Predicting Human Brain Activity Associated with the Meanings of Nouns

Tom M. Mitchell,[1]* Svetlana V. Shinkareva,[2] Andrew Carlson,[1] Kai-Min Chang,[3,4] Vicente L. Malave,[5] Robert A. Mason,[3] Marcel Adam Just[3]

The question of how the human brain represents conceptual knowledge has been debated in many scientific fields. Brain imaging studies have shown that different spatial patterns of neural activation are associated with thinking about different semantic categories of pictures and words (for example, tools, buildings, and animals). We present a computational model that predicts the functional magnetic resonance imaging (fMRI) neural activation associated with words for which fMRI data are not yet available. This model is trained with 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.

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 tools, buildings, and animals. 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 effects of brain damage have found deficits that are specific to given semantic categories (such as animals) (*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 nonliving objects (*18*, *24*). Although 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 make no attempt to predict the specific brain activation that will be produced when a human subject reads a particular word or views a drawing of a particular object.

We present a computational model that makes directly testable predictions of the fMRI activity associated with thinking about arbitrary concrete nouns, including many nouns for which no fMRI data are currently available. The theory underlying this computational model is that the neural basis of the semantic representation of concrete nouns is related to the distributional properties of those words in a broadly based corpus of the language. We describe experiments training competing computational models based on different assumptions regarding the underlying features that are used in the brain for encoding of meaning of concrete objects. We present experimental evidence showing that the best of these models predicts 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 those expected by chance. These results establish a direct, predictive relationship between the statistics of word co-occurrence in text and the neural activation associated with thinking about word meanings.

Approach. We use a trainable computational model that predicts the neural activation for any given stimulus word w using a two-step process, illustrated in Fig. 1. Given an arbitrary 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 very large 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

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

where $f_i(w)$ is the value of the ith 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 ith 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 component images that model the brain activation associated with different semantic components of the input stimulus words.

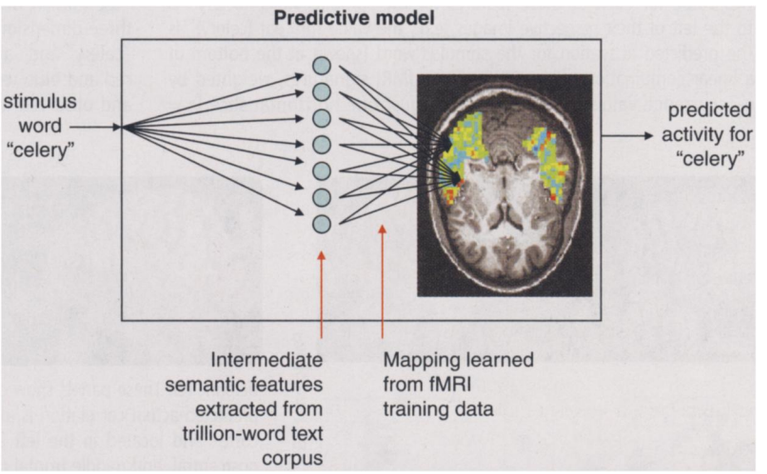


Fig. 1. Form of the model for predicting fMRI activation for arbitrary noun stimuli. fMRI activation is predicted in a two-step process. The first step encodes the meaning of the input stimulus word in terms of intermediate semantic features whose values are extracted from a large corpus of text exhibiting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

[1]Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [2]Department of Psychology, University of South Carolina, Columbia, SC 29208, USA. [3]Center for Cognitive Brain Imaging, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [4]Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [5]Cognitive Science Department, University of California, San Diego, La Jolla, CA 92093, USA.

