NeuroImage 56 (2011) 716–727

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Quantitative modeling of the neural representation of objects: How semantic feature norms can account for fMRI activation

Kai-min Kevin Chang [a,c,*], Tom Mitchell [b], Marcel Adam Just [c]

[a] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[b] Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA
[c] Center for Cognitive Brain Imaging, Carnegie Mellon University, Pittsburgh, PA, USA

ARTICLE INFO

Article history:
Received 30 November 2009
Revised 20 April 2010
Accepted 30 April 2010
Available online 5 May 2010

ABSTRACT

Recent multivariate analyses of fMRI activation have shown that discriminative classifiers such as Support Vector Machines (SVM) are capable of decoding fMRI-sensed neural states associated with the visual presentation of categories of various objects. However, the lack of a generative model of neural activity limits the generality of these discriminative classifiers for understanding the underlying neural representation. In this study, we propose a generative classifier that models the hidden factors that underpin the neural representation of objects, using a multivariate multiple linear regression model. The results indicate that object features derived from an independent behavioral feature norming study can explain a significant portion of the systematic variance in the neural activity observed in an object-contemplation task. Furthermore, the resulting regression model is useful for classifying a previously unseen neural activation vector, indicating that the distributed pattern of neural activities encodes sufficient signal to discriminate differences among stimuli. More importantly, there appears to be a double dissociation between the two classifier approaches and within- versus between-participants generalization. Whereas an SVM-based discriminative classifier achieves the best classification accuracy in within-participants analysis, the generative classifier outperforms an SVM-based model which does not utilize such intermediate representations in between-participants analysis. This pattern of results suggests the SVM-based classifier may be picking up some idiosyncratic patterns that do not generalize well across participants and that good generalization across participants may require broad, large-scale patterns that are used in our set of intermediate semantic features. Finally, this intermediate representation allows us to extrapolate the model of the neural activity to previously unseen words, which cannot be done with a discriminative classifier.

Introduction

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

There have been a variety of approaches from different scientific communities trying to capture the intermediate semantic attributes and organization underlying object- and word-representation. Linguists have tried to characterize the meaning of a word with feature-based approaches, such as semantic roles (Kipper 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 and Hanks, 1990). Psychologists have studied word meaning in many ways, one of which is through feature norming studies (Cree and McRae, 2003) in which human participants are asked to list the features they associate with various words. There are also approaches that treat the intermediate semantic representation as hidden (or latent) variables and use techniques like the traditional PCA and factor analysis, or the

* Corresponding author. Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3891, USA. Fax: +1 412 268 6298.
E-mail address: kkchang@cs.cmu.edu (K.K. Chang).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.04.271

more recent LSA (Landauer and Dumais, 1997) and topic models (Blei et al., 2003) to recover these latent structures from text corpora. Kemp et al. (2007) have also presented a Bayesian model of inductive reasoning that incorporates both knowledge about relationships between objects and knowledge about relationships between object properties. The model is useful to infer some properties of previously unseen stimuli, based on the learned relationships between objects. Finally, connectionists have long employed *hidden layers* in their neural networks to mediate non-linear correspondences between input and output. Hanson et al. (2004) proposed a neural network classifier with hidden units to account for brain activation patterns, but the learned hidden units are difficult to interpret in terms of an intermediate semantic representation.

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

Here we model the intermediate semantic knowledge with features from an independently performed feature norming study (Cree and McRae, 2003), where participants were explicitly asked to list features of 541 words. Our results suggest that (1) object features derived from a behavioral feature norming study can explain a significant portion of the systematic variance in the neural activity observed in our object-contemplation task. Moreover, we demonstrate how a generative classifier[1] that includes an intermediate semantic representation (2) generalizes better across participants, compared to a discriminative classifier that does not utilize such an intermediate semantic representation, and (3) enables a predictive theory that is capable of predicting fMRI neural activity well enough that it can successfully match words it has not yet encountered to their previously unseen fMRI images with accuracies far above chance levels, which simply cannot be done with a discriminative classifier.

Materials and methods

The fMRI data acquisition data and signal processing methods were previously reported in another publication (Mitchell et al., 2008). Some central information about the data is repeated here.

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

Participants

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

Experimental paradigm

The stimuli were line drawings and noun labels of 60 concrete objects from 12 semantic categories with 5 exemplars per category. Most of the line drawings were taken or adapted from the Snodgrass and Vanderwart (1990) set and others were added using a similar drawing style. Table 1 lists the 60 stimuli.

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

The entire set of 60 stimuli was presented 6 times during the scanning session, in a different random order each time. Participants silently viewed the stimuli and were asked to think of the same item properties consistently across the 6 presentations. Each stimulus was presented for 3 s, followed by a 7 s rest period, during which the participants were instructed to fixate on an X displayed in the center of the screen. There were two additional presentations of the fixation, 31 s each, at the beginning and at the end of each session, to provide a baseline measure of activity. A schematic representation of the design is shown in Fig. 1.

