Predicting Human Brain Activity Associated with Noun Meanings

Tom M. Mitchell[1], Svetlana V. Shinkareva[1,2], Andrew Carlson[2], Kai-Min Chang[2,3], Vicente L. Malave[2], Robert A. Mason[2], Marcel Adam Just[2]

[1] Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213
[2] Center for Cognitive Brain Imaging, Carnegie Mellon University, Pittsburgh, PA 15213
[3]Language Technologies Institute, School of Computer Science, Carnegie Mellon University

Abstract:

Recent brain imaging studies have shown that different spatial patterns of neural activation are associated with thinking about different semantic categories of words and pictures (e.g., tools, buildings, animals). As a next step we seek a general theory capable of predicting the neural activity associated with arbitrary words not yet included in experiments. We present here the first such predictive theory, in the form of a computational model that is trained using a combination of data from a trillion-word text corpus, and observed fMRI data associated with viewing several dozen concrete nouns. Once trained, the model predicts fMRI activation for thousands of other concrete nouns in the text corpus, with highly significant accuracies over the 60 nouns for which we currently have fMRI data.

One sentence summary:

We present the first computational model capable of predicting observed fMRI activity produced when humans think about an arbitrary concrete noun, along with experimental results showing strong prediction accuracy over the 60 nouns for which we have fMRI data.

The question of how the human brain represents and organizes conceptual knowledge has been studied by many scientific communities. Neuroscientists using brain imaging studies (*1-9*) have shown that distinct spatial patterns of fMRI activity are associated with viewing pictures of certain semantic categories including animals, tools, and buildings. Linguists have characterized different semantic roles associated with individual verbs, as well as the types of nouns that can fill those semantic roles (e.g., VerbNet (*10*) and WordNet (*11, 12*)). Computational linguists have analyzed the statistics of very large text corpora and 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 (*13-17*). Psychologists have studied word meaning through feature norming studies (*18*) in which participants are asked to list the features they associate with various words, revealing a consistent set of core features across individuals and suggesting a possible grouping of features by sensory/motor modalities. Researchers studying semantic deficits associated with brain damage have found that people who lose the ability to name a particular animal will often lose the ability to name other animals as well, but will not lose the ability to name specific artifacts or fruits/vegetables, and that more generally the loss of ability to name items in one of these three categories does not imply naming difficulties in the other two categories (*19-21*).

This variety of experimental results has led to competing theories of how the brain encodes meanings of words and knowledge of objects, including theories that meanings are encoded in sensory-motor cortical areas (*22, 23*), and theories that they are instead organized by semantic categories such as living and non-living objects (*18, 24*). While these competing theories sometimes lead to different predictions (e.g., of which naming disabilities will co-occur in brain damaged patients), they are primarily *descriptive* theories that do not attempt to *predict* the specific brain activation that will be produced when a human subject reads a particular word or views an image of a particular object.

We present here the first theory that makes directly testable predictions of the fMRI activity associated with thinking about arbitrary concrete nouns, including nouns for which no fMRI data are currently available. More generally, we present a paradigm for representing such theories in the form of computational models, and for training them using a combination of fMRI data and data from a trillion-token corpus of text that captures typical use of English words. We describe the use of this approach to train several competing computational models based on different assumptions regarding the primitive features that underlie the encoding of meaning. We present experimental evidence showing that the best of these models is capable of predicting fMRI neural activity well enough that it can successfully match words it has not yet encountered to their previously unseen fMRI images with accuracies far above chance levels. These results establish for the first time a direct relationship between the statistics of word co-occurrence in text, and the neural activation associated with thinking about word meanings.

APPROACH

We employ a trainable computational model that predicts the neural activation for any given stimulus word *w* in a two-step process shown in Figure 1. Given a stimulus word, *w*, the first step encodes the meaning of *w* as a vector of intermediate semantic features computed from the occurrences of stimulus word *w* within a trillion-token text corpus (*25*) that captures the typical use of words in English text. For example, one intermediate semantic feature might be the frequency with which *w* co-occurs with the verb "hear." The second step predicts the neural fMRI activation at every voxel location in the brain, as a weighted sum of neural activations contributed by each of the intermediate semantic features. More precisely, the predicted activation y_v at voxel *v* in the brain for word *w* is given by

$$y_v = \sum_{i=1}^{n} c_{vi} f_i(w)$$

where $f_i(w)$ is the value of the i^{th} intermediate semantic feature for word *w*, *n* is the number of semantic features in the model, and c_{vi} is a learned scalar parameter that specifies the degree to which the i^{th} intermediate semantic feature activates voxel *v*. This equation can be interpreted as predicting the full fMRI image across all voxels for stimulus word *w* as a weighted sum of images, one per semantic feature f_i. These semantic feature images, defined by the learned c_{vi}, constitute a basis set of primitive images that model the brain activation associated with different semantic components of the input stimulus words.

Insert Figure 1 here

To fully specify a model within this computational modeling framework, one must first define a set of intermediate semantic features $f_1(w)\ f_2(w)\ \ldots\ f_n(w)$ to be extracted from the corpus statistics. In this paper we define intermediate semantic features in terms of corpus co-occurrence statistics of the input stimulus word *w* with a particular other word (e.g., "taste") or set of words (e.g., "taste," "tasted" or "tastes"). Once the semantic features $f_i(w)$ are specified, one must also specify the parameters c_{vi} that define the neural signature contributed by the i^{th} semantic feature to the v^{th} voxel. This is accomplished by training the model using a set of observed fMRI images associated with known stimulus words. Each training stimulus w_t is first re-expressed in terms of its feature vector $< f_1(w_t)\ \ldots\ f_n(w_t) >$, and multiple regression is then used to obtain maximum likelihood estimates of the c_{vi} values; that is, the set of c_{vi} values that minimize the sum of squared errors in reconstructing the training fMRI images

Once trained, the resulting computational model can be used to predict the full fMRI activation image for any other word found in the trillion-token text corpus, as shown in Figure 2A. Given an arbitrary new word w_{new} the model first extracts the intermediate semantic feature values $< f_1(w_{new})\ \ldots\ f_n(w_{new}) >$ from the corpus statistics database, then applies the above formula using the previously learned values for the parameters c_{vi}. The

trained computational model thus embodies a precise theory that predicts the fMRI activation associated with arbitrary words found in the trillion-token corpus. The computational model and corresponding theory can be directly evaluated by comparing their predictions for words outside the training set to observed fMRI images associated with those words. Different predefined sets of intermediate semantic features can be directly compared by training competing models and evaluating their prediction accuracies.