*To whom correspondence should be addressed. E-mail: Tom.Mitchell@cs.cmu.edu

To fully specify a model within this computational modeling framework, one must first define a set of intermediate semantic features $f_1(w)$ $f_2(w)$...$f_n(w)$ to be extracted from the text corpus. In this paper, each intermediate semantic feature is defined in terms of the co-occurrence statistics of the input stimulus word w with a particular other word (e.g., "taste") or set of words (e.g., "taste," "tastes," or "tasted") within the text corpus. The model is trained by the application of multiple regression to these features $f_i(w)$ and the observed fMRI images, so as to obtain maximum-likelihood estimates for the model parameters c_{vi} (*26*). Once trained, the computational model can be evaluated by giving it words outside the training set and comparing its predicted fMRI images for these words with observed fMRI data.

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 statistical word distributions are frequently used to approximate the meaning of documents and words (*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. Although the correctness of this linearity assumption is debatable, it is consistent with the widespread use of linear models in fMRI analysis (*27*) and with the 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. Instead, it considers all cortical voxels and allows the training data to determine which locations are systematically modulated by which aspects of word meanings.

Results. We evaluated this computational model using fMRI data from nine healthy, college-age participants who viewed 60 different word-picture pairs presented six times each. Anatomically defined regions of interest were automatically labeled according to the methodology in (*28*). The 60 randomly ordered stimuli included five items from each of 12 semantic categories (animals, body parts, buildings, building parts, clothing, furniture, insects, kitchen items, tools, vegetables, vehicles, and other man-made items). A representative fMRI image for each stimulus was created by computing the mean fMRI response over its six presentations, and the mean of all 60 of these representative images was then subtracted from each [for details, see (*26*)].

To instantiate our modeling framework, we first chose a set of intermediate semantic features. To be effective, the intermediate 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 com-

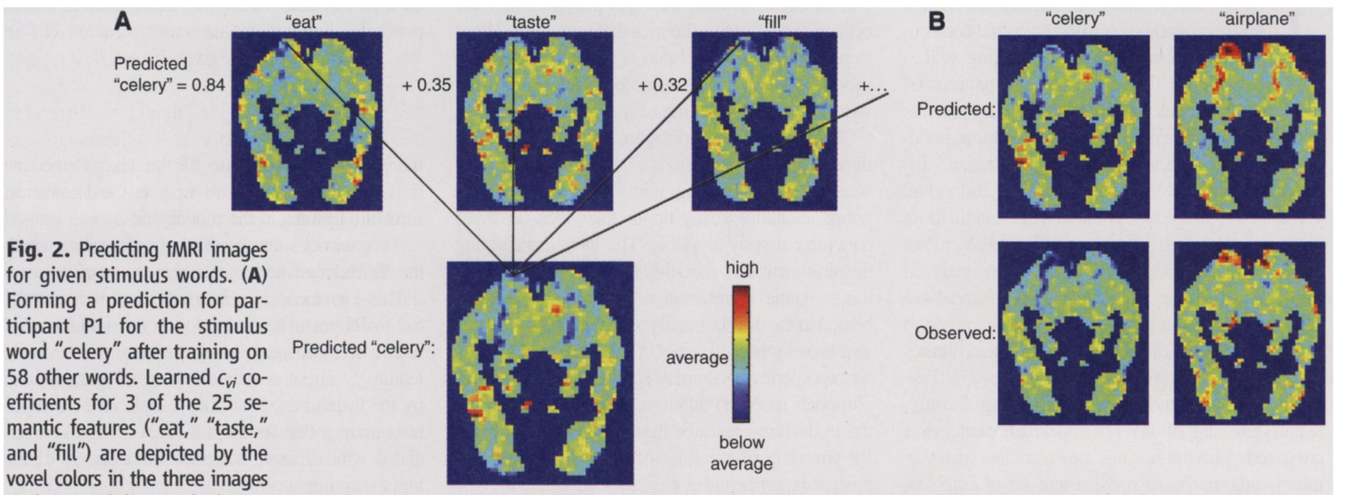
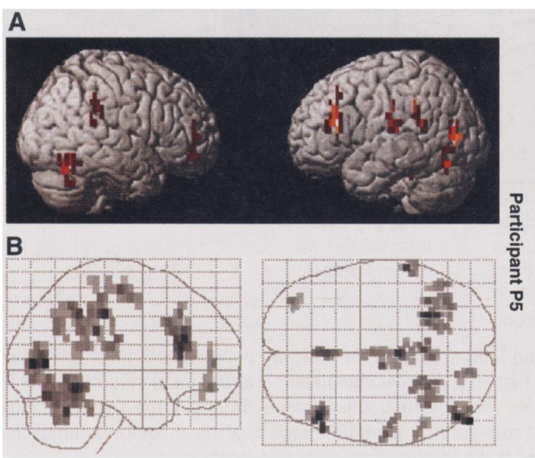