Data acquisition

Functional images were acquired on a Siemens Allegra 3.0 T 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 64×64 with $3.125 \times 3.125 \times 5$ mm voxels.

Data processing and analysis

Data processing and statistical analysis were performed with Statistical Parametric Mapping software (SPM99, Wellcome Department of Cognitive Neurology, London, UK). The data were corrected for slice timing, motion, linear trend, and were temporally smoothed with a high-pass filter using 190 s cutoff. The data were normalized to the MNI template brain image using 12-parameter affine transformation.

Table 1
List of 60 words.

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

The asterisks mark the words that were not part of the Cree and McRae (2003) study.

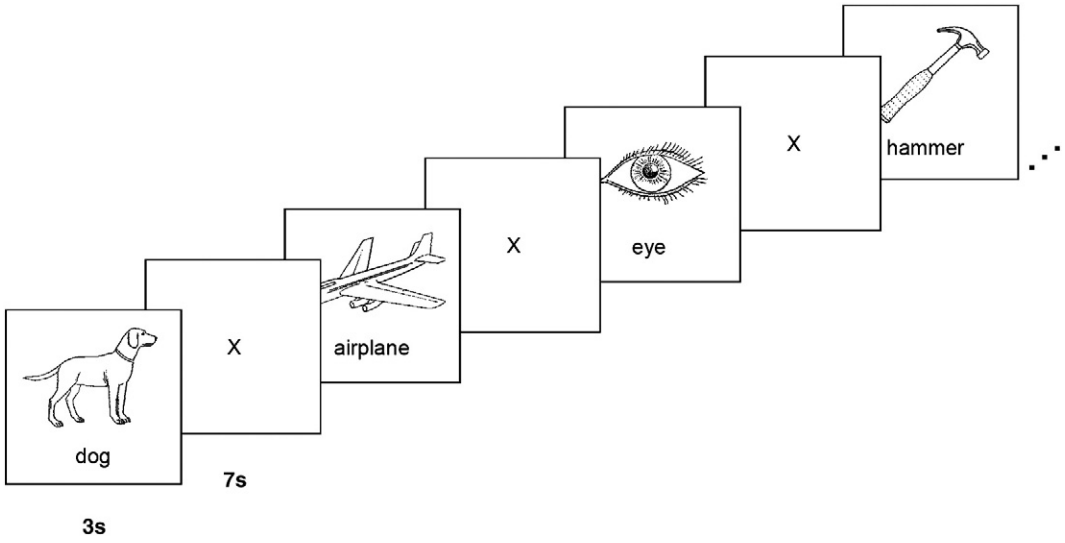

Fig. 1. Schematic representation of experimental design for the 60-word object-thinking experiment.

The data were prepared for regression and classification analysis by being spatially normalized into MNI space and resampled to $3\times3\times6$ mm^3 voxels. We try to keep approximately the same acquisition voxel size which has been used in many of our previous studies and is adequate for a list of different cognitive tasks. Voxels outside the brain or absent from at least one participant were excluded from further analysis. The percent signal change (PSC) relative to the fixation condition was computed for each object presentation at each voxel. The mean of the four images (mean PSC) acquired within a 4 s window, offset 4 s from the stimulus onset (to account for the delay in hemodynamic response) provided the main input measure for subsequent analysis. The mean PSC data for each word or picture presentation were further normalized to have mean zero and variance one to equate the variation between participants over exemplars.

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

The stable voxels were located in multiple areas of the brain. Fig. 2 shows voxel clusters from the union of stable voxels from all nine participants. As shown, many of these locations are in occipital, occipital-temporal, and occipital-parietal areas, with more voxels in the left hemisphere. Table 2 lists the distribution of the 120 voxels selected by the stability measure for each participant, sorted by major brain structures and size of clusters.

For classifier analysis, voxel stability was computed using only the training set within each fold in the cross-validation paradigm. For within-participants analysis, where the training data consist of 5 of the 6 presentations and the testing data consist of the remaining presentation, the voxel stability was computed using only the training data for that particular participant. For between-participants analysis, where the training data consists of 8 of the 9 participants and the testing data consist of the remaining participant, the voxel stability was computed using only the training data for the 8 participants. The

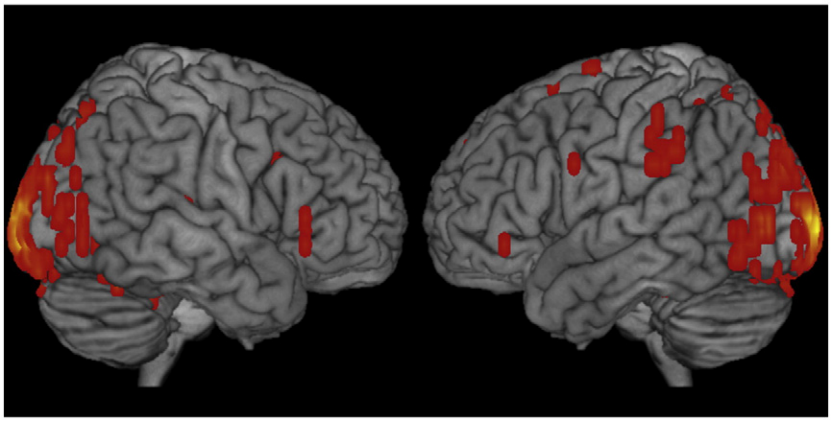

Fig. 2. Voxel clusters from the union of stable voxels from all nine participants.

