Quantitative modeling of the neural representation of adjective-noun phrases to account for fMRI activation

Kai-min K. Chang[1] Vladimir L. Cherkassky[2] Tom M. Mitchell[3] Marcel Adam Just[2]
Language Technologies Institute[1]
Center for Cognitive Brain Imaging[2]
Machine Learning Department[3]
Carnegie Mellon University
Pittsburgh, PA 15213, U.S.A.
{kkchang,cherkassky,tom.mitchell,just}@cmu.edu

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 this study, we investigate how humans comprehend adjective-noun phrases (e.g. *strong dog*) while their neural activity is recorded. Classification analysis shows that the distributed pattern of neural activity contains sufficient signal to decode differences among phrases. Furthermore, vector-based semantic models can explain a significant portion of systematic variance in the observed neural activity. Multiplicative 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.

1 Introduction

How humans represent meanings of individual words and how lexical semantic knowledge is combined to form complex concepts are issues fundamental to the study of human knowledge. 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).

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 1 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). Furthermore, 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

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 638–646,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

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.

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 representation of adjective-noun phrases. In this study, we applied the vector-based models of semantic composition used in computational linguistics to model neural activation patterns obtained while subjects comprehended adjective-noun phrases. In an object-contemplation task, human participants were presented with 12 text labels of objects (e.g. *dog*) and were instructed to think of the same properties of the stimulus object consistently during multiple presentations of each item. The participants were also shown adjective-noun phrases, where adjectives were used to modify the meaning of nouns (e.g. *strong dog*).

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.

In section 2, we describe the experiment and how functional brain images were acquired. In section 3, we apply classifier analysis to see if the distributed pattern of neural activity contains sufficient signal to discriminate among phrases. In section 4, 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 5 distinguishes between two types of adjectives that are used in our stimuli: attribute-specifying adjectives and object-modifying 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.

2 Brain Imaging Experiments on Adjective-Noun Comprehension

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

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. 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.

2.2 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 here only the spatial distribution of the neural activity after the stimuli are comprehended and do not attempt to model the cogntive process of comprehension.

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

3.1 Classifier Analysis

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.

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 24 different items were most stable across presentations. Voxel stability was computed as the average pairwise correlation between 24 item vectors 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. Many of our previous analyses have indicated that 120 voxels is a set size suitable for our purposes.

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.

3.2 Results and Discussion

Table 2 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 2. 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.

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

4.1 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, Mitchell et al. (2008) encoded 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 (Caramazza and Shelton, 1998), they 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*. 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 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 3. The lexical semantic representation for *strong* and *dog*.

4.2 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 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. To foreshadow our results, we found the modifier-head interpretation of the multiplicative model to best account for the neural activity observed in adjective-noun phrase data.

Table 4 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 4. 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).

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.

4.3 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 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).

4.4 Results and Discussion

The second column of Table 5 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 5. 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 5 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 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 6 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 6. Separate regression-based classification rank accuracy for phrases with attribute-specifying or object-modifying adjectives.

In light of this observation, 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 exists an extensive literature on the conceptual combination of noun-noun phrases. Costello and Keane (1997) provide extensive studies on the *polysemy* of conceptual combination. More importantly, they outline different rules of combination, including property mapping, relational mapping, hybrid mapping, etc. It will be interesting to see if different composition models better account for neural activation when different kinds of combination rules are used.

6 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. 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 investigating if the compositional models also operate in the learned latent semantic space.

The premise of our research relies on advancements in the fields of computational linguistics and cognitive neuroimaging. Indeed, we are at an especially opportune time in the history of the study of language, when linguistic corpora allow word meanings to be computed from the distribution of word co-occurrence in a trillion-token text corpus, and brain imaging technology allows us to directly observe and model neural activity associated with the conceptual combination of lexical items. 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 human representation of semantic knowledge.

Acknowledgements

This research was supported by the National Science Foundation, Grant No. IIS-0835797, and by the W. M. Keck Foundation. We would like to thank Jennifer Moore for help in preparation of the manuscript.

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.