Fig. 2. Predicting fMRI images for given stimulus words. (**A**) Forming a prediction for participant P1 for the stimulus word "celery" after training on 58 other words. Learned c_{vi} coefficients for 3 of the 25 semantic features ("eat," "taste," and "fill") are depicted by the voxel colors in the three images at the top of the panel. The co-occurrence value for each of these features for the stimulus word "celery" is shown to the left of their respective images [e.g., the value for "eat (celery)" is 0.84]. The predicted activation for the stimulus word [shown at the bottom of (A)] is a linear combination of the 25 semantic fMRI signatures, weighted by their co-occurrence values. This figure shows just one horizontal slice [z = −12 mm in Montreal Neurological Institute (MNI) space] of the predicted three-dimensional image. (**B**) Predicted and observed fMRI images for "celery" and "airplane" after training that uses 58 other words. The two long red and blue vertical streaks near the top (posterior region) of the predicted and observed images are the left and right fusiform gyri.



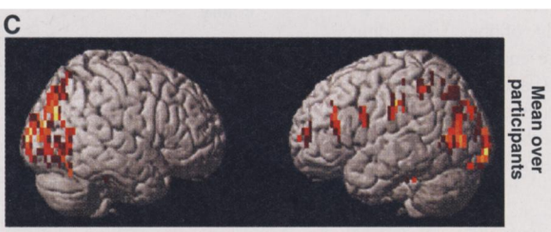


Fig. 3. Locations of most accurately predicted voxels. Surface (**A**) and glass brain (**B**) rendering of the correlation between predicted and actual voxel activations for words outside the training set for participant P5. These panels show clusters containing at least 10 contiguous voxels, each of whose predicted-actual correlation is at least 0.28. These voxel clusters are distributed throughout the cortex and located in the left and right occipital and parietal lobes; left and right fusiform, postcentral, and middle frontal gyri; left inferior frontal gyrus; medial frontal gyrus; and anterior cingulate. (**C**) Surface rendering of the predicted-actual correlation averaged over all nine participants. This panel represents clusters containing at least 10 contiguous voxels, each with average correlation of at least 0.14.

ponents 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 (*18*, *29*), 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." 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 for a given input stimulus word w is the normalized co-occurrence count of w with any of three forms of the verb (e.g., "taste," "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.

We trained a separate computational model for each of the nine participants, using this set of 25 semantic features. Each trained model was evaluated by means of a "leave-two-out" cross-validation approach, in which the model was repeatedly trained with only 58 of the 60 available word stimuli and associated fMRI images. Each trained model was tested by requiring that it first predict the fMRI images for the two "held-out" words and then match these correctly to their corresponding held-out fMRI images. The process of predicting the fMRI image for a held-out word is illustrated in Fig. 2A. The match between the two predicted and the two observed fMRI images was determined by which match had a higher cosine similarity, evaluated over the 500 image voxels with the most stable responses across training presentations (*26*). The expected accuracy in matching the left-out words to their left-out fMRI images is 0.50 if the model performs at chance levels. An accuracy of 0.62 or higher for a single model trained for a single participant was determined to be statistically significant ($P < 0.05$) relative to chance, based on the empirical distribution of accuracies for randomly generated null models (*26*). Similarly, observing an accuracy of 0.62 or higher for each of the nine independently trained participant-specific models would be statistically significant at $P < 10^{-11}$.