Table 2
Locations (MNI centroid coordinates) and sizes of the voxel clusters selected by the stability measure.

Participant	Label	X	Y	Z	Voxels[a]	Radius
P1	*Occipital*					
	R fusiform gyrus	31.5	−50.4	−10	24	7.02
	L fusiform gyrus	−26.9	−50.9	−11.7	21	6.13
	L occipital middle	−20.1	−98.7	6	21	6.03
	L occipital inferior	−15.1	−91.1	−10.2	13	5.22
	R occipital middle	34.9	−76	13	6	4.72
	R calcarine	6.2	−91.1	4	6	4.17
P2	*Medial temporal*					
	L parahippocampal gyrus	−25	−42.2	−15	6	3.79
	Occipital					
	R calcarine	15.5	−96	−0.9	70	9.73
	L calcarine	−16.6	−98.6	−4.1	22	7.1
	L cuneus	−20.6	−60	15.6	5	3.51
P3	*Parietal*					
	L precuneus	−5.6	−57.5	24	5	2.65
	Occipital					
	R calcarine	18.2	−93.5	2.8	75	11.26
	L occipital middle	−17.1	−98.3	−1.5	28	7.73
P4	*Temporal*					
	R fusiform gyrus	36.5	−40.1	−23	6	5.72
	Parietal					
	L supramarginal gyrus	−53.8	−33.1	33	10	4.56
	L parietal inferior	−35.4	−39.6	43	6	3.31
	R parietal superior	19.4	−63.7	56.4	5	3.51
	Occipital					
	L fusiform	−28.6	−53.1	−14	12	6.59
	R occipital middle	32	−86.7	19.5	12	5.36
	L occipital superior	−13.2	−84.7	40	9	5.69
	L occipital middle	−31.6	−87.5	24	9	5.4
	R lingual	13.3	−101.2	−7.5	8	4.02
P5	*Temporal*					
	L fusiform gyrus	−31.5	−42.9	−18.8	15	5.3
	R fusiform gyrus	34.4	−41.6	−16.2	13	4.51
	Occipital					
	L lingual	−14.9	−89.7	−2.3	44	7.75
	R calcarine	20.5	−94.6	−2.9	35	8.45
P6	*Medial temporal*					
	R parahippocampal gyrus	25.9	−47.5	−13.2	10	6.47
	Occipital					
	R calcarine	17.3	−96.6	−1.1	51	10.92
	L occipital middle	−19.4	−97.8	−3.1	23	8.56
	L fusiform gyrus	−23.8	−49.5	−9.2	13	5.68
	R fusiform gyrus	30	−71.9	−9.6	5	3.76
P7	*Temporal*					
	L fusiform gyrus	−28.8	−46.1	−16.5	20	5.96
	Occipital					
	R calcarine	8.8	−96.1	−2.1	35	7.93
	R fusiform gyrus	31.2	−49.9	−14.9	21	5.65
	L calcarine	−16	−98.8	−5.2	8	3.97
	L lingual	−9.8	−88.8	−11.1	7	4.17
P8	*Temporal*					
	L temporal inferior	−45.5	−67.2	−7.7	14	5.09
	Occipital					
	R lingual	7.7	−87.9	−6.4	43	9.64
	L occipital middle	−18.2	−97.4	−1.9	28	8.48
	R calcarine	11.9	−100.3	−0.6	10	6.9
P9	*Temporal*					
	L fusiform gyrus	−31.8	−39.8	−21.3	11	5.04
	R temporal inferior	45	−64.4	−3.6	5	3.61
	Medial temporal					
	R parahippocampal gyrus	23.8	−42	−15	16	5.05
	Occipital					
	R calcarine	20.6	−98	−2.5	19	5.42
	L occipital middle	−16.4	−102	4.5	8	5.61
	L occipital middle	−26.8	−88.4	35.1	7	4.15
	L lingual	−20.3	−44.8	−10	6	4.16
	R occipital middle	37.5	−78.8	38.4	5	3.68

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

focus on the most stable voxels effectively increased the signal-to-noise ratio in the data and also served as a dimensionality reduction tool that facilitated further analysis by classifiers.

Approach

In this study, we model hidden factors that underpin semantic representation of object knowledge with a multivariate multiple linear regression model. We adopt a feature-based representation of semantic knowledge, in which a word's meaning is determined by a vector of features. Two competing models based on Cree and McRae (2003)'s feature norming study were developed and evaluated using three types of criteria. The three types of evaluation criteria are a regression fit to the fMRI data, the ability to decode mental states given a neural activation pattern, and the ability to distinguish between the activation of two previously unseen objects. Fig. 3 depicts the flow chart of our approach.

