visual or physical aspect (e.g. muscular) of a *dog*, as opposed to the behavioral aspects (e.g. playable, shy).

Table 3.1 Word stimuli. Asterisks mark the object-changing adjectives, as opposed to the ordinary attribute-specifying adjectives.

Adjective	**Noun**	**Category**
Soft	Bear	Animal
Large	Cat	Animal
Strong	Dog	Animal
Plastic	Bottle	Utensil
Small	Cup	Utensil
Sharp	Knife	Utensil
Hard	Carrot	Vegetable
Cut	Corn	Vegetable
Firm	Tomato	Vegetable
Paper*	Airplane	Vehicle
Model*	Train	Vehicle
Toy*	Truck	Vehicle

Notice that the last three adjectives in Table 3.1 are marked by asterisks to denote they are *object-changing adjectives*. These adjectives appear to behave differently from the ordinary *attribute-specifying adjectives*. Section 6 discusses the different adjective types in more detail.

To ensure that participants had a consistent set of properties to think about, they were asked to generate and write a set of properties for each stimulus item in a session prior to the scanning session (such as “4 legs, house pet, fed by me” for *dog*). However, nothing was done to elicit consistency across participants.

The entire set of 24 stimuli was presented 6 times during the scanning session, in a different random order each time. Participants silently viewed the stimuli and were asked to think of the same item properties consistently across the 6 presentations of the items. Each stimulus was presented for 3s, followed by a 7s rest period, during which the participants were instructed to fixate on an X displayed in the center of the screen. There were two additional presentations of fixation, 31s each, at the beginning and end of each session, to provide a baseline measure of activity. A b representation of the design is shown in Figure 3.1.

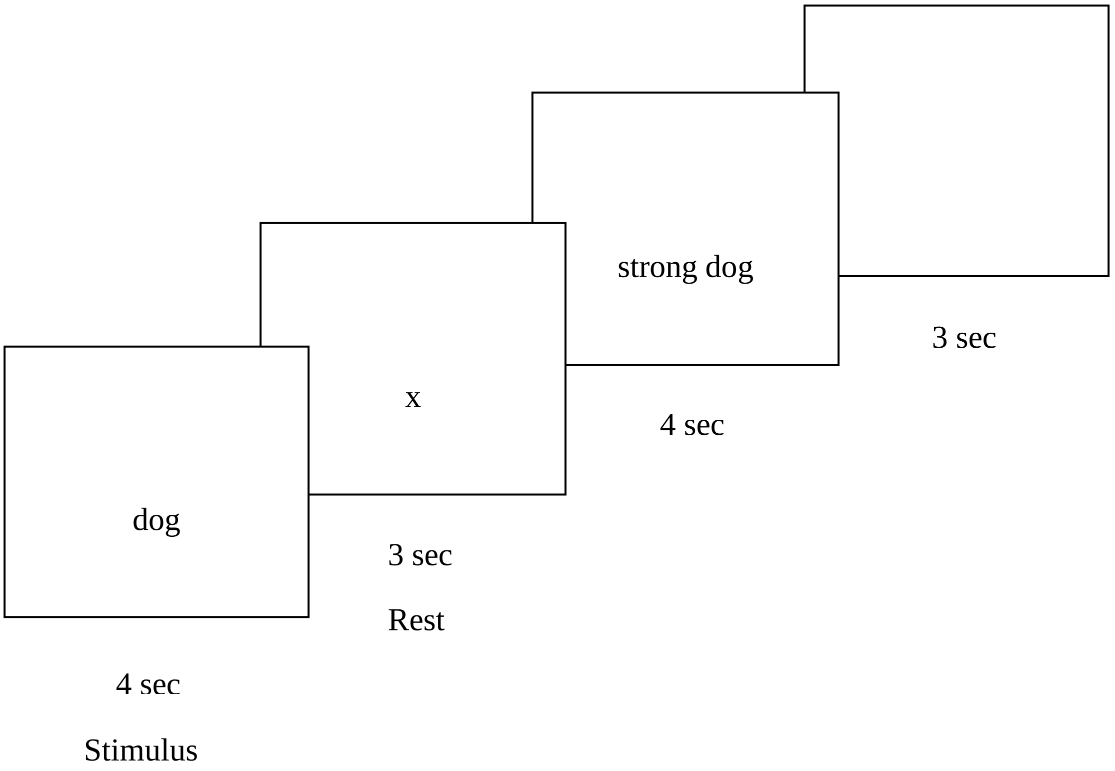


Figure 3.1 Schematic representation of experimental design for the adjective-noun experiment.

Data acquisition. Functional images were acquired on a Siemens Allegra 3.0T scanner (Siemens, Erlangen, Germany) at the Brain Imaging Research Center of Carnegie Mellon University and the University of Pittsburgh using a gradient echo EPI pulse sequence with TR = 1000 ms, TE = 30 ms, and a 60° flip angle. Seventeen 5-mm thick oblique-axial slices were imaged with a gap of 1-mm between slices. The acquisition matrix was 64 x 64 with 3.125 x 3.125 x 5-mm voxels.

Data processsing and analysis. Data processing and statistical analysis were performed with Statistical Parametric Mapping software (SPM2, Wellcome Department of Cognitive Neurology, London, UK; Friston, 2005). The data were corrected for slice timing, motion, and linear trend, and were temporally smoothed with a high-pass filter using a 190s cutoff. The data were normalized to the MNI template brain image using a 12-parameter affine transformation.

The data were prepared for classification analysis by being spatially normalized into MNI space and resampled to 3 x 3 x 6-mm^3 voxels. We try to keep approximately the same acquisition voxel size which has been used in many of our previous studies and is adequate for a list of different cognitive tasks. Voxels outside the brain were excluded from further analysis. The percent signal change (PSC) relative to the fixation condition was computed for each item presentation at each voxel. The mean of the four images (mean PSC) acquired within a 4s window, offset 4s from the stimulus onset (to account for the delay in hemodynamic response), provided the main input measure for subsequent analysis. The mean PSC data for each word presentation were further normalized to have mean zero and variance one to equate the variation between participants over exemplars. Due to the inherent limitations in the temporal properties of fMRI data, we consider here only the spatial distribution of the neural activity.

3.3 fMRI Analyses for SPM Group Contrasts

To compare the distribution of activation across experimental conditions, group t-test analyses were performed using a random-effects model (Friston et al., 2005). All t-maps in each contrast were calculated across the entire cortical volume, thresholded at an uncorrected height threshold of p < .001 and an extent threshold of 5 voxels. Statistical maps were superimposed on the high-resolution, normalized, T1-weighted, SPM2 individual template image for viewing. Labels for coordinates of activation were confirmed in MNI space (Tzourio-Mazoyer et al., 2002).

The brain activation for stimulus minus fixation was mostly left-lateralized, and included the areas of the inferior frontal gyrus, supplementary motor area, fusiform, middle temporal, hippocampus, inferior parietal, inferior occipital areas, as well as right middle frontal, insula, angular gyrus and Calcarine (Figure 3.2; Table 3.2).

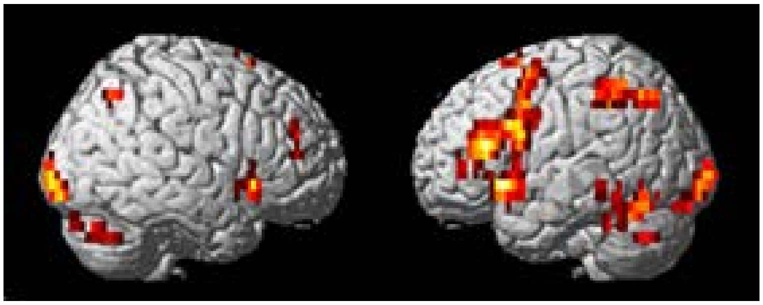


Figure 3.2 Brain activation for all stimuli contrasted with fixation, p<0.001 uncorrected; T=3.61; extent threshold voxels=5. The brain activation for stimulus minus fixation was mostly left-lateralized, and included the areas of the inferior frontal gyrus, supplementary motor area, fusiform, middle temporal, hippocampus, inferior parietal and inferior occipital areas.

Table 3.2 Locations (MNI centroid coordinates) and sizes of the voxel clusters of activation for all stimuli contrasted with fixation, p<0.001 uncorrected; T=3.61; extent threshold voxels=5.

Label	X	Y	Z	Voxels	Radius
Frontal					
L Frontal Inf Oper	-43.9	15.9	21.3	354	22.2
L Supp Motor Area	-1.1	18.6	48	198	13.98
R Frontal Mid	47.1	43	21	12	7.37
Temporal					
L Fusiform	-46.7	-55.8	-18.9	76	10.87
R Insula	42.3	17.6	-6	33	7.42
L Temporal Mid	-53.1	-39.1	-8	6	3.96
L Hippocampus	-17.5	-33.8	-2.4	5	4.35
Parietal					
L Parietal Inf	-35.1	-52.5	44.5	158	12.6
R Angular	34.9	-61.2	45	22	6.29
Occipital					
R Calcarine	20.6	-97.6	-4.6	69	8.66
L Occipital Inf	-22.6	-96.9	-6.2	60	10.95
Subcortical					
R Cerebelum Crus1	32.2	-70.7	-33.5	44	8.61
R Caudate	19	8.9	11.1	35	8.84
R Cingulum Mid	0.5	-27.4	28.9	32	6.91
R Cerebelum Crus2	10.9	-83.6	-28.4	19	6.3
L Putamen	-15.6	-0.8	12.4	16	5.82
L Cerebelum Crus1	-44.1	-65.3	-36	10	3.84
L Putamen	-19.1	10.2	-3	8	4.29

The isolated noun contrast minus adjective-noun phrase reveals brain activation at left insula (Figure 3.3). There was no activation for the adjective-noun phrase minus isolated noun contrast. Subsequently, we perform this contrast separately for each of the four categories: animal, utensil, vegetable and vehicles. Notice that because a threshold of p < .001 yields no positive activation, we hereby relax the threshold to p < .01. The brain activation for animal phrases minus animals and utensil phrases minus utensils was mostly bi-lateralized in the occipital lobe, and included the areas of the middle occipital gyrus, lingual and Cuneus (Figure

3.4; Table 3.3). This pattern is consistent with the fact that most of the adjectives in animal and utensil phrases are used to emphasize the visual or physical aspect of the noun (e.g. *strong dog* has a muscular appearance). Though, the occipital activation may also be a consequence of the difference in word length (adjective-noun phrases are always longer than isolated nouns in word length). Moreover, the brain activation for vegetable phrases minus vegetables is left-lateralized in the inferior and middle occipital gyrus, as well as the left supra marginal gyrus in the parietal lobe. This pattern is consistent with the fact that most of the adjectives in vegetable phrase are used to emphasize the tactile aspects of the noun (e.g. *hard carrot* is a dense or firm carrot). Finally, the brain activation for vehicle phrase minus vehicles include left-lateralized activity in the middle occipital area, Cuneus, and supra marginal gyrus in the parietal lobe and right-lateralized activity in the superior temporal gyrus. Notice that adjectives used in vehicle phrase tend to change the meaning of the noun. The activation in the Wernicke’s area may correspond to the cognitive process of experiencing concept combinations.

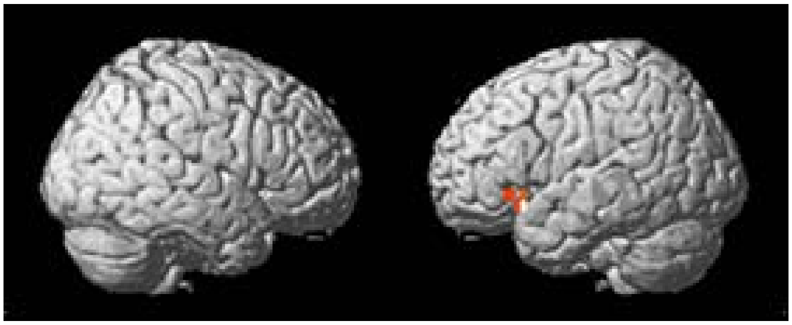


Figure 3.3 Brain activation for isolated noun stimuli contrasted with adjective-noun phrase stimuli, p<0.001 uncorrected; T=3.61; extent threshold voxels=5. The isolated noun contrast minus adjective-noun phrase reveals brain activation at left insula.

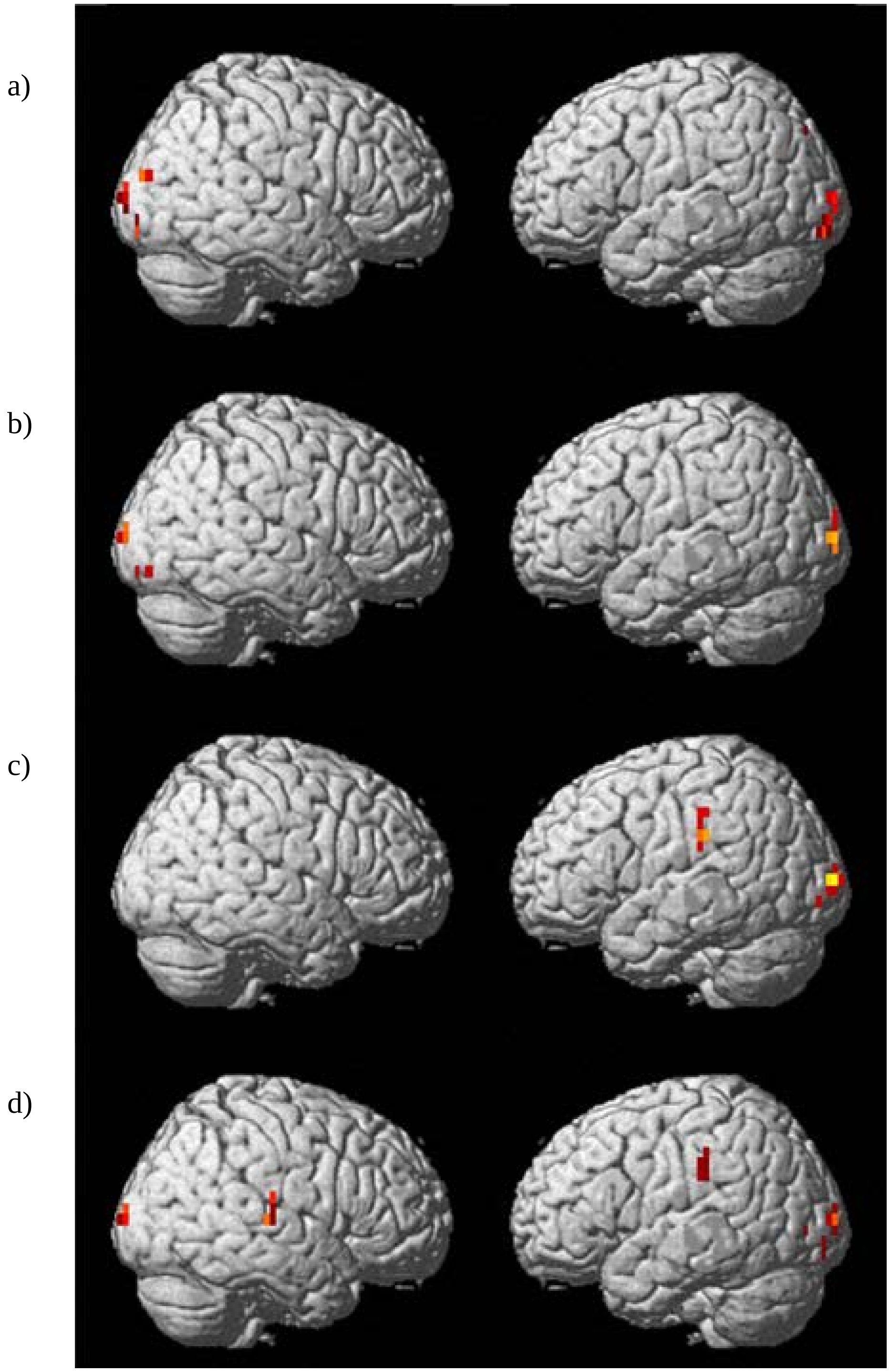


Figure 3.4 Brain activation for adjective-noun phrase stimuli contrasted with isolated-noun stimuli, for each of the four categories: a) animal; b) utensil; c) vegetable; d) vehicle, p<0.01 uncorrected T=2.55; extent threshold voxels=5. The brain activation for animal phrases minus animals and utensil phrases minus utensils was mostly bi-lateralized in the occipital lobe, suggesting an emphasis on the visual aspects. The brain activation for vegetable phrases minus vegetables is left-lateralized and include activation in the left supra marginal gyrus in the parietal lobe, suggesting an emphasis on the tactile aspects. The brain activation for vehicle phrase minus vehicles include right-lateralized activity in the superior temporal gyrus, suggesting the participants may be experiencing concept combination.

Table 3.3 Locations (MNI centroid coordinates) and sizes of the voxel clusters of activation for phrase stimuli contrasted with isolated word stimuli, p<0.01 uncorrected; T=2.55; extent threshold voxels=5.

Category	Label	X	Y	Z	Voxels	Radius
Animal	*Occipital*					
	L Occipital Mid	-17.7	-96	-3.7	18	7.33
	R Lingual	18.8	-92.4	-6.3	18	6.36
	R Cuneus	19.8	-101	9	6	3.65
	R Occipital Mid	29.4	-89.4	18	5	3.24
	Subcortical					
	L Caudate	-11.6	20.1	16.3	7	5.43
Utensil	*Occipital*					
	R Lingual	17.5	-90.6	-12	5	3.37
	L Occipital Mid	-16.9	-98.4	5.4	10	5.51
	R Cuneus	19.8	-101	9	6	3.65
Vegetable	*Parietal*					
	L SupraMarginal	-61.3	-29.3	33	8	6.26
	Occipital					
	L Occipital Inf	-21.2	-88.1	-6	5	4.11
	L Occipital Mid	-16.7	-97.5	3	14	5.29
Vehicle	*Temporal*					
	R Temporal Sup	40	-41.2	3.6	5	4.33
	R Temporal Sup	58.3	-22.9	11	6	5.82
	Parietal					
	L SupraMarginal	-62.5	-30	34.8	5	4.28
	Occipital					
	L Occipital Mid	-18.8	-96.9	1.8	10	7.07
	L Occipital Mid	-24.1	-80.8	-2.6	7	4.44
	R Cuneus	20	-100.6	8.4	5	3.51

In short, the contrast of phrase versus word revealed occipital activation for modifiers that emphasizes visual aspects, parietal activation for modifiers that emphasizes tactile aspects, right superior temporal activation for modifiers that changes the meaning of noun.