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

Visual inspection of the predicted fMRI images produced by the trained models shows that these predicted images frequently capture substantial aspects of brain activation associated with stimulus words outside the training set. An example is shown in Fig. 2B, where the model was trained on 58 of the 60 stimuli for participant P1, omitting "celery" and "airplane." Although the predicted fMRI images for "celery" and "airplane" are not perfect, they capture substantial components of the activation actually observed for these two stimuli. A plot of similarities between all 60 predicted and observed fMRI images is provided in fig. S3.

The model's predictions are differentially accurate in different brain locations, presumably more accurate in those locations involved in encoding the semantics of the input stimuli. Figure 3 shows the model's "accuracy map," indicating the cortical regions where the model's predicted activations for held-out words best correlate with the observed activations, both for an individual participant (P5) and averaged over all nine participants. These highest-accuracy voxels are meaningfully distributed across the cortex, with the left hemisphere more strongly represented, appearing in left inferior temporal, fusiform, motor cortex, intraparietal sulcus, inferior frontal, orbital frontal, and the occipital cortex. This left hemisphere dominance is consistent with the generally held view that the left hemisphere plays a larger role than the right hemisphere in semantic representation. High-accuracy voxels also appear in both hemispheres in the occipital cortex, intraparietal sulcus, and some of the inferior temporal regions, all of which are also likely to be involved in visual object processing.

It is interesting to consider whether these trained computational models can extrapolate to make accurate predictions for words in new semantic categories beyond those in the training set. To test this, we retrained the models but this time we excluded from the training set all examples belonging to the same semantic category as either of the two held-out test words (e.g., when testing on "celery" versus "airplane," we removed every food and vehicle stimulus from the training set, training on only 50 words). In this case, the cross-validated prediction accuracies were 0.74, 0.69, 0.67, 0.69, 0.64,

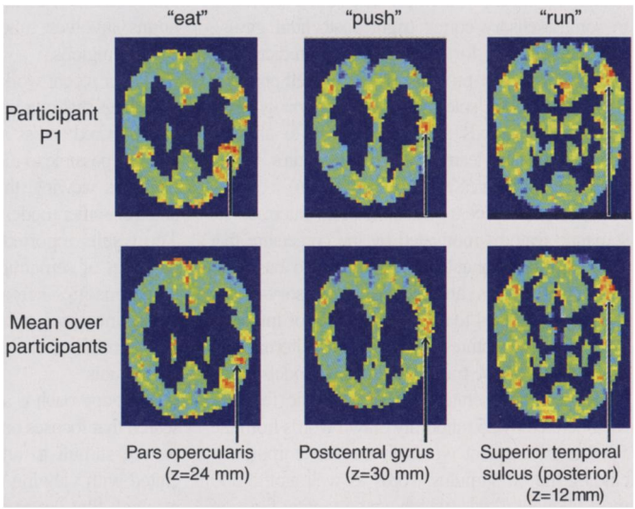


Fig. 4. Learned voxel activation signatures for 3 of the 25 semantic features, for participant P1 (top panels) and averaged over all nine participants (bottom panels). Just one horizontal z slice is shown for each. The semantic feature associated with the verb "eat" predicts substantial activity in right pars opercularis, which is believed to be part of the gustatory cortex. The semantic feature associated with "push" activates the right postcentral gyrus, which is believed to be associated with premotor planning. The semantic feature for the verb "run" activates the posterior portion of the right superior temporal sulcus, which is believed to be associated with the perception of biological motion.

