NeuroImage 157 (2017) 511–520

Contents lists available at ScienceDirect

NeuroImage

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

Neural representations of the concepts in simple sentences: Concept activation prediction and context effects

Marcel Adam Just*,[1], Jing Wang[1], Vladimir L. Cherkassky

Center for Cognitive Brain Imaging, Psychology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

ARTICLE INFO

Keywords:
Neural representations of concepts
Predictive modeling
Multi-concept sentences
FMRI
Sentence context effects

ABSTRACT

Although it has been possible to identify individual concepts from a concept's brain activation pattern, there have been significant obstacles to identifying a proposition from its fMRI signature. Here we demonstrate the ability to decode individual prototype sentences from readers' brain activation patterns, by using theory-driven regions of interest and semantic properties. It is possible to predict the fMRI brain activation patterns evoked by propositions and words which are entirely new to the model with reliably above-chance rank accuracy. The two core components implemented in the model that reflect the theory were the choice of intermediate semantic features and the brain regions associated with the neurosemantic dimensions. This approach also predicts the neural representation of object nouns across participants, studies, and sentence contexts. Moreover, we find that the neural representation of an agent-verb-object proto-sentence is more accurately characterized by the neural signatures of its components as they occur in a similar context than by the neural signatures of these components as they occur in isolation.

Introduction

Concepts may be the basic building blocks of thought, but the minimally composed structure of human thought is a proposition consisting of multiple concepts. We report here the capability of predicting the brain activation patterns evoked by the reading of an agent-verb-object proto-sentence. We develop a model that estimates the activation pattern of component words from the mappings learned in context-sensitive environments, and combines them to produce predictions of the resulting activation.

The types of concepts that have previously been most amenable to a mapping between a stimulus item and a brain activation pattern have been concrete object concepts. This type of mapping was initially demonstrated in studies in which the objects were presented visually or were being recalled (Carlson et al., 2003; Connolly et al., 2012; Cox and Savoy, 2003; Eger et al., 2008; Hanson et al., 2004; Haxby et al., 2001; Ishai et al., 1999; Mitchell et al., 2004; O'Toole et al., 2005; Polyn et al., 2005; Shinkareva et al., 2008), and subsequently in studies where the concept was evoked by the word that named it (Just et al., 2010; Peelen and Caramazza, 2012; Shinkareva et al., 2011). Predictive modeling of brain activity associated with concepts was enabled by the postulation of a mediating layer of perceptual and semantic features of the objects, resulting in the decoding from their fMRI signature pictures or text concerning objects (Anderson et al., 2015; Mitchell et al., 2008; Pereira et al., 2011), natural images (Naselaris et al., 2009), faces (Cowen et al., 2014), objects and actions in video clips (Huth et al., 2012; Nishimoto et al., 2011) and in speech (Huth et al., 2016). Other studies have found distinct activation patterns associated with the neural representations of concepts of varying degrees of semantic abstractness (Anderson et al., 2014; Ghio et al., 2016; Wang et al., 2013). A few studies have further demonstrated the ability to associate brain activation patterns with inter-concept relations in a proposition (Frankland and Greene, 2015; Wang et al., 2016). However, characterizing the neural representations of sentences and the effect of contexts has remained a considerable challenge.

The major advance attempted in the current study was to characterize the neural representation of simplified prototype sentences and the effect of context within a theory-driven computational framework. The model was theory-driven, built on the previous knowledge of the neural representations of objects in several ways. First, the stimulus sentence prototypes (e.g., *Plumber grabs pliers*) described a scenario associated with a general theme pertaining to one of three known semantic dimensions of neural representation, namely *shelter, manipulation, or eating*. Second, the meaning components of the stimulus word-concepts were defined as the concepts' relatedness to each of the three dimensions. These three semantic properties were used to predict

* Corresponding author.
E-mail address: just@cmu.edu (M.A. Just).
[1] These authors made equal contributions to the work.

http://dx.doi.org/10.1016/j.neuroimage.2017.06.033
Received 21 November 2016; Accepted 15 June 2017
Available online 17 June 2017

Table 1
Thirty-six stimulus sentences. Each set includes 9 sentences that are composed from 27 content words.

	Set 1	Set 2	Set 3	Set 4
Shelter	Explorer enters car	Explorer repairs church	Explorer exits house	Tourist enters car
	Hiker exits church	Hiker enters house	Hiker repairs car	Explorer exits church
	Tourist repairs house	Tourist exits car	Tourist enters church	Hiker repairs house
Manipulation	Plumber drops hammer	Plumber lifts knife	Plumber grabs pliers	Mechanic drops hammer
	Carpenter grabs knife	Carpenter drops pliers	Carpenter lifts hammer	Plumber grabs knife
	Mechanic lifts pliers	Mechanic grabs hammer	Mechanic drops knife	Carpenter lifts pliers
Eating	Diner bites carrot	Diner tastes celery	Diner chews tomato	Picnicker bites carrot
	Glutton chews celery	Glutton bites tomato	Glutton tastes carrot	Diner chews celery
	Picnicker tastes tomato	Picnicker chews carrot	Picnicker bites celery	Glutton tastes tomato

the neural representation of word concepts and the proto-sentences that they composed. Third, the analysis of the activation patterns focused on the *a priori* specified brain regions that were associated with these three dimensions. The approach was computational in the sense that it established the mapping between the concepts' semantic properties on the three dimensions and the voxel activation patterns associated with reading the sentences. This mapping was sufficiently detailed to characterize the behavior of individual voxels with respect to the neurosemantic properties, and sufficiently robust to be generalized to predict the neural signatures of new sentences with similar contexts but composed of new words.

A brief terminological and theoretical disclaimer is warranted about our use of the word *sentence* in this article. The stimuli were proto-sentences, in that they were lacking articles for the nouns: *Hiker enters house* was presented instead of *The hiker enters the house*. The stimuli would be more accurately referred to as *proto-sentences*, but for brevity following this disclaimer we use the term *sentences*. The theoretical disclaimer concerns the fact that our analysis and model focuses on the neural representation of the three content words of each sentence, and not their thematic or syntactic structure. In fact, all of the stimuli were identical with respect to thematic and syntactic structure (agent-verb-object), and our model does not characterize the representation of these identical aspects of the sentences. Thus our reference to the neural representation of the sentence stimuli refers to the neural representation of the three content words, and not to thematic or syntactic aspects of the representation.