Insert Figure 2 (A and B) here

This computational modeling framework is based on two key theoretical assumptions. First, it assumes the semantic features that distinguish the meanings of arbitrary concrete nouns are reflected in the statistics of their use within a very large text corpus. This assumption is drawn from the field of computational linguistics where the distribution of words that co-occur with word *w* is often used to approximate the meaning of *w* (e.g., *14-17*). Second, it assumes that the brain activity observed when thinking about any concrete noun can be derived as a weighted linear sum of contributions from each of its semantic features. While the correctness of this linearity assumption is debatable, it is consistent with the widespread use of the general linear model in fMRI analysis (e.g., *26*), and with its underlying assumption that fMRI activation often reflects a linear superposition of contributions from different sources. Our theoretical framework does not take a position on whether the neural activation encoding meaning is localized in particular cortical regions – it considers the entire cortex and allows the training data to determine whether and how neural activation is localized.

EXPERIMENTS

To train and evaluate this computational model, fMRI data were collected from a set of 11 participants who viewed 60 different word-picture pairs (Figure 3) presented six times each, with the stimulus sequence permuted on each presentation. Participants were asked to think about the properties of the item they were viewing. Data were acquired on a Siemens Allegra 3.0T scanner at the Brain Imaging Research Center (BIRC), Carnegie Mellon University and the University of Pittsburgh. The study was performed with a gradient echo, EPI sequence with TR = 1000 ms, TE = 30 ms and a 60° flip angle. Seventeen oblique-axial slices were imaged; each slice was 5-mm thick with a gap of 1-mm between slices. The acquisition matrix was 64 x 64 with 3.125-mm x 3.125 x 5-mm voxels Images were corrected for slice acquisition timing, motion-corrected and normalized to the Montreal Neurological Institute (MNI) template within the SPM2 package (SPM2 (Wellcome Department of Cognitive Neurology, London, UK), and anatomically defined regions of interest (ROIs) were automatically labeled (*27*). Data from two participants were rejected due to excessive head motion. Separate models were trained for each of the other 9 participants. To train a model, the data were first processed to create a single image of mean activity for each of the 60 stimulus items, by averaging over the images collected at 4,5,6 and 7 seconds following stimulus onset for

each of the six presentations of the item. The resulting 60 mean images were then normalized by subtracting from each the mean of all 60 images.

Insert Figure 3 here

Alternative computational models were trained based on different sets of intermediate semantic features. Each model was trained and evaluated using a cross validation approach, in which the model was repeatedly trained using only 58 of the 60 available stimulus items, then tested using the two items that had been left out. On each iteration, the trained model was tested by giving it the two word stimuli it had not yet seen (w_1 and w_2), plus their observed fMRI images (i_1 and i_2), then requiring it to predict which of the two novel images was associated with which of the two novel word stimuli. The trained model was first used to create predicted image p1 for word w1 and predicted image p2 for word w2. It then decided which was a better match: (p1=i1 and p2=i2) or (p1=i2 and p2=i1), by choosing the image pairing with the best similarity score. The similarity score between a predicted and observed image was calculated as the cosine similarity between the two images (the dot product of the images represented as vectors normalized to unit length), and the similarity for the two image pairs was taken to be the sum of the two similarity scores. This leave-two-out train-test procedure was iterated 1770 times, leaving out each of the possible word pairs. The expected accuracy in matching the two left-out words to their left-out fMRI images is 0.50 if the matching is performed at chance levels. A label permutation test was performed to determine that an accuracy of 0.612 for a single participant is statistically significant at $p<0.05$.

We trained and tested a variety of computational models based on different sets of intermediate semantic features. To be effective, the set of semantic features must simultaneously encode the wide variety of semantic content of the input stimulus words, and factor the observed fMRI activation into more primitive components that can be linearly recombined to successfully predict the fMRI activation for arbitrary new stimuli. Motivated by existing conjectures regarding the centrality of sensory-motor features in neural representations of objects (e.g., *18, 28*), we designed 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*. Notice these verbs generally correspond to basic sensory and motor activities, actions performed on objects and actions involving changes to spatial relationships. For each verb, the value of the corresponding intermediate semantic feature is the normalized co-occurrence count of the input stimulus word *w* with any of three forms of the verb (e.g., “taste” or “tastes” or “tasted”) over the text corpus. One exception was made, for the verb "see." Its past tense was omitted because "saw" is one of our 60 stimulus nouns. Normalization consists of scaling the vector of 25 feature values to unit length.

A separate computational model was trained for each of the nine participants, using the above set of 25 semantic features. The cross-validated accuracies in matching two unseen word stimuli to their unseen fMRI images for these nine trained models were

0.773, 0.707, 0.679, 0.619, 0.616, 0.584, 0.571, 0.542, and 0.473. Thus, five of the nine participant-specific models exhibited accuracies significant at $p<0.05$, and the most accurate of these models succeeded in distinguishing pairs of previously unseen words in over three quarters of the 1770 cross-validated test cases. The discussion below focuses in greater depth on the three most accurate trained models, for participants P1, P2 and P3.