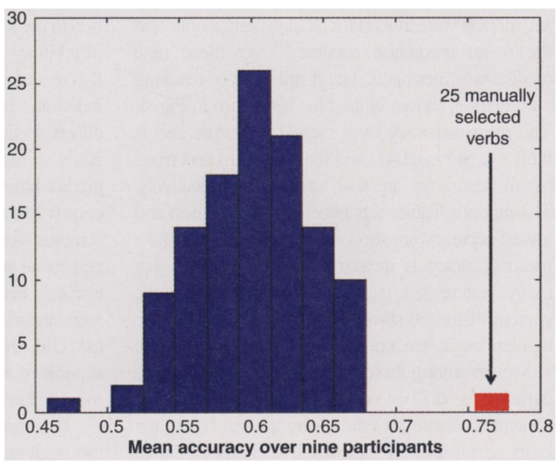


Fig. 5. Accuracies of models based on alternative intermediate semantic feature sets. The accuracy of computational models that use 115 different randomly selected sets of intermediate semantic features is shown in the blue histogram. Each feature set is based on 25 words chosen at random from the 5000 most frequent words, excluding the 500 most frequent words and the stimulus words. The accuracy of the feature set based on manually chosen sensory-motor verbs is shown in red. The accuracy of each feature set is the average accuracy obtained when it was used to train models for each of the nine participants.

0.78, 0.68, 0.64, and 0.78 (mean = 0.70). This ability of the model to extrapolate to words semantically distant from those on which it was trained suggests that the semantic features and their learned neural activation signatures of the model may span a diverse semantic space.

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

In order to test the ability of the model to distinguish among an even more diverse range of words, we tested its ability to resolve among 1000 highly frequent words (the 1300 most frequent tokens in the text corpus, omitting the 300 most frequent). Specifically, we conducted a leave-one-out test in which the model was trained using 59 of the 60 available stimulus words. It was then given the fMRI image for the held-out word and a set of 1001 candidate words (the 1000 frequent tokens, plus the held-out word). It ranked these 1001 candidates by first predicting the fMRI image for each candidate and then sorting the 1001 candidates by the similarity between their predicted fMRI image and the fMRI image it was provided. The expected percentile rank of the correct word in this ranked list would be 0.50 if the model were operating at chance. The observed percentile ranks for the nine participants were 0.79, 0.71, 0.74, 0.67, 0.73, 0.77, 0.70, 0.63, and 0.76 (mean = 0.72), indicating that the model is to some degree applicable across a semantically diverse set of words [see (*26*) for details].

A second approach to evaluating our computation model, beyond quantitative measurements of its prediction accuracy, is to examine the learned basis set of fMRI signatures for the 25 verb-based signatures. These 25 signatures represent the model's learned decomposition of neural representations into their component semantic features and provide the basis for all of its predictions. The learned signatures for the semantic features "eat," "push," and "run" are shown in Fig. 4. Notice that each of these signatures predicts activation in multiple cortical regions.

Examining the semantic feature signatures in Fig. 4, one can see that the learned fMRI signature for the semantic feature "eat" predicts strong activation in opercular cortex (as indicated by the arrows in the left panels), which others have suggested is a component of gustatory cortex involved in the sense of taste (*30*). Also, the learned fMRI signature for "push" predicts substantial activation in the right postcentral gyrus, which is widely assumed to be involved in the planning of complex, coordinated movements (*31*). Furthermore, the learned signature for "run" predicts strong activation in the posterior portion of the right superior temporal lobe along the sulcus, which others have suggested is involved in perception of biological motion (*32*, *33*). To summarize, these learned signatures cause the model to predict that the neural activity representing a noun will exhibit activity in gustatory cortex to the degree that this noun co-occurs with the verb "eat," in motor areas to the degree that it co-occurs with "push," and in cortical regions related to body motion to the degree that it co-occurs with "run." Whereas the top row of Fig. 4 illustrates these learned signatures for participant P1, the bottom row shows the mean of the nine signatures learned independently for the nine participants. The similarity of the two rows of signatures demonstrates that these learned intermediate semantic feature signatures exhibit substantial commonalities across participants.

The learned signatures for several other verbs also exhibit interesting correspondences between the function of cortical regions in which they predict activation and that verb's meaning, though in some cases the correspondence holds for only a subset of the nine participants. For example, additional features for participant P1 include the signature for "touch," which predicts strong activation in somatosensory cortex (right postcentral gyrus), and the signature for "listen," which predicts activation in language-processing regions (left posterior superior temporal sulcus and left pars triangularis), though these trends are not common to all nine participants. The learned feature signatures for all 25 semantic features are provided at (*26*).