We hypothesized that the neural activation pattern evoked by a simple proto-sentence can be predicted by the semantic properties of the sentence's component words. The theoretical background for construing the neural representations of concrete objects is a neural/semantic account of concrete object representation (Just et al., 2010; Mitchell et al., 2008) that postulates that the neural signature of a concept is composed of component activations in the various brain subsystems that come into play during the consideration of, or interaction with, the concept. These component activations are associated with different dimensions of the meaning of the concept. For example, thinking about a *knife* evokes motor and premotor areas, which have been associated with action verb processing in previous studies (Hauk et al., 2004; Pulvermüller et al., 2005). These regions are found to respond to various manipulable objects such as hand tools, and to not respond to non-manipulable objects, indicating their role in representing the meaning dimension of *manipulation* or body-object interaction. The three main neurosemantic dimensions underlying the representation of 60 concrete nouns were *shelter, manipulation,* and *eating* (Just et al., 2010). Such sets of dimensions and their corresponding brain subsystems are proposed to constitute part of the basis set for neurally representing concrete objects. Thus, the activation pattern associated with a particular concrete concept should be predictable, based on the concept's semantic properties that are encoded by these brain subsystems.

Of course, a sentence context is very likely to modulate the neural representation of its component concepts, but the neural modulatory principles by which such context effects operate have not been determined. We propose a hypothesis and a method to decode the multiple concepts embedded in a sentence context from a particular form of their fMRI activation signature. One key to the method is to base the estimate of a concept's neural signature on its instantiation in a set of roughly similar sentence contexts (excluding the sentence whose activation is being predicted). This approach assumes that the neural signature of a component concept in context is modulated by the context, and thus differs systematically from the signature of the word when it is processed in isolation, an assumption that we tested. The estimate of the neural signature of a concept was obtained by averaging the activation patterns of several different sentences containing the concept, on the assumption that the neural signals contributed by the other concepts in the sentences would be cancelled out by the averaging. The implication underlying this method is that the neural representation of a simple proto-sentence can be predicted by the sum of the neural representations of its content word concepts, as estimated from approximately similar contextual environments.

Materials and methods

Materials

Each of the 36 simplified three-word sentences (of the form *Agent-verb-object*) described a scenario associated with a general theme pertaining to one of three semantic factors *shelter, manipulation, or eating*, previously shown to underlie the representations of concrete nouns (Just et al., 2010). The objects were selected from among the 60 concrete objects whose associated activation patterns were factor analyzed in a previous study (Just et al., 2010); the selected objects were those with some of the highest factor scores for their dominant semantic factor. The agents and verbs were chosen for consistency with the theme or factor. There were 12 sentences per theme, for a total of 36 sentences, as shown in Table 1. The sentences were composed of triplets of words from among 27 content words: 9 words of each word class (agent, verb, and object). Within each class, 3 words were associated with each neurosemantic factor. The 36 sentences were assigned to four sets, so that each set contained nine sentences, three per theme, and sentences in each set used all 27 words, each occurring in one sentence. Each of the four blocks of trials presented the 36 sentences in a different order.

Participants

Ten healthy, right-handed, native speaking adults (7 females) from the Carnegie Mellon community participated in the fMRI experiment. All participants gave written informed consent approved by the Carnegie Mellon Institutional Review Board. The data from all the participants were included in the analyses below.

Procedure

Participants were asked to read a set of 36 sentences that was presented 4 times. The three words of each sentence were presented one at a time and they accumulated on the screen, with each additional word appearing 500 ms after the onset of the preceding word. After all the words had been displayed (1500 ms after sentence onset), the screen was blanked for 3500 ms, followed by a 7000 ms presentation of the letter "X" in the center of the screen. The total duration of a sentence trial was thus 12 s.

As the participants read each word, they were to think of the intrinsic properties of that concept, and integrate their representation of the concept with the accumulating context of the sentence. After the whole sentence had been presented, they were to finish their thought of the sentence during the blank screen period. Participants were instructed to relax and clear their mind during the fixation interval (indicated by a central "X") that followed the blank screen period. This instruction was intended to evoke a rest state and allow the hemodynamic response associated with the task to decrease toward baseline. The participant was instructed to consistently think of the same properties of a concept across the four presentations. This instruction facilitated the identification of the voxels that had a stable set of responses to the stimuli across multiple presentations. At the beginning of each presentation block and after the last block, the fixation "X" was presented for 17 s, to obtain baseline activation measures. These instructions and general procedures have been used in several previous fMRI studies of semantic decoding (Buchweitz et al., 2012; Damarla and Just, 2013; Just et al., 2014, 2010; Mason and Just, 2016; Shinkareva et al., 2011; Wang et al., 2013).

In the last scan session, participants were asked to read the 27 content words presented in isolation. Each word was presented for 3 s, followed by a 7 s fixation. Participants were instructed to keep thinking about the properties of the concept during the 3 s presentation, and clear their mind when the fixation cross appeared. The 27 words were presented 5 times in 5 blocks of presentations. At the beginning of each presentation block and after the last block, the fixation cross was presented for 17 s, to obtain baseline activation measures.

fMRI imaging protocol

Data were collected using a Siemens Verio 3 T at the Scientific Imaging and Brain Research (SIBR) Center at Carnegie Mellon University. Functional images were acquired using a gradient echo EPI pulse sequence with TR = 1000 ms, TE = 25 ms and a 60° flip angle. Twenty 5-mm thick, AC-PC aligned slices were imaged with a gap of 1 mm between slices. The acquisition matrix was 64 × 64 with 3.125 × 3.125 × 5 mm^3 voxels.