3.4 Does the distribution of neural activity encode sufficient signal to classify adjective-noun phrases?

Given the observed neural activity when participants processed the adjective-noun phrases, Gaussian Naïve Bayes classifiers were trained to identify cognitive states associated with processing nouns and phrases from the evoked patterns of functional activity (mean PSC). For instance, the classifier would predict which of the 24 exemplars the participant was viewing and thinking about. Separate classifiers were also trained for classifying the isolated nouns, the phrases, and the 4 semantic categories.

3.4.1 Classifier Model

Classifiers were trained to identify cognitive states associated with viewing stimuli from the evoked pattern of functional activity (mean PSC). Classifiers were functions f of the form: f: mean_PSC → Yi, i=1,…n, where Yi were the sixty exemplars, and mean_PSC was a vector of mean PSC voxel activation level, as described above. The Gaussian Naïve Bayes (GNB) pooled variance classifier was used (Mitchell 1997). It is a generative classifier that models the joint distribution of a class Y (exemplar or category) and attributes X (voxels), and assumes the attributes X1,…,Xn are conditionally independent given Y. The classification rule is:

$$Y \leftarrow \arg\max_{y_j} \mathrm{P(Y = y_j)} \prod_{\mathrm{i}}^{\mathrm{n}} P(X_i \mid Y = y_j),\ j = 1{,}2\ .$$

Classification results were evaluated using 6-fold cross validation, where one of the 6 repetitions was left out for each fold. The voxel selection procedure (described below) was performed separately inside each fold, using only the training data. Since multiple classes were involved, rank accuracy was used (Mitchell et al., 2004) to evaluate the classifier. Given a new fMRI image to classify, the classifier outputs a rank-ordered list of possible class labels from most to least likely. The rank accuracy is defined as the percentile rank of the correct class in this ordered output list. Rank accuracy ranges from 0 to 1. Classification analysis was performed separately for each participant, and the mean rank accuracy was computed over the participants.

3.4.2 Voxel Selection

Since fMRI acquires the neural activity in the entire brain (15,000 – 20,000 distinct voxel locations, in our parcellation), many locations might not exhibit neural activity that encodes word or phrase meaning. Thus, the classifier analysis selected the voxels whose responses to the different items were most stable across presentations. Voxel stability was computed as the average pairwise correlation between 24 item vectors across presentations, using only the training set within each fold in the cross-validation paradigm. The focus on the most stable voxels effectively increased the signal-to-noise ratio in the data and also served as a dimensionality reduction tool that facilitated further analysis by classifiers. Many of our previous analyses have indicated that 120 voxels is a set size suitable for our purposes (Just et al., 2010). Our theoretical framework considers all cortical voxels and allows the training data to determine which locations are systematically modulated by which aspects of word meanings.

3.4.3 Results and Discussion

Table 3.4 shows the results of the exemplar-level classification analysis. All classification accuracies were significantly higher than chance ($p < 0.05$), where the chance level for each classification is determined based on the empirical distribution of rank accuracies for randomly generated null models. One hundred null models were generated by permuting the class labels.

The classifier was able to distinguish among the 24 exemplars with mean rank accuracies close to 70%. The classification accuracies were also determined separately for nouns only and phrases only. Distinct classifiers were trained. Classification accuracies were significantly higher (p < 0.05) for the nouns than the phrases, calculated with a paired *t*-test. For 3 participants, the classifier did not achieve reliable classification accuracies for the phrase stimuli. Moreover, the classification accuracies were determined separately for each semantic category of stimuli. There were no significant differences in accuracy across categories, except for the difference between vegetables and vehicles.

Table 3.4 Rank accuracies for classifiers. Distinct classifiers were trained to distinguish all 24 examples, nouns only, phrases only, and only words within each of the 4 semantic categories.

Categories of concepts to train the classifier	**Rank accuracy**
All 24 exemplars	0.69
Nouns	0.71
Phrases	0.64
Animals	0.67
Tools	0.66
Vegetables	0.65
Vehicles	0.69

High classification accuracies indicate that the distributed pattern of neural activity does encode sufficient signal to discriminate differences among stimuli. The classification accuracy for the nouns was comparable to previous research, providing a replication of previous findings (Mitchell et al, 2004). The classifiers performed better on the nouns than the phrases, consistent with the expectation. It is easier for participants to recall properties associated with a familiar object than to comprehend a noun whose meaning is further modified by an adjective. The classification analysis also helps to identify participants whose mental representations for phrases are consistent across phrase presentations. Subsequent regression analysis on phrase activation was based on subjects who performed the phrase task well.

3.5 Using vector-based models of semantic representation to account for the systematic variances in neural activity

3.5.1 Lexical Semantic Representation

Computational linguistics has demonstrated that a word's meaning is captured to some extent by the distribution of words and phrases with which it commonly co-occurs (Church and Hanks, 1990). Consequently, Mitchell et al. (2008) coded the meaning of a word as a vector of intermediate semantic features computed from the co-occurrences with stimulus words within the Google trillion-token text corpus that captures the typical use of words in English text. Motivated by existing conjectures regarding the centrality of sensory-motor features in neural representations of objects (e.g. Caramazza and Shelton, 1998), they selected a set of 25 semantic features defined by co-occurrence with 25 verbs: *see, hear, listen, taste, smell, eat, touch, rub, lift, manipulate, run, push, fill, move, ride, say, fear, open, approach, near, enter, drive, wear, break,* and *clean*. These verbs generally correspond to basic sensory and motor activities, actions performed on objects, and actions involving changes in spatial relationships.

Because there are only 12 stimuli in our experiment, we consider only 5 sensory verbs (*see hear, smell, eat* and *touch*) to avoid overfitting with the full set of 25 verbs. Following the work of Bullinaria and Levy (2007), we consider the "basic semantic vector" which normalizes *n(c,t)*, the count of times context word *c* occurs within a window of 5 words around the target word *t*. The basic semantic vector is thus the vector of conditional probabilities,

$$p(c \mid t) = \frac{p(c,t)}{p(t)} = \frac{n(c,t)}{\sum_{c} n(c,t)}$$

where all components are positive and sum to one. Table 3.5 shows the semantic representation for *strong* and *dog*. Notice that *strong* is heavily loaded on *see* and *smell*, whereas *dog* is heavily loaded on *eat* and *see*, consistent with the intuitive interpretation of these two words.

Table 3.5 The lexical semantic representation for strong and dog.

Concept	**See**	**Hear**	**Smell**	**Eat**	**Touch**
Strong	0.63	0.06	0.26	0.03	0.03
Dog	0.34	0.06	0.05	0.54	0.02

3.5.2 Semantic Composition

We adopt the vector-based semantic composition models discussed in Mitchell and Lapata (2008). Let *u* and *v* denote the meaning of the adjective and noun, respectively, and let *p* denote the composition of the two words in vector space. We consider two non-composition models, the adjective model and the noun model, as well as two composition models, the additive model and the multplicative model.

The adjective model assumes that the meaning of the composition is the same as the adjective:

$$p = u$$

The noun model assumes that the meaning of the composition is the same as the noun:

$$p = v$$

The adjective model and the noun model correspond to the assumption that when people comprehend phrases, they focus exclusively on one of the two words. This serves as a baseline for comparison to other models.

The additive model assumes the meaning of the composition is a linear combination of the adjective and noun vector:

$$p = A \cdot u + B \cdot v$$

where A and B are scalars of weighting coefficients.

The multiplicative model assumes the meaning of the composition is the element-wise product of the two vectors:

$$p = C \cdot u \cdot v$$

Mitchell and Lapata (2008) fitted the parameters of the weighting vectors A, B, and C, though we assume A = B = C = 1, since we are interested in the model comparison. Also, there are no model complexity issues, since the number of parameters in the four models is the same.

More critically, the additive model and multiplicative model correspond to different cognitive processes. On one hand, the additive model assumes that people concatenate the meanings of the two words when comprehending phrases. On the other hand, the multiplicative model assumes that the contribution of *u* is scaled to its relevance to *v*, or vice versa. The assumption of the multiplicative model corresponds to the modifier-head interpretation where adjectives are used to modify the meaning of nouns. Notice that as a result of a symmetric operator, the multiplicative model is insensitive to word order. Yet, the modifier-head relationship is not symmetric. For instance, reversing the word order of a phrase (e.g. *dog strong*) may result in a different syntactic structure (e.g. adverb-adjective), and can mean something very different (e.g. strong like a dog). Although Mitchell and Lapata (2008) described how these composition models may be extended to relax the symmetric assumption, the simplified models suffice for our purposes since we only consider adjective-noun phrases and not other syntactic structures. To foreshadow the results, the modifier-head interpretation of the multiplicative model provided the best account for the neural activity observed in adjective-noun phrase data.

Table 3.6 shows the semantic representation for *strong dog* under each of the four models. Although the multiplicative model appears to have small loadings on all features, the relative distribution of loadings still encodes sufficient information, as our later analysis will show. Notice how the additive model concatenates the meaning of two words and is heavily loaded on see, eat, and smell, whereas the multiplicative model zeros out unshared features like eat and smell. As a result, the multiplicative model predicts that the visual aspects will be emphasized when a participant is thinking about strong dog, while the additive model predicts that, in addition, the behavioral aspects (e.g., eat, smell, and hear) of dog will be emphasized.

Table 3.6 The semantic representation for strong dog under the adjective, noun, additive, and multiplicative models.

Semantic Composion	See	Hear	Smell	Eat	Touch
Adjective model	0.63	0.06	0.26	0.03	0.03
Noun model	0.34	0.06	0.05	0.54	0.02
Additive model	0.96	0.12	0.31	0.57	0.04
Multiplicative model	0.21	0.00	0.01	0.01	0.00

Notice that these 4 vector-based semantic composition models ignore word order. This corresponds to the bag-of-words assumption, such that the representation for *strong dog* will be the same as that of *dog strong*. The bag-of-words model is used as a simplifying assumption in several semantic models, including LSA (Landauer & Dumais, 1997) and topic models (Blei, Ng, & Jordan, 2003).

There were two main hypotheses that we have tested. First, people usually regard the noun in the adjective-noun pair as the linguistic head. Therefore, meaning associated with the noun should be more evoked. Thus, we predicted that the noun model would outperform the adjective model. Second, people make more interpretations that use adjectives to modify the meaning of the noun, rather than disjunctive interpretations that add together or take the union of the semantic features of the two words. Thus, we predicted that the multiplicative model would outperform the additive model.

3.5.3 Regression Fit

In this analysis, we train a regression model to fit the activation profile for the 12 phrase stimuli. We focused on subjects for whom the classifier established reliable classification accuracies for the phrase stimuli. The regression model examined to what extent the semantic feature vectors (explanatory variables) can account for the variation in neural activity (response variable) across the 12 stimuli. All explanatory variables were entered into the regression model simultaneously. More precisely, the predicted activity *a*v at voxel *v* in the brain for word *w* is given by

$$a_v = \sum_{i=1}^{n} \beta_{vi} f_i(w) + \varepsilon_v$$

where $f_i(w)$ is the value of the i^{th} intermediate semantic feature for word *w*, β_{vi} is the regression coefficient that specifies the degree to which the i^{th} intermediate semantic feature activates voxel *v*, and ε_v is the model's error term that represents the unexplained variation in the response variable. Least squares estimates of β_{vi} were obtained to minimize the sum of squared errors in reconstructing the training fMRI images. An L2 regularization with lambda = 1.0 was added to prevent overfitting given the high parameter-to-data-points ratios. A regression model was trained for each of the 120 voxels and the reported R^2 is the average across the 120 voxels. R^2 measures the amount of systematic variance explained by the model. Regression results were evaluated using 6-fold cross validation, where one of the 6 repetitions was left out for each fold.

Linear regression assumes a linear dependency among the variables and compares the variance due to the independent variables against the variance due to the residual errors. While the linearity assumption may be overly simplistic, it reflects the assumption that fMRI activity often reflects a superposition of contributions from different sources, and has provided a useful first order approximation in the field (Mitchell et al., 2008).

3.5.4 Results and Discussion

The second column of Table 3.7 shows the R^2 regression fit (averaged across 120 voxels) of the adjective, noun, additive, and multiplicative model to the neural activity observed in adjective-noun phrase data. The noun model significantly ($p < 0.05$) outperformed the adjective model, estimated with a paired *t*-test. Moreover, the difference between the additive and adjective models was not significant, whereas the difference between the additive and noun models was significant ($p < 0.05$). The multiplicative model significantly ($p < 0.05$) outperformed both of the non-compositional models, as well as the additive model.

More importantly, the two hypotheses that we were testing were both verified. Notice Table 3.7 supports the hypothesis that the noun model should outperform the adjective model based on the assumption that the noun is generally more central to the phrase meaning than is the adjective. Table 3.7 also supports our hypothesis that the multiplicative model should outperform the additive model, based on the assumption that adjectives are used to emphasize particular semantic features that will already be represented in the semantic feature vector of the noun. Our findings here are largely consistent with Mitchell and Lapata (2008).

Table 3.7 Regression fit and regression-based classification rank accuracy of the adjective, noun, additive, and multiplicative models for phrase stimuli.

Semantic Composition	**R^2**	**Rank accuracy**
Adjective model	0.34	0.57
Noun model	0.36	0.61
Additive model	0.35	0.60
Multiplicative model	0.42	0.62

Following Mitchell et al. (2008), the regression model can be used to decode mental states. Specifically, for each regression model, the estimated regression weights can be used to generate the predicted activity for each word. Then, a previously unseen neural activation vector is identified with the class of the predicted activation that had the highest correlation with the given observed neural activation vector. Notice that, unlike Mitchell et al. (2008), where the regression model was used to make predictions for items outside the training set, here we are just showing that the regression model can be used for classification purposes.

The third column of Table 3.7 shows the rank accuracies classifying concepts using the predicted activation from the adjective, noun, additive, and multiplicative models. All rank accuracies were significantly higher ($p < 0.05$) than chance, where the chance level for each classification is again determined by permutation testing. More importantly, here we observe a ranking of these four models similar to that observed for the regression analysis. Namely, the noun model performs significantly better ($p < 0.05$) than the adjective model, and the multiplicative model performs significantly better ($p < 0.05$) than the additive model. However, the difference between the multiplicative model and the noun model is not statistically significant in this case.

3.6 Comparing the attribute-specifying adjectives with the object-changing adjectives

Some of the phrases contained adjectives that changed the meaning of the noun. In the case of vehicle nouns, adjectives were chosen to modify the manipulability of the nouns (e.g., to make an *airplane* more manipulable, *paper* was chosen as the modifier). This type of modifier raises two issues. First, these modifiers *(e.g. paper, model, toy)* more typically assume the part of speech (POS) tag of nouns, unlike our other modifiers *(e.g., soft, large, strong)* whose typical POS tag is adjective. Second, these modifiers combine with the noun to denote a very different object from the noun in isolation *(e.g. paper airplane, model train, toy truck),* in comparison to other cases where the adjective simply specifies an attribute of the noun *(e.g. large cat, strong dog).* In order to study this difference, we performed classification analysis separately for the attribute-specifying adjectives and the object-changing adjectives.

Our hypothesis is that the phrases with attribute-specifying adjectives will be much more difficult to distinguish from the original nouns than the adjectives that change the referent. For instance, we hypothesize that it is much more difficult to distinguish the neural representation for *strong dog* versus *dog* than it is to distinguish the neural representation for *paper airplane* versus *airplane*. To verify this, Gaussian Naïve Bayes classifiers were trained to discriminate between each of the 12 pairs of nouns and adjective-noun phrases. The average classification for phrases with object-changing adjectives is 0.76, whereas classification accuracies for phrases with attribute-specifying adjectives are 0.68. The difference is statistically significant at $p < 0.05$. This result supports our hypothesis.

Furthermore, we performed regression-based classification separately for the two types of adjectives. Notice that the number of phrases with object-changing adjectives is much less than the number of phrases with attribute-specifying adjectives (3 vs. 9). This affects the parameter-to-data-points ratio in our regression model. Consequently, an L2 regularization with lambda = 10.0 was used to prevent overfitting. Table 3.8 shows a pattern similar to that seen in section 4 is observed for the attribute-specifying adjectives. That is, the noun model outperformed the adjective model and the multiplicative model outperformed the additive model when using attribute-specifying adjectives. However, for the object-changing adjectives, the noun model no longer outperformed the adjective model. Moreover, the additive model performed better than the noun model. Although neither difference is statistically significant, this clearly shows a pattern different from the attribute-specifying adjectives. This result suggests that when interpreting phrases like *paper airplane*, it is more important to consider contributions from the adjectives, compared to when interpreting phrases like *strong dog*, where the contribution from the adjective is simply to specify a property of the item typically referred to by the noun in isolation.

Table 3.8 Separate regression-based classification rank accuracy for phrases with attribute-specifying or object-changing adjectives.

Semantic Composion	Attribute-specifying adjective	Object-changing adjective
Adjective model	0.57	0.65
Noun model	0.62	0.64
Additive model	0.61	0.65
Multiplicative model	0.63	0.67

3.7 Contribution and Conclusion

Experimental results have shown that the distributed pattern of neural activity while people are comprehending adjective-noun phrases does contain sufficient information to decode the stimuli with accuracies significantly above chance. Furthermore, vector-based semantic models can explain a significant portion of systematic variance in observed neural activity. Multiplicative composition models outperform additive models, a trend that is consistent with the assumption that people use adjectives to modify the meaning of the noun, rather than conjoining the meaning of the adjective and noun.

In this study, we represented the meaning of both adjectives and nouns in terms of their co-occurrences with 5 sensory verbs. While this type of representation might be justified for concrete nouns (hypothesizing that their neural representations are largely grounded in sensory-motor features), it might be that a different representation is needed for adjectives. Further research is needed to investigate alternative representations for both nouns and adjectives. Moreover, the composition models that we presented here are overly simplistic in a number of ways. We look forward to future research to extend the intermediate representation and to experiment with different modeling methodologies.

Due to the inherent limitations in the temporal properties of fMRI data, in most of this thesis work we consider only the spatial distribution of the brain activity after the stimuli are comprehended and do not attempt to model the cognitive process of comprehension. One extension is to see if the temporal resolution of fMRI encodes sufficient signal to model the process of combination and not just the comprehended concepts.