Visual inspection of the predicted fMRI images produced by the models trained for P1, P2 and P3 shows that these predicted images frequently capture significant aspects of brain activation associated with stimulus words outside the training set. An example is shown in Figure 2B, where the model was trained on 58 of the 60 stimuli for participant P1, omitting "celery" and "airplane." Note that although the predicted fMRI images for "celery" and "airplane" are imperfect, they capture substantial components of the activation actually observed for these two stimuli.

Given that the 60 stimuli are grouped into 12 semantic categories, it is interesting to ask whether the successful predictions follow solely from the ability of the model to distinguish words from different categories (e.g., "celery" belongs to the category "food" whereas "airplane" belongs to "vehicles."), or whether the model can also predict nuances in brain activation that distinguish semantically similar words belonging to the same category (e.g., "celery" versus "corn"). One way to answer this question is to measure the prediction accuracy over only those pairs of held-out test words belonging to the same semantic category. These within-category prediction accuracies for participants P1, P2 and P3 are 0.667, 0.667, and 0.533 (mean 0.622), above the 0.566 mean accuracy that corresponds to $p<0.05$ according to a label permutation test. For two of these three participants, the trained model distinguishes highly similar word pairs it has not previously observed, in two thirds of the test cases. In contrast, the cross-class accuracies (considering only word pairs from distinct classes) for these three participants are 0.781, 0.710, and 0.689 (mean 0.727).

It is also interesting to ask whether this approach can learn to make predictions for words in new semantic categories not included at all in the training set. We tested this by retraining the models for participants P1, P2 and P3, this time removing 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 0.614, 0.631 and 0.555. Thus, the trained model can to some degree generalize to words semantically distant from those on which it was trained, suggesting that the semantic features and their learned neural activation signatures span a diverse semantic space.

A second method for evaluating the model is to examine the learned basis set of fMRI signatures for the 25 verb-based semantic features. The learned signatures for the semantic features "eat," "listen," and "touch" for participant P1 are shown in Figure 4. As shown there, the learned fMRI signature for the semantic feature "eat" exhibits strong activation in gustatory cortex, the signature for "listen" exhibits activation in cortical regions associated with audition, and the signature for "touch" exhibits activation in

somatosensory regions. More generally, averaging across participants P1, P2 and P3, several features exhibit activation in cortical regions associated with the sensory input or activity they describe: The signature for “listen” exhibits activation in auditory and language processing cortical areas (Left pars triangularis, insula, and posterior superior temporal gyrus). The signatures for “touch,” “push,” and “rub” each exhibit activation in somatosensory and motor regions (Right postcentral and precentral). The signature for “fear” generates activation in areas including anterior and posterior cingulate, and to a lesser degree the amygdala and hippocampus. The signature for “move” generates significant activation in many areas of extrastriate cortex, calcarine sulcus, and superior parietal and right intraparietal sulcus. The signature for “open” exhibits activation in the left and right fusiform gyrus. Despite these correspondences, some feature signatures do not predict activation in cortical regions associated with their purported function (e.g., the signature for “smell” does not predict strong activation in olfactory cortex).

Insert Figure 4 here

Consider next the degree of similarity of the learned feature signatures for different participants. Figure 5 shows the signatures for "ride" and “near” for participants P1, P2 and P3. Despite the fact that these three signatures are components of models trained independently for each participant, there is a striking similarity across participants, as there is for many of the learned feature signatures. Notice that for each of the three participants, the learned signature for “ride” shows activation in extrastriate cortex (at the top of the image, corresponding to the posterior cortex) whereas the signature for “near” does not. The signatures for “ride” and “near” both exhibit activity in left and right fusiform in each of the three models, but in all three participants the fusiform activation is more posterior for “ride” and more anterior for “near.” The discovery of this pattern independently in each of the three participants suggests that this regularity captures some aspect of semantic representations that may hold across many individuals. More generally, the similarity in many of the learned feature signatures across models provides evidence in support of the conjecture that semantic representations are similar across individuals, and that our trained theory captures some of these similarities.

Insert Figure 5 here

Given the success of this set of 25 intermediate semantic features motivated by the conjecture that semantic primitives are related to sensory-motor verbs, it is natural to ask how this set of intermediate semantic features compares to alternatives. To explore this we trained and tested models using 300 different randomly generated sets of semantic features, each defined by 25 randomly drawn words from the 5000 most frequent words in the text corpus, excluding the 60 stimulus words as well as the 500 most frequent words (which contain many function words and words without much specific semantic content, such as "the" and “have”). A total of 300 random features sets were generated,

and for each feature set models were trained for each of the three participants P1, P2 and P3. The prediction accuracy for held-out words averaged across these three participants was measured for each of the 300 sets of semantic features, and the distribution of resulting accuracies is shown in the blue histogram in Figure 6. The mean accuracy over these 300 feature sets is 0.553, the standard deviation is 0.039, and the maximum accuracy achieved is 0.655. The fact that the mean accuracy is greater than 0.50 suggests that there are many feature sets that can capture some of the semantic content of the 60 stimulus words and some of the regularities in the corresponding brain activation. However, among these 300 feature sets, none came close to the 0.72 mean accuracy of our manually generated feature set (shown by the red item in the histogram figure). This result suggests the set of features defined by our sensory-motor verbs are somewhat unique in capturing regularities in the neural activation encoding the semantic content of words in the brain.

Insert Figure 6 here

DISCUSSION

To our knowledge, this is the first work to establish a direct relationship between the fMRI activity observed when a person thinks about a concrete object, and the statistical properties of the corresponding word in a very large text corpus. It is also the first work to offer a testable, generative theory of brain activation associated with thousands of concrete nouns. This paradigm is based on the assumption that the meaning of a word can be approximated by a set of semantic features corresponding to statistics describing the distribution of that word in a large corpus, and that these semantic features correspond to stable neural signatures whose weighted linear combinations accurately predict the neural activations associated with arbitrary nouns. The success of the model based on 25 sensory-motor verbs (compared to alternative models based on randomly sampled sets of 25 semantic features) lends credence to the conjecture that neural representations of concrete nouns are in part grounded in sensory-motor features. However, the learned signatures associated with the 25 intermediate semantic features also exhibit significant activation in brain areas not directly associated with sensory-motor function, including frontal regions. Thus, it appears that the basis set of features that underlie neural representations of concrete nouns involves much more than sensory-motor cortical regions.

