Carnegie Mellon University

Language Technologies Institute

Thesis Proposal

Doctor of Philosophy

Quantitative modeling of the neural representation of nouns and phrases

Date of Submission:

Submitted by: Kai-min Kevin Chang

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

Professor Tom Mitchell
Machine Learning Department, Carnegie Mellon University

Professor Charles Kemp
Department of Psychology, Carnegie Mellon University

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

Abstract

Recent advances in functional Magnetic Resonance Imaging (fMRI) offer 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. In the proposed work, we used fMRI to study the cortical systems that underpin semantic representation while people comprehended linguistic concepts like concrete objects, adjective-noun phrases, or noun-noun concept combinations. The thesis of this research is that the distributed pattern of neural activity encodes the meanings of linguistic concepts and intermediate semantic representation can be used to model how brain composes the meaning of words or phrases in terms of more primitive semantic features. 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. By learning 1) the correspondence between features in semantic models and neural activity, and 2) the semantic composition model that governs how words are combined to form phrases, this work enables a predictive theory that is capable of extrapolating the model of the neural activity to previously unseen words and phrases.

1 Introduction

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 neural 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. Recent 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 neural activity

(Cox and Savoy, 2003; O'Toole et al., 2005; Haynes and Rees, 2006; Mitchell et al., 2004). Given these early succesess in using fMRI to discriminate categorial information, we like to find out if a similar approach can be used to study the representation of linguistic concepts like noun and phrases. Thus, the first question we ask is:

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

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 dataset, 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. 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 hidden 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 brain composes the meaning of words or phrases in terms of more primitive semantic features?

To answer these questions, we designed an object-contemplation task where participants view and think about words or multi-words phrases while their activities are recorded by fMRI devices. To date, two brain imaging experiments were conducted to record neural activation patterns obtained while subjects comprehended concrete objects or adjective-noun phrases.

2 Thesis Statement

The thesis of this research is that the distributed pattern of neural activity encodes the meanings of linguistic concepts and intermediate semantic representation can be used to model how brain composes the meaning of words or phrases in terms of more primitive semantic features. Our goal is to build a computational model of the neural activity when people contemplate noun and phrases. Our model postulates that the brain activity is based on a hidden intermediate semantic level of representation. By learning the correspondence between features in the semantic model and neural activity, this work enables a predictive theory that is capable of extrapolating the model of the neural activity to previously unseen words. By learning the semantic composition model that governs how words are combined to form phrases, this work furthers enables the predictive theory to extrapolate the model of the neural activity to previously unseen phrases.

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?

3 Related Work

Haxby et al. (2001) was one of the first studies to apply multivariate analysis to study patterns of fMRI activity. They showed distinct pattern of response in ventral temporal cortex could be found while participants viewed faces and objects. Their result supported a distributed and overlapping representations 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 neural activity (Cox and Savoy, 2003; O'Toole et al., 2005; Haynes and 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), orientation of invisible stimuli (Haynes & Rees, 2005), lie detection (Davatzikos, 2005), stream of consciousness (Haynes & Rees, 2005), etc.

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 noun and phrases. Though, not only are we interested in classifiers that can decode the mental states of participants in a particular dataset, we wish to characterize the underlying principles that would allow for extensibility to novel stimuli. To achieve this, we propose a generative model that represents word meaning with a vector of primitive features and learns the mapping between feature and neural activation. 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.

To account for the neural activity observed while participants contemplate phrases, our work relied on successful characterization of lexical semantic representation and semantic composition models that governs how words are combined to form phrases. There have been a variety of approaches from different scientific communities trying to characterize semantic representations. Linguists have tried to characterize the meaning of a word with feature-based approaches, such as semantic roles (Kipper et al., 2006), 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 through feature-norming studies (Cree & McRae, 2003) in which human participants are asked to list the features they associate with various words. There are also efforts to recover the latent semantic structure from text corpora using techniques such as LSA (Landauer & Dumais, 1997) and topic models (Blei et al., 2003). In the proposed work, we adopted the vector-based approach to semantic representation with word cooccurrence data. One advantage of using word cooccurrence data is that semantic features can be computed for any word in the corpus, unlike behavioral feature norming studies or other feature-based approaches where semantic features can only be computed for words included in the experiment.