4 QUANTITATIVE MODELING OF THE NEURAL REPRESENTATION OF NOUN-NOUN CONCEPT COMBINATION

4.1 Introduction

Conceptual combination is the process in which complex concepts *(e.g. coffee shop)* are constructed from basic concepts (e.g. *coffee* and *shop*). An improved understanding of semantic composition in multi-word phrases is an important step toward accounts of sentence processing. There has been extensive study of the two different types of combination rules that people used when interpreting noun-noun concept combination, namely the property-based interpretation and relation-based interpretations. On one hand, in property-based interpretation, one property (e.g., shape, color, size) of the modifier object is applied to modify the head object. For example, the interpretation that *corn coat* is a coat that is bright yellow is a type of property-based interpretation. On the other hand, in relation-based interpretation, the modifier object is realized in its entirety and related to the head object as a whole. For example, the interpretation that *corn coat* is a coat that is used to protect corn is a type of relation-based interpretation.

Baron et al. (in press) used a categorization task to evoke patterns of neural activiation for complex concepts *(e.g. young man)* as well as the constituents *(e.g. young, man)*. They found that the superimposition of activity for constituents at left anterolateral temporal lobe reliably predicted activation pattern for the complex concepts. Though, they used computer generated faces to represent the combined concepts, which could potentially reflect attention to distinctive visual features rather than true conceptual meaning. Graves et al. (2010) studied familiar, highly meaningful phrases *(e.g. lake house)* and unfamiliar, minimally meaningful phrases created by reversing the word order of the familiar phrases *(e.g. house lake)*. They found a hemispheric dissociation between levels of semantic representation: lexical processing is more correlated with activation in the left hemisphere, whereas combinatorial semantic processing is more correlated with activation in the right hemisphere.

In the present work, fMRI data is used to study the brain activity when people comprehend noun-noun concept combinations with a property-based or a relation-based interpretation. In an object-contemplation task, participants were shown the 10 noun-noun phrases with accompanying contexts that either bias toward property-based or relation-based interpretations. That is, the participant was expected to contemplate a property-based interpretation in one context, but a relation-based interpretation when the same noun-noun phrase was presented in another context. They were instructed to think of the same properties of the stimulus object consistently during each presentation. Given the brain activity signatures evoked by this visual presentation, multivariate machine learning classifier is estimated to decode whether a participant is thinking about a property-based interpretation or a relation-based interpretation. The setup of the experiment poses a challenge for classifiers that obtain its discriminative power from distinguishing the brain activity of low-level visual perceptions. Since the visual stimuli are identical, the discrimination must be made on the semantic differences between the two types of interpretations.

In section 2, we describe the experiment and how functional brain images were acquired. In section 3, we show that the distributed pattern of brain activity encodes sufficient signal to discriminate among different interpretations of the same phrase. In section 4, we perform group-level analysis to determine brain regions that are activated in different experimental conditions. In section 5, we compare the brain activation of isolated concepts and concept combinations to study the neural underpinning of the semantic composition. Finally, we discuss some of the implications of our work and suggest some future studies.

4.2 Material and Methods

Participants. Ten right-handed adults (5 female, age between 18 and 32) from the Carnegie Mellon community participated and gave informed consent approved by the University of Pittsburgh and Carnegie Mellon Institutional Review Boards.

Experimental paradigm. The stimuli were text labels of 10 noun-noun phrases: *window cup, cow chair, corn coat, bell dress, bee, airplane, pliers hand, dog beetle, refrigerator house, celery table*, and *tomato ant*. The objects in these phrases were chosen from Mitchell et al. (2008) where the fMRI neural signatures of these objects have been found to elicit different brain activity. The participants were shown the noun phrases with accompanying contexts that either bias toward property-based or relation-based interpretations. For example, a context like *"Sally was known for always choosing clothes that matched her light blond hair, like that eye-catching ..."* will lead the participant to interpret a *corn coat* as a coat that is bright yellow (a property-based interpretation where the color of a corn is mapped to a coat). On the other hand, a context like "*A severe thunderstorm was expected, so the farmer protected each of his crops with their own ..."* will lead the participant to interpret a *corn coat* as a coat that is used to protect corn (a relation-based interpretation where the modifier object is realized in its entirety and related to the head object as a whole). Table 4.1 and Table 4.2 show the contextual passages that were used to induce property and relation-based interpretations, respectively. The length of the contextual sentence is controlled and has an average of 17.5 words in contexts that bias toward property-based interpretations and 17.9 words in contexts that bias toward relation-based interpretations.

Table 4.1 Contextual passage to induce property-based interpretations

Property-based Context	**Phrase**
The mug has panels of clear glass that allow light to pass through; it is called a	window cup

...	
The furniture store was successful in selling all of its animal-print pieces, except for one last ...	cow chair
Sally was known for always choosing clothes that matched her light blond hair, like that eye-catching ...	corn coat
The little girl admired the wedding gown, with its dramatic, puffy skirt that she called a ...	bell dress
The villagers could hear the buzzing engine before they saw the black and yellow wings of the ...	bee airplane
Mark is so strong that he can saw through a copper pipe while grasping it with his bare ...	pliers hand
After discovering a new insect, Ann thought that due to its wagging tail, she would name it the ...	dog beetle
Molly's parents keep the thermostat set to such a low temperature that Jerry calls it a ...	refrigerator house
Her living room was decorated with modern pieces, in green colors with long, straight lines, especially her ...	celery table
Ricky screamed and got goosebumps when he found, crawling on his arm, a big, fat, red ...	tomato ant

Table 4.2 Contextual passage to induce relation-based interpretation

Relation-based Context	**Phrase**
John often wakes up thirsty, and since he doesn't have a bedside table, he keeps water in a ...	window cup
The rancher was notorious for lavishly pampering his herd, going so far as to build a ...	cow chair
A severe thunderstorm was expected, so the farmer protected each of his plants with their own ...	corn coat
The church's call to services was not as loud as usual due to the muffling effect of the ...	bell dress
When Sara had to move her honey farm across the country, she needed to rent a ...	bee airplane
Mark makes his living as a plumber, so he is very careful not to injure his ...	pliers hand
Fido has been sad lately because of the many itchy bites that he got from the ...	dog beetle
Before being sent to the store to be purchased, the large electrical appliances are kept in a ...	refrigerator house
Ralph's kitchen is so orderly because every food has its own place, like the fruit bowl and the ...	celery table
The gardener got angry when he saw that his prize-winning plants were overrun by a nasty kind of ...	tomato ant

To ensure that participants had a consistent set of properties to think about, they were each asked to generate and write a set of properties for each exemplar in a session prior to the scanning session. They were asked to describe the object in the given context in one sentence and also answer three questions: what does it look like (appearance), how do you physically interact with it (interaction), and for what purpose is it used (purpose)? However, nothing was done to elicit consistency across participants. The entire set of 10 stimuli was presented 6 times under each context during the scanning session, in a different random order each time. The contextual sentence is presented for 4s, followed by a 3s rest. Then, the participant is presented with the noun-noun phrase for 4s. Participants silently viewed the stimuli and were asked to think about the object in the given context and mentally go over the same set of properties (appearance, interaction, purpose) consistently across the 6 presentations of the items. There is a 7s rest period before the next stimulus item is presented, during which the participants were instructed to fixate on an X displayed in the center of the screen. We also record the brain activity when each noun in the noun-noun phrases is presented in isolation, which we call the "word" condition. There were two additional presentations of fixation, 31s each, at the beginning and end of each session, to provide a baseline measure of activity. A schematic representation of the design is shown in Figure 4.1.

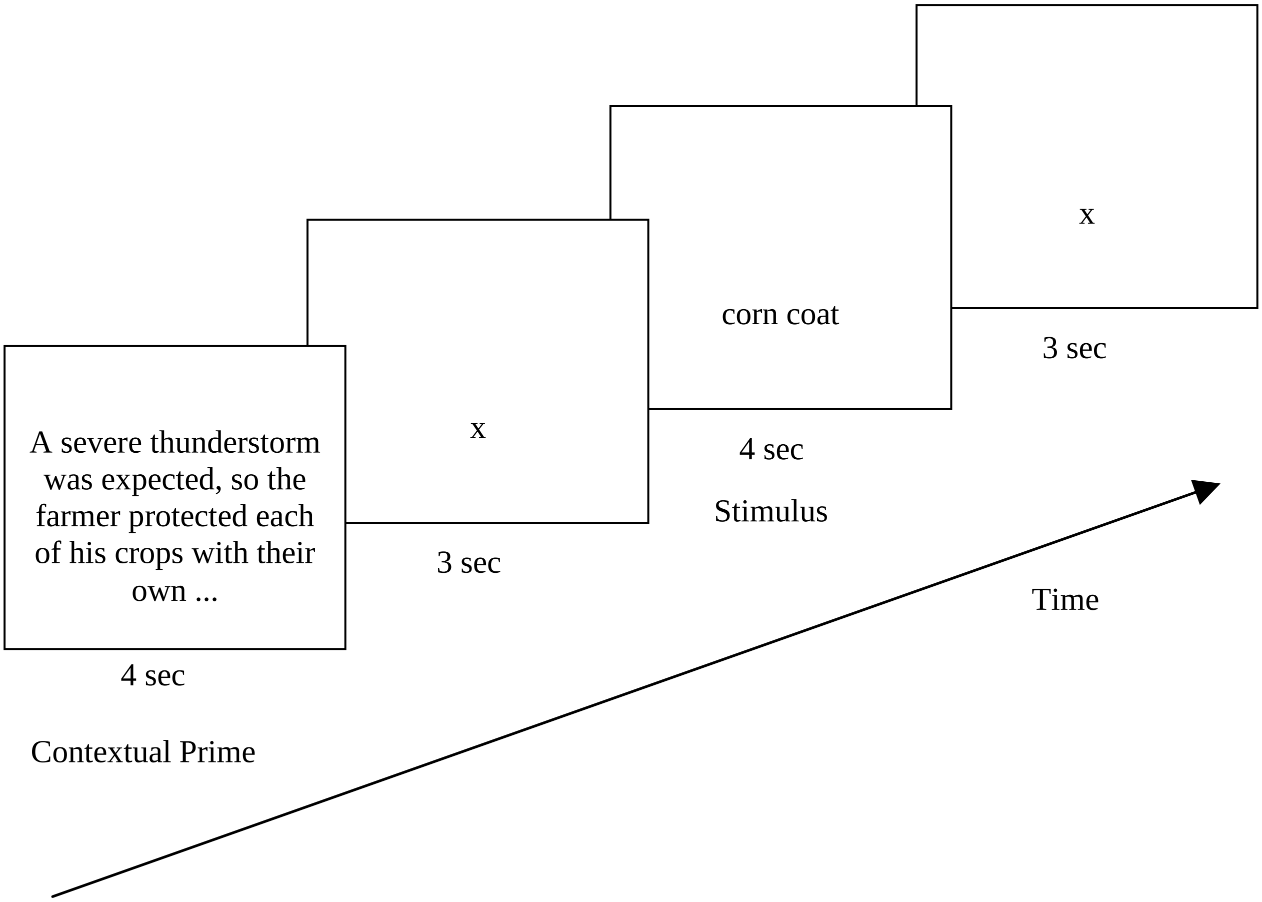

A severe thunderstorm was expected, so the farmer protected each of his crops with their own ...
4 sec
Contextual Prime
x
3 sec
corn coat
4 sec
Stimulus
x
3 sec
Time

Figure 4.1 Schematic representation of experimental design for the noun-noun concept combination experiment.

Data acquisition. Functional images were acquired on a Siemens Allegra 3.0T scanner (Siemens, Erlangen, Germany) at the Brain Imaging Research Center of Carnegie Mellon University and the University of Pittsburgh using a gradient echo EPI pulse sequence with TR = 1000 ms, TE = 30 ms and a 60° flip angle. Seventeen 5-mm thick oblique-axial slices were imaged with a gap of 1-mm between slices. The acquisition matrix was 64 x 64 with 3.125 x 3.125 x 5-mm voxels.

Data processing and analysis. Data processing and statistical analysis were performed with Statistical Parametric Mapping software (SPM2, Wellcome Department of Cognitive

Neurology, London, UK; Friston, 2005). The data were corrected for slice timing, motion, linear trend, and were temporally smoothed with a high-pass filter using 190s cutoff. The data were normalized to the MNI template brain image using 12-parameter affine transformation.

The data were prepared for classification analysis by being spatially normalized into MNI space and resampled to 3x3x6 mm^3 voxels. We try to keep approximately the same acquisition voxel size which has been used in many of our previous studies and is adequate for a list of different cognitive tasks. Voxels outside the brain or absent from at least one participant were excluded from further analysis. The percent signal change (PSC) relative to the fixation condition was computed for each object presentation at each voxel. The mean of the four images (mean PSC) acquired within a 4s window, offset 4s from the stimulus onset (to account for the delay in hemodynamic response) provided the main input measure for subsequent analysis. The mean PSC data for each word or picture presentation were further normalized to have mean zero and variance one to equate the variation between participants over exemplars.

4.3 Does the distribution of neural activity encode sufficient signal to classify noun-noun concept combination?

We are interested in whether the distribution of brain activity encodes sufficient signal to decode the mental state associated with viewing and contemplating about the object. Given the observed brain activity when participants contemplated one of the 40 presented objects, classifiers were trained to identify cognitive states associated with viewing stimuli from the evoked patterns of functional activity. The classification analysis can be performed to decode exemplars or categories. In exemplar classification, the classifiers are trained to decode which of the 40 exemplars a participant is thinking about. In category classification, binary classifiers are trained to decode which of the property-based or relation-based interpretation a participant is thinking about.

4.3.1 Classifier Model

Classifiers were trained to identify cognitive states associated with viewing stimuli from the evoked pattern of functional activity (mean PSC). Classifiers were functions f of the form: f: mean_PSC → Yi, i=1,…n, where Yi were the sixty exemplars, and mean_PSC was a vector of mean PSC voxel activation level, as described above. The Gaussian Naïve Bayes (GNB) pooled variance classifier was used (Mitchell 1997). It is a generative classifier that models the joint distribution of a class Y (exemplar or category) and attributes X (voxels), and assumes the attributes X1,…,Xn are conditionally independent given Y. The classification rule is:

$$Y \leftarrow \arg\max_{y_j} \mathrm{P(Y = y_j)} \prod_{\mathrm{i}}^{\mathrm{n}} P(X_i \mid Y = y_j),\ j = 1{,}2\ .$$

Classification results were evaluated using 6-fold cross validation, where one of the 6 repetitions was left out for each fold. The voxel selection procedure (described below) was performed separately inside each fold, using only the training data. Since multiple classes were involved, rank accuracy was used Mitchell et al. (2004) to evaluate the classifier. Given a new fMRI image to classify, the classifier outputs a rank-ordered list of possible class labels from most to least likely. The rank accuracy is defined as the percentile rank of the correct class in this ordered output list. Rank accuracy ranges from 0 to 1. Classification analysis was performed separately for each participant, and the mean rank accuracy was computed over the participants.

4.3.2 Voxel Selection

Since fMRI acquires the neural activity in the entire brain (15,000 – 20,000 distinct voxel locations, in our parcellation), many locations might not exhibit neural activity that encodes word or phrase meaning. Thus, the classifier analysis selected the voxels whose responses to the different stimuli were most stable across presentations. Voxel stability was computed as the average pair wise correlation between 40-item vectors across presentations, using only the training set within each fold in the cross-validation paradigm. The focus on the most stable voxels effectively increased the signal-to-noise ratio in the data and also served as a dimensionality reduction tool that facilitated further analysis by classifiers. Many of our previous analyses have indicated that 120 voxels is a set size suitable for our purposes (Just et al., 2010).

The locations of stable voxels for the pattern-based classification of brain activity within participants are reported in Figure 4.2. Table 4.4 and Table 4.4 list the stable voxels collected from all stimuli and only noun-noun stimuli, resepectively. The stable voxels were located in multiple areas of the brain that are consistent with the group-level activation. The overall characteristics of stable voxel locations were: (1) inferior and middle frontal gyri, inferior and middle temporal gyri, the Fusiform areas, inferior parietal gyrus, the Precuneus area in the parietal lobe, the Lingual area in the occipital lobe; (2) bi-lateral; and (3) primarily located in the frontal and temporal lobes.

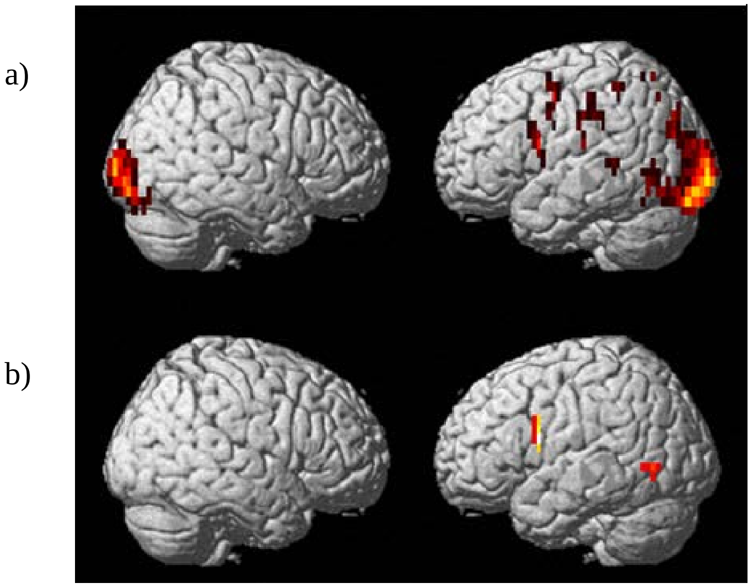


Figure 4.2 Voxel clusters from the union of stable voxels from all nine participants for a) all isolated noun and noun-noun concept combination stimuli, b) noun-noun concept combination stimuli only; extent threshold voxels=5. The overall characteristics of stable voxel locations were: (1) inferior and middle frontal gyri, inferior and middle temporal gyri, the Fusiform areas, inferior parietal gyrus, the Precuneus area in the parietal lobe, the Lingual area in the occipital lobe; (2) bi-lateral; and (3) primarily located in the frontal and temporal lobes.

Table 4.3 Locations (MNI centroid coordinates) and sizes of the voxel clusters selected by the stability measure of all noun and phrase stimuli.

Label	X	Y	Z	Voxels	Radius
Temporal					
L Temporal Inf	-46.2	-67.2	-8.4	27	8.81
L Fusiform	-32.4	-43.8	-13.6	11	6.15
R Fusiform	32.3	-39.6	-16	6	4.85
L Fusiform	-23.1	-48.1	-13.2	5	4.57
L Temporal Mid	-54.4	-40.6	4.8	5	3.87
Frontal					
L Frontal Inf Oper	-52.4	7.5	19.3	18	6.82
Parietal					
L Precentral	-39.4	-2.5	48	27	10.67
L SupraMarginal	-54.9	-24.9	34.3	21	9.64
L Parietal Inf	-36.6	-43.4	48.4	17	6.75
Occipital					
L Calcarine	-3.9	-86.9	2.4	729	24.98

Table 4.4 Locations (MNI centroid coordinates) and sizes of the voxel clusters selected by the stability measure of property and relation phrases.