This research is analogous in some ways to recent research analyzing fMRI activation associated with picture stimuli, in terms of visual features of the pictures *(9,29)*. Our work differs in that we employ text corpus features to capture semantic aspects of the stimulus, rather than visual features that capture perceptual aspects. It also differs in that our computational model is capable of predicting fMRI activation for stimuli beyond the training set. In future work it may be interesting to jointly analyze perceptual image features and semantic corpus-derived features. It may also be productive to employ more sophisticated text corpus properties than simple word co-occurrence, and to develop algorithms to automatically derive optimal basis sets of semantic features.

Although at first glance it may seem enigmatic that brain activity and language corpora should have much to say about each other, they are inherently linked by the brain system that generates both types of data. This new research approach opens the possibility of exploring a large range of human semantic representation issues by combining brain imaging with language corpus studies.

REFERENCES AND NOTES

1. J. V. Haxby, M. Gobbini, M. Furey, A. Ishai, J. Schouten & P. Pietrini, *Science* **293**, 2425-2430 (2001).

2. A. Ishai, L. G. Ungerleider, A. Martin, J. L. Schouten, J. V. Haxby, *PNAS, USA* **96**, 9379-9384 (1999).

3. N. Kanwisher, J. McDermott, M. M. Chun, *J. Neurosci.* **17**, 4302-4311 (1997).

4. T. A. Carlson, P. Schrater, S. He, *J. Cog. Neurosci* **15**, 704-717 (2003).

5. D. D. Cox, R. L. Savoy, *NeuroImage* **19**, 261-270 (2003).

6. T. Mitchell, R. Hutchinson, R.S. Niculescu, F. Pereira, X. Wang, M. Just & S. Newman, *Machine Learning* **57**, 145-175 (2004).

7. S. J. Hanson, T. Matsuka, J. V. Haxby, *NeuroImage* **23**, 156-166 (2004).

8. S. M. Polyn, V. S. Natu, J. D. Cohen, K. A. Norman, *Science* **310**, 1963-1966 (2005).

9. A. J. O'Toole, F. Jiang, H. Abdi, J. V. Haxby, *J. Cog. Neurosci.* **17**, 580-590 (2005).

10. K. Kipper, A. Korhonen, N. Ryant, M. Palmer, *Proc. Fifth Int. Conf. on Language Resources and Evaluation*, Genoa, Italy, May 2006.

11. G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller, *International Journal of Lexicography* **3**, 235-244 (1990).

12. C. Fellbaum, *WordNet: An electronic lexical database.* (MIT Press, Cambridge, MA, 1998).

13. K. W. Church, P. Hanks, *Computational Linguistics* **16**, 22-29 (1990).

14. T. K. Landauer, S. T. Dumais, *Psychological Review* **104**, 211-240 (1997).

15. D. Lin, S. Zhao, L. Qin, M. Zhou, *Proc. of the Int. Joint Conf. on Artificial Intelligence*, 2003.

16. D. M. Blei, A. Y. Ng, M. I. Jordan, *Journal of Machine Learning Research* **3**, 993-

1022 (2003).

17. R. Snow, D. Jurafsky, A. Ng, *Proc. of the 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, July 2006.

18. G. S. Cree, K. McRae, *J. Exp. Psychol. Gen.* **132**, 163-201 (2003).

19. A. Caramazza, J. R. Shelton, *J. Cog. Neurosci.* **10**, 1-34 (1998).

20. S. J. Crutch, E. K. Warrington, *Brain* **126**, 1821-1829 (2003).

21. D. Samson, A. Pillon, *Brain Lang* **91**, 252-264 (2004).

22. A. Martin, L. L. Chao, *Current Opinion in Neurobiology* **11**, 194-201 (2001).

23. R. F. Goldberg, C. A. Perfetti, W. Schneider, *J. Neurosci.* **26**, 4917-4921 (2006).

24. B. Z. Mahon, A. Caramazza, in *The Encyclopedia of language and linguistics*, K. Brown, Ed. (Elsevier Science, Amsterdam, ed. 2, 2005).

25. T. Brants, A. Franz. [URL] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13 (Linguistic Data Consortium, Philadelphia, PA, 2006).

26. K. J. Friston *et al.*, *Human Brain Mapping* **2**, 189-210 (1995).

27. N. Tzourio-Mazoyer *et al.*, *NeuroImage* **15**, 273-289 (2002).

28. A. Martin, L. G. Ungerleider, J. V. Haxby, in *The new cognitive neurosciences*, M. S. Gazzinga, Ed. (MIT Press, Cambridge, MA, ed. 2, 2000), pp.1023-1036.

29. D. R. Hardoon, J. Mourao-Miranda, M. Brammer, J. Shawe-Taylor, *NeuroImage* **37**, 1250-1259 (2007).

This research was funded in part by the W. M. Keck Foundation, and by a Yahoo! Fellowship to Andrew Carlson.

Supporting online material:

Text corpus data. The text corpus data was originally provided by Google Inc., and is available online at http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13. It consists of a set of n-grams (sequences of words and other text tokens) ranging from unigrams (single tokens) up to five-grams (sequences of five tokens), along with counts giving the number of times each n-gram appeared in a large corpus containing over a trillion total tokens. The corpus consisted of publicly available English text web pages. N-grams occurring fewer than 40 times were not provided. We used this data to calculate co-occurrence counts for words occurring within five tokens of one another. These are the co-occurrence counts used in all experiments reported in this paper.