Given the success of this set of 25 intermediate semantic features motivated by the conjecture that the neural components corresponding to basic semantic properties are related to sensory-motor verbs, it is natural to ask how this set of intermediate semantic features compares with alternatives. To explore this, we trained and tested models based on 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 115 random feature sets was generated. For each feature set, models were trained for all nine participants, and the mean prediction accuracy over these nine models was measured. The distribution of resulting accuracies is shown in the blue histogram in Fig. 5. The mean accuracy over these 115 feature sets is 0.60, the SD is 0.041, and the minimum and maximum accuracies are 0.46 and 0.68, respectively. The random feature sets generating the highest and lowest accuracy are shown at (*26*). The fact that the mean accuracy is greater than 0.50 suggests that many feature sets capture some of the semantic content of the 60 stimulus words and some of the regularities in the corresponding brain activation. However, among these 115 feature sets, none came close to the 0.77 mean accuracy of our manually generated feature set (shown by the red bar in the histogram in Fig. 5). This result suggests the set of features defined by our sensory-motor verbs is somewhat distinctive in capturing regularities in the neural activation encoding the semantic content of words in the brain.

Discussion. The results reported here establish a direct, predictive relationship between the statistics of word co-occurrence in text and the neural activation associated with thinking about word meanings. Furthermore, the computational models trained to make these predictions provide insight into how the neural activity that represents objects can be decomposed into a basis set of neural activation patterns associated with different semantic components of the objects.

The success of the specific model, which uses 25 sensory-motor verbs (as compared with 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.

Other recent work has suggested that the neural encodings that represent concrete objects are at least partly shared across individuals, based on evidence that it is possible to identify which of several items a person is viewing, through only their fMRI image and a classifier model trained from other people (*34*). The results reported here show that the learned basis set of semantic features also shares certain commonalities across individuals and may help determine more directly which factors of neural representations are similar and different across individuals.

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 (*9*, *35*, *36*) and to research that compares perceived similarities between object shapes to their similarities based on fMRI activation (*37*). Recent work (*36*) has shown that it is possible to predict aspects of fMRI activation in parts of visual cortex based on visual features of arbitrary scenes and to use this predicted activation to identify which of a set of candidate scenes an individual is viewing. Our work differs from these efforts, in that we focus on encodings of more abstract semantic concepts signified by words and predict brain-wide fMRI activations based on text corpus features that capture semantic aspects of the stimulus word, rather than visual features that capture perceptual aspects. Our work is also related to recent research that uses machine learning algorithms to train classifiers of mental states based on fMRI data (*38*, *39*), though it differs in that our models are capable of extrapolating to predict fMRI images for mental states not present in the training set.

This research represents a shift in the paradigm for studying neural representations in the brain,

moving from work that has cataloged the patterns of fMRI activity associated with specific categories of words and pictures to instead building computational models that predict the fMRI activity for arbitrary words (including thousands of words for which fMRI data are not yet available). This is a natural progression as the field moves from pretheoretical cataloging of data toward development of computational models and the beginnings of a theory of neural representations. Our computational models can be viewed as encoding a restricted form of predictive theory, one that answers such questions as "What is the predicted fMRI neural activity encoding word *w*?" and "What is the basis set of semantic features and corresponding components of neural activation that explain the neural activations encoding meanings of concrete nouns?" Although we remain far from a causal theory explaining how the brain synthesizes these representations from its sensory inputs, answers even to these questions promise to shed light on some of the key regularities underlying neural representations of meaning.