Behavioral ratings

To independently assess the strength of the relation of the 27 content words to each of the three neurosemantic factors, behavioral ratings of the semantic relatedness (on a 1–7 scale) of each of the 27 stimulus words to each of the 3 semantic factors were obtained from 17 participants who did not participate in the fMRI studies. The mean ratings across these participants were then used as the latent semantic features in the modeling. The mean score of *shelter*-related words on the 3 factors: *shelter* = 5.08 (SD = 0.83), *manipulation* = 3.57 (SD = 1.16), *eating* = 2.43 (SD = 0.88). For *manipulation*-related words, the score of *shelter* = 2.40 (SD = 0.80), *manipulation* = 6.27 (SD = 0.59), *eating* = 2.56 (SD = 0.84). For *eating*-related words, the score of *shelter* = 1.52 (SD = 0.52), *manipulation* = 2.84 (SD = 0.69), *eating* = 6.83 (SD = 0.30). The behavioral ratings for all the content words are shown in Table S1.

fMRI data preprocessing

The fMRI data were corrected for slice timing, head motion and linear trend, and were normalized into MNI space with a voxel size of 3.125 × 3.125 × 6 mm^3, using SPM2 (Wellcome Department of Cognitive Neurology, London). The percent signal change (PSC) in signal intensity during each presentation of a sentence was computed at each voxel relative to a baseline activation level measured during fixation intervals and averaged over sixteen such intervals, each one 17 s long (but excluding the first 4 s from the measurement to account for the hemodynamic response delay). The fMRI data from the sentence presentation that were used for analysis consisted of the mean of 5 s of brain images (5 images with a TR of 1 s), the first starting after 6 s from the sentence onset. This same window was optimal for decoding the agent, verb, and object. (This temporal window was determined in an independent pilot study, as described in Supplementary Materials). The PSC of this mean image was then normalized to a mean of 0 and variance of 1 across voxels for each image, to equate the overall intensities of the mean PSC images.

Predicting neural signatures of sentences

The predictions of the neural activity associated with a sentence within a participant were generated using the following steps, depicted schematically in Fig. 1. The model first predicts the brain image for each word (based on the mapping between the word's semantic properties and the activation patterns evoked by those properties in other words) and then combines these predicted images to produce predicted images of a large number of sentences. The predicted sentence images are compared to the observed image, and the model accuracy (based on the ranked similarity between the predicted and observed activation of a sentence) is computed.

The predictions used a cross-validation procedure that partitioned the data into a training set and a test set during the prediction generation and testing of each sentence. The brain images associated with a given sentence, including all four sentence presentations, were left out of the training set in each cross-validation fold. When each of the word classes of agent, verb, or object was modeled, the sentences that contained the word of that class in the test sentence were also excluded from training. For example, when the sentence *Explorer enters car* was tested, the model of agent was trained without using data from sentences containing the word *explorer*, and the model of verb was trained using sentences without *enter*, etc.

1. The 120 voxels distributed in the three semantic factors areas (12 spheres; Fig. 2 and Table S2) with the most stable profiles over presentations for the words in the training set were selected as neural features. A voxel's activation profile is its vector of activation levels evoked by the set of stimulus items, in this case, the images evoked by the set of sentences containing any of the 8 training words within a class. The stability of a voxel was measured by the mean of the pairwise correlations between the pairs of activation profiles across the 4 presentations. In each of the 12 spheres, the 3 most stable voxels in that sphere (175.8 mm^3 per sphere) were first selected. Then the next 84 most stable voxels were selected regardless of which sphere they were in. Such selection criteria obtain contributions from all *a priori* locations and the contribution of voxels with highest overall stability. Stability-based voxel selection is a common practice in fMRI decoding studies (Pereira et al., 2009). The rationale for selecting voxels within the *a priori* clusters, instead of using all the voxels in these regions, was that these spherical clusters represent the approximate locations of the neural substrates that contribute most to the factors. Selecting particular voxels within the spheres counteracts the spatial and coregistration variations across scans and across individuals. The choice of the number of voxels to be