Feature norming study

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

Because participants in the feature norming study were free to recall any feature that came to mind, the norms had to be coded to enable further analysis. Two encoding schemes, Cree and McRae's (2003) brain region (BR) scheme and Wu and Barsalou's (2002) detailed taxonomic (DT) encodings, were compared. BR encoding was based on a knowledge taxonomy that adopts a modality-specific view of semantic knowledge. That is, the semantic representation of an object is assumed to be distributed across several cortical processing regions known to process related sensory input and motor output. BR encoding therefore groups features into knowledge types according to their relations to some sensory/perceptual or functional processing regions of the brain. For example, features for *cow* like *eats grass* would be encoded as visual-motion, *is eaten as beef* as function, and *is animal* as taxonomic in this scheme. By contrast, DT encoding captures features from four major perspectives: entity, situation, introspective, and taxonomic, which are further categorized into 37 hierarchically-nested specific categories. For example, features for *cow* like *eats grass* would be encoded as entity-behavior, *is eaten as beef* as function, and *is an animal* as superordinate. Adapted from Cree and McRae (2003), Table 3 lists the features and the corresponding BR and DT encodings for the words *house* and *cow*. Also, Tables 4 and 5 list all the classes and knowledge types in BR and DT encodings that are relevant to our stimulus set.

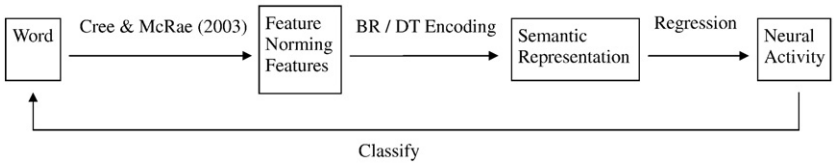

Fig. 3. The flow chart of the generative model. First, the feature norming features associated with the word are retrieved from Cree and McRae (2003). Secondly, the feature norming features are encoded into BR or DT knowledge types, which constitute the semantic representation. Then, a linear regression model learns the mapping between the semantic representation and fMRI neural activity. Finally, a nearest neighbor classifier uses the predicted neural activity generated by the regression model to decode the mental state (word) associated with an observed neural activity.

The analyses below are applied only to those 43 of the 60 words in our study that also occurred in Cree and McRae's study. The missing stimuli are marked with asterisks in Table 1. A matrix was thus constructed for each of the two types of encodings of the feature norms, of size 43 exemplars by the number of knowledge types (10 for BR encoding and 27 for DT encoding, which have non-zero entries). A row in the matrix corresponds to the semantic representation for an exemplar, where elements in the row correspond to the number of features (for that exemplar) categorized as particular knowledge types. Normalization consists of scaling the row vector of feature values to unit length. Consequently, these matrix representations encoded the meaning of each exemplar in terms of the pattern distributed across different knowledge types. For example, the word *house* would have a higher value in the *visual form and surface properties* knowledge type, as opposed to *sound* or *smell*, because people tended to recall more features that described the appearance of a house rather than its sound or smell.

Table 3
Example of Concepts from Feature Norms.

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

Regression model

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

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

where $f_i(w)$ is the value of the 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

Table 4
Cree and McRae (2003)'s Brain Region (BR) Encoding Scheme.

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

Table 5
Wu and Barsalou (2002)'s Detailed Taxonomic (DT) Encoding Scheme.

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

fMRI images. This least squares estimate of the β_{vi} yields the maximum likelihood estimate under the assumption that ε_v follows a Noormal distribution with zero mean. A small L2 regularization with lambda = 0.5 was added to avoid rank deficiency.

The use of a linear regression model to model the hidden factors is not new to analysis of neural activity. Indeed, both linear regression analysis and Statistical Parametric Mapping (SPM)–the most commonly used technique for fMRI data analysis–belong to the more general mathematical paradigm called Generalized Linearized Models (GLM). GLM is a statistical inference procedure that models the data to partition the observed neural response into components of interest, confounds, and error (Friston, 2005). Specifically, GLM assumes a linear dependency among the variables and compares the variance due to the independent variables against the variance due to the residual errors. While the linearity assumption underlying the general linearized model may be overly simplistic, it reflects the assumption that fMRI activity often reflects a superimposition of contributions from different sources, and has provided a useful first order approximation in the field.

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

Classifier model

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

In this study, two classifiers were compared: a Support Vector Machine (SVM) classifier that does not utilize a hidden layer representation and a nearest neighbor classifier that utilizes a hidden layer representation learned in the regression analysis. The SVM classifier (Boser et al., 1992) is a widely-used discriminative classifier that maximizes the margin between exemplar classes. The SVM classifier is implemented in a software package called SVM-light, which is an efficient implementation of SVM by Thorsten Joachims and can be obtained from http://svmlight.joachims.org. On the other hand, the nearest neighbor classifier proposed here uses the estimated regression weights to generate predicted activity for each word. The regression model first estimates a predicted activation vector for each of the 60 objects. Then, a previously unseen observed neural activation vector is identified with the class of the predicted activation that had the highest correlation with the given observed neural activation vector.