Label	X	Y	Z	Voxels	Radius
Temporal					
L Fusiform	-32	-42.8	-18.9	20	6.23
L Occipital Mid	-39.6	-76.7	27.2	15	8.94
R Fusiform	31.9	-39.3	-18	14	6.2
L Temporal Inf	-58.3	-55.6	-6.7	9	5.36
R Temporal Inf	56.2	-60.4	-5	6	4.3
L Insula	-44.4	13.1	-6	5	5.04
Frontal					
L Frontal Mid	-38.4	48.7	10.3	7	5.06
R Frontal Inf Oper	44.8	10.4	27	6	4.41
R Frontal Inf Oper	53.8	13.8	27.6	5	3.51
Parietal					
R Precuneus	8.9	-46.9	52	6	5.66
L Parietal Inf	-43.1	-48.1	45.6	5	4.31
R Precuneus	3.1	-60	55.2	5	3.62
Occipital					
L Lingual	-6.2	-73.1	-4.8	5	4.61
R Lingual	13.8	-69.4	-3.6	5	5.75
Subcortical					
L Calcarine	-13.5	-60.4	13.6	15	6.16
L Pallidum	-19.4	3.8	-4.8	5	3.79
R Cerebelum Crus2	11.2	-83.1	-31.2	5	4.66

Our theoretical framework considers all cortical voxels and allows the training data to determine which locations are systematically modulated by which aspects of word meanings. In addition to selecting voxel from the whole brain, we perform the voxel selection separately for each brain region. Distinct classifiers were trained for the frontal, temporal, parietal, and occipital lobe.

4.3.3 Results

Which of the 40 exemplars was the participant thinking about? The first and second row in Table 4.5 shows the results of the exemplar classification analysis. All classification accuracies were significantly higher than chance ($p < 0.05$), where the chance level for each classification is determined based on the empirical distribution of rank accuracies for randomly generated null models. One hundred null models were generated by permuting the class labels. The classifier was able to distinguish among the 40 exemplars and 20 nouns with mean rank accuracies close to 77%, and 67% respectively. We also determined the classification accuracies separately for each brain region. Distinct classifiers were trained for the frontal, temporal, parietal, and occipital lobe. Occipital lobe yields the best exemplar classification at 74% and 64%, respectively.

Was the participant thinking of a property or relation-based interpretation? The third and fourth row in Table 4.5 shows the results of the category classification analysis. All classification accuracies were significantly higher than chance ($p < 0.05$), where the chance level for each classification is determined based on the empirical distribution of rank accuracies for randomly generated null models. One hundred null models were generated by permuting the class labels. The classifier was able to distinguish between the two types of interpretations with mean rank accuracies close to 66%. More importantly, the discriminability is not due to the differences in contextual prime as the classifier was not able to distinguish between the contextual primes that are used to induce property-based and relation-based interpretations (mean rank accuracies at 53%, which is not statistically different from chance). We also determined the classification accuracies separately for each brain region. Unlike exemplar classification, the parietal lobe and frontal lobe yields the best category classification at 0.66 and 0.63%, respectively.

Table 4.5 Classification analysis

	Stimuli	**All**	**Frontal**	**Temporal**	**Parietal**	**Occipital**	**Non-Occipital**
Exemplar	All	0.77	0.67	0.69	0.71	0.74	0.76
	Noun	0.67	0.58	0.59	0.62	0.64	0.64
Category	Context	0.53	0.52	0.50	0.54	0.60	0.50
	Phrase	0.66	0.63	0.62	0.66	0.59	0.66

4.4 fMRI Analyses for SPM Group Contrasts

To compare the distribution of activation across experimental conditions, group t-test analyses were performed using a random-effects model (Friston et al., 1995). All t-maps in each contrast were calculated across the entire cortical volume, thresholded at an uncorrected height threshold of $p < .001$ and an extent threshold of 5 voxels. Statistical maps were superimposed on the high-resolution, normalized, T1-weighted, SPM2 individual template image for viewing. Labels for coordinates of activation were confirmed in MNI space (Tzourio-Mazoyer et al., 2002).

On one hand, the brain activation for isolated noun stimuli contrasted with noun-noun concept combination stimuli was mostly left-lateralized, and included the cortical areas that are part of the language network of the brain (inferior and superior frontal gyrus), in additions to the occipital areas (Calcarine, the middle and superior occipital areas). There was also right-lateralized activation in Cuneus and Precuneus areas. On the other hand, the brain activation for noun-noun concept combination stimuli contrasted with isolated noun stimuli included left-lateralized activation in the middle temporal gyrus, the angular gyrus, as well as right-lateralized activation in the fusiform gyrus and Precuneus. (Figure 4.3; Table 4.6).

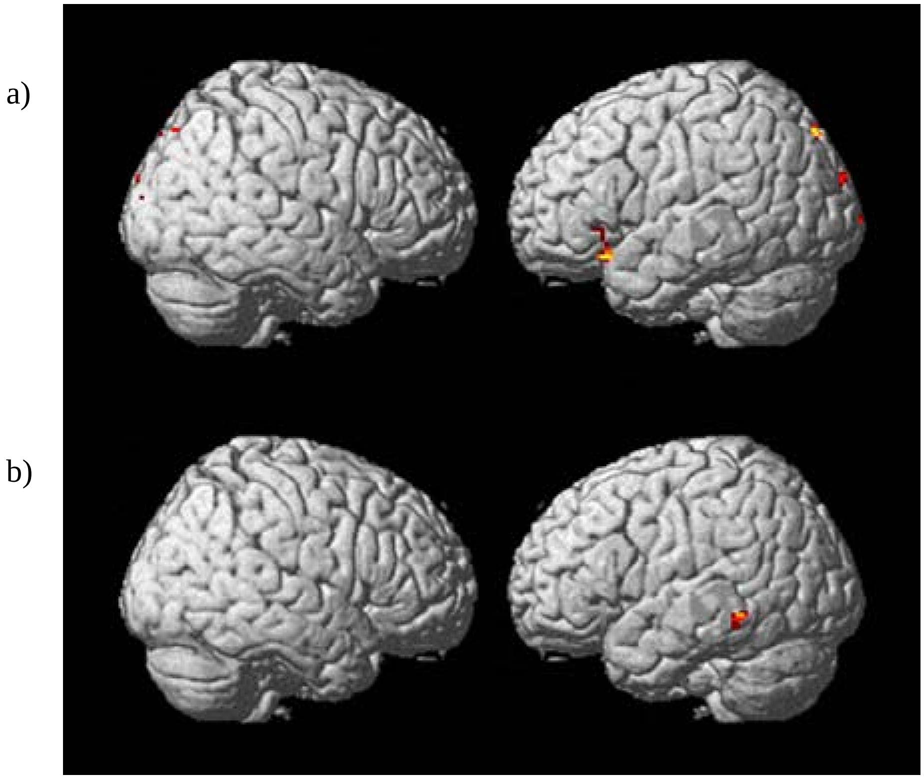


Figure 4.3 Brain activation for a) isolated word stimuli contrasted with noun-noun concept combination stimuli, b) noun-noun concept combination stimuli contrasted with isolated word stimuli, p<0.001 uncorrected; T=4.30; extent threshold voxels=5. The brain activation for isolated noun stimuli contrasted with noun-noun concept combination stimuli was mostly left-lateralized in inferior and superior frontal gyrus. The brain activation for noun-noun concept combination stimuli contrasted with isolated noun stimuli included left-lateralized activation in the middle temporal gyrus.

Table 4.6 Locations (MNI centroid coordinates) and sizes of the voxel clusters of activation for isolated noun stimuli contrasted with noun-noun concept combination stimuli, p<0.001 uncorrected; T=4.30; extent threshold voxels=5.

Contrast	Label	X	Y	Z	Voxels	Radius
Word - Phrase	*Frontal*					
	L Frontal Inf Orb	-34	23.4	-9.5	50	6.94
	L Frontal Sup Orb	-15.5	40.6	-13.2	23	3.05
	R Frontal Inf Orb	29	30	-9	8	1.93
	Occipital					
	L Calcarine	-2.1	-85	1.3	113	8.19
	L Occipital Sup	-9	-83.4	46.6	29	2.97
	R Cuneus	10.3	-92.7	13.6	23	2.99
	L Cuneus	-7.2	-95.9	22.4	17	2.9
	L Occipital Mid	-9.7	-106	3.7	6	1.82
	Parietal					
	R Precuneus	15	-78	47	8	1.93
Phrase - Word	*Temporal*					
	L Temporal Mid	-63	-44.7	-9.3	21	3.08
	R Fusiform	35.6	-37.2	-25.6	5	2.16
	Parietal					
	R Precuneus	21	-39.6	13.8	72	4
	L Angular	-32.4	-53.2	23.9	63	4.27

The brain activation for property minus relation was mostly bi-lateralized, and included the Broca’s areas (inferior frontal gyrus), the inferior and middle temporal gyrus, the fusiform area, as well as the sensorimotor areas in the parietal lobe (precentral and postcentral gyri). There was no positive activity for relation minus property contrast (Figure 4.4; Table 4.7).

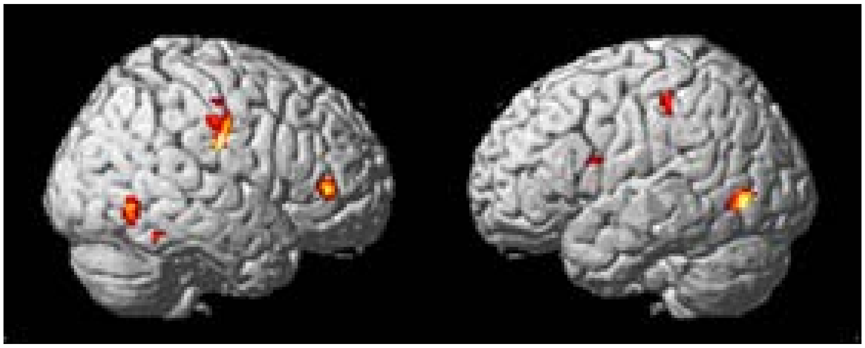


Figure 4.4 Brain activation for property-based interpretation of the concept combination stimuli contrasted with relation-based interpretation of the isolated word stimuli, p<0.001 uncorrected; T=4.30; extent threshold voxels=5. The brain activation for property minus relation was mostly bi-lateralized, and included the Broca's areas (inferior frontal gyrus), the inferior and middle temporal gyrus, the fusiform area, as well as the sensorimotor areas in the parietal lobe.

Table 4.7 Locations (MNI centroid coordinates) and sizes of the voxel clusters of activation for property-based interpretation of the concept combination stimuli contrasted with related-based interpretation of the concept combination stimuli, p<0.001 uncorrected; T=4.30; extent threshold voxels=5.

Label	**X**	**Y**	**Z**	**Voxels**	**Radius**
Temporal					
L Occipital Mid	-38.6	-68.7	-1.4	140	5.54
R Temporal Inf	46.2	-61.9	-6.2	78	4.75
R Temporal Inf	49.3	-48.9	-19.3	11	2.14
Frontal					
R Frontal Inf Tri	41.4	38.9	4.6	63	3.84
L Frontal Inf Oper	-44.7	6.5	18.7	12	2.3
Parietal					
R Postcentral	52.8	-16.2	34.6	133	8.14
L Postcentral	-40.3	-30.6	48.6	37	3.41
R Postcentral	55.2	-17.5	49	8	1.87

4.5 Neural Composition of noun-noun concept combination

We now study the neural composition of noun-noun concept combination. In our experiment, we have recorded the brain activity for noun-noun phrases, as well as the corresponding nouns in isolation. One direct way of assessing compositionality is to compare the brain activity for phrases to individual words. Our hypothesis is the brain activity for property-based interpretation should be more similar to the head word (since only one property of the modifier word is extracted to modify the head word), whereas the brain activity for relation-based interpretation should be similar to both the modifier and head word (since the modifier object is realized in its entirety to the head object as a whole).

Figure 4.5 and rows 1 and 2 in Table 4.8 show the correlation analysis for all stimuli. Each of the 40 stimulus items is represented by a vector of brain activity measured at the 120 most stable voxels whose responses to the 20 different nouns were most stable across presentations. Unlike our hypothesis, brain activity for property-based interpretation is more similar to the modifier word than the head word (r = 0.24 > 0.13), whereas brain activity for relation-based interpretation is more similar to the head word than the modifier word (r = 0.19 > 0.16). The difference in correlation with the modifier and head word is statistically significant (p = 0.05) for the property-based interpretations, but not for relation-based interpretations. One possible explanation is that property-based interpretations are less accessible / intuitive to people; as a result, people think more about the modifier word to find a fitting property. This pattern of result occurs despite the fact that the noun-noun concept combinations were shown during a pre-training phase. Although people were not struggling to make sense of the combination for the first time, property-based interpretations require people to pay more attention to the modifier words. On the other hand, relation-based interpretations are made more easily and people can move on to focus on the head word (linguistic head).

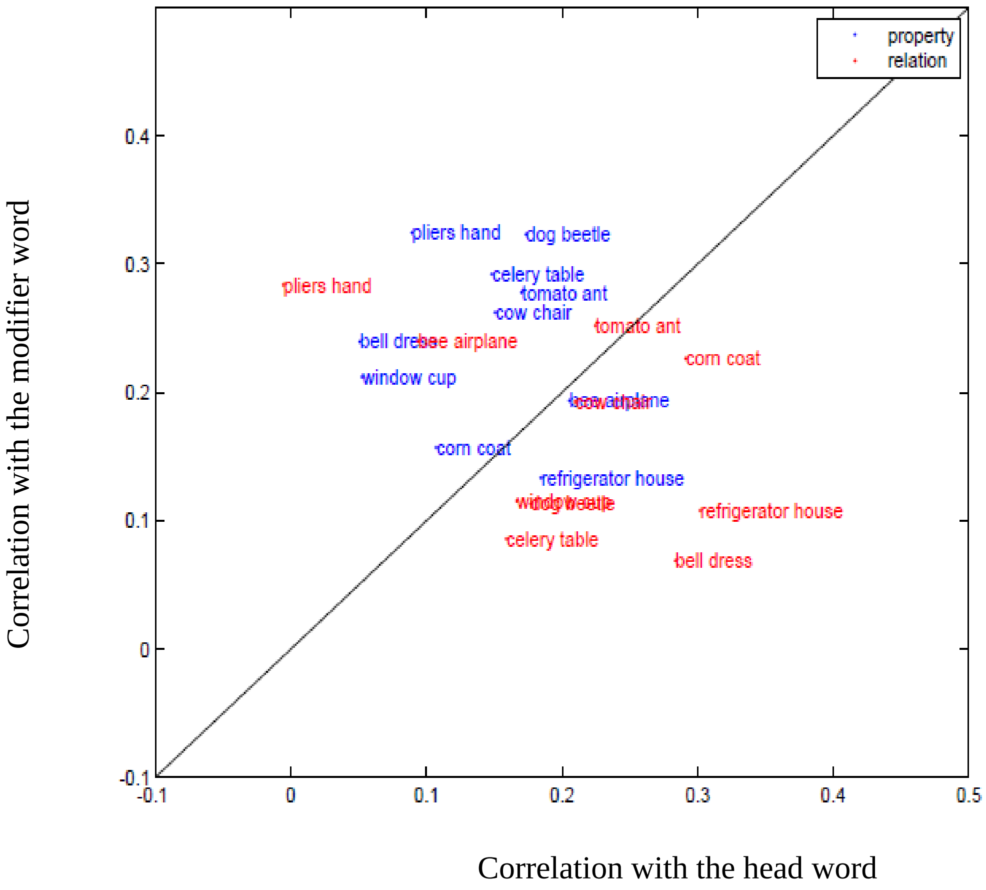


Figure 4.5 Correlation between noun-noun concept combination and isolated noun in the brain space. Brain activity for property-based interpretation is more similar to the modifier word than the head word (r = 0.24 > 0.13), whereas brain activity for relation-based interpretation is more similar to the head word than the modifier word (r = 0.19 > 0.16).

Table 4.8 Correlations between noun-noun concept combination and isolated noun

Space	**Type**	**Mod**	**Head**	**Property**	**Relation**
Brain activity	Property	0.24	0.13	1.00	0.35
	Relation	0.16	0.19	0.35	1.00
Data-driven (ILFM)	Property	0.23	0.16	1.00	0.26
	Relation	0.18	0.20	0.36	1.00

Figure 4.6 shows the brain activation for *corn*, *coat*, the property-based interpretation of *corn coat*, and the relation-based interpretation of *corn coat* at postcentral and parahippocampal gyrus. As seen in Figure 4.6,

1. A property-based interpretation is more similar to the modifier word: both *corn* and a property-based interpretation of *corn coat* have activation in the postcentral gyrus.
2. A relation-based interpretation is more similar to the head word: both *coat* and a relation-based interpretation of *corn coat* have activation at the parahippocampal gyrus.
3. A relation-based interpretation is also similar to the modifier word: both *corn* and a relation-based interpretation of *corn coat* have activation in the postcentral gyrus.