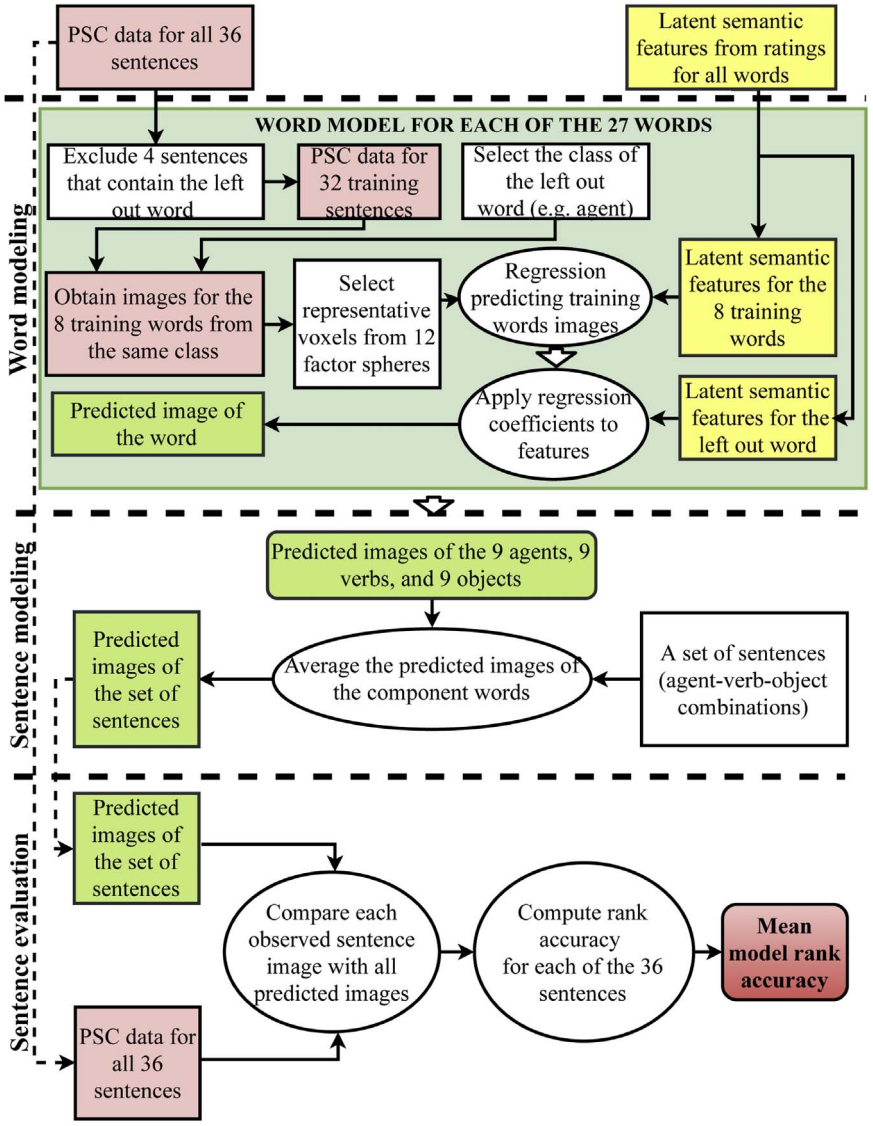

Fig. 1. Block diagram of the processing pipeline for the main analysis.

selected (3 voxels per sphere, 120 voxels in total) was arbitrary. Post hoc analysis showed that the decoding accuracy was robust to the number of voxels being chosen (Table S3).

2. The relations between PSCs at each voxel and the latent semantic features of the 8 training words were mapped using a linear multiple regression model. The semantic properties of each content word on the three dimensions of *shelter, manipulation,* and *eating* were ratings from an independent group of 17 participants. The PSCs associated with a word were estimated by the mean PSCs of training sentences that contain the word. The assumption underlying this

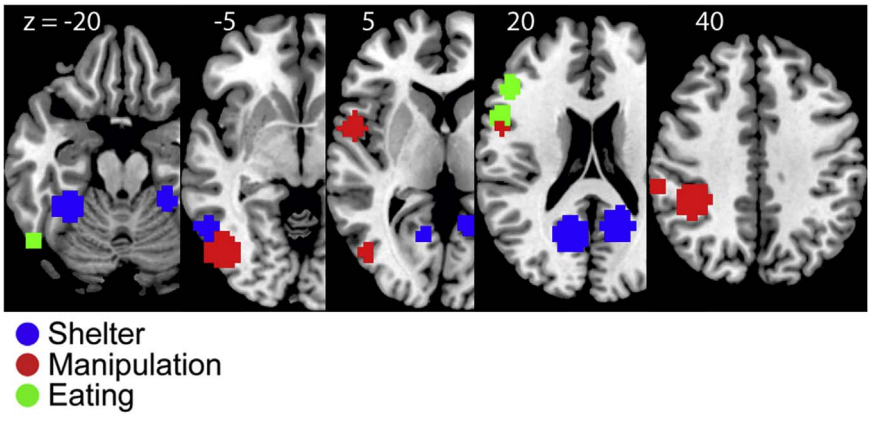

Fig. 2. Twelve brain locations associated with representation of the three factors of *shelter*, *manipulation* and *eating*. Five of the spheres are postulated to code various aspects of the *shelter* factor, four for the *manipulation* factor, and three for *eating*.

procedure was that the neural representation of each sentence contained a representation of its content words; averaging the representation of a set of sentences all containing a given content word would bring out the representation of that shared content word while averaging out the representations of the other words. The weight b_{vi} that specifies the contribution of the i th semantic feature to the activation level is learned by training the regression model and then the predicted images for the three content words in a sentence are combined

$$PSC_v = \sum_{i=1}^{3} b_{vi} f_i$$

where PSC_v is the percent signal change in the representation of sentence at voxel v, and f_i is the ith feature. The trained model should predict the neural signatures of all 9 words using the corresponding semantic features (i.e., each separate model based on 8 words should predict PSC_v for the 9th, left-out word).

3. The predicted activation patterns for a given sentence were generated as follows. Predicted brain images of 351 sensible sentences (out of 729 possible agent-verb-object combinations, excluding nonsense sentences such as *Carpenter chews car*) consisting of one of the 9 agents, one of the 9 verbs, and one of the 9 objects were generated by averaging the predicted images of each of the combinations of words that formed a meaningful sentence. The predicted sentence image consisted of the union of the images of the selected voxels in each word-class model.
4. The accuracy of the sentence prediction was assessed by comparing each predicted image to the observed test sentence image using cosine similarity. The measure of the prediction's accuracy was the normalized rank of its similarity among the 351 (i.e., the normalized rank of the correct label in the posterior-probability-ordered list). The rank accuracy in Fig. 3 was averaged across all cross-validation folds, in which each of the 36 sentences was tested.
5. The statistical significance of a model's accuracy was evaluated by random permutation testing. The model's accuracy was compared with the accuracy of a random classifier with the same design parameters: the same number of classes, cross-validation folds, and items with arbitrary labels. Given the distribution of accuracies of such a random classifier in a large (100,000 or 10,000) number of permutations, the critical levels of accuracy were computed for a given probability level.

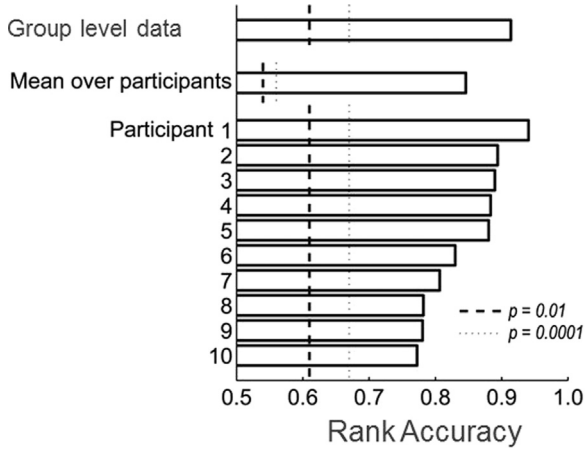

Fig. 3. Mean rank accuracies of sentence prediction at the group level and individual level. The *Group-level* indicates the model accuracy on fMRI images aggregated across participants. The *Mean over participants* indicates the mean model accuracy within individual subjects, i.e., the mean of the next 10 bars. The dashed and dotted lines indicate the critical values of accuracy being significantly different from chance (dashed: $p < 0.01$; dotted: $p < 0.0001$). The critical values of accuracy were determined by 100,000-iteration random permutation tests of the corresponding classification paradigms, taking into account the number of cross-validation folds, number of test items per fold, and number of classes. All the other analyses were performed at the individual participant level unless otherwise specified.