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

Results

Using feature norms to explain the variance in neural activity

The regression models were assessed in terms of their ability to explain the variance in neural activity patterns. A multivariate multiple linear regression was run for each participant, using either BR or DT encoding as explanatory variables, and average neural activity (mean PSC) across 120 most stable voxels as response variables. Specifically, DT encoding (with its 27 independent variables) accounted for an average of 58% of the variance in neural activity, whereas BR encoding (with its 10 independent variables) accounted for an average of 35% of the variance. R^2 is higher for DT than for BR for all 9 of the participants, as shown in Table 6. Notice that DT encoding outperforms BR encoding in explaining the variance in neural activity pattern, even though Cree and McRae (2003) found that the two encodings produce similar results in their hierarchical clustering analysis of behavioral data and that they both can be used to explain the tripartite impairment pattern in category-specific deficit studies. This difference may, however, simply be due to the different number of parameters (explanatory variables) that the two regression models use. Akaike information criterion (AIC) is a measure of the goodness of fit that accounts for the tradeoff between the accuracy and complexity of different models and is invariant to the number of parameters. The relative values of AIC scores are used for *model selection* among a class of parametric models with different numbers of parameters, with the model with lowest AIC being

Table 6
Regression analysis R^2.

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

preferred. The BR decoding yields an average AIC score of −37.18, whereas the DT encoding yields an average AIC score of −23.93. Thus, it appears that the difference in regression fit may be due to the different number of parameters that the two regression models use. We further explore this issue in the discussion section.

The regression models produce a predicted neural activity pattern for each word, which can be compared to the observed pattern. For example, Fig. 4 shows one slice of both the observed and the predicted neural activity pattern for the words *house* and *cow*. In each case, the predicted activity is more similar to the observed activity of the target word than to the other word.

Classifying mental states

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

For the within-participants analysis, a regression model was developed from the data from 4 out of 6 presentations of a participant and applied to the average activation of the two remaining presentations of the same participant, using a nearest neighbor classifier to classify the neural activity pattern. A regression model using BR or DT encoding classified the items from the held-out presentations with an average of 72% and 78% rank accuracy, respectively. Since multiple classes were involved, *rank accuracies* are reported, which measure the percentile rank of the correct word within a list of predictions made by the classifier (Mitchell et al., 2004). The rank accuracy for each participant, along with the 95% confidence interval, estimated by 10,000 bootstrapped samples, is reported in Fig. 5. All classification accuracies were significantly ($p<0.05$) different from a chance level of 50% determined by permutation testing of class labels. DT encoding performed significantly better ($p<0.05$) than BR encoding for 7 out of 9 participants. Furthermore, the generative classifiers were compared with the SVM classifier which does not utilize a hidden layer representation. The SVM classifier, which achieved an average of 84% rank accuracy, performed significantly ($p<0.05$) better than the two generative classifiers for 7 out of 9 participants.

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

Distinguishing between the activation of two unseen stimuli

Can the predictions from the regression model be used to classify the mental states of participants on words that were never seen