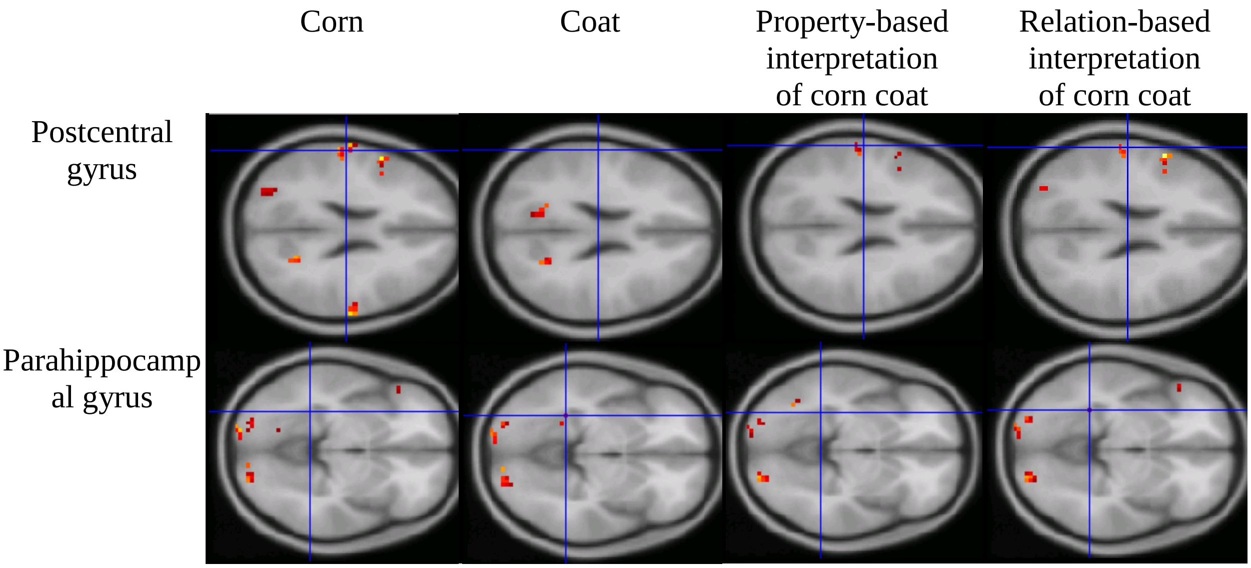

Corn
Coat
Property-based interpretation of corn coat
Relation-based interpretation of corn coat
Postcentral gyrus
Parahippocampal gyrus

Figure 4.6 The image colors codes brain activity at 500 most stable voxels, only clusters of size 5 voxels or up. Red circles indicate there are brain activities at the postcentral gryus ((MNI -62.29, -20.14, 21.73) . Blue circles indicate there are brain activities at the parahippocampal gryus (MNI (MNI -31.25, -43.75, -6.00).

4.6 Discussion

The locations of the stable voxels indicate that the fMRI data used for discriminating property-based vs. relation-based interpretations may be reflecting cognitive process – novelty, ambiguity, syntax, semantic, compositionality. The right hemisphere activation may reflect selective attention to meaning (Dapretto & Bookheimer, 1999), the use of linguistic context to disambiguate particular interpretation, or prosody.

To test the hypothesis that relation-based interpretations are more easily accessible, a step-wise analysis was performed to gradually enter stable voxels in the correlation analysis, instead of entering all of them simultaneously. As seen in Figure 4.7, for both property and relation-based interpretations, the initial stable voxels correlate more with the head word than the modifier word. As the number of voxels included increases, the property-based interpretations started to correlate more with the modifier word, whereas the relation-based interpretations continued to correlate more with the head word.

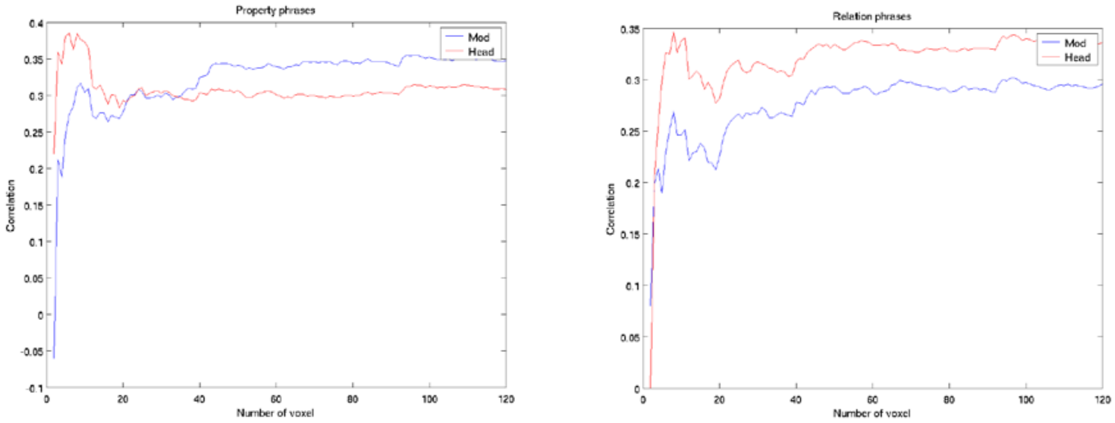


Figure 4.7 Correlation between phrase and word by number of voxels. For both property and relation-based interpretations, the initial stable voxels correlate more with the head word than the modifier word. As the number of voxels included increases, the property-based interpretations started to correlate more with the modifier word, whereas the relation-based interpretations continued to correlate more with the head word.

To examine the temporal sequence when people pay more attention to the modifier or the head word, we performed the correlation analysis with a moving window. As seen in Figure 4.8, for both property and relation-based interpretations, the brain activity for the phrase correlate more with the modifier word earlier in time – likely to reflect the fact that modifier words are read first. However, as time moves on, the property-based interpretations continue to correlate more with the modifier word, whereas the relation-based interpretations started to correlate more with the head word. The pattern for the property and relation-based interpretations are clearly different.

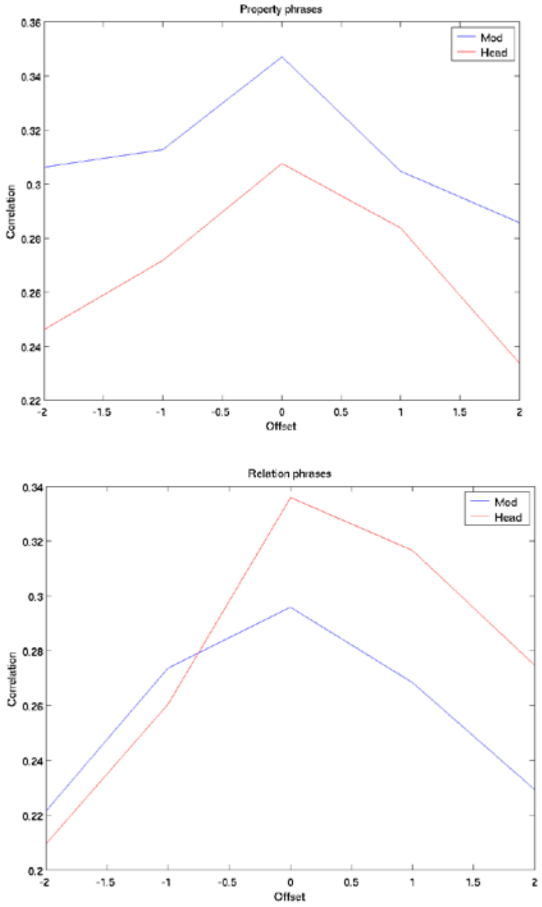


Figure 4.8 Correlation between phrase and words by offset. For both property and relation-based interpretations, the brain activity for the phrase correlate more with the modifier word earlier in time. As

time moves on, the property-based interpretations continue to correlate more with the modifier word, whereas the relation-based interpretations started to correlate more with the head word

4.7 Conclusions and Contributions

In this study, we use contextual prime to induce one of the two interpretations and record fMRI-measured brain activity while people think about properties associated with the stimulus. Classification analysis shows that the distributed pattern of brain activity contains sufficient signal to decode the semantic differences between property-based vs. relation-based interpretations despite the identical visual stimuli. More importantly, by employing contextual primes to induce different interpretations of the same visual stimuli, we were able to train classifiers that discriminate semantic distinctions, instead of differences in visual perception. This is evident from the observations that parietal and temporal lobe yield better classification than the occipital lobe for decoding different interpretations of the same visual stimuli.

Moreover, we study the compositionality in the meaning of phrases. An improved understanding of semantic composition in multi-word phrases is an important building step toward neural accounts of sentence processing. We find that brain activity for property-based interpretation is more similar to the modifier word, whereas brain activity for relation-based interpretation is more similar to the head word. One possible explanation is that property-based interpretations are less accessible / intuitive to people; as a result, people think more about the modifier word to find a fitting property.

Some of the future directions include investigating how context dependent conceptual combinations are. In this study, we assume interpretations of conceptual combinations are strongly conditioned by the situations and communicative task at hand. Thus, we used contextual conditions to induce either relation- or property-based interpretation. We did not include control stimuli where the compound phrases are presented in the absence of contextual primes, which would allow us to study to what extent do conceptual representations have a "default" and/or context-independent form. We also did not include frequently encountered compound phrases *(e.g. coffee shop)*, that may be lexicalized and involved a different cognitive process.

Furthermore, in this work we considered the brain activity after the stimuli are comprehended and did not attempt to model *how* the combinations are derived. For instance, some relational interpretations may be derived by deductive reasoning (e.g. a knowing that a coat is used to guard against cold weather, one can deduct that a *corn coat* is a coat used to protect corn from thunderstorm), whereas some may emerge spontaneously from regularities in the world (e.g., knowing that animals are often named after their appearances, like *sword fish*, one can derive that a *dog beetle* is named after the beetle's wigging tails). One future direction is to implement a computational model that both derives the meaning of compound noun and models the observed brain activity. For instance, we could implement PUNC (Costello and Keane, 1997), a computational model that is based on the constraint theory of conceptual combination and the C3 model. PUNC assumes that meaning of a compound noun can be derived from all possible combinations of the modifier and head noun, where the acceptability of the each interpretation is subsequently ranked by three constraints of diagnosticity, plausibility, and informativeness. PUNC has been shown to be capable of deriving the meaning of familiar, similar, and novel word combinations that mirror human behavior.

5 A LATENT FEATURE ANALYSIS OF THE NEURAL REPRESENTATION OF OBJECT KNOWLEDGE

5.1 Introduction

Mitchell et al. (2008) showed that word features computed from the occurrences of stimulus words (within a trillion-token Google text corpus that captures the typical use of words in English text) can predict the brain activity associated with the meaning of these words. The advantage of using word co-occurrence data is that semantic features can be computed for any word in the corpus – in principle any word in existence. Nonetheless, despite the success of this model, the work leaves open the question about how to determine the optimal set of semantic features. Mitchell et al. (2008) hand-picked a set of 25 semantic features defined by 25 verbs: *see, hear, listen, taste, smell, eat, touch, rub, lift, manipulate, run, push, fill, move, ride, say, fear, open, approach, near, enter, drive, wear, break*, and *clean*. This selection process was motivated by existing conjectures regarding the centrality of sensory-motor features in neural representations of objects (Caramazza & Shelton, 1998). However, are there other sets of semantic features that better characterize the brain activity? Is 25 a sufficient or more than necessary number of features to characterize the semantic representation? One can exhaustively search for the optimal set of features, but such an approach is computationally intractable and certainly not a satisfying approach.

In this study, we address the question by taking a bottom-up approach. Instead of searching for the optimal set of features that can account for the brain activity, we try to infer the most likely feature structure directly from the patterns of brain activity. We take a generative approach and model the semantic representation as some hidden variables in the probabilistic Bayesian framework. A generative process is used to describe how brain activity is generated from this latent semantic representation[c]. The basic proposition of the model is that the human semantic knowledge system is capable of producing an unbounded number of features associated with a concept; however, only a subset of them are actively recalled and reflected in brain activity during any given task. Moreover, the same set of features is not recalled by a group of people. There may be an overlap of features that people commonly recall and people may come up with features that no one has thought of before.

Thus, a set of latent indicator variables is introduced to indicate whether a feature is actively recalled. By describing the prior distribution of these latent indicator variables and the distribution of the observed brain activity given the assignment of these latent variables, standard Bayesian inference procedure can be used to infer the recalled features. More specifically, we used the infinite latent feature model (ILFM) with an Indian Buffet Process (IBP) prior (Griffiths & Ghahramani, 2005) to derive a binary feature representation of object knowledge from the brain activity. ILFM is especially suited for our task because it automatically determines the number of features that are manifested in the data. This data-driven feature representation is neurologically-informed and may better capture what people were thinking. To foreshadow our results, the ILFM is able to capture a latent semantic representation that is consistent with human ratings of three semantic factors recovered by factor analysis. Furthermore, we show that the recovered latent features are consistent with some existing conjectures regarding the role of different brain areas in processing different psycholinguistics features.

[c] We use the term *latent semantic* to refer to some intermediate semantic representations that are modeled by with hidden variables in a probabilistic Bayesian framework. It is not to be confused with latent semantic analysis (LSA), although LSA can also be casted in the same Bayesian framework (Hofmann, 1999).

In the study, we used the data from the two experiments in Chapter 2 and 4, where participants were asked to think about properties associated with a visually-presented object or noun-noun phrase, respectively. The fMRI data acquisition data and signal processing methods were reported in previous chapters. In section 2, we discuss the infinite latent feature model and show how it can be used to recover the latent semantic representation encoded by brain activity. In section 3, we try to interpret the recovered latent features by correlating the latent features with the human ratings of the shelter, manipulation, and eating factors, as well as some psycholinguistic word features. Finally, we discuss some of the implications of our work and suggest some future studies.

5.2 Latent Feature Analysis

To characterize the semantic content that is encoded in the brain activity, we take a generative approach and model the semantic representation as some hidden variables in a probabilistic Bayesian framework. A generative process is used to describe how brain activity is generated from this latent semantic representation. Given the brain activity associated with people viewing and contemplating different objects, we then apply Bayesian inference procedures to infer the most likely latent structure that gives rise to the observed brain activity pattern.

Griffiths & Ghahramani (2005) described a non-parametric Bayesian approach to latent variable modeling in which the number of latent variables is unbounded. They defined a probability distribution over equivalence classes of binary matrices and derived a generative process called Indian buffet process that results in the same distribution. The distribution can be used to define probabilistic models that represent objects with an unbounded number of binary features. They further derive an infinite latent feature model (ILFM) with an IBP prior on the latent semantic structure. The ILFM is an appealing approach to model the semantic content that is encoded in the brain activity. The basic proposition of the model is that the human semantic knowledge system is capable of producing an unbounded number of features associated with a concept; however, only a subset is actively recalled during any given task. Thus, a set of latent indicator variables is introduced to indicate whether a feature is actively recalled at any given task.

Let *X* denotes the brain activity recorded in our object-contemplating task and *Z* denotes the latent semantic representation that underlies the brain activity pattern, the infinite latent feature model is specified by 1) a prior over the feature vectors *P(Z)*, and 2) a distribution over the brain activity matrices conditioned on the feature assignments, *p(X|Z)*.

In a linear-Gaussian infinite latent feature model, the distribution of *Z* is modeled with an IBP prior, and the distribution of *X*|*Z* is assumed to be matrix Gaussian with mean *ZA* and variance $\sigma_X I$. The following equations summarize the linear-Gaussian infinite latent feature model. For more details regarding the derivation of *P(Z)* and *p(X|Z)*, please see Griffiths & Ghahramani (2005).

$$\begin{aligned}
&Z \sim IBP(\alpha, \beta) \\
&A \sim Gaussian\left(0, \sigma_Z^2 I\right) \\
&X \mid Z, A, \sigma_X \sim Gaussian\left(ZA, \sigma_X^2 I\right)
\end{aligned}$$

In the context of the 60-words experiment, *X* is a matrix of size *N* × *V*, where x_{nv} is the brain activity for object *n* at voxel *v*. *N* = 60 and *V* = 120 since our stimulus set consists of 60 objects and the voxel selection procedure selects the 120 most stable voxels. Notice that each object was presented 6 times in our experiment; a representative fMRI image for each object was created by computing the mean fMRI response over the 6 presentations, and the mean of all 60 of these representative images was then subtracted from each brain activity vector.

Z is a matrix of size $N \times K$, where z_{nk} is a binary value indicating if the feature is actively recalled. By assuming an IBP prior on the distribution of Z, the number of K is unbounded. The hyper parameters α and β for the IBP controls the number of features per object and the total number of features in the matrix, respectively.

A is matrix of size $K \times V$, where a_{kv} denote the feature-to-brain activity mapping, such that $X = Z \times A$. By assuming that the distribution of A is matrix Gaussian with mean 0 and variance $\sigma_A I$, we can easily integrate out A when computing the full distribution of $P(Z) \cdot p(X|Z)$.

We used Gibbs Sampling (Geman & Geman, 1984) to inference Z. The Gibbs sampler was initialized with $K_+ = 1$, with a random assignment to the first column by setting $z_{i1} = 1$ with probability 0.5. The model parameters, α, β , σ_A, and σ_X were all initially set to 0.5, and then sampled by adding Metropolis-Hastings (Metropolis et al., 1953) steps to the MCMC algorithm. A separate ILFM is estimated for each participant and each brain region. Figure 5.1 shows the trace plots for the 1000 iterations of MCMC for the temporal lobe of the first participant. The parameters converge after approximately 100 iterations.

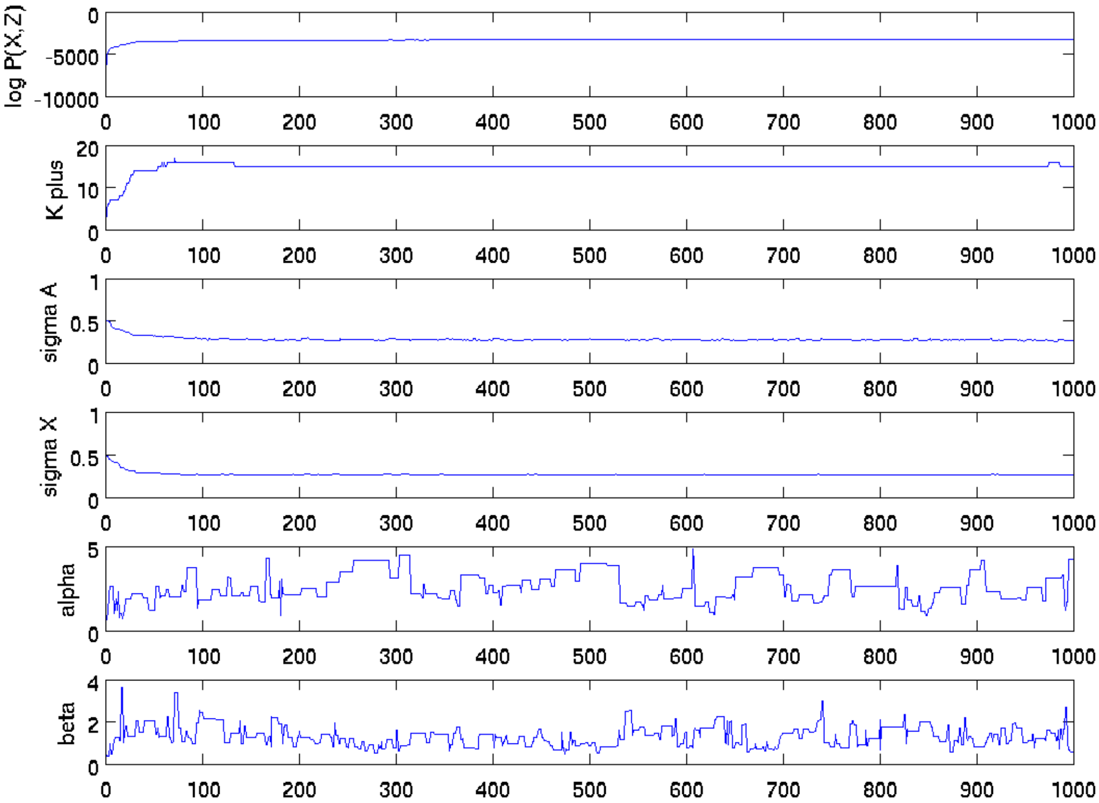


Figure 5.1 Trace plots for the MCMC simulation for the temporal lobe of the first participant. The parameters converge after approximately 100 iterations.