Statistical significance. Statistical significance of cross-validated predication accuracies were calculated using a permutation test in which the labels of the words were randomly permuted. For each of 188 permutations, models were trained for participants P1, P2 and P3. The mean accuracy of these three models, μ_{avg}, was measured using our cross validation method for each permutation. The resulting empirical distribution over μ_{avg} had a mean of 0.4996 and standard deviation of 0.0403. Modeling this as a normal distribution, values of μ_{avg} above 0.566 correspond to $p<0.05$. Similar permutation tests on the average *within-class* accuracies show three-participant average accuracies above 0.567 are significant with $p<0.05$. Similarly, permutation tests show that single-participant accuracies above 0.612 (0.587) and single-participant within-class accuracies above 0.623 (0.595) are significant at $p<0.05$ ($p<0.10$).

Figure Legends

Figure 1. **Form of the theory for predicting fMRI activation for arbitrary noun stimuli.** fMRI activation is predicted in a two-step process. The first step encodes the meaning of the input stimulus word in terms of intermediate semantic features whose values are extracted from a large corpus of text exhibiting typical word use. The second step predicts the fMRI image as a linear combination of the fMRI signatures associated with each of the intermediate semantic features.

Figure 2. **Predicting fMRI images for given stimulus words. (A) Forming a prediction for the stimulus word "celery" after training on 58 other words**. Data from the text corpus is used to assign coefficients to the 25 verb-based semantic features; the coefficient of "eat" for stimulus word "celery" is large (0.84) because "eat" co-occurs frequently with "celery" in the corpus. The predicted activation for the stimulus word is a linear combination of the learned fMRI signatures for each semantic feature, weighted by its corpus-derived coefficient. Figure 2 shows just one horizontal slice (z=7 in MNI space) of the predicted 3-dimensional image. **(B) Predicted and observed fMRI images for "celery" and "airplane" after training using 58 other words**. Though imperfect, predictions capture aspects of observed activity. The long red and blue streaks near the top (posterior region) of the predicted and observed images are the left and right fusiform gyri.

Figure 3. **Presentation and set of exemplars used in the experiment.** Participants were presented 360 word-picture pairs of common concrete nouns, consisting of 60 distinct objects from 12 categories, each presented six times. A slow event-related paradigm was followed in which the stimulus was presented for 3s, followed by a 7s fixation period.

Figure 4. **Learned voxel activation signatures for 3 of the 25 semantic features, for participant P1**. Notice the semantic feature associated with the verb "eat" activates Right pars opercularis (arrow), part of gustatory cortex. The semantic feature associated with "listen" activates the left posterior superior temporal sulcus, superior extrastriate, insula and pars triangularis (arrows) associated with audition and language processing. The semantic feature for the verb "touch" activates right post-central sulcus and right inferior parietal lobule (arrow) the location of primary somatosensory cortex.

Figure 5. **fMRI signatures for the semantic features "ride" and "near" for three participants, and the mean signature over all three (for slice z=7 in MNI space).** Notice a consistent pattern in the three independently trained models: "ride" shows more activation in posterior portions of left and right fusiform, whereas "near" shows more activation in anterior fusiform. In addition, "ride" consistently shows more activation in left and right inferior extrastriate. Top of image is posterior.

Figure 6. **Histogram shows in blue the accuracy of 300 trained computational models utilizing different intermediate semantic features.** Accuracy is the mean accuracy of models trained independently for participants P1, P2 and P3. Each model is

based on 25 words chosen at random from the 5000 most frequent words, excluding the 500 most frequent words and the stimulus words. The best of these 300 random model has accuracy substantially below the 0.720 accuracy of our 25 manually chosen sensory-motor verbs (shown in red).