The effects of sentence context on the neural and semantic representations of component concepts

PSCs evoked by the words presented in isolation were computed using the same approach as sentence PSC, except that the images were based on a 4 s window starting after 4 s from the word onset (Pereira et al., 2009). The same approach as described above used these images to predict the neural signature of a sentence, except that the training data were the PSCs of words in isolation. The sentence predictions used the same procedure as in the main analysis. In addition, a non-predictive sentence classification task was performed without modeling the association to the intermediate semantic features latent factors, by directly averaging the word PSCs to estimate sentence activation.

The comparison of activation differences between words-in-context and words-in-isolation was performed within the set of spheres associated with each neurosemantic factor. The activation data of words-in-context were the mean PSC across all the sentences containing the word. The activation levels of the voxels within the set of spheres associated with each neurosemantic factor were averaged. No additional image normalization (aside from setting each image to have mean of 0 and variance of 1) was applied.

Predicting activation patterns of concrete nouns across studies and across participants

In a previous study (Just et al., 2010), 11 participants were scanned while being presented with 60 concrete nouns (in a paradigm similar to the presentation of the words in isolation in the current study). Nine of these nouns were the same as the 9 objects of the action in the current study. Factor analysis of the data revealed the presence of three main semantic factors underpinning the neural representation of nouns naming physical objects, factors labeled as manipulation, shelter, and eating. The interpretation of these factors was supported by the finding that behavioral ratings of the salience of each of the three factors to each of the nouns were highly correlated with the nouns' fMRI-based factor scores. These factors were represented in a total of 12 brain locations. The goal of the analysis here was to predict activation patterns across studies and across participants.

A Gaussian Naïve Bayes classifier was trained on the fMRI data of the 11 participants in the previous study (Just et al., 2010) for the 9 nouns, denoting 9 objects, viewed in isolation. The classifier was then tested on data acquired in the current experiment in which the same 9 nouns appeared in sentence contexts, and the classifier attempted to identify these 9 nouns. In this cross-study classification, using the exact same voxels across participants as features is not a viable approach. Instead, the locations of the features were specified in terms of the 12 factor-related-spheres (where each sphere was characterized by the mean activation of its representative voxels). In the training set data, the most representative voxels were the 10 voxels per sphere with the highest product of their stability scores times the correlation between the voxel's set of mean activation levels for the 9 nouns and the independent ratings of these nouns with respect to the factor associated with the sphere. This definition of representativeness was chosen *a priori*, based on the analyses reported in the previous study (Just et al., 2010). For example, a representative voxel in the premotor sphere would have a stable profile across presentations and would have higher activation levels for the more manipulation-salient (defined by the independent ratings) objects like *pliers* and *hammer*. The 12 features for the model's training set were thus the means of the activation levels of the 10 most representative voxels in each sphere for each of 54 observations (9 nouns presented 6 times each) for each of the 11 participants in the previous study, as well as the 9 labels of the concepts.

The model was then tested on the corresponding test data from each participant in the current study, averaging over the 4 presentations of the nouns. The test brain images from the current study for the nine nouns were obtained by averaging the fMRI images of all of the sentences in the current study that included that noun. The 12 features of the test set were the mean activation levels of the 10 voxels in each sphere that were most stable across 4 presentation blocks. As mentioned above, this approach allowed for cross-scan and cross-individual variation. Note that the stability was measured by using only the nominal labels of the test data without knowing which label was which. No information from outside of the test set entered into the processing of the test set.

The choice of specific parameters (total number of voxels, minimum number of voxels per sphere) did not substantially affect the accuracy of within-participant sentence classification (as shown in Table S3). The specific number (10) of representative voxels per sphere for cross-participant analysis was chosen to equate the total number of voxels in the two main analyses. The statistical significance of model accuracy was evaluated by random permutation testing.

The following is a summary of the main analyses that were performed, which assess the accuracy of the machine learning predictions of sentence activation, and which compare accuracies across different models, including random models.

- The main sentence activation model was applied 1. to group-level data (fMRI data averaged across participants to minimize noise) and 2. to individual participants' data.
- The contributions of semantic features and *a priori* selected brain locations to prediction accuracy were assessed in terms of alternative models that used randomized semantic features and brain locations. The contribution of the voxel (feature) selection procedure was assessed by using voxels outside the factor locations.
- The differences in the neural representations (activation levels in factor locations) between words-in-context and words-in-isolation were assessed by:

  - comparing the prediction accuracies of the models that use the data acquired from words-in-sentences versus words-in-isolation, and also by comparing the similarities between the observed and predicted sentence images for the two models;
  - comparing the neural representations for words-in-3-different-contexts with words-in-isolation in the brain regions associated with each of the three semantic factors.

- The generality of the neural representation of objects is evaluated by a cross-study, cross-context, cross-participant classification of the 9 stimulus nouns.