Figure 5.2 shows the Z matrix inferred from the temporal lobe of the first participant. As can be seen in the figure, the Z matrix is quite dense. Each latent feature is possessed by a number of different objects. Conversely, the meaning of each word is distributed across many latent features.

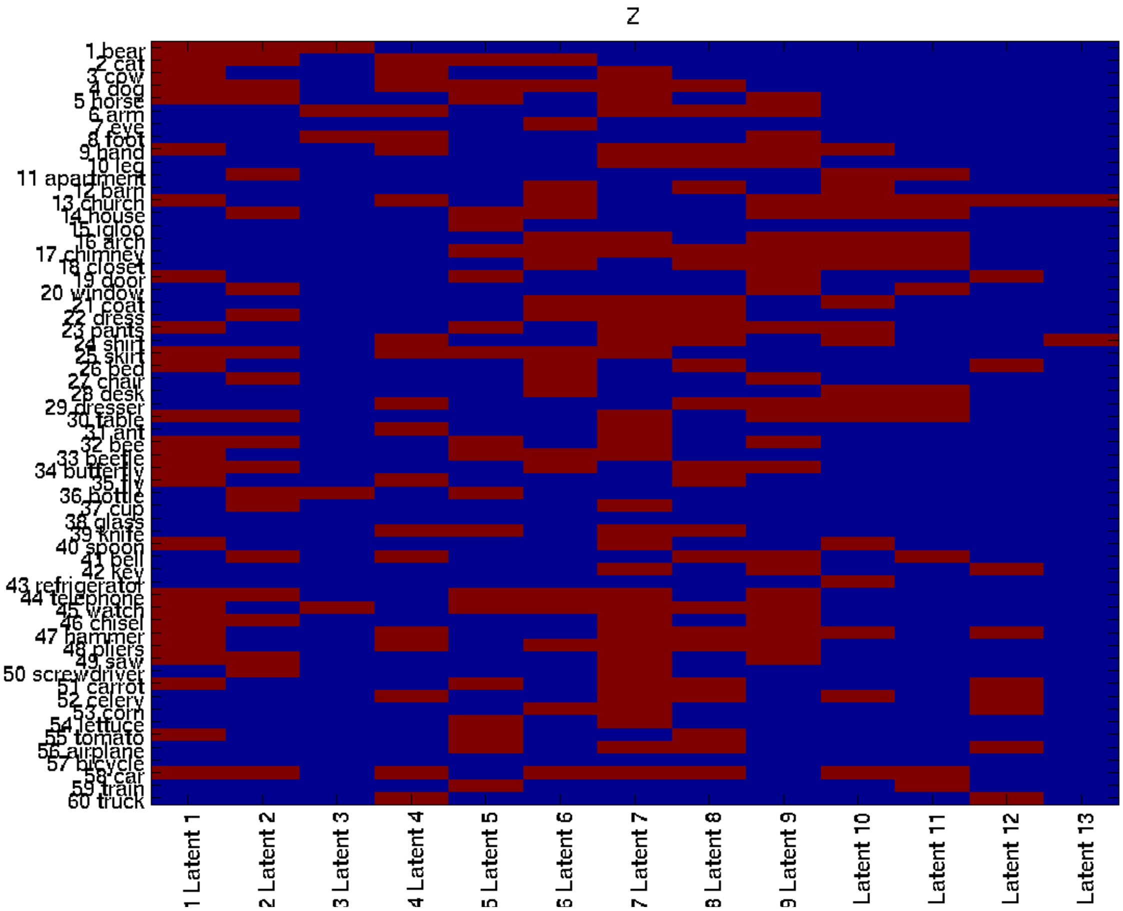


Figure 5.2 The Z matrix inferred from the temporal lobe of the first participant. The meaning of each word is distributed across many latent features.

5.2.1 Independent Human Rating

Just et al. (2010) used factor analysis to identify three semantic factors: manipulation, eating, and shelter that provide a good basis for the representation of the 60 objects. The manipulation factor assigns high scores to objects that are held and manipulated with one's hands *(e.g. pliers, screwdriver)*. The eating factor assigns high scores to objects that are edible *(e.g. vegetables)* or are instruments for eating or drinking *(e.g. glass, cup)*. The shelter factor assigns high scores to objects that provide shelter *(e.g. house, apartment)* or entry to a sheltering enclosure *(e.g. airplane)*. They collected an independent set of ratings of each word with respect to each of the three semantic factors from a separate set of 14 participants. For example, for the eating-related factor, participants were asked to rate each word on a scale from 1 (completely unrelated to eating) to 7 (very strongly related). Table 5.1 shows the collected ratings for 10 participants in the experiment.

Table 5.1 Independent human rating of the 60 words

Category	Words	Shelter	Manipulation	Eating	Length
animal	bear	1.1	1.3	1.6	4
animal	cat	1.1	2.4	1.8	3
animal	cow	1.1	1.9	4.9	3
animal	dog	1.4	2.1	2.3	3
animal	horse	1.2	2.9	2.4	5
bodypart	arm	2.2	3.1	3.3	3
bodypart	eye	1.3	2.2	2.6	3
bodypart	foot	1.3	1.6	1.3	4
bodypart	hand	1.8	5.6	4.2	4
bodypart	leg	1.4	1.6	1.4	3
building	apartment	6.9	2.7	3	9
building	barn	6.6	1.7	2.3	4
building	church	6.1	1.6	1.7	6
building	house	6.9	2.9	3.5	5
building	igloo	6.6	2.6	2.9	5
buildpart	arch	4.2	1.5	1	4
buildpart	chimney	3.1	1.3	1.1	7
buildpart	closet	4.5	2.4	1.2	6
buildpart	door	4.9	4.3	1.2	4
buildpart	window	3.6	3.6	1.1	6
clothing	coat	2.9	3.9	1.1	4
clothing	dress	1.9	3.4	1.1	5
clothing	pants	2.1	3.4	1.1	5
clothing	shirt	1.8	3.6	1.3	5
clothing	skirt	1.5	3.1	1.1	5
furniture	bed	3.3	2.2	1.3	3
furniture	chair	2.6	3.1	2.8	5
furniture	desk	3.1	2.9	1.9	4
furniture	dresser	2.4	3.1	1	7
furniture	table	3.2	2.8	4.9	5

Category	**Words**	**Shelter**	**Manipulation**	**Eating**	**Length**
insect	ant	1.1	1.6	3	3
insect	bee	1.1	1.5	2.9	3
insect	beetle	1	1.7	2.6	6
insect	butterfly	1.1	1.4	1.4	9
insect	fly	1.1	1.4	1.4	3
kitchen	bottle	1.1	4.5	4.5	6
kitchen	cup	1.1	6.1	4.7	3
kitchen	glass	1.7	3	3.6	5
kitchen	knife	1.5	6.9	4.5	5
kitchen	spoon	1.1	6.6	5	5
manmade	bell	1.7	6.2	1.4	4
manmade	key	2.9	6.4	1.1	3
manmade	refrigerator	2.7	3.9	5.5	12
manmade	telephone	1.8	5.5	2.4	9
manmade	watch	1.1	4.6	1.1	5
tool	chisel	1.7	6.4	1	6
tool	hammer	1.9	6.7	1	6
tool	pliers	1.6	6.8	1	6
tool	saw	1.7	6.3	1.1	3
tool	screwdriver	1.8	6.7	1.1	11
vegetables	carrot	1.1	3.9	6.6	6
vegetables	celery	1	4.1	7	6
vegetables	corn	1.1	3.5	7	4
vegetables	lettuce	1	2.6	6.6	7
vegetables	tomato	1	3.9	7	6
vehicles	airplane	4.6	3.9	1.5	8
vehicles	bicycle	1.3	5.1	1.2	7
vehicles	car	4	5.3	1.6	3
vehicles	train	5	2.1	1.8	5
vehicles	truck	4.4	4.5	1.2	5

We also collect human ratings of the three semantic factors on the 40 exemplars in the noun-noun experiment. Table 5.2 shows the collected ratings for 5 participants in the experiment.

Table 5.2 Independent human rating of stimuli in the noun-noun experiment

Type	Stimuli	Shelter	Manipulation	Eating
noun	bee	1.1	1.5	2.9
noun	bell	1.7	6.2	1.4
noun	celery	1	4.1	7
noun	corn	1.1	3.5	7
noun	cow	1.1	1.9	4.9
noun	dog	1.4	2.1	2.3
noun	pliers	1.6	6.8	1
noun	refrigerator	2.7	3.9	5.5
noun	tomato	1	3.9	7
noun	window	3.6	3.6	1.1
noun	airplane	4.6	3.9	1.5
noun	dress	1.9	3.4	1.1
noun	table	3.2	2.8	4.9
noun	coat	2.9	3.9	1.1
noun	chair	2.6	3.1	2.8
noun	beetle	1	1.7	2.6
noun	hand	1.8	5.6	4.2
noun	house	6.9	2.9	3.5
noun	ant	1.1	1.6	3
noun	cup	1.1	6.1	4.7
property	Bee airplane	3.6	2.4	2
property	Bell dress	2	2.8	1
property	Celery table	3.4	2	3.6
property	Corn coat	2.2	2.4	1.4
property	Cow chair	2.2	1.4	1.2
property	Dog beetle	1	1.8	1
property	Pliers hand	1.2	7	1.4
property	Refrigerator house	6.8	1.4	1.8
property	Tomato ant	1	3	2.8
property	Window cup	1.8	5	4
relation	Bee airplane	4.8	2.6	3.2
relation	Bell dress	2.8	3	1
relation	Celery table	3	3.2	5.2
relation	Corn coat	4.4	4	3.8
relation	Cow chair	2.2	1.8	1.4
relation	Dog beetle	1	2.2	1.8
relation	Pliers hand	1.2	6.8	1.4
relation	Refrigerator house	6	1.4	3
relation	Tomato ant	1	2.2	3.8
relation	Window cup	2	5	4.6

5.2.2 MRC Psycholinguistics Database

The MRC Psycholinguistic Database (Coltheart, 1981) is a dictionary that contains 150837 words with up to 26 linguistic and psycholinguistic attributes for each word. While linguistic measures are defined for most of the words, psychological measures are recorded for only about 2500 words. Some of the psycholinguistic measures that are of interest to us include meaningfulness (cmean), familiarity (fam), concreteness (cnc), imaginability (img), number of letters (nlet), number of phonemes (nphn), and frequency (t-lfrq).

5.3 Results

5.3.1 60 Word picture study

Table 5.3 shows the amount of systematic variance (R^2) accounted by the latent semantic structure and the average number of latent features (K_+) inferred from the brain activity in each brain region. The amount of variance explained correlates almost perfectly (r = 0.98) with the classification rank accuracy. Moreover, there is a strong negative correlation (r = -0.78) between the number of latent features and classification rank accuracy. One possible explanation is that the more number of features a participant is contemplating about an object, the more variance there is to the word representation and the worse classification performance.

Table 5.3 Classification and infinite latent feature analysis

Metric	**All**	**Frontal**	**Temporal**	**Parietal**	**Occipital**
Rank accuracy	0.81	0.58	0.70	0.66	0.80
R^2	0.77	0.66	0.69	0.69	0.76
K_+	14.44±3.09	16.67±4.47	14.22±3.67	15.44±6.13	14.89±4.81

The question now is what does each latent feature mean? Do different brain areas encode different types of word features? In the following sections, we try to interpret the recovered latent features with human ratings of the shelter, manipulation, and eating factors that are recovered by factor analysis, as well as some psycholinguistics features.

We show that the latent features recovered by ILFM are consistent with the human ratings of the shelter, manipulation, and eating factors that are recovered by the factor analysis. For each latent feature inferred, we correlate the latent feature vector (column vector describing which objects possess this feature) with human ratings of the three semantic factors (column vector describing how human rate the relatedness between the 60 objects and the specified factor). For each brain region and each of the three semantic factors, we identify the maximum correlation between the semantic factors with any one of the latent semantic feature. Figure 5.3 shows the maximum correlation between the latent feature vector and human rating vector, averaged across subjects. The error bar indicates 95% confidence interval, where the distribution of statistic is estimated from the 900 Gibbs samples (excluding the first 100 burn in samples). Different brain regions infer different latent features: the frontal lobes tend to infer latent features that correlate with human ratings of manipulation vector, whereas temporal and parietal lobes tend to infer latent features that correlate with human ratings of shelter and eating factor, respectively. This pattern of result is consistent with contemporary conjecture that the precentral area in the frontal lobe is involved with motor planning, the fusiform and parahippocampal place areas that are included in our temporal lobe are involved with thought about places, and parietal area is involved in aggregation of sensory input.

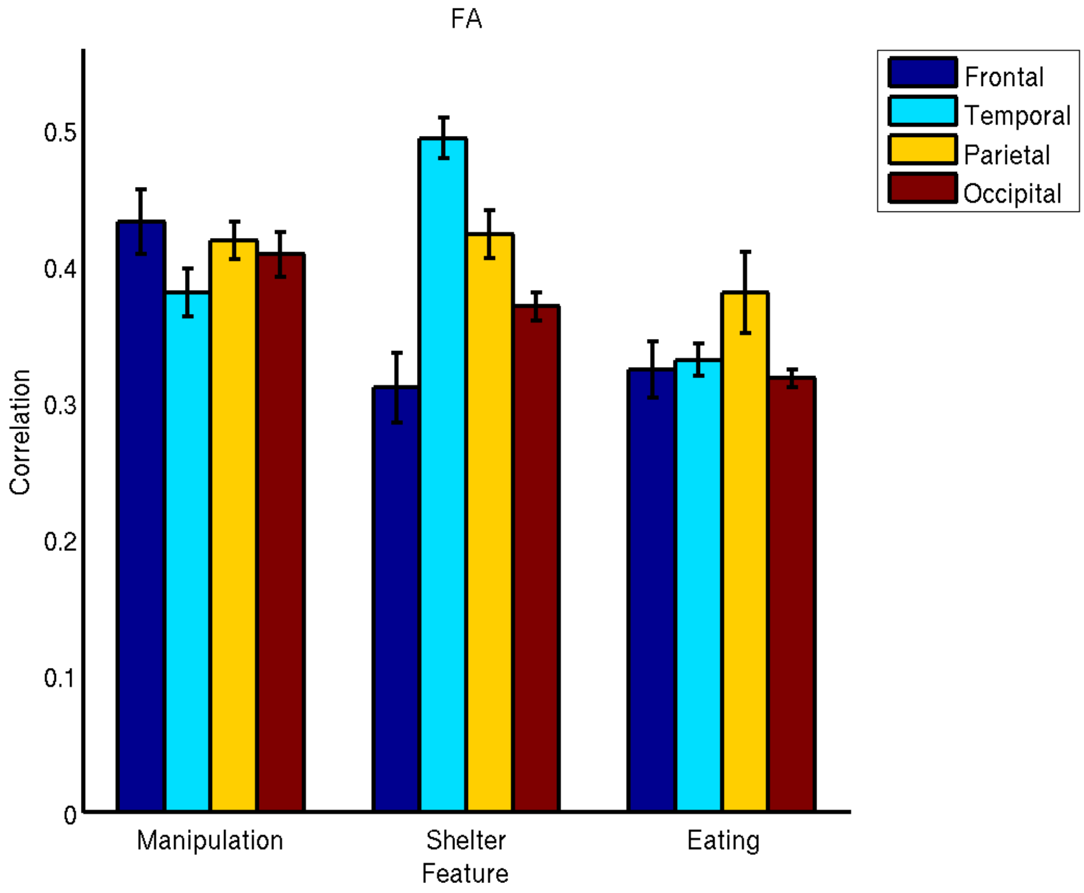


Figure 5.3 Correlating the latent features with human ratings of shelter, manipulation, and eating factor. Different brain regions infer different latent features: the frontal lobes tend to infer latent features that correlate with human ratings of manipulation vector, whereas temporal and parietal lobes tend to infer latent features that correlate with human ratings of shelter and eating factor, respectively.

For each latent feature inferred, we also correlate the latent feature vector (column vector describing which objects possess this feature) with each of the MRC psycholinguistic measure (column vector describing the psycholinguistic score of the 60 objects). Figure 5.4 shows the maximum correlation between the latent feature vector and MRC feature vector, averaged across subjects. The error bar indicates 95% confidence interval, where the distribution of statistic is estimated from the 900 Gibbs samples (excluding the first 100 burn in samples). Again, different brain regions infer different latent features: the frontal lobes tend to encode features that correlate with meaningfulness, although the correlation is not significantly different from that of the temporal and parietal lobe. The parietal lobe tends to encode features that correlate with concreteness and imaginability feature, compared to the other brain regions. The temporal lobe tends to encode features that correlate with number of phonemes in a word, consistent with the existing conjecture that the temporal lobe is involved in speech production. Notice that the occipital lobe tends to encode features that correlate with the number of letters, but not the number of phonemes.