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 multiplicative and additive models can be used to model human behavior in similiarity judgements when human participants were presented with a reference containing a subject-verb phrase (e.g., *horse ran*) and two landmarks (e.g., *galloped* and *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. 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.

4 Approach

In this section, we discuss the general experiment paradigm and modeling methodology used throughout our brain imaging studies. To date, two brain imaging experiments were conducted to record neural activation patterns obtained while subjects comprehended concrete objects or adjective-noun phrases. While the individual experiment design differs (e.g. list of stimuli, presentation onset, duration), similar signal processing methods, classifier analysis, and vector-based semantic composition models are used. Thus, some central information about the brain imaging experiments and modeling methodology is presented here.

4.1 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 of the same properties of the stimulus object consistently during multiple 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 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 schematic representation of the design used in the 60 concrete objects experiment is shown in Figure 1.

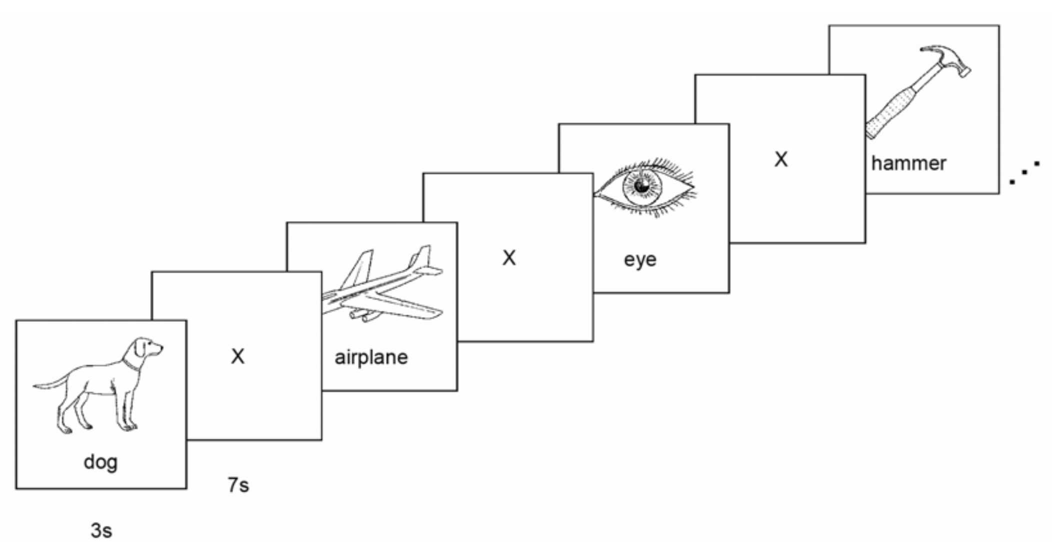

Figure 1 Schematic representation of experimental design for the 60 concrete object experiment.

4.2 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. 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-mm^3 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 cogntive process of comprehension.

4.3 Decoding mental states

To find out if the distribution of neural activity encode sufficient signal to decode linguistic concepts like noun 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: $mean\_PSC \rightarrow Y_i$, 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.

Three classifiers were compared: a Support Vector Machine (SVM) classifier, a Gaussian Naïve Bayes (GNB) classifier, 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. The GNB classifier is a generative classifier that models the joint distribution of class Y and attributes. It makes a conditional independence assumption of the attributes $X_1,\ldots, X_n$ given Y. The classification rule is then:

$$Y \leftarrow \arg\max_{y_k} \prod_i P(X_i \mid Y = y_k)$$

Furthermore, we proposed a nearest neighbor classifier that 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 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.

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 pairwise correlation between all stimuli across presentations. The focus on the most stable voxels effectively increased the signal-to-noise ratio in the data and facilitated further analysis by classifiers.

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 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.

4.4 Modeling intermediate semantics

To find out if models of semantic representation be used to model how 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 2 depicts the modeling framework for multi-words phrases. In the following sections, we will discuss the three steps in detail.