Results

The prediction of the neural signature of a sentence was based on a computational model of the association between the sum of the brain activation patterns of the contextualized word concepts and the semantic properties of the component concepts of the sentence. The semantic-to-neural mapping was learned by the model at a concept-in-context level. When the model was evaluated at the group level ((by averaging the PSC data of all 10 participants and normalizing each image to a mean of 0 and variance of 1 across voxels, to equate the overall intensities of the images), the neural representations of sentences were accurately predicted with a mean rank accuracy of 0.91 across the 36 sentences. (This accuracy is far above the chance level accuracy of 0.5 ($p < 0.0001$ based on a 100,000-iteration random permutation test; see Table S3 for similar accuracy from models using different numbers of voxels from each location). When the list of alternative sentences was expanded to include all 729 possible agent-verb-object combinations, including nonsense combinations, the rank accuracy of sentence prediction on the group aggregated fMRI data increased from 0.91 to 0.94 (paired-sample *t*-test of the accuracy over sentences, $t_{35} = 3.6$, $p = 0.006$), suggesting the nonsense sentences were not very confusable by the model with the sensible ones. When the list of alternative sentences was restricted to only the 36 stimulus sentences, the rank accuracy was 0.86, significantly lower than among all sensible sentences (paired sample *t*-test over sentences, $t_{35} = 2.3$, $p = 0.03$) but still far above chance level ($p < 0.0001$). Further decoding analyses computed rank accuracy based only on the sensible set of sentences.

To evaluate the model at the individual participant level, the decoding was performed on each participant's data. The mean rank accuracy of the predictions across participants was 0.83. The accuracy of each individual participant was reliably above chance level, with $p < 0.0001$, as shown in Fig. 3 (the analyses below evaluated the models on an individual participant basis, unless otherwise specified). These results show that the neural representation of a sentence that was previously unseen by the model and containing previously unseen words can be predicted by using an additive neurosemantic model of the contextualized content word representations. Moreover, this high level of sentence prediction accuracy was not due to any single word class; the prediction accuracies for the agent, verb, and object classes were 0.81, 0.85, and 0.84 respectively.

Comparison against three alternative models: randomizing semantic features, randomizing brain locations, and selecting voxels outside the factor spheres

The choice of semantic properties and the corresponding *a priori* brain regions, which reflected the core theory underlying the model, were further tested against random models. First, to test whether the choice of semantic properties, i.e., the intermediate variables, contributed to the reliable modeling of the neural signature of semantic representations, a 10,000-iteration random permutation test was performed in which the semantic variables were randomly assigned to each word (i.e. the semantic vectors associated with the different words were interchanged) while all the other aspects of the analysis remained unchanged. This analysis was performed on the data aggregated across participants. The neurosemantic variables performed significantly better than the random variables ($p < 0.0001$). Second, to test whether the choice of *a priori* regions contributed to the reliable modeling of the neural signature, a 10,000-iteration random permutation test was performed on the aggregated data across participants in which the locations of the spheres in the brain were randomly chosen while all the other procedures remained unchanged. The *a priori* neurosemantic regions performed significantly better than randomly selected regions ($p = 0.0001$). These analyses demonstrate the ability of the key components of the theory to account for the data in the modeling of the neural representations of multi-concept sentences. While future models may well outperform the current one, the current model provides a first account that outperforms random models.

An additional exploratory analysis showed that voxels selected from brain regions other than the factor-related spheres also contained information about the sentences, resulting in a mean rank accuracy of sentence prediction of 0.80, which is significantly lower than the accuracy based on factor-related spheres ($p < 0.05$; see Supplementary Materials). This finding suggests that relevant semantic information is also present in brain areas other than the regions that are most closely associated with the factors. That is not to say, however, that the hypothesis of a commonality of semantically organized brain locations in the brain is unnecessary, because (a) the locations of the voxels selected outside the factor-related regions were inconsistent across participants (Fig. S1; Table S4), (b) the *a priori* neurosemantic regions performed better than this exploratory analysis in which the voxels were chosen specifically based on the current dataset, (c) the neurosemantic regions outperformed randomly selected locations, and

(d) it was the intermediate semantic variables that provided the hypothesis of *what to model* given the voxel activations. In addition, the difference of the voxel selection methods between the main analysis and this exploratory analysis empirically and intentionally was not biased toward the former. The main analysis forced the selection of a minimal number of voxels from each of the spheres, while this exploratory analysis did not have such a restriction. The purpose of constraining the distributed voxel selection in the main analysis was to test the contributions from all factor-related locations rather than to produce the highest possible decoding accuracy. When the selection of voxels from factor-related locations was solely based on stability, without the constraint of distributed selection across these locations, the mean rank accuracy remained at 0.83. This result indicates that the accuracy difference between using voxels within and outside the factor-related locations was not due to the constraint of distributed voxel selection.

Differences between the neural representation of words in context versus words in isolation

The main findings reported above entailed the estimation of the neural representations of individual concepts by averaging over the brain images evoked by three sentences that contained that concept. However, the success of this additivity-based modeling does not imply that the neural representation of a sentence is no more than the sum of the activation patterns of isolated words. In our estimation procedure, the neural representations of the words were acquired while the words were being processed in a sentence context. To directly examine this issue, the fMRI data associated with reading individual word concepts were also acquired from the same participants, when each of the concepts was presented in isolation. The comparison of the concept representations obtained in these two ways provided an illuminating insight into the neural and semantic nature of sentence contexts effects.

One way to assess the two types of concept representations is to compare their ability to predict new sentences. We found that the sentence prediction was significantly less accurate based on the activation patterns associated with the reading of single words in isolation. The same approach as the main analysis was used to predict the neural signature of a sentence, except that the estimation of the neural signatures of each word concept was the activation patterns evoked by reading and thinking about that word in isolation. The mean accuracy of sentence prediction based on concept representations obtained in isolation was 0.68 across participants, ranging from 0.60 to 0.81, significantly lower than the mean accuracy of 0.83 based on the representations of component words estimated over other sentences (paired-sample *t*-test, $t_9 = -8.73$, $p < 0.0001$). The relatively poor fit provided by the words in isolation remained poor even when the sentence modeling involved no predictive mapping: when the activation patterns of individual words were directly combined to predict the sentence they constituted, the mean accuracy was 0.74, again significantly lower than using the learned mapping between semantic properties and neural signatures of words in context (paired-sample *t*-test over participants, $t_9 = -3.9$, $p = 0.004$). In addition to the two approaches being compared in terms of prediction accuracies, they were also compared using the (Fisher-transformed) correlation between the predicted and actual sentence images in the two cases. This comparison showed a clear advantage in prediction accuracy of words in sentences. The mean correlation was 0.56 for words-in-sentences and 0.17 for words-in-isolation ($F(1,9) = 188.4$, $p < 0.001$).