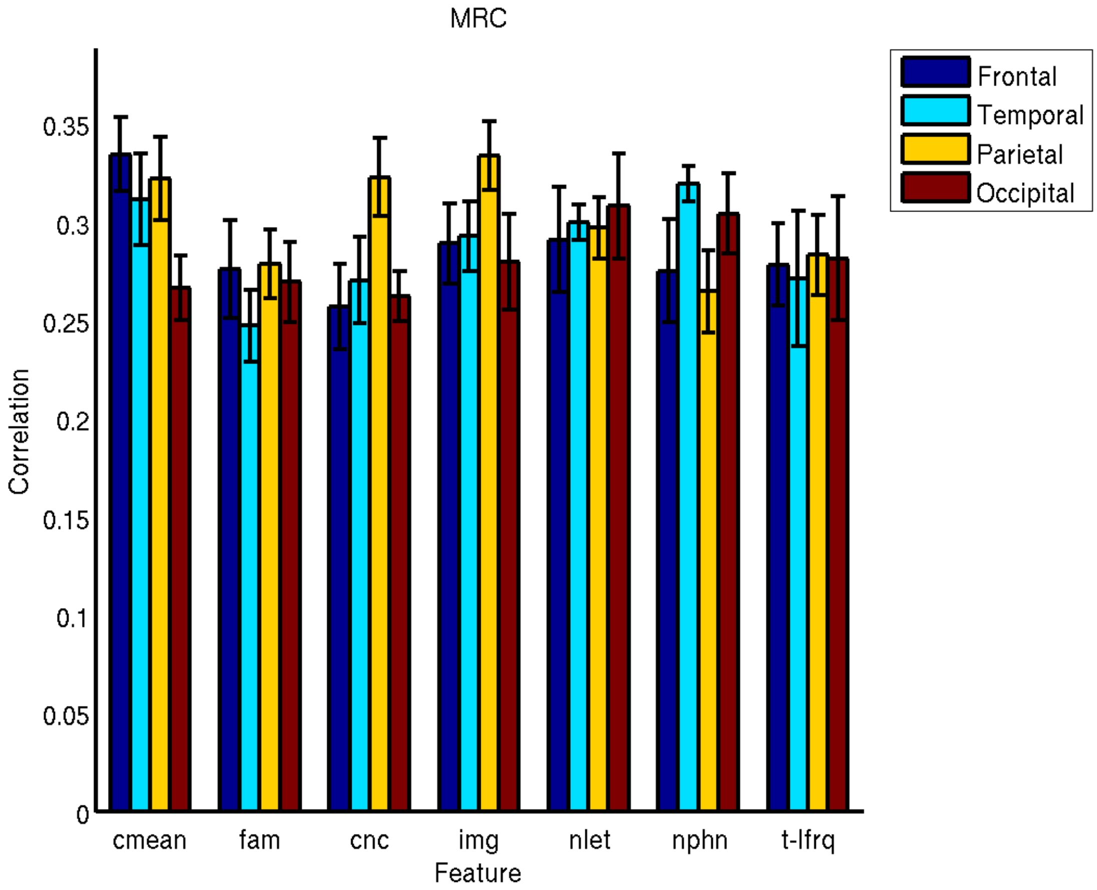


Figure 5.4 Correlating the latent features with MRC psycholinguistics features. Different brain regions infer different latent features: the frontal lobes tend to encode features that correlate with meaningfulness. The parietal lobe tends to encode features that correlate with concreteness and imaginability feature. The temporal lobe tends to encode features that correlate with number of phonemes in a word.

5.3.2 Noun-noun Concept Combination

We now study the neural composition of noun-noun concept combination. In our experiment, we have recorded the brain activity for noun-noun phrases, as well as the corresponding nouns in isolation. One direct way of assessing compositionality is to compare the brain activity for phrases to individual words. Our hypothesis is the brain activity for property-based interpretation should be more similar to the head word (since only one property of the modifier word is extracted to modify the head word), whereas the brain activity for relation-based interpretation should be similar to both the modifier and head word (since the modifier object is realized in its entirety to the head object as a whole).

In this analysis, we measure the similarity between the brain activity for the phrases and the corresponding modifier noun and the head noun. We used correlation as our similarity measure. Each of the 40 stimulus items is represented a vector of brain activity measured at 120 most stable voxels whose responses to the 20 different nouns were most stable across presentations. Rows 1 and 2 in Table 5.4 show the correlation analysis. Unlike our hypothesis, brain activity for property-based interpretation is more similar to the modifier word than the head word ($r = 0.24 > 0.13$), whereas brain activity for relation-based interpretation is more similar to the head word than the modifier word ($r = 0.19 > 0.16$). The difference in correlation with the modifier and head word is statistically significant ($p = 0.05$) for the property-based interpretations, but not for relation-based interpretations. One possible explanation is that property-based interpretations are less intuitive; as a result, people think more about the modifier word to find a fitting property.

In addition, we also measure the similarity between the human ratings for the phrases and the corresponding modifier noun and the head noun. Each of the 40 stimulus items is represented by a vector of human ratings on the three semantic factors. Rows 3 and 4 in Table 5.4 show the correlation analysis. The human ratings for both property-based and relation-based interpretations correlate highly with the head word (p = 0.62, 0.67, respectively). Human ratings for relation-based interpretations correlates weakly with the modifier word (p = 0.10), whereas there is almost no correlation between human ratings for property-based interpretation and the modifier word (p = -0.04). This pattern is clearly different from the brain activity. This may be a result of high correlation between the human ratings of property-based and relation-based interpretation.

Finally, we measure the similarity between the latent semantic feature vector for the phrases and the corresponding modifier noun and the head noun. Each of the 40 stimulus items is represented by a vector of latent semantic features. Latent semantic feature vector for relation-based interpretations correlates. Rows 5 and 6 in Table 5.4 show the correlation analysis. The data-driven feature representation preserves the pattern in brain activity. Namely, the data-driven features for property-based interpretation are more similar to the modifier word (p = 0.24), whereas the data-driven features for relation-based interpretations are more similar to the head word (p = 0.21). The difference in correlation with the modifier and head word is statistically significant for the property-based interpretations, but not for relation-based interpretations.

Table 5.4 Correlations between phrases and nouns

Space	Type	Mod	Head	Property	Relation
Brain activity	Property	0.24	0.13	1.00	0.35
	Relation	0.16	0.19	0.35	1.00
Human rating	Property	0.03	0.62	1.00	0.83
	Relation	0.10	0.66	0.83	1.00
Data-driven (ILFM)	Property	0.23	0.16	1.00	0.26
	Relation	0.18	0.20	0.36	1.00

5.4 Discussion and Conclusion

In this study we use a generative probabilistic model to describe how fMRI-measured brain activity is generated from some latent semantic representation. More specifically, a linear-Gaussian infinite latent feature model with an Indian Buffet Process prior can be used to derive a binary feature representation of object knowledge from the brain activity recorded when people view and contemplate about properties associated with an object.

Compared to the more traditional factor analysis or multi-dimensional scaling, there are several advantages of using ILFM to model the semantic representation that underlie brain activity: ILFM 1) offers a formal probabilistic account of the brain activity, 2) automatically determines the number of features that are manifested in the data, and 3) allows different number of features to be inferred per words. In this study, we use a binary representation of the feature matrices, but it can be easily extended to a continuous representation. Griffiths & Ghahramani (2005) showed that the binary matrix Z can be combined with a continuous matrix V to define a richer representation.

There are several extensions of this work. First, in this study we try to interpret the learned latent semantic features by comparing the vectors to human ratings of three semantic factors and MRC psycholinguistic word features, but one shouldn't stop here. One obvious direction is to compare the feature vector with other word feature vectors, such as behavior feature-norming features Cree & McRae (2003) and word co-occurrence statistics (Church & Hanks, 1990). Second, ILFM can be used to find an optimal set of features in word co-occurrence based representation. For instance, we can find the word features that have the highest correlation with the data-driven features recoverd by ILFM and use these features as the basis set for the word co-occurrence based feature representation. We can compare the performance of a biologically-informed word representation and the manually selected 25 verbs in the leave-two-words-out classification task Mitchell et al., (2008) and see if a biologically-informed model of semantic representation yield better classification.

6 THESIS CONTRIBUTION AND FUTURE WORK

6.1 From mass univariate analysis to multivariate analysis

One of our major contribution is to shift the focus of fMRI analysis from characterizing the location of brain activity (traditional univariate approaches) toward understanding how patterns of brain activity differentially encode information in a way that distinguishes among different stimuli. Functional neuroimaging research has been mostly focused on attempting to identify the functions of cortical regions. In particular, language-related brain imaging research has been limited to relatively coarse analyses (e.g. high-level features such as animacy or part-of-speech). Here we present one of the first studies to investigate cortex-wide representations of semantic knowledge and further apply it in a finer classification task (e.g. identifying a concept among other concepts). Machine learning classifiers were trained to decode which linguistic concepts a person is contemplating. The distributed pattern of brain activity encodes the meanings of linguistic concepts.

The debate of localist vs. distributed processing can be directly verified and grounded by the observed patterns of brain activity. Contemporary leaders in computational models of reading are divided over whether a localist or distributed processing account is more appropriate. The former assumes processing the meaning of a concept to be localized in isolated voxels, whereas the latter assumes such to be a pattern of activation distributed over a number of brain regions. Whereas mass univariate analysis aims to reveal focal areas that are responsible for processing, multivariate analysis aims to reveal a network of units that are responsible for processing. The advantage of multivariate analysis is that it can detect cases where a cognitive process involves simultaneous activation in multiple voxels / areas, which simply cannot be done with a mass univariate analysis. For instance, mental state decoders of multiple classes can only be constructed from the distributed pattern of brain activity that encodes the meanings of linguistic concepts, but not from brain activity in isolated voxels, which is often limited to one or two classes. The success of multivariate classifiers supports a distributed processing account.

Nonetheless, voxel-scale activities and neuro-scale activities do inform multivariate analysis. For instance, various mass univariate studies show that parahippocampus place areas are involved in the processing of places; in-depth neuronal studies show how pre-motor areas are involved in processing motion planning. Indeed, in the sixty-words experiment, activities at parahippocampus place and motor areas yield the best discrimination of buildings and vehicles, respectively. Activation at the parahippocampus place area may correspond to thoughts about location of the buildings, whereas activation at the pre-motor areas may correspond to thoughts about operating a vehicle or movement of a vehicle.

6.2 From classifier analysis to intermediate semantic analysis

By postulating that brain activity is based on an intermediate semantic level of representation (derived from word co-occurrence statistics or feature norming studies), this work enables a computational model that can help predict brain activity for a new stimulus, based on its relation to the semantic level of representation. Compared to a discriminative classifier like SVM, a generative model that utilizes an intermediate semantic representation generalizes better across people. The intermediate representation allows us to extrapolate the neural activity for previously unseen words. Akin to the recent multivariate fMRI analysis, which shifted the focus from localizing brain activity toward understanding how patterns of brain activity encode information in an intermediate semantic representation, we take one step further and ask 1) what information might be encoded to enable such discrimination? and 2) what is the nature of this semantic representation?

In this work, we have utilized word co-occurrence statistics and feature norming features as the intermediate semantic representation. An extension of this work is to utilize different sources of linguistic knowledge or different linguistic corpuses, like the Brown corpus (Kucera & Francis, 1967) or BNC (Burnard, 1995). A comparison of the performance of a feature representation derived from one source of linguistic knowledge versus another source in the leave-two-words-out classification task (Mitchell et al., 2008) may reveal which model of semantic representation better accounts the semantic representation in humans, or if information derived from one particular source better reflects the set of features that are recalled at a particular task. Palatucci et al. (2009) and Pereira, Botvinick, & Detre (2010) have done this by utilizing features derived from norming studies of 20 characteristic questions collected over Mechanical Turk, and features derived from definitive articles in Wikipedia, respectively.

There are several advantages to working with an intermediate semantic representation. In this study, we have demonstrated how learning the mapping between feature and neural activation enables a predictive theory that is capable of extrapolating the model of neural activity to previously unseen words, which cannot be done with a discriminative classifier. Another advantage of working with an intermediate semantic representation is that features in the intermediate semantic representation are more likely to be shared across experiments. For example, in one experiment, the participant may be presented with the word *dog*, while the word *cat* is shown in another experiment. Even though the individual category differs, there are many features that are shared (e.g. is a pet, has 4 legs, etc.) between the two words. Learning the mapping between features and voxel activation, instead of the mapping between categories and voxel activation, may facilitate data to be shared across experiments. This is especially important when brain imaging data are relatively more expensive to acquire and that many classifier techniques would perform significantly better if more training data were available. Rustandi et al. (2009) used canonical correlation analysis (CCA) to find the common dimension among multiple fMRI datasets. By learning the common dimension (a form of intermediate representation), they were able to better predict brain activations than when each subject's data is analyzed separately. Furthermore, by utilizing a knowledge base of semantic properties, Palatucci et al. (2009) were able to train classifiers in the zero-shot learning problem, where classifiers must learn to predict novel classes that were omitted from the training set.

Although we propose that brain activity is based on an intermediate semantic level of representation and propose a specific implementation of intermediate semantic representation, we do not necessarily suggest that these are serious psychological proposals. These semantic representations are not intended to reflect the actual representation in the brain. Instead, they are capturing some of the same information as the representations in the brain. For instance, even though corpus co-occurrence statistics provide a useful semantic representation in our classification task, the brain does not necessarily store or represent these statistics. However, the patterns of brain activity when contemplating about different concepts do reflect aspects of these statistics. Cognitive psychologists are encouraged to extend the intermediate representation and experiment with different modeling methodologies.

6.3 The nature of semantic representation

Nonetheless, despite the success of this model, the work leaves open the question of how to determine the optimal set of semantic features and the nature of semantic representation. Bayesian probabilistic analysis offers a new approach to characterize semantic presentation by inferring the most likely feature structure directly from the patterns of brain activity. In this study, we use an infinite latent feature model to infer a binary representation of the feature matrices. Compared to the more traditional factor analysis or multi-dimensional scaling, there are several advantages of using ILFM to model the semantic representation that underlie brain activity: ILFM 1) offers a formal probabilistic account of the brain activity, 2) automatically determines the number of features that are manifested in the data, and 3) allows different numbers of features to be inferred per words. More importantly, the neurally-inspired semantic representation is consistent with some existing conjectures regarding the role of different brain areas in processing different psycholinguistics features, and suggests a multimodal semantic representation.

One of the open questions regarding the nature of semantic representation is the repertory of concepts: are concepts hard-wired due to some genetic/evolutionary constraints, or environmentally determined? For instance, to what extent do individual languages/cultures make a difference? Pereira (2007) was able to train cross-language classifiers to predict brain activity from viewing stimuli that are in one of two different languages, Portuguese or English. His result suggested that there are certain aspects of semantic knowledge that can be generalized across languages.

Another question to consider is the characterization of the category and prototypes of concepts. Why do we store the particular concepts that we do, and group them into the particular categories that we do? In this work, we showed that category classification (identifying the category of a concept) is more difficult than exemplar[d] classification (identifying a concept among other concepts). Mitchell et al. (2008) also showed that within-category exemplar classification (identifying a concept among other concepts in the same category) is much harder

[d] Notice, the term “instance” is used more often in the prototype theory (Rosch, 1970) to denote a concept in a category. In this work we use the term concept, object, exemplar, and instance interchangeably.

than between-category exemplar classification (identifying a concept among other concepts in a different category). The difficulty in discriminating categorical aspects of concepts may be a consequence of the experimental task, where we asked participants to think specifically of the stimulus object (the appearance, the purpose of the object, and how one physically interacts with the object), and not its relation to other objects or the category of the objects that people may normally think of. Future experiments may design tasks to investigate the effect of the category and prototypes of objects. For instance, how do people distill many varied instances of a concept into some compact aggregate representation? What is the nature of a prototype, and how are prototypes realized neurologically?

Furthermore, is the representation amodal (grounded in symbolic conceptual entries) or modal (embodied in different sensory modalities)? Or, do both types of representations use and activate selectively depending on the task? If both types of representations are present, do they involve different cognitive processes (e.g. dual-route processing), or can they both emerge from the same cognitive process (e.g. connectionist account)? Our results from multivariate analysis (word meaning encoded in patterns of distributed brain activity) and latent semantic analysis (identification of modality-specific word features) suggest a modal representation. However, future studies are required to fully address this question.

Finally, is the representation localized or distributed? ILFM has shown that the meaning of each word is distributed across many latent features, supporting a distributed representation account. Notice that in this work we distinguish between semantic representation and semantic processing. Whereas the former describes how the meaning of a concept is represented (is the semantic content of a word atomic or compounded?), the latter describes how the processing of a concept is distributed spatially (does the processing of a concept involve brain activity localized in a few voxels or distributed across a range of brain regions?). Both representation and processing can be either localized or distributed – the two need not be mutually exclusive. In our work, we have shown distributed accounts for both the semantic representation and semantic processing.

6.4 From single nouns to compound phrases

One of the ultimate goals in computational neurolinguistics is to account for the human language that enables communication. The milestones to achieving this goal include a better understanding of how the human brain processes single nouns, phrases, and sentences. Each chapter in this thesis work fills in a piece of the puzzle. We started with lexical semantics, and then proceeded to combinatorial semantics. In particular, our work has shown that the difference in brain activity when contemplating an isolated noun *(e.g. dog)* vs. the same noun modified by an adjective *(e.g. strong dog)* can be detected by machine learning classifiers. The distributed pattern of brain activity contains sufficient signal to decode between a property-based interpretation *(e.g. a coat that is bright yellow)* and a relation-based interpretation *(e.g. a coat that is used to protect corn)* of the identical visual stimuli *(e.g. corn coat)*.

Due to the inherent limitations in the temporal properties of fMRI data, in most of this thesis work we consider only the spatial distribution of the brain activity after the stimuli are comprehended, and do not attempt to model the cognitive process of comprehension. One extension is to model the process of combination and not just the comprehended concepts. Does the temporal resolution of fMRI encode sufficient signal? Polyn et al. (2005) analyzed the time-series data of fMRI to test the contextual reinstatement hypothesis, which postulates that when asked to recall memories, people use reinstated activity in a top-down fashion to cue for additional details. They showed that category-specific brain activity during a free-recall period correlated more with brain activity of matching categories during a prior study period. We can adopt an approach similar to Polyn et al. (2005) and correlate the brain activity of the noun phrases to the brain activity of each word in the phrase. For instance, time-series analysis of the activity pattern may reveal if participants first recall features associated with each word in the phrase and then combine them to interpret the phrase as a whole. Time-series analysis may also reveal to what extent do the patterns we see represent processing (dynamic, transitory), and what extent representations (static, enduring)?

Alternatively, there are other types of brain imaging techniques that offer better temporal (albeit worse spatial) resolutions than fMRI, such as electroencephalography (EEG), magnetoencephalography (MEG), and functional near-infrared spectroscopy (fNIRS). In Mostow, Chang, & Nelson (2011), we demonstrated how a single-channel EEG headset can be used in schools to measure students' mental states while reading. Using its signal from adults and children reading text and isolated words, both aloud and silently, we trained and tested classifiers to tell easy from hard sentences, and to distinguish among easy words, hard words, pseudo-words, and unpronounceable strings. We also identified which EEG components appear sensitive to which lexical features. Better-than-chance performance shows promise for tutors to use EEG at school.