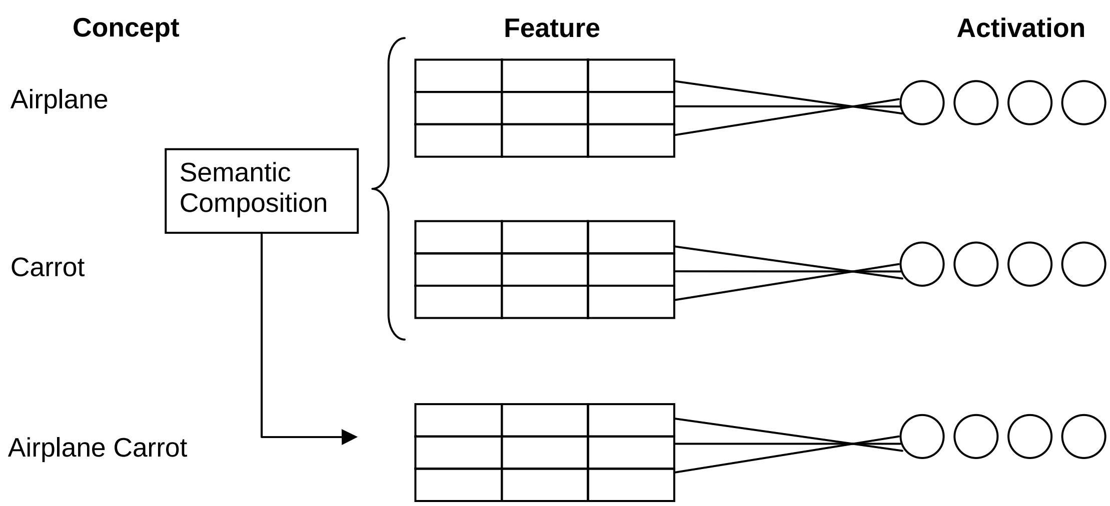

Figure 2 Modeling framework of the intermediate semantic representation.

4.5 Lexical Semantic Representation

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 and Hanks, 1990). Consequently, we encoded the meaning of a word as a vector of intermediate semantic features computed from the cooccurrences 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 (Caramazza and Shelton, 1998), we selected 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* . In addition, a subset of 5 sensory verbs: *see*, *hear*, *smell*, *eat*, and *touch* was selected. These verbs generally correspond to basic sensory and motor activities, actions performed on objects, and actions involving changes in spatial relationships.

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 1 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.

	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

Table 1 The lexical semantic representation for strong and dog.

4.6 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 vectors 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 the 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. Notice that the former assumption of the multiplicative model corresponds to the modifier-head interpretation where adjectives are used to modify the meaning of nouns.

Table 2 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.

	See	Hear	Smell	Eat	Touch
Adj	0.63	0.06	0.26	0.03	0.03
Noun	0.34	0.06	0.05	0.54	0.02
Add	0.96	0.12	0.31	0.57	0.04
Multi	0.21	0.00	0.01	0.01	0.00

Table 2 The semantic representation for strong dog under the adjective, noun, additive, and multiplicative models.

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 et al., 2003).

4.7 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 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 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 ith intermediate semantic feature for word w, β_{vi} is the regression coefficient that specifies the degree to which the ith 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 superimposition of contributions from different sources, and has provided a useful first order approximation in the field (Mitchell et al., 2008).

5 Results to Date

5.1 Concrete Object Comprehension

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. They developed a generative model 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 level. The distributed pattern of neural activity encodes the meanings of words, and the model's success indicates some initial access to the encoding.

5.1.1 *Experimental Paradigm*

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.

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 3 lists the 60 stimuli.

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

Table 3 List of 60 words used in conceete object contemplation task.

5.1.2 *Does the distribution of neural activity encode sufficient signal to classify concrete objects?*

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 classifier 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. The SVM classifier achieved an average of 84% rank accuracy. For the between-participants analysis, a classifier 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. The SVM classifier achieved an average of 63% rank accuracy. All classification accuracies were significantly ($p < 0.05$) different from a chance level of 50%.

5.1.3 Distinguishing between the activation of two unseen stimuli

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. However, discriminative classification provides a characterization of only a particular dataset, 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. Then the model can predict the activation for a new stimulus, based on its relation to the semantic level of representation.