Thus, the neural representations of isolated words provide a significantly less accurate prediction of the neural representation of the sentence that contains them, compared to the neural representation of concepts obtained from a sentence context. One possible explanation was that the contexts were similar in the training and testing sentences, and so the contexts of all the training sentences may have modulated the concept representations in a similar way, which was also similar to the modulation effect occurring in the test sentence. The evidence below was consistent with this explanation.

How context selectively modulates the meaning components of a concept

The difference between the activation of a word encountered in a sentence vs. in isolation was then directly assessed separately for each semantic factor and each context type. The activation pattern of a word encountered in sentences was estimated by the mean activation of the sentences that contain the word, on the assumption that the representations of the other words are averaged out. This comparison focused on the activations within the *a priori* cortical regions associated with the three semantic factors, as the decoding models did. Because the content words of each sentence had been selected to be associated with one of the factors, the theme of the sentences was thus expected to be strongly associated with that factor. For example, the words *picnicker, bite,* and *carrot* all have various meaning components, but the sentence they formed, *Picnicker bites carrot*, conveyed a salient theme of *eating*, the meaning component shared across the concepts. Such a context might increase the activation of a component word in the brain regions that represent the theme, relative to the activation of the same word when presented in isolation. In the example above, because the sentence context highlighted the *eating* property of *carrot*, the neural representation in the eating-related regions was expected to be increased when *carrot* was processed in context as opposed to in isolation.

As hypothesized, the mean of the percent signal change (PSC) of words-in-context was significantly greater than words-in-isolation in the regions that were identified *a priori* as representing the semantic properties emphasized by the context (paired-sample one-tailed $t_9 = 5.9$, $p = 0.0001$). (All p-values in this section are Bonferroni-corrected for the 3 factors). This effect was present for each of the three categories of concepts in the brain locations associated with each semantic factor respectively (as shown in Fig. 4); *shelter*: $t_9 = 8.8$, $p = 0.000015$; *manipulation*: $t_9 = 2.5$, $p = 0.0493$; *eating*: $t_9 = 3.1$, $p = 0.02$. Moreover, words in some categories showed lower activation when they were presented within a context than in isolation in regions

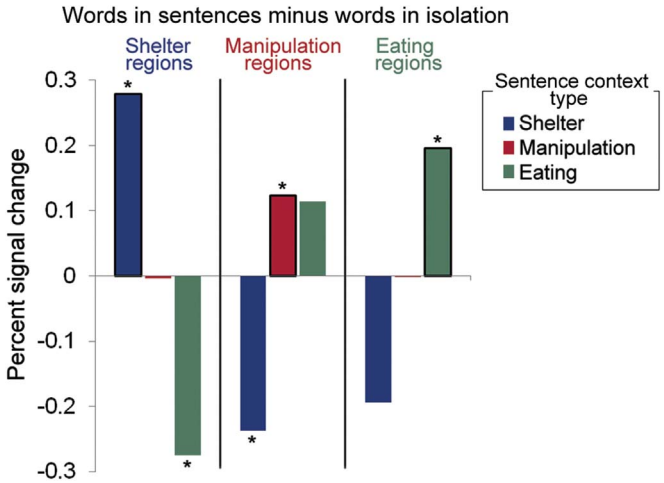

Fig. 4. Differences between the activation levels of words in sentences (obtained by averaging the activation of all sentences that contain the word) and the activation of words presented in isolation are plotted by the three semantic categories of contexts (indicated by color) in three types of regions (indicated within vertical panels). The activation levels were averaged across voxels in the regions associated with each semantic factor. The mean percent signal change (PSC) of words in context was significantly greater than the PSC of words in isolation in the home regions of the concepts for each semantic category (indicated by asterisks on the upward-directed bars). The mean percent signal change of words in context was also significantly lower than words in isolation in the some of the non-home regions of some concepts.

specialized for factors other than their home factor, such as the eating-related concepts in *shelter* regions ($t_9 = -7.14$, $p = 0.0001$, paired-sample two-tailed *t*-test), suggesting the representation of the *shelter* component in a word (e.g., a *diner* may sit at table in an indoor restaurant) was suppressed when the sentence (e.g., *Diner bites carrot*) concerned eating and defocused the spatial setting. Similarly, *shelter*-related concepts in *manipulation* regions showed less activation when presented in context ($t_9 = -7.15$, $p = 0.0001$). There was also a non-significant increase for *eating*-related concepts in *manipulation* regions ($t_9 = 2.76$, $p = 0.13$), suggesting that eating contexts may highlight manipulation as well as eating.

Overall, these results indicate how context modulates the neural representation of a word concept: context highlights the context-relevant semantic component and downplays the irrelevant components. *Highlight* and *downplay* are neurally realized as modulation of the activation levels of the relevant components of the representation. The neural signatures of context-derived concepts are dilated with respect to the semantic factor that is shared by the other words in the sentence, as proposed by a recent computational account of context effects (Mitchell and Lapata, 2010). More generally, this finding indicates the neural nature of semantically-tuned context effects. According to this account, the context selectively elevates and enhances the neural representation of the context-relevant component of the concept's meaning, thus making it fit better into a similar predicted context.

Predicting activation patterns of concrete nouns across studies

The generality of the findings was tested by decoding the word concepts across two independent studies with two different sets of participants. In the previous fMRI study, 11 participants read the 60 nouns, including the 9 object concepts in the current study, presented in isolation (Just et al., 2010). A Gaussian Naïve Bayes classifier was trained on data from the participants in the previous study for the 9 nouns, and tested on data of these nouns from each participant in the current experiment, using only the 12 critical brain locations associated with the three underlying factors. In the previous study, each of three factors considered alone (corresponding to 3–5 of the 12 spheres) made a similar contribution to the classification of the 60 concrete nouns.