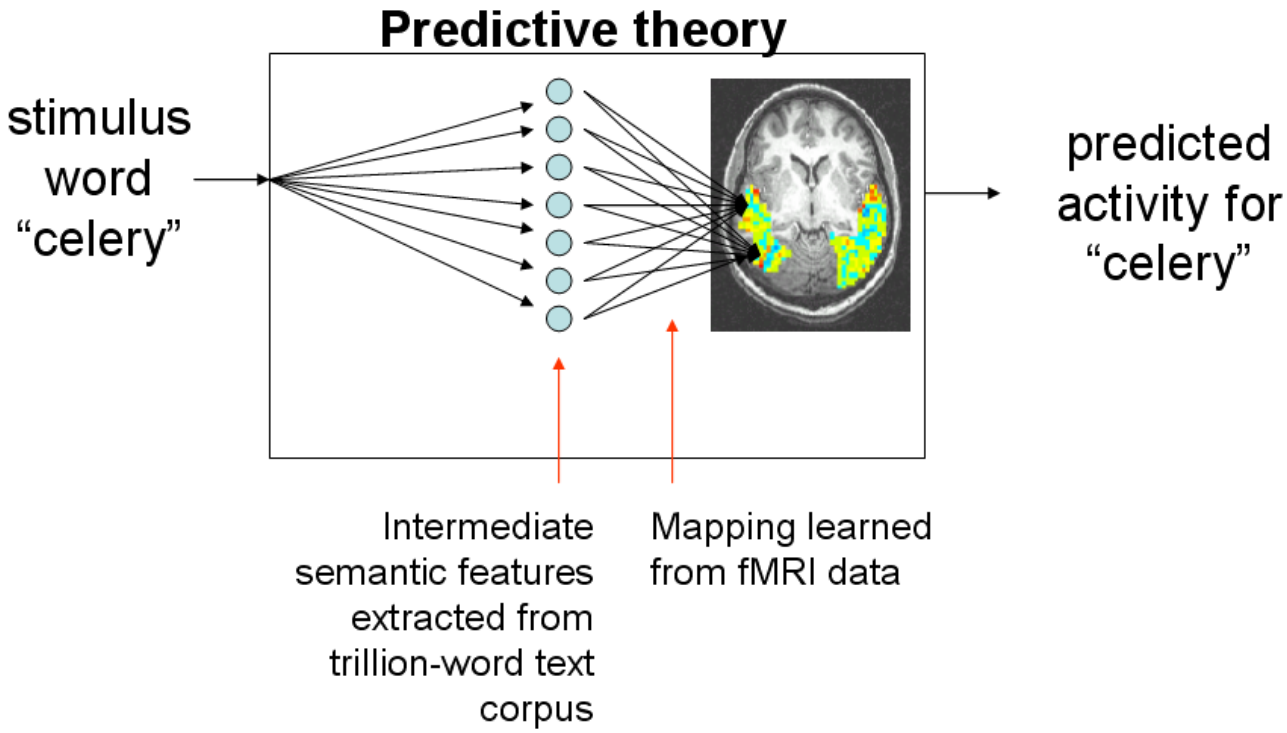


Figure 1

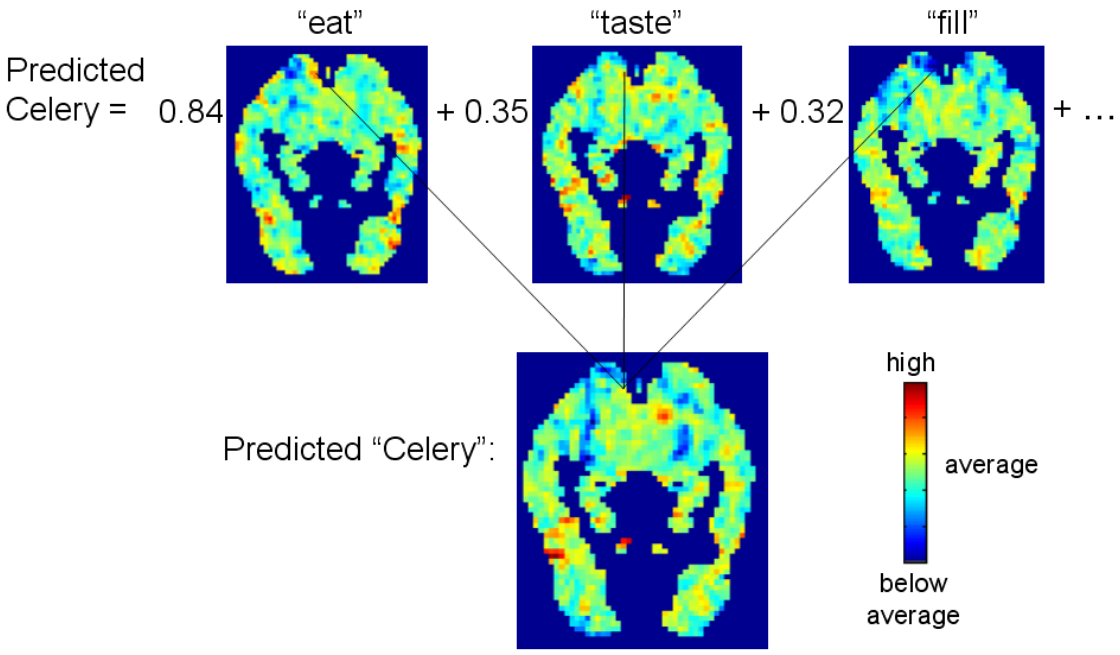