We trained a separate regression model for each of the nine participants, using the set of 25 sensory-motor semantic features described earlier. Each trained model was evaluated by means of a "leave-two-out" cross-validation approach, in which the model was repeatedly trained with only 58 of the 60 available word stimuli and associated fMRI images. Each trained model was tested by requiring that it first predict the fMRI images for the two "held-out" words and then match these correctly to their corresponding held-out fMRI images. The match between the two predicted and the two observed fMRI images was determined by which match had a higher cosine similarity, evaluated over the 500 image voxels with the most stable responses across training presentations. The expected accuracy in matching the left-out words to their left-out fMRI images is 0.50 if the model performs at chance levels. An accuracy of 0.62 or higher for a single model trained for a single participant was determined to be statistically significant ($P < 0.05$)

relative to chance, based on the empirical distribution of accuracies for randomly generated null models. Similarly, observing an accuracy of 0.62 or higher for each of the nine independently trained participant-specific models would be statistically significant at $P < 10^{-11}$.

The cross-validated accuracies in matching two unseen word stimuli to their unseen fMRI images for models trained on participants P1 through P9 were 0.83, 0.76, 0.78, 0.72, 0.78, 0.85, 0.73, 0.68, and 0.82 (mean = 0.77). Thus, all nine participant-specific models exhibited accuracies significantly above chance levels. The models succeeded in distinguishing pairs of previously unseen words in over three-quarters of the 15,930 cross-validated test pairs across these nine participants. Accuracy across participants was strongly correlated ($r = -0.66$) with estimated head motion (i.e., the less the participant's head motion, the greater the prediction accuracy), suggesting that the variation in accuracies across participants is explained at least in part by noise due to head motion.

5.1.4 Semantic Category Classification

It is interesting to consider whether these trained computational models can extrapolate to make accurate predictions for words in new semantic categories beyond those in the training set. To test this, we re-trained the models but this time we excluded from the training set all examples belonging to the same semantic category as either of the two held-out test words (e.g., when testing on "celery" versus "airplane," we removed every food and vehicle stimulus from the training set, training on only 50 words). In this case, the cross-validated prediction accuracies were 0.74, 0.69, 0.67, 0.69, 0.64, 0.78, 0.68, 0.64, and 0.78 (mean = 0.70). This ability of the model to extrapolate to words semantically distant from those on which it was trained suggests that the semantic features and their learned neural activation signatures of the model may span a diverse semantic space.

Given that the 60 stimuli are composed of five items in each of 12 semantic categories, it is also interesting to determine the degree to which the model can make accurate predictions even when the two held-out test words are from the same category, where the discrimination is likely to be more difficult (e.g., "celery" versus "corn"). These within-category prediction accuracies for the nine individuals were 0.61, 0.58, 0.58, 0.72, 0.58, 0.77, 0.58, 0.52, and 0.68 (mean = 0.62), indicating that although the model's accuracy is lower when it is differentiating between semantically more similar stimuli, on average its predictions nevertheless remain above chance levels.

5.1.5 Discussion

The results indicated that features from 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. Previous studies have been limited to binary, categorical classification and here we have demonstrated that it is possible to make finer discriminations, even at the exemplar level. Furthermore, 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.

5.2 Adjective-Noun Comprehension

Given the succesess in using fMRI to discriminate categorial information in 60 concrete objects experiments, one natural direction to follow is to see if similiar approach can be applied to multi-words phrases. We began our analysis with adjective-noun phrases, where adjectives were used to modify the meaning of nouns (e.g. s*trong dog*). We compare two composition models, namely the *additive* and *multiplicative* model, as well as two non-composition models, namely the *adjective* and noun model discussed in Mitchell et al (2008). There were two main hypotheses that we 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.

5.2.1 Experimental Paradigm

Nineteen right-handed adults (aged 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. Four additional participants were excluded from the analysis due to head motion greater than 2.5 mm.

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 1). The fMRI neural signatures of these objects have been found in previous studies to elicit different 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. play, eat, petted) that people more often associate with the term. Notice that the last three adjectives in Table 4 are marked by asterisks to denote they are *object-modifying adjectives*. These adjectives appear to behave differently from the ordinary *attribute-specifying adjectives*. Section 5 is devoted to discussing the different adjective types in more detail.

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

Table 4. Word stimuli. Asterisks mark the object-modifying adjectives, as opposed to the ordinary attribute-specifying adjectives.