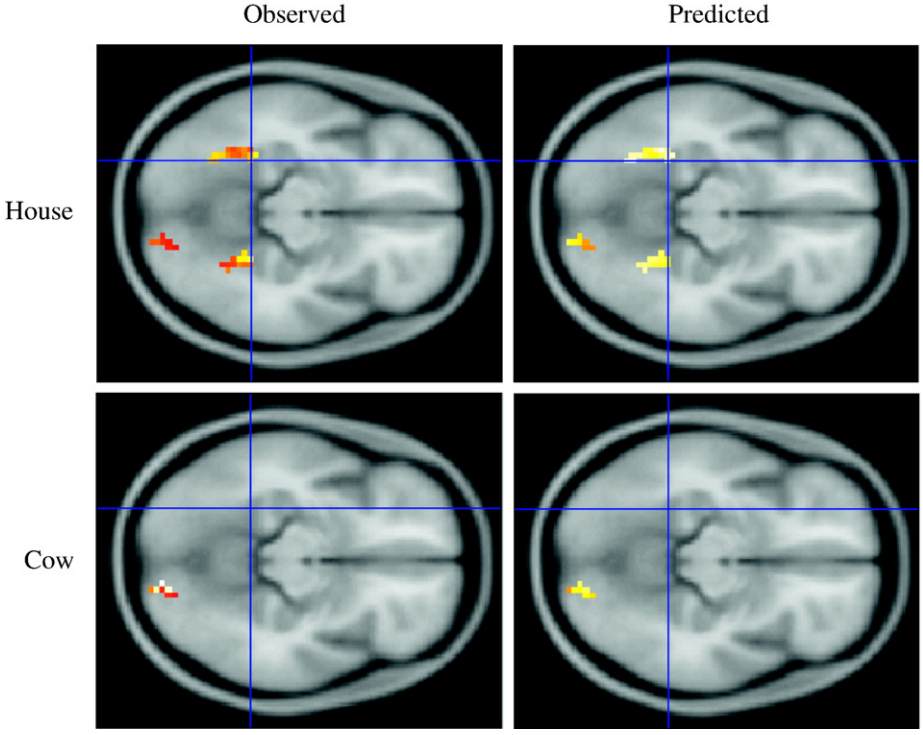

Fig. 4. Observed vs. predicted neural activities at left parahippocampal gyrus (Brodmann area 37, coordinates −28.125, −43.75, −12) for the stimulus words *house* and *cow*. The observed neural activity vector is taken from participant P1, whereas the predicted neural activity vector is estimated by the regression model with BR encoding as explanatory variables and 120 most stable voxels as response variables. In each case, the predicted activity is more similar to the observed activity of the target word than to the other word, suggesting that the predicted activity may be useful to classify words.

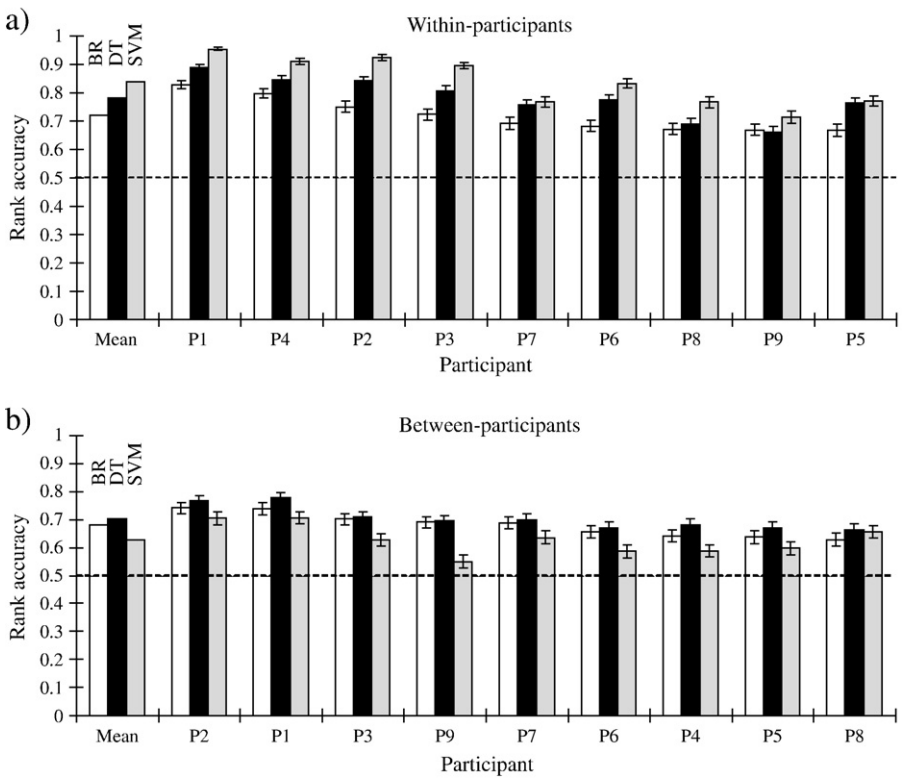

Fig. 5. Decoding mental states given neural activation pattern. A discriminative SVM classifier, which utilizes no hidden layer representation, is compared to two generative nearest neighbor classifiers which extend the regression model, with BR or DT as the explanatory variables. The dashed line indicates chance level at 50%. Participants are sorted according to rank accuracy of the BR model. (a) Within-participants analysis, (b) between-participants analysis. Whereas the discriminative SVM classifier performs the best in the within-participants classification, the generative classifiers generalize better in the between-participants classification.

before by the model? In other words, can the regression model generalize to make predictions for a previously unseen word, given the values of the independent variables (the semantic features) for that word? To test this possibility, all possible pairs of the 43 words were held out (one pair at a time) from the analysis, and a multivariate multiple linear regression model was developed from the data of the remaining 41 words, with semantic feature vectors (either the BR or DT encoding) as the explanatory variables, and observed neural activity vectors (mean PSC across 120 most stable voxels) as the response variables. The estimated regression weights were then used to generate the predicted activation vector for the two unseen words, based on the feature encodings of those two words. Then, the observed neural activation vector for the two unseen words was identified with the class of the predicted activation vector with which it had the higher correlation.

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

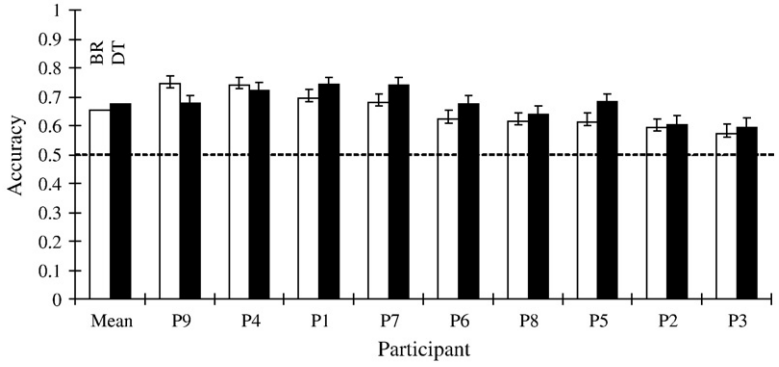

Fig. 6. Distinguishing between two unseen words. Two generative nearest neighbor classifiers which extend the regression model, with BR or DT encoding as explanatory variables, are shown. The dashed line indicates chance level at 50%. Participants are sorted according to accuracy of BR encoding.