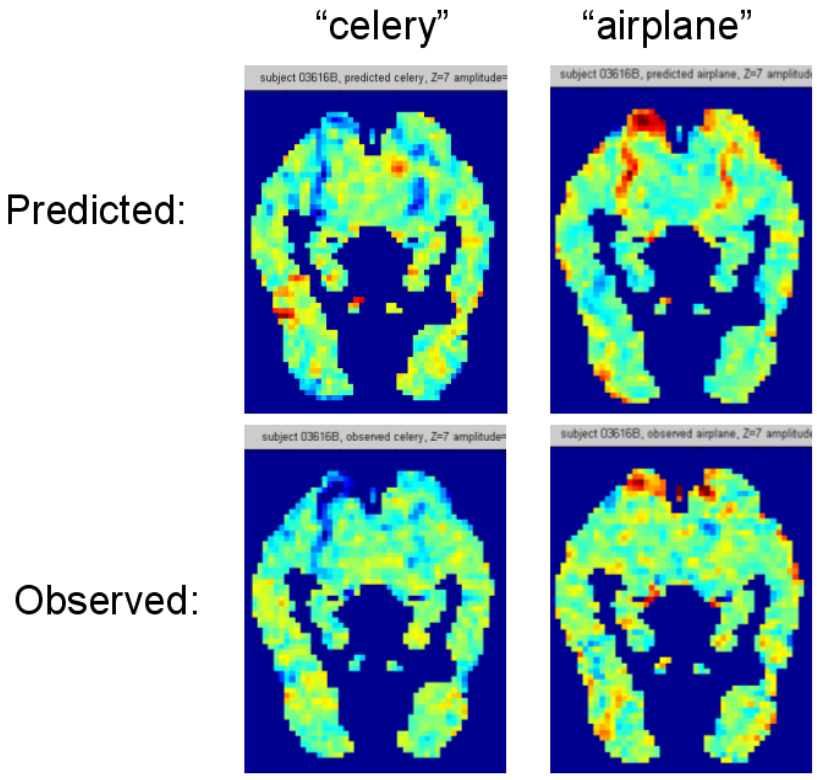
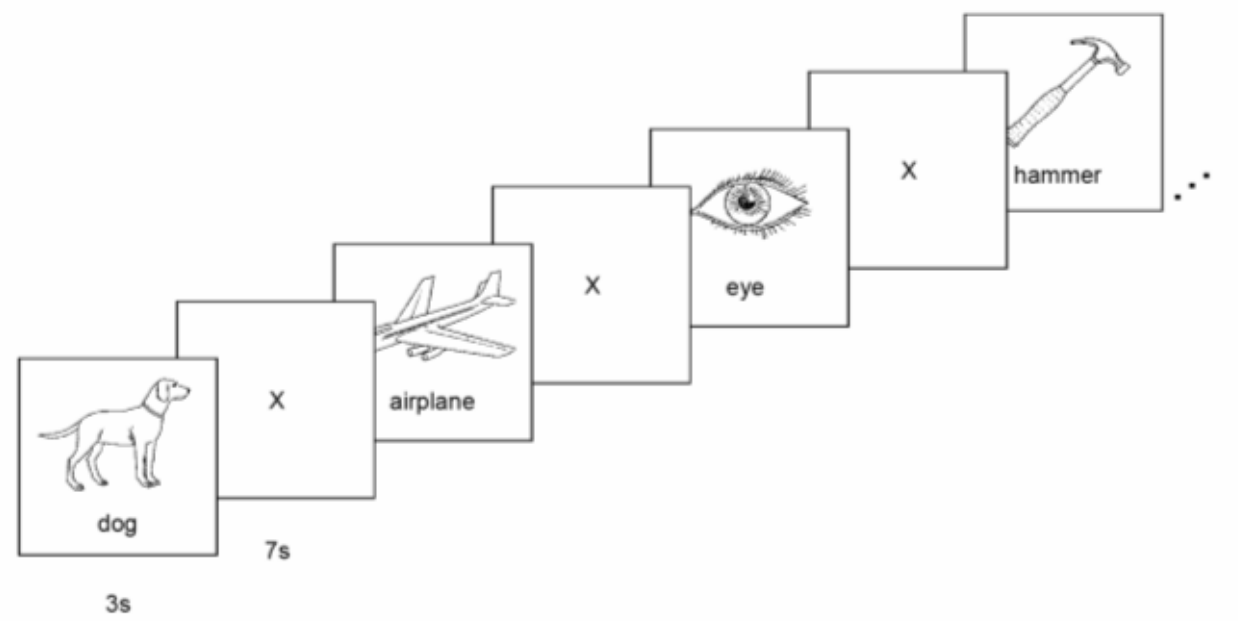