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

We are interested in whether the distribution of neural activity encodes sufficient signal to decode both nouns and adjective-noun phrases. Given the observed neural activity when participants comprehended the adjective-noun phrases, Gaussian Naïve Bayes classifiers were trained to identify cognitive states associated with viewing stimuli 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.

Table 5 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%. We also determined the classification accuracies separately for nouns only and phrases only. Distinct classifiers were trained. Classification accuracies were significantly higher ($p < 0.05$) for the nouns, calculated with a paired *t*-test. For 3 participants, the classifier did not achieve reliable classification accuracies for the phrase stimuli. Moreover, we determined the classification accuracies separately for each semantic category of stimuli. There were no significant differences in accuracy across categories, except for the difference between vegetables and vehicles.

Classifier	Racc
All 24 exemplars	0.69
Nouns	0.71
Phrases	0.64
Animals	0.67
Tools	0.66
Vegetables	0.65
Vehicles	0.69

Table 5. 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.

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 on par with previous research, providing a replication of previous findings (Mitchell et al, 2004). The classifiers performed better on the nouns than the phrases, consistent with our expectation that characterizing phrases is more difficult than characterizing nouns in isolation. 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 us to identify participants whose mental representations for phrases are consistent across phrase presentations. Subsequent regression analysis on phrase activation will be based on subjects who perform the phrase task well.

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

The second column of Table 6 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 5 supports our 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 5 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).

	R^2	Racc
Adjective	0.34	0.57
Noun	0.36	0.61
Additive	0.35	0.60
Multiplicative	0.42	0.62

Table 6. Regression fit and regression-based classification rank accuracy of the adjective, noun, additive, and multiplicative models for phrase stimuli.

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 6 shows the rank accuracies classifying mental 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.

5.2.4 Comparing the attribute-specifying adjectives with the object-modifying 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 (*paper airplane*, *model train*, *toy truck*), in comparison to other cases where the adjective simply specifies an attribute of the noun (*soft bear*, *large cat*, *strong dog*, etc.). In order to study this difference, we performed classification analysis separately for the attribute-specifying adjectives and the object-modifying 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-modifying 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-modifying 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 7 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-modifying 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.

	Attribute-specifying	Object-modifying
Adjective	0.57	0.65
Noun	0.62	0.64
Additive	0.61	0.65
Multiplicative	0.63	0.67

Table 7. Separate regression-based classification rank accuracy for phrases with attribute-specifying or object-modifying adjectives.

5.2.5 Discussion

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. Noun model outperforms adjective models, consistent with the assumption that the noun is generally more central to the phrase meaning than is the adjective. Moreover, 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. Further investigation into the different types of adjectives reveals the above phenomenon appears to hold for the attribute-specifying adjectives, but not object-modifying adjectives.

6 Proposed Work

Our results to date have shown that the distributed pattern of neural activity does contain sufficient information to decode the mental state of participants while they contemplate noun and adjective-noun phrases. Furthermore, a generative model that leverages word cooccurrence statistics as intermediate semantic representation enables a predictive theory that can predict neural activity for previously unseen stimuli. There are three ways to extend this work:

- Run more multi-words neurosemantic experiments (e.g. noun-noun concept combination)
- Extend the semantic representation and semantic composition model
- Explore the time series data

6.1 Noun-noun concept combination experiment

We plan to extend our analysis of adjective-nouns phrases to noun-noun phrases, where participants will be shown noun phrases (e.g. *carrot knife*) and instructed to think of a likely meaning for the phrases.

Unlike adjective-noun phrases, where a single interpretation often dominates, noun-noun combinations allow multiple interpretations (e.g., *carrot knife* can be interpreted as a knife that is specifically used to cut carrots or a knife carved out of carrots). There has been extensive literature on the two different types of combination rules that people used when interpreting 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 extracted to modify the head object. For example, the interpretation that *mushroom cloud* is a cloud which shaped like a mushroom 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 *mushroom sauce* is a sauce made of mushroom is a type of relation-based interpretation.