There are no significant differences between BR and DT encoding for the remaining 6 participants.

Discussion

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

Comparing the generative classifier and discriminative classifier

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

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

Encoding feature norming features into knowledge types

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

Comparing BR and DT encoding

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

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

To compare the two encoding schemes while equating the number of independent variables, a step-wise analysis was performed to gradually enter additional variables in the regression model, instead of entering all of them simultaneously. As the number of knowledge types included in the DT encoding increases, the regression fit keeps increasing, as shown in Fig. 7a, but the classification accuracy on novel stimuli, shown in Fig. 7b, increases at first but peaks and gradually decreases—clear evidence of overfitting. With fewer knowledge types, the BR encoding overfits less to the data and generalizes better to

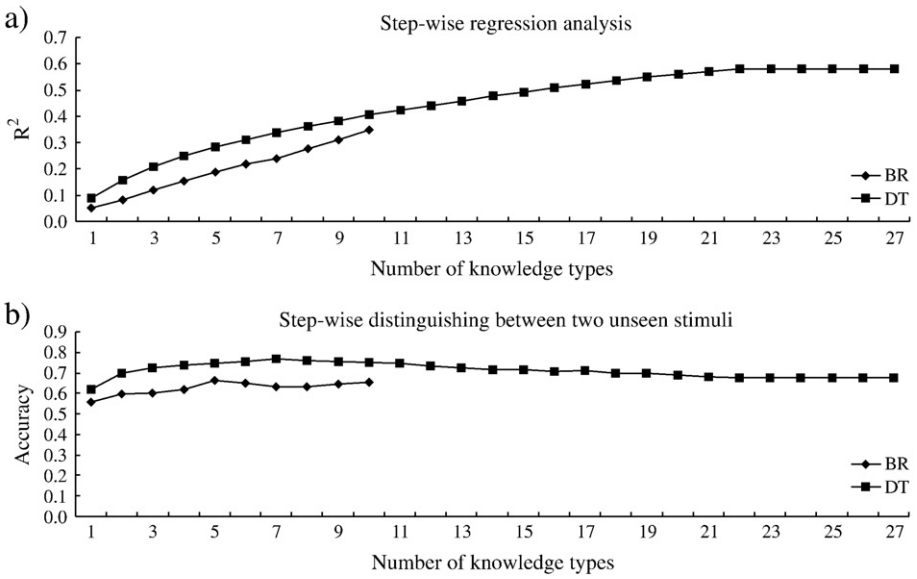

Fig. 7. Step-wise analysis. (a) Step-wise regression analysis, (b) step-wise distinguishing between two unseen stimuli. With finer distinction of knowledge types, DT encoding is more prone to overfitting than BR encoding. As the number of knowledge types in DT encoding is increased, the regression fit keeps increasing, but classification accuracy on unseen stimuli increases at first but peaks and gradually decreases—clear evidence of overfitting. With fewer knowledge types, BR overfits to a lesser extent.

unseen words. Moreover, the performance of the BR encoding peaks when about 6 knowledge types are entered into the regression model, reaching an average accuracy of 68%, whereas the performance of the DT encoding peaks when about 8 knowledge types are used, reaching an average accuracy of 77%. Notice that, although the BR and DT encodings are constructed subject to different criteria, the features of the two encoding schemes that are found to be the most important in the step-wise analysis are similar. The underlying semantic features that provide the best account of the neural activation data consist of taxonomic and visual features (e.g. *visual color*, *visual motion,* and *function* for the BR encoding and *internal component*, *entity behavior,* and *associated entity* for the DT encoding). Tables 7 and 8 show the ranked order list of each of the BR knowledge type and each of the DT knowledge type's ability to classify mental state (within-participants analysis, averaged over participants), respectively. Thus the superficial differences between BR and DT feature encoding schemes lessen or disappear in the light of more sensitive assessments, and the modeling converges on some core encoding features that provide a good converging account of the data.

Comparing feature norming features and word–co-occurrence features

The various models described here were compared to a similar analysis that used features derived from word co-occurrence in a text corpus (Mitchell et al., 2008). In that model, the features of each word were its co-occurrence frequencies with each of 25 verbs of sensorimotor interaction with physical objects, such as *push* and *see*. The model using co-occurrence features produced an average R^2 of 0.71 when accounting for the systematic variance in neural activity, an average rank accuracy of 0.82 when classifying mental states within-participants, an average rank accuracy of 0.75 when classifying mental states across-participants, and an average accuracy of 0.79 when distinguishing between two previously unseen stimuli. While the performance in rank accuracy when classifying mental states is not statistically different ($p<0.05$) from that of DT encoding, the advantage of the co-occurrence model in distinguishing between two unseen stimuli is statistically significant ($p<0.05$). One explanation may be that the encoded object-by-knowledge-type matrices are sparse and heavily weighted in a handful of knowledge types (e.g.