References and Notes

1. J. V. Haxby *et al.*, *Science* **293**, 2425 (2001).
2. A. Ishai, L. G. Ungerleider, A. Martin, J. L. Schouten, J. V. Haxby, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 9379 (1999).
3. N. Kanwisher, J. McDermott, M. M. Chun, *J. Neurosci.* **17**, 4302 (1997).
4. T. A. Carlson, P. Schrater, S. He, *J. Cogn. Neurosci.* **15**, 704 (2003).
5. D. D. Cox, R. L. Savoy, *Neuroimage* **19**, 261 (2003).
6. T. Mitchell *et al.*, *Mach. Learn.* **57**, 145 (2004).
7. S. J. Hanson, T. Matsuka, J. V. Haxby, *Neuroimage* **23**, 156 (2004).
8. S. M. Polyn, V. S. Natu, J. D. Cohen, K. A. Norman, *Science* **310**, 1963 (2005).
9. A. J. O'Toole, F. Jiang, H. Abdi, J. V. Haxby, *J. Cogn. Neurosci.* **17**, 580 (2005).
10. K. Kipper, A. Korhonen, N. Ryant, M. Palmer, *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 24 to 26 May 2006.
11. G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller, *Int. J. Lexicography* **3**, 235 (1990).
12. C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database* (Massachusetts Institute of Technology Press, Cambridge, MA, 1998).
13. K. W. Church, P. Hanks, *Comput. Linguist.* **16**, 22 (1990).
14. T. K. Landauer, S. T. Dumais, *Psychol. Rev.* **104**, 211 (1997).
15. D. Lin, S. Zhao, L. Qin, M. Zhou, *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, August 2003 (Morgan Kaufmann, San Francisco, 2003), pp. 1492–1493.
16. D. M. Blei, A. Y. Ng, M. I. Jordan, *J. Mach. Learn. Res.* **3**, 993 (2003).
17. R. Snow, D. Jurafsky, A. Ng, *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 17 to 21 July 2006.
18. G. S. Cree, K. McRae, *J. Exp. Psychol. Gen.* **132**, 163 (2003).
19. A. Caramazza, J. R. Shelton, *J. Cogn. Neurosci.* **10**, 1 (1998).
20. S. J. Crutch, E. K. Warrington, *Brain* **126**, 1821 (2003).
21. D. Samson, A. Pillon, *Brain Lang.* **91**, 252 (2004).
22. A. Martin, L. L. Chao, *Curr. Opin. Neurobiol.* **11**, 194 (2001).
23. R. F. Goldberg, C. A. Perfetti, W. Schneider, *J. Neurosci.* **26**, 4917 (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, www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13 (Linguistic Data Consortium, Philadelphia, PA, 2006).
26. See Supporting Online Material.
27. K. J. Friston *et al.*, *Hum. Brain Mapp.* **2**, 189 (1995).
28. N. Tzourio-Mazoyer *et al.*, *Neuroimage* **15**, 273 (2002).
29. A. Martin, L. G. Ungerleider, J. V. Haxby, in *The New Cognitive Neurosciences*, M. S. Gazzainga, Ed. (Massachusetts Institute of Technology Press, Cambridge, MA, ed. 2, 2000), pp. 1023–1036.
30. B. Cerf, D. LeBihan, P. F. Van de Moortele, P. Mac Leod, A. Faurion, *Ann. N.Y. Acad. Sci.* **855**, 575 (1998).
31. K. A. Pelphrey, J. P. Morris, C. R. Michelich, T. Allison, G. McCarthy, *Cereb. Cortex* **15**, 1866 (2005).
32. L. M. Vaina, J. Solomon, S. Chowdhury, P. Sinha, J. Belliveau, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 11656 (2001).
33. K. Sakai *et al.*, *Magn. Reson. Med.* **33**, 736 (1995).
34. S. V. Shinkareva *et al.*, *PLoS One* **3**, e1394 (2008).
35. D. R. Hardoon, J. Mourao-Miranda, M. Brammer, J. Shawe-Taylor, *Neuroimage* **37**, 1250 (2007).
36. K. N. Kay, T. Naselaris, R. J. Prenger, J. L. Gallant, *Nature* **452**, 352 (2008).
37. S. Edelman, K. Grill-Spector, T. Kushnir, R. Malach, *Psychobiology* **26**, 309 (1998).
38. J. D. Haynes, G. Rees, *Nat. Rev. Neurosci.* **7**, 523 (2006).
39. K. A. Norman, S. M. Polyn, G. J. Detre, J. V. Haxby, *Trends Cogn. Sci.* **10**, 424 (2006).