To date, six right-handed adults from the Carnegie Mellon community participated and gave informed consent approved by the University of Pittsburgh and Carnegie Mellon Institutional Review Boards. 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* (see Table 1). 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 neural 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 "*The mug has panels of glass that allow light to pass through; it is called a ...*" will lead the participant to interpret a *window cup* as a cup that are transparent (a property-based interpretation where the transaprent property of a window is mapped to a cup). On the other hand, a context like "*John often wakes up thirsty, and since he doesn't have a bedside table, he keeps water in a ...*" will lead the participant to interpret a *window cup* as a cup that is kept on the window frame (a relation-based interpretation where the modifier object is realized in its entirety and related to the head object as a whole. 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.

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 (appearnce), 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 particant 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 (appearnce, 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 object in the noun-noun phrases is presented in isolation. There were two additional presentations of fixation, 31s each, at the beginning and end of each session, to provide a baseline measure of activity.

6.2 Extend the semantic representation and semantic composition model

The composition models that we have utilized so far are overly simplistic in a number of ways. For example, none of the composition models considered can differentiate between different types of interpretation of the same stimulus. That is, the same feature vector will be provided when the *mushroom cloud* is interpreted as a cloud that shaped like a mushroom and when it is interpreted as a sauce made of mushroom. In order to account for these differences, the semantic composition model must be able to differentiate the different combination rules. We might need to consider different semantic representations.

6.2.1 Semantic composition model that utilizes feature-norming features

Another 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. 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). In our noun-noun concept combination experiment, we included a behavioral response session where participants were asked to describe the object in a sentence, as well as listing properties associated with the object's appearance, interaction, and purpose. We can leverage information in participants' behavioral data or Cree and McRae's feature norming studies. For instance, we could code participants' behavioral response for the modifier noun, the head noun, and the compound noun. Then, we could check if the compound noun inherits features more from the modifier or head noun and if the pattern differs when the compound noun is interpreted with property-based interpretation or relation-based interpretation. Moreover, we could implement PUNC (Costello and Keane, 1997), a computational model that is based on the constraint theory of conceptual combination and the C^3 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.

6.2.2 Infinite latent semantic models

An alternative approach is to model the semantic representation as a hidden variable using a generative probabilistic model that describes how neural activity is generated from some latent semantic representation. We are currently exploring the infinite latent semantic feature model (ILFM; Griffiths & Ghahramani, 2005), which assumes a non-parametric Indian Buffet prior to the binary feature vector and models neural activation with a linear Gaussian model. The basic proposition of the model is that the human semantic knowledge system is capable of storing an infinite list of features (or semantic components) associated with a concept; however, only a subset is actively recalled during any given task (context-dependent). Thus, a set of latent indicator variables is introduced to indicate whether a feature is actively recalled at any given task. We are currently investigating if the compositional models also operate in the learned latent semantic space. For instance, we are checking if the compound noun share more latent feature with the modifier or head noun and if the pattern differs when the compound noun is interpreted with property-based or relation-based interpretation.

6.3 Explore the time-series data

One goal of applying language technologies to brain imaging studies is to be able to decode brain activities when participants are speaking in their mind, that is, a real-time thought-to-text system akin to speech-to-text system in automatic speech recognizers. Thus, it is interesting to study the time series data. To model the time-series of neural activity, the slow-reacting hemodynamic bold response must be considered. The temporal resolution is indeed a weakness of fMRI images. Consequently, it is important to ask the right question. It may be a difficult task to decode each individual word that a participant is thinking, but it may be a relatively simple task to decode what are the actively recalled features or words in a given duration, where the model can leverage the contextual information.

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.

7 Thesis Goals and Timetable

Task	**Time**
Thesis Proposal	Jan, 2010

Concrete-noun experiment	Complete
Adjective-noun experiment	Complete
Noun-noun experiment	Dec 2009 - Feb, 2010
Explore feature norms	Feb, 2010 (already started)
Explore latent feature models	Mar, 2010 (already started)
Explore time series data	Apr, 2010 (already started)
Thesis Writing	May, 2010
Thesis Defense	June, 2010

References

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

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

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

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

Cree, G. S., and 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.

Costello, F., and 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., and 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.

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

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

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

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

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.

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

Mitchell, J., and Lapata, M. 2008. Vector-based models of semantic composition. *Proceedings of ACL-08: HLT*, 236-244.

Mitchell, T., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M. A., and 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., and Just, M.A. 2008. Predicting human brain activity associated with the meanings of nouns. *Science* 320, 1191-1195.

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