Table 7
Each BR knowledge type's ability to classify mental states.

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

Table 8
Each DT knowledge type's ability to classify mental state.

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

visual knowledge types). Feature norming may have fared better if the features corresponded more closely to the types of interactions with objects that are suggested by the 25 sensorimotor verbs. The shortcoming of feature norming in accounting for participants' thoughts when they think about an object is that participants may fail to retrieve a characteristic but psychologically unavailable feature of an object. For example, for an item like *celery*, the attribute of *taste* may be highly characteristic but relatively unavailable. By contrast, using a fixed set of 25 verbs ensures that all 25 will play a role in the encoding. One way to bring the two approaches together is to ask participants in a feature norming study to assess 25 features of an object that correspond to the verbs.

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

Voxel selection method

One property of this study is that it focused on only the most stable voxels, which may have biased the findings in favor of encodings of visual attributes of the items. The voxel selection procedure increases the signal-to-noise ratio and serves as an effective dimensionality reduction tool that empirically derives regions of interest by assuming that the most informative voxels are those that have activation patterns that are stable across multiple presentations of the set of stimuli. The ability of our models to perform classification across previously unseen words suggests we have, to some extent, successfully captured this intermediate semantic representation. Whether the voxels extracted by this procedure correspond to the human semantic system may be task-dependent. For instance, in our task where the stimulus presentations consist of line drawings with text labels, the voxels extracted by this procedure are mostly in the posterior and occipital regions, since our stimuli consist of easily depicted objects and the visual properties of the stimuli are the most invariant part of the stimuli. Indeed, visual features are among the most important features that account for our neural activation data. If the stimulus presentation consists of only line drawings or text labels, different sets of voxels might be selected. Shinkareva et al. (2007) studied the exact question of the neural representation of pictures versus words. They applied similar machine learning methods on fMRI data to identify the cognitive state associated with viewings of 10 words (5 tools and 5 dwellings) and, separately, with viewings of 10 pictures (line drawings) of the objects named by the words. In addition to selecting voxels from the whole brain, they also identified single brain regions that consistently contained voxels used in identification of object categories across participants. We performed a similar analysis to restrict the analysis space to some predetermined regions of interests. That is, instead of selecting 120 voxels from the whole brain, the voxel selection is applied separately to the frontal lobe, temporal lobe, parietal lobe, occipital lobe, fusiform gyrus, and hippocampus. When only a single region of interest is considered, the highest category identification in the within-participant mental state decoding task is achieved when analysis space is restricted within the occipital lobe, as shown in Table 9. However, other regions of interests like the parietal lobe and the fusiform gyrus also carry important information to decode mental state between participants and to distinguish between the activation of two previously unseen words. Indeed, selecting voxels from the whole brain yields the best category identification in the classifier analysis.

Table 9
Restricting analysis space through ROIs.

Model	All	Frontal	Temporal	Parietal	Occipital	Fusiform	Hippocampus
(a) Regression fit to the fMRI data (R^2)							
BR	0.35	0.27	0.27	0.32	0.30	0.38	0.24
DT	0.58	0.55	0.55	0.58	0.56	0.61	0.52
(b) Ability to decode mental states, within participants (rank accuracy)							
BR	0.72	0.57	0.60	0.64	0.70	0.67	0.52
DT	0.78	0.58	0.62	0.66	0.77	0.69	0.53
(c) Ability to decode mental states, within participants (rank accuracy)							
BR	0.68	0.47	0.47	0.57	0.59	0.61	0.50
DT	0.70	0.46	0.47	0.56	0.60	0.60	0.49
(d) Ability to distinguish between the activation of two previously unseen words (accuracy)							
BR	0.65	0.60	0.57	0.66	0.62	0.69	0.49
DT	0.68	0.61	0.60	0.69	0.64	0.70	0.51

Conclusions and contributions

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

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

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

expensive to acquire and that many classifier techniques would perform significantly better if more training data were available.

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

Acknowledgments

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. Also, we would like to thank Dr. Matthew Harrison and Dr. Larry Wasserman for advice on the statistical analysis.

References

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

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. The 5th Annual ACM Workshop on COLT, ACM Press, 144–152.

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

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

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

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

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

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

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

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

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

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

McRae, K., Cree, G.S., Westmacott, R., de Sa, V.R., 1999. Further evidence for feature correlations in semantic memory. Canadian Journal of Experimental Psychology 53, 360–373.

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

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

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

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

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

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

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

Shinkareva, S.V., Mason, R.A., Malave, V.L., Wang, W., Mitchell, T.M., Just, M.A., 2008. Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. PLoS ONE 3 (1), e1394. doi:10.1371/journal.pone.0001394.

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