Another extension is to work toward sentence-level analysis and develop neural accounts of sentence processing. This will enable many BCI applications, such as thought-to-text systems that are akin to speech-to-text systems. The first step toward this goal is to collect broader types of concepts, such as abstract words, verbs, and adjectives. Thus, I am motivated to collect brain activity for hundreds, if not thousands, of words with different semantic categories and part-of-speech tags. There will be many research venues that stem out of this effort. For instance, this database will be useful for building a neurologically-informed ontology similar to WordNet. Moreover, reliable recordings for some nouns, verbs, and pronouns will help in developing neural accounts of sentence processing and enable basic level thought identification. This also raises the question of whether our models' organization/representation of single concepts extends to sentence-level analysis?

6.5 From visual perception to semantic cognition

In our task, where the stimulus presentations consist of line drawings with text labels, the voxels extracted by this procedure are mostly in the posterior and occipital regions, since our stimuli consist of easily depicted objects and the visual properties of the stimuli are the most invariant part of the stimuli. Indeed, visual features were among the most important features that account for our neural activation data. If the stimulus presentation consists of only line drawings or text labels, different sets of voxels might be selected. As a result, there is reservation regarding the claim of "mental state" decoding. Are classifiers really decoding the mental state of the higher-level cognition? Or are the classifiers decoding the brain activity that is the result of lower-level perception? The goal of computational neurolinguistics is to get at language, not vision.

We addressed this issue by asking "can we discriminate different interpretations of the same stimuli?" In the noun-noun concept combination experiment, the participant was expected to contemplate a property-based interpretation in one context, but a relation-based interpretation when the same noun-noun phrase was presented in another context. The setup of the experiment poses a challenge for classifiers that obtain its discriminative power from distinguishing the brain activity of low-level visual perceptions. Since the visual stimuli are identical, the discrimination must be made on the semantic differences between the two types of interpretations. An extension of this work is to build a mental state decoder that is capable of word-sense disambiguation of polysemous words *(e.g. bank* can be *financial bank* or *river bank).*

6.6 Computational models of language processing

Marr put forth the idea that one must understand information processing systems at three levels of analysis: 1) computational level (what does the system do), 2) algorithmic level (how does the system represent and perform its task), and 3) implementational level (how is the system physically realized?). Most of the contemporary linguistic or cognitive science research has focused either on the computational or algorithmic level of analysis. Recent advances in brain imaging and machine learning technologies offer a significant new approach to studying language processing in humans. For the first time, theories regarding how linguistic concepts are represented and processed can be grounded by the brain activity while people comprehend words and phrases – the implementational level of analysis. On one hand, theories from computational linguistics can be used to help predict brain activity. For instance, in this work, word co-occurrence features were used in a regression model to help predict brain activity. Moreover, vector-based semantic composition was used to provide a neural account of how people use adjectives to modify the meaning of the noun.

On the other hand, patterns of brain activity can be used to verify linguistic or psycholinguistic theories. In computational linguistics, the cognitive plausibility of language models has primarily been evaluated against collections of subjective intuitions (e.g. semantic feature norms, grammaticality judgments, corpus annotations, dictionaries). Evaluation of the large body of computational linguistics work, based on data-driven distributional approaches, has also relied on hand-crafted resources such as WordNet or data sets manually tagged with a predefined list of categories. Comparison with neural data may provide a more objective yardstick for both models and resources. For instance, in this work, multiplicative semantic composition models of the two-word phrase outperform additive models, consistent with the assumption that people use adjectives to modify the meaning of the noun, rather than conjoining the meaning of the adjective and noun. Moreover, concept combination experiments provide a neural account of how relation-based interpretations are more accessible to humans.

6.7 Contribution to compuational neurolinguistics

Over recent years, machine learning methods have become a crucial analytical tool in cognitive neuroscience. Decoding techniques have dramatically increased the sensitivity of experiments, and therefore the subtlety of cognitive questions that can be asked. At the same time the mental phenomena being studied are moving beyond lower-level perceptual and motor processes which are directly grounded in external measurable realities. However, decoding higher cognition and interpreting the learned behavior of the classifiers used pose unique challenges, as these psychological states are complex, fast-changing and often ill-defined. Furthermore, for the cognitive scientists who use these methods, the primary question is often not "how much" but rather "how" and "why" the patterns of neural activity identified by a machine learning algorithm encode particular cognitive processes.

In this work, I have shown how we can leverage theories from computational linguistics to help decoding language and interpret the learned behavior of the classifiers. Conversely, machine learning methods and brain imaging techniques can also help verify and ground existing theories regarding the nature of semantic representation. This interdisciplinary work has motivated a new research area, computational neurolinguistics. Computational neurolinguistics is an emerging research area which integrates recent advances in computational linguistics and cognitive neuroscience, with the objective of developing cognitively plausible models of language and gaining a better understanding of the human language system. It builds on research in decoding cognitive states from recordings of neural activity, and computational models of lexical representations and sentence processing. Together with Dr. Brian Murphy and Dr. Anna Korhonen, I helped pioneer the field of computational neurolinguistics and co-organized workshops on Computational Neurolinguistics (NAACL-HLT 2010, NIPS 2011 submitted).

Though the field is still in its infancy, many simplifying assumptions were made (e.g. only focus on spatially-distributed patterns, assume a feature-based semantic representation and linearity assumption of feature-voxel mapping, etc.). The field requires techniques that are capable of taking advantage of spatially-distributed patterns in the brain that are separated in space but coordinated in their activity. Methods should also be sensitive to the fine-grained temporal patterns of multiple processes which may proceed in a serial fashion, overlapping or in parallel with each-other, or in multiple passes with bidirectional information flows. Different recording modalities have distinctive advantages: fMRI provides very fine millimeter-level localization in the brain but poor temporal resolution, while EEG and MEG have millisecond temporal resolution at the cost of spatial resolution. Ideally machine learning methods would be able to meaningfully combine complementary information from these different neuroimaging techniques (see e.g. De Martino et al., 2010). Moreover, as the processes underlying higher cognition are so complex, methods should be able to disentangle even tightly linked and confounded subprocesses. Finally, general use algorithms that could induce latent dimensions from neural data, and reveal the "hidden" psychological states, would be a dramatic advance on current hypothesis-driven analytical paradigms.

Advances in computational neurolinguistics require close collaboration between neuroscience, language technology, cognitive psychology, and machine learning. To this end, my thesis work helps advance existing work and initiates new research. Here I have listed several topics of interests within each subfield that are either connected to my thesis work or are motivated by my thesis work. By stimulating discussions among experts in the different fields, I hope to help generating novel insights and new directions for research.

Computational Linguistic Focus

- Contribution:
    - Describe a framework to ground linguistic theories by the patterns of brain activity.
- Future work:
    - Word-level analyses (e.g. corpus semantic models, lexica, lexical relations and ontology, parts-of-speech, word senses, morphology)
    - Phrase-level analyses (e.g. word compounds, meaning composition in multi-word expressions)

Machine Learning Focus

- Contribution:
    - Distributed patterns of brain activity contain sufficient signal to decode differences among nouns and phrases.
    - The generative classifiers that utilize an intermediate semantic representation are applicable to many other problems that involve high-dimensional sparse data.
- Future work:
    - Decoding of cognitive states from neural activity

  - Feature selection and data mining techniques for decoding linguistic information

Neural Science Focus

- Contribution:
  - Present a quantitative model of multiple-word phrases like adjective-noun and noun-noun phrases, which is an important building step toward neural accounts of sentence processing.
- Future work:
  - Brain imaging techniques: fMRI, EEG, MEG, NIRS, including cross-modality analysis (e.g. combining fMRI and EEG)
  - Localizing Regions of Interest (e.g. identify the roles / functions of brain regions)

Cognitive Science Focus

- Contribution:
  - Describe a framework to study the nature of semantic representation and how it is grounded neurologically.
- Future work:
  - Comparisons with behavioral (e.g. priming experiments, eye-tracking, self-paced reading) and elicited data (e.g. semantic feature norms)
  - Biologically plausible connectionist approaches

6.8 Conclusion

To conclude, we are at an especially opportune time in the history of the study of language, when machine learning methods allow us to analyze and model the brain activity when people view and contemplate different objects. An improved understanding of language processing in the brain could yield a more biologically-informed model of semantic representation of lexical knowledge. We therefore look forward to further brain imaging studies shedding new light on the nature of the human representation of semantic knowledge.

REFERENCES

Bandettini, P., Ungerleider, L. G., 2011. From neuron to BOLD: new connections, *Nature Neuroscience* 4 (9), 864–866.

Baron, S., Thompson-Schill, S., Weber, M., Osherson, D., in press. The neural basis of conceptual combination. *Journal of Cogninitive Neuroscience*.

Bemis, D. K., Pylkkanen, L., 2011. Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *Journal of Neurocience* 32 (8), 2801-2814.

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., Prieto, T., 1997. Human Brain Language Areas Identified by Functional Magnetic Resonance Imaging. *The Journal of Neuroscience* 17 (1), 353–362.

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3, 993-1022.

Bookheimer, S., 2002. Functional MRI of language: New approaches to understanding the cortical organization of semantic processing. *Annual Review of Neuroscience* 25, 151–88.

Bullinaria, J., Levy, J., 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavioral Research Methods* 39, 510-526.

Burnard, L., 1995. The Users Reference Guide for the British National Corpus.

Caramazza, A., Shelton, J. R., 1998. Domain-specific knowledge systems in the brain the animate-inanimate distinction. *Journal of Cognitive Neuroscience* 10 (1), 1-34.

Chang, K. M., Cherkassky, V. L., Mitchell, T. M., & Just M. A., 2009. Quantitative modeling of the neural representation of phrases: How vector-based models of semantic composition can account for fMRI activation. *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, Suntec, Singapore, 638-646.

Chang, K. M., Mitchell, T. M., & Just M. A., 2010. Quantitative modeling of the neural representations of objects: How semantic feature norms can account for fMRI activation. *NeuroImage* 56, 716-727.

Church, K. W., Hanks, P., 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16, 22-29.

Coltheart, M., 1981. The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, 33A, 497-505.

Costello, F., Keane, M., 2001. Testing two theories of conceptual combination: Alignment versus diagnosticity in the comprehension and production of combined concepts. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 27 (1): 255-271.

Cox, D. D., Savoy, R. L., 2003. Functioning magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* 19, 261-270.

Cree, G. S., McRae, K., 2003. Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General* 132 (2), 163-201.

Dapretto, M., Bookheimer, S. Y., 1999. Form and content: Dissociating syntax and semantics in sentence comprehension. *Neuron* 24, 427–432.

Davatzikosa, C., Ruparelb, K., Fana, Y., Shena, D. G., Acharyyaa, M., Lougheadb, J. W., Gurb, R. C., Langlebenb, D. D., 2005. Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage* 15, 663-668.

Formisano, E., De Martino, F., Bonte, M., Goebe, R., 2008. “Who” Is Saying “What”? Brain-Based Decoding of Human Voice and Speech. *Science* 32, 970-973.

Friston, K., Ashburner, J., Frith, C., Poline, J.-B., Heather, J., Frackowiak, R., 1995. Spatial registration and normalization of images. *Human Brain Mapping* 2, 165-189.

Friston, K.J., 2005. Models of brain function in neuroimaging. *Annual Review of Psychology* 56, 57-87.

Gagne, C. L., 2000. Relation-based combinations versus property-based combinations: A test of the CARIN theory and the dual-process theory of conceptual combination. *Journal of Memory and Language* 42, 365–389.

Geman, S., Geman, D., 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (6): 721–741.

Graves, W. W., Binder, J. R., Desai, R. H., Conant, L. L., Seidenberg, M. S., 2010. Neural correlates of implicit and explicit combinatorial semantic processing. *NeuroImage* 53, 638–646.

Griffiths, T. L., Ghahramani, Z., 2005. Infinite latent feature models and the Indian buffet process. *Gatsby Unit Technical Report* GCNU-TR-2005-001.

Guyon, I., Boser, B., & Vapnik, V., 1993. Automatic capacity tuning of very large VCdimension classifiers. *Advances in Neural Information Processing Systems* 5, 147–155.

Hampton, J. A., 1997. Conceptual combination: Conjunction and negation of natural concepts. *Memory and Cognition* 25, 888–909.

Hanson, S. J., Matsuka, T., Haxby, J. V., 2004. Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: Is there a “face” area? *NeuroImage* 23 (1), 156–166.

Hardoon, D. R., Mourao-Miranda, J., Brammer, M., Shawe-Taylor, J., 2007. Using image stimuli to drive fMRI analysis. *NeuroImage* 37, 1250–1259.

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

Haxby, J.V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2429.

Haynes, J. D., Rees, G., 2005. Predicting the stream of consciousness from activity in human visual cortex. *Current Biology* 15 (14), 1301-1307.

Haynes, J. D., Rees, G., 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience* 8, 686 - 691.

Haynes, J. D., Rees, G., 2006. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience* 7, 523-534.

Heeger, D. J., Ress, D., 2002. What does fMRI tell us about neuronal activity? *Nature Reviews Neuroscience* 3, 142–151.

Hofmann, T., 1999, Probabilistic Latent Semantic Analysis, *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, Stockholm, Sweden, 289-296.

Ishai, A., Ungerleider, L. G., Martin, A., Haxby, J. V., 2000. The representation of objects in the human occipital and temporal cortex. *Journal of Cognitive Neuroscience* 12 (Suppl 2), 35-51.

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., Thulborn, K. R., 1996. Brain activation modulated by sentence comprehension. *Science* 274, 114-116.

Just, M. A., Cherkassky, V. L., Aryal, S., Mitchell, T. M., 2010. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS ONE*, 5, e8622.

Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience* 8, 679 - 685.

Kay, K. N., Naselaris, T., Prenger, R. J., Gallant, J. L., 2008. Identifying natural images from human brain activity. *Nature*, 452, 352-355.

Kemp, C., Shafto, P., Berke, A., Tenenbaum, J. B., 2007. Combining causal and similarity-based reasoning. *Advances in Neural Information Processing Systems* 19.

Kintsch, W., 2001. Prediction. *Cognitive Science*, 25 (2), 173-202.

Kipper, K., Dang, H. T., & Palmer, M., 2000. Class-based construction of a verb lexicon. *Proceedings of the 17th National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, Austin, Texas, 691–696.

Kucera, H., Francis, W.N., 1967. Computational analysis of present-day American English. *Brown University Press*, Providence, Rhode Island.

Landauer, T.K., and Dumais, S. T., 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104 (2), 211-240.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., Oeltermann, A., 2001. Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412, 150–157.

Lund, K., Burgess, C., 1996. Producing high-dimensional semantic spaces from lexical cooccurrence. *Behavior Research Methods, Instruments Computers* 28, 203–220.

Mashal, N., Faust, M., Hendler, T., Jung-Beeman, M., 2007. An fMRI investigation of the neural correlates underlying the processing of novel metaphoric expressions. *Brain and Language* 100, 115-126.

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* 53, 360–373.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E., 1953. Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21 (6): 1087–1092.

Miller, G. A,, 1995. WordNet: A lexical database for English. *Communications of the ACM* 38, 39-41.

Mitchell, T., 1997. Machine Learning. McGraw Hill Higher Education, Columbus.

Mitchell, J., Lapata, M., 2008. Vector-based models of semantic composition. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Columbus, Ohio, 236-244.

Mitchell, T., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M. A., Newman, S. D., 2004. Learning to decode cognitive states from brain images. *Machine Learning* 57, 145-175.

Mitchell, T., 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.

Mostow, J., Chang, K. M., Nelson, J., 2011. Toward Exploiting EEG Input in a Reading Tutor. *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, Auckland, New Zealand.

Norman, K. A., Polyn, S. M., Detre, G. J., Haxby, J. V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences* 10 (9), 424-430.

O'Toole, A. J., Jiang, F., Abdi, H., Haxby, J. V., 2005. Partially distributed representations of objects and faces in ventral temporal cortex. *Journal of Cognitive Neuroscience* 17, 580-590.

Palatucci, M., Pomerleau, D., Hinton, G., Mitchell, T. M., 2009. Zero-Shot Learning with Semantic Output Codes. *Advances in Neural Information Processing Systems* 21.

Pereira, F., 2007. Beyond Brain Blobs: Machine Learning Classifiers as Instruments for Analyzing Functional Magnetic Resonance Imaging Data. *Ph.D. thesis*.

Pereira, F., Botvinick, M., Detre, G., 2010. Learning semantic features for fMRI data from definitional text. *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, 1–9.

Polyn, S. M., Natu, V. S., Cohen, J. D., Norman, K. A., 2005. Category-Specific Cortical Activity Precedes Retrieval During Memory Search. *Science* 310 (5756), 1963-1966.

Rodd, J. M., Davis, M. H., Johnsrude, I. S., 2005. The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex* 15, 1261-1269.

Rosch, E., 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3), 1975, 192-233.

Rosch, E., Mervis, C. B., 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7, 573–695.

Rustandi, I., Just, M. A., Mitchell, T. M., 2009. Integrating multiple-study multiple-subject fMRI datasets using canonical correlation analysis. *Proceedings of the MICCAI 2009 Workshop: Statistical modeling and detection issues in intra- and inter-subject functional MRI data analysis.*

Saykin, A. J., Johnson, S. C., Flashman, L. A., McAllister, T. W., Sparling, M., Darcey, T. M., Moritz, C. H., Guerin, S. J., Weaver, J., Mamourian, A., 1999. Functional differentiation of medial temporal and frontal regions involved in processing novel and familiar words: an fMRI study. *Brain* 122 (10), 1963-1971.

Shinkareva, S. V., Malave, V. L., Mason, R. A., Mitchell, T. M., Just, M. A., 2007. Cross-modal identification of semantic categories in words and pictures from fMRI brain activation. Poster presentation. *Cognitive Neuroscience Society*, New York, NY.

Shinkareva, S. V., Mason, R. A., Malave, V. L., Wang, W., Mitchell, T. M., Just, M. A., 2008. Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS ONE* 3 (1): e1394.

Snodgrass, J. G., Vanderwart, M., 1990. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 174-215.

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J. B., Lebihan, D., Dehaene, S., 2006. Inverse retinotopy: Inferring the visual content of images from brain activation patterns. *NeuroImage* 33 (4), 1104-1116.

Tzourio-Mazoyera, N., Landeaub, B., Papathanassioua, D., Crivelloa, F., Etarda, O., Delcroixa, N., Mazoyerc, B., Joliota, M., 2002. Automated anatomical labeling of activations in

SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* 15, 273-289.

Wu, L. L., Barsalou, L. W., 2009. Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica* 132 (2), 173-189.