40. This research was funded by grants from the W. M. Keck Foundation, NSF, and by a Yahoo! Fellowship to A.C. We acknowledge Google for making available its data from the trillion-token text corpus. We thank W. Cohen for helpful suggestions regarding statistical significance tests.

Supporting Online Material

www.sciencemag.org/cgi/content/full/320/5880/1191/DC1
Materials and Methods
SOM Text
Figs. S1 to S5
References

12 November 2007; accepted 3 April 2008
10.1126/science.1152876

REPORTS

The Cassiopeia A Supernova Was of Type IIb

Oliver Krause,[1]* Stephan M. Birkmann,[1] Tomonori Usuda,[2] Takashi Hattori,[2] Miwa Goto,[1] George H. Rieke,[3] Karl A. Misselt[3]

Cassiopeia A is the youngest supernova remnant known in the Milky Way and a unique laboratory for supernova physics. We present an optical spectrum of the Cassiopeia A supernova near maximum brightness, obtained from observations of a scattered light echo more than three centuries after the direct light of the explosion swept past Earth. The spectrum shows that Cassiopeia A was a type IIb supernova and originated from the collapse of the helium core of a red supergiant that had lost most of its hydrogen envelope before exploding. Our finding concludes a long-standing debate on the Cassiopeia A progenitor and provides new insight into supernova physics by linking the properties of the explosion to the wealth of knowledge about its remnant.

The supernova remnant Cassiopeia A is one of the most-studied objects in the sky, with observations from the longest radio waves to gamma rays. The remnant expansion rate indicates that the core of its progenitor star collapsed around the year 1681 ± 19, as viewed from Earth (*1*). Because of its youth and proximity of $3.4^{+0.3}_{-0.1}$ kpc (*2*), Cas A provides a unique opportunity to probe the death of a massive star and to test theoretical models of core-collapse supernovae. However, such tests are compromised because the Cas A supernova showed at most a faint optical display on Earth at the time of explosion. The lack of a definitive sighting means that there is almost no direct information about the type of the explosion, and the true nature of its progenitor star has been a puzzle since the discovery of the remnant (*3*).

The discovery of light echoes due both to scattering and to absorption and re-emission of the outgoing supernova flash (*4*, *5*) by the interstellar dust near the remnant raised the possibility of conducting a postmortem study of the last historic Galactic supernova by observing its scattered light. Similarly, the determination of a supernova spectral type long after its explosion using light echoes was recently demonstrated for an extragalactic supernova (*6*).

We have monitored infrared echoes around Cas A at a wavelength of 24 μm with use of the multiband imaging photometer (MIPS) instrument aboard the Spitzer Space Telescope (*4*). The results confirm that they arise from the flash emitted in the initial explosion of Cas A (*5*). An image taken on 20 August 2007 revealed a bright (flux density $F_{24\mu m} = 0.36 \pm 0.04$ Jy, 1 Jy $\equiv 10^{-26}$ W m^{-2} Hz^{-1}) and mainly unresolved echo feature located 80 arc min northwest of Cas A (position angle 311° east of north). It had not been detected ($F_{24\mu m} < 2$ mJy; 5-σ) on two previous images of this region obtained on 2 October 2006 and 23 January 2007 (Fig. 1).

An image obtained on 7 January 2008 shows that the peak of the echo has dropped in surface brightness by a factor of 18 and shifted toward the west. Transient optical emission associated with the infrared echo was detected in an *R*-band image obtained at a wavelength of 6500 Å at the Calar Alto 2.2-m telescope on 6 October 2007

[1]Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany. [2]National Astronomical Observatory of Japan, 650 North A'ohoku Place, Hilo, HI 96720, USA. [3]Steward Observatory, 933 North Cherry Avenue, Tucson, AZ 85721, USA.

*To whom correspondence should be addressed. E-mail: krause@mpia.de