Figure 2a (above) and 2B (below)





Category	Exemplar 1	Exemplar 2	Exemplar 3	Exemplar 4	Exemplar 5
body parts	leg	arm	eye	foot	hand
furniture	chair	table	bed	desk	dresser
vehicles	car	airplane	train	truck	bicycle
animals	horse	dog	bear	cow	cat
kitchen utensils	glass	knife	bottle	cup	spoon
tools	chisel	hammer	screwdriver	pliers	saw
buildings	apartment	barn	house	church	igloo
building parts	window	door	chimney	closet	arch
clothing	coat	dress	shirt	skirt	pants
insects	fly	ant	bee	butterfly	beetle
vegetables	lettuce	tomato	carrot	corn	celery
man made objects	refrigerator	key	telephone	watch	bell

Figure 3

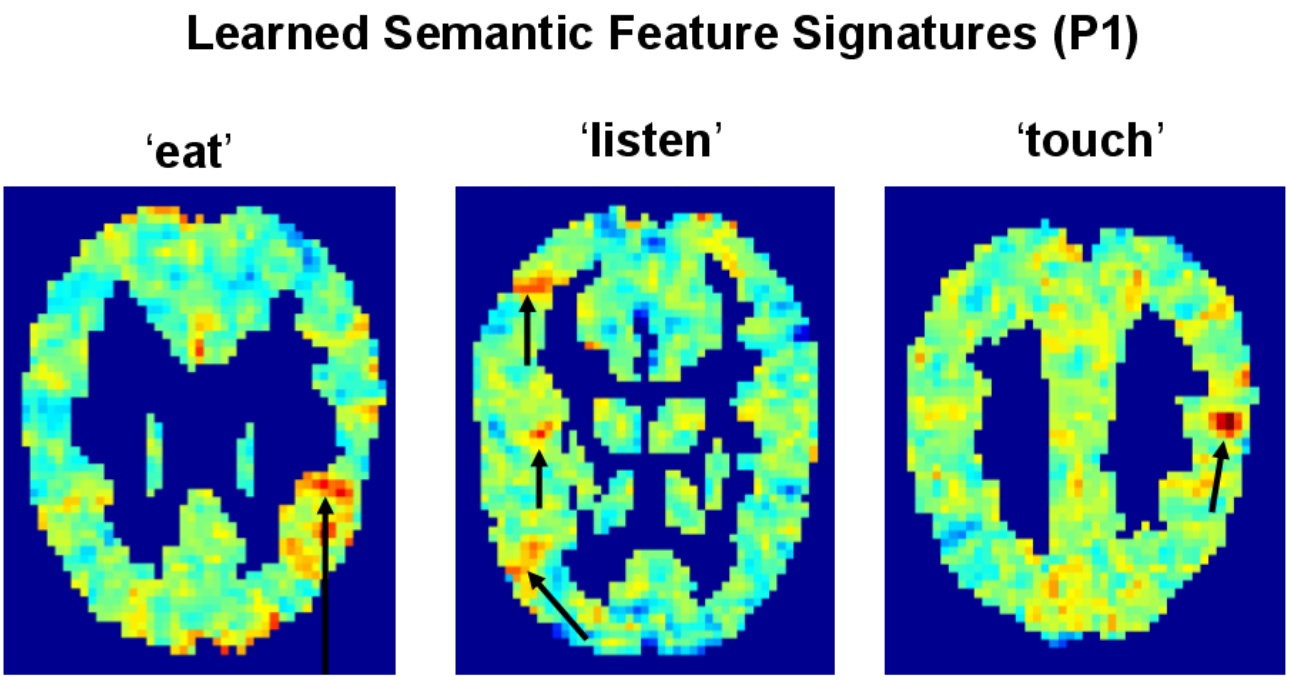


Figure 4

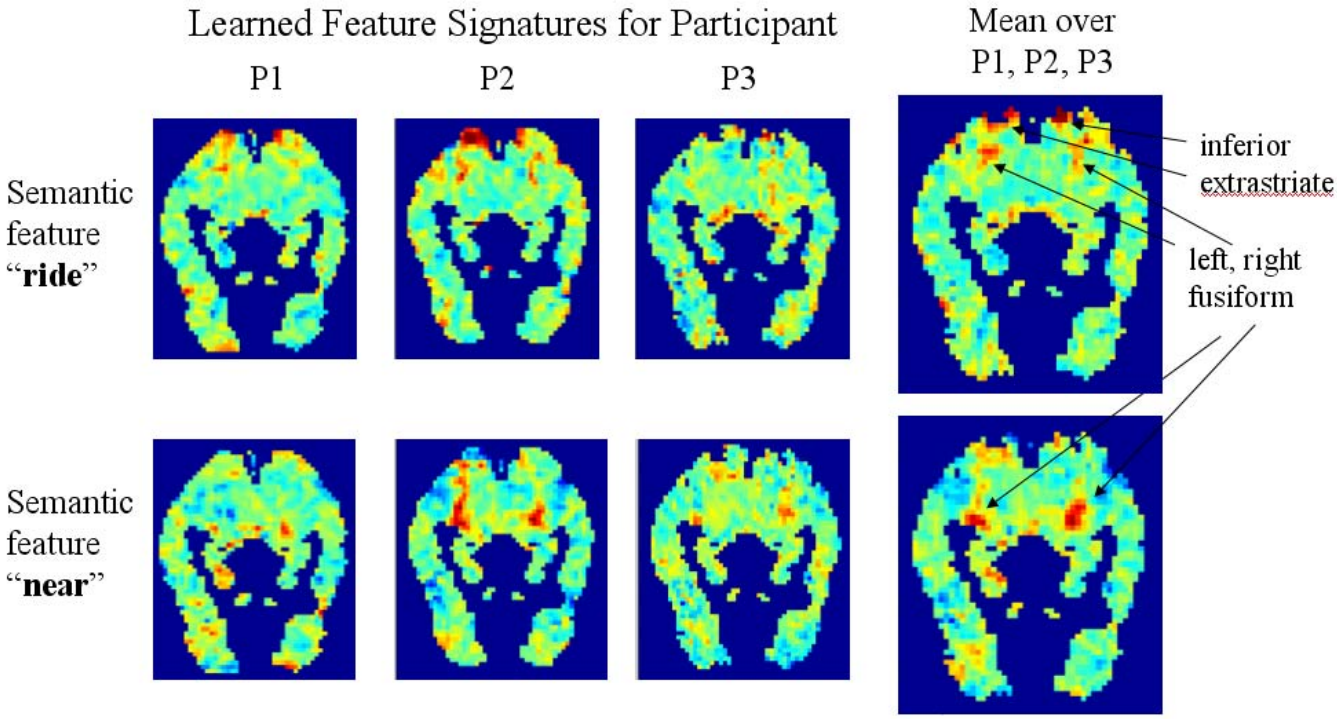


Figure 5

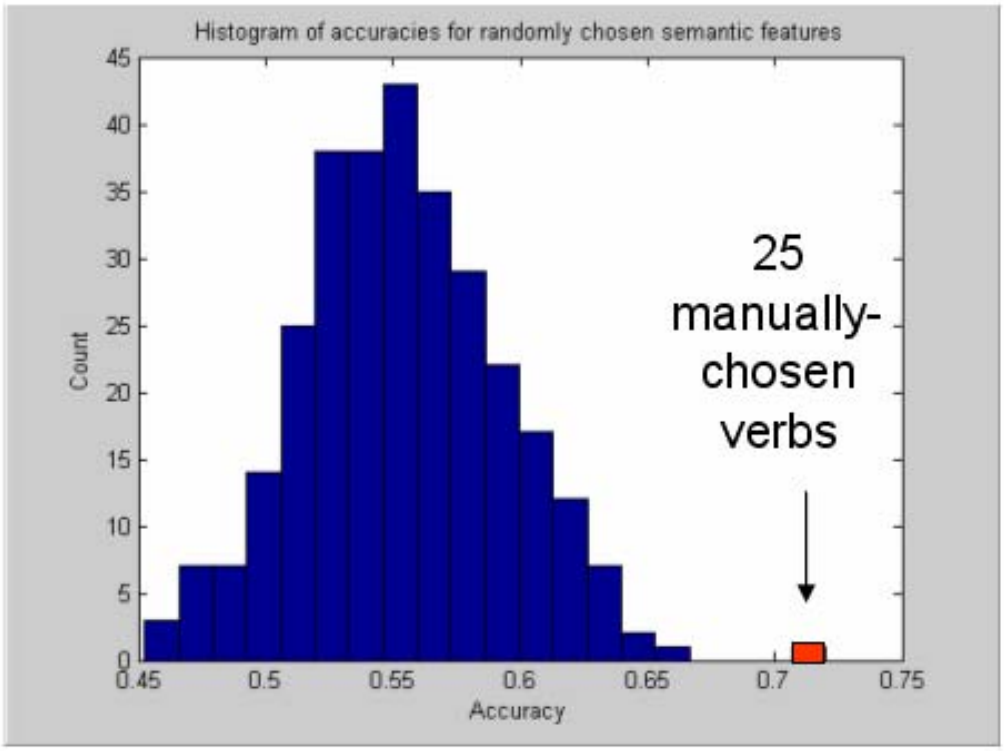


Figure 6