This cross-study, cross-context, cross-participant classification of the 9 nouns resulted in a mean rank accuracy of 0.76 (range over participants = 0.56 – 0.85). The accuracy was reliably above chance level for 8 out of the 10 participants in the current study; the significance level was $p < 0.01$ for 6 of these participants and $p < 0.05$ for 2 of them. Despite the vast difference in experimental protocols and the fact that contexts modulate the neural representation of concepts, the neural signature of word concepts was identifiable by a pattern that resides in brain areas defined *a priori* using a theoretically-grounded analysis and independent fMRI training data obtained from independent participants. This result speaks to the existence of the generalizability of the neural representations of concepts in the *a priori* specified brain regions across contexts, people, and studies.

Discussion

The main contributions of this study include (1) a generalizable mapping that incorporates neurally driven and semantically interpretable properties of concepts in sentences, (2) the finding that words presented in context characterize neural representations of concepts in sentences with a similar context better than words presented in isolation, and this seems to be due to (3) the finding that the semantic context neurally highlights the context-relevant semantic component of a concept and down-modulates the irrelevant component of a concept.

The capability of predicting the neural representations of multiple concepts in a sentence, without using prior information about the activation evoked by its content words, is general across word classes and participants. The accurate predictions are based on learning the activation patterns within a small set of brain locations that are independently identified as encoding the neurosemantic factors underlying object (concrete noun) representations. The activation patterns are modeled solely by a set of three intermediate features that encode the semantic relatedness of a word concept to the three factors. Most strikingly, this predictive theory generalizes beyond the concrete objects to transitive verbs and nouns referring to occupations which have never previously been decoded. Although the neural representation of concrete objects has been characterized with increasing granularity over the past fifteen years in various modalities of stimuli at various processing levels or across species (Carlson et al., 2003; Eger et al., 2008; Haxby et al., 2001; Huth et al., 2012; Kriegeskorte et al., 2008; Mitchell et al., 2008; Reddy et al., 2010; Shinkareva et al., 2011), the characterization of neural representations of individual action verbs or profession nouns has been meager (Kemmerer et al., 2008). These representations may have been enriched in particular ways in the current study by virtue of their roles in the sentences. This contextualization may have contributed to the accurate prediction of the neural representations of verbs and profession nouns, reflecting that at least part of the meaning of a word is its use in language (Wittgenstein, 1953).

Another interesting finding was that a word's neural signature as it occurred in a sentence was better predicted by that word's signature in another similar sentence context than by that word's signature when it occurred in isolation. Many previous studies have made excellent progress in characterizing the rules of semantic composition, by operating in abstract vector or matrix spaces that are derived from the use of words, i.e., text corpora (Baroni and Zamparelli, 2010; Coecke et al., 2010; Kintsch, 2001; Mitchell and Lapata, 2010; Smolensky, 1990; Socher et al., 2012). The current model, operating on fMRI images of sentences, utilized the observed neural data that contained the context information and applied a simple additive model, to deconstruct contextualized concepts and rebuild representations of new sentences.

According to this account, a sentence context alters the neural representation of a component of a concept, and that altered representation provides a better prediction of that concept's instantiation in a sentence that has a similar context. This phenomenon may be a special case of semantic priming. Very many behavioral studies have shown that the time to process a word (say, in a lexical decision task) is decreased when it is preceded by a related word (e.g. *dog* is processed faster if it is preceded by *cow* than by an unrelated word), and this is referred to as a semantic priming effect. But fMRI studies have shown that the semantic priming phenomenon is more complex than a simple overall priming effect. The priming effect can be an increase or a decrease in the brain activation, depending on the brain region and the nature of the semantic relation (associative (e.g. *key - chain*) versus categorical) (Kotz et al., 2002; Rissman et al., 2003; Rossell et al., 2003; Sachs et al., 2011; Vuilleumier et al., 2002). In the current study, the effect of a sentence context was a reliable increase in the activation of only that part of the concept's representation that pertains to the context. One can refer to this as a special case of a semantic priming effect, distinguished by its content-selective effect.

Although the current study does not rule out the possibility that the context effect could have arisen due to the simple juxtaposition of the concepts in each sentence (as opposed to additionally being due to the thematic structure of the sentence), previous studies have shown that the same set of words arranged in different thematically structured sentences are neurally differentiable (Frankland and Greene, 2015). The dog in *Dog chases cat* is neurally distinguishable from the dog in *Cat chases dog*, indicating that the neural encoding is more than just the sum of the individual concept representations. In another study using the same approach as the current one, a classifier was able to reliably discriminate *Monkey pats rabbit* and *Rabbit pats monkey* and

to correctly assign thematic roles to the two animal concepts (Wang et al., in press). Thus, the thematic role encoding could be part of a structured sentence representation that plays a role in determining how the components of a concept are selectively modulated.

This study used sentences highly associated with the semantic factors of *shelter*, *manipulation* and *eating* to investigate multi-concept representations, so the generalization of the findings to a much larger semantic space awaits further research. Although these three factors can also characterize other frequent and typical concrete concepts such as animals (Bauer and Just, 2015), there are surely other semantic dimensions that may account for the neural representations of concrete concepts to a comparable or greater degree, such animacy (Capitani et al., 2003) or intentionality. A related limitation of the current study is that the three dimensions are not sufficient to characterize the difference between the concepts from the same category. Because the granularity of the model was constrained by the intermediate variables (semantic properties), when two concepts received very similar ratings on all the three dimensions, such as *explorer* and *hiker* (r = 0.986), the model was not able to capture the neural representational difference between these concepts. Future studies are needed to expand and deepen the model to apply to a broader and richer semantic space.

Perhaps the greatest significance of this research is its potential for providing an approach for neurally characterizing not just isolated concepts but more complex thoughts composed of multiple concepts. The findings here demonstrate the capability of identifying the components of a simple proposition from its fMRI signature. As these lines of research make further advances, the neural structure of thoughts of increasing complexity may become understood as they never have before.

Funding

This work was supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Air Force Research Laboratory (AFRL) (contract number FA8650-13-C-7360), and by the National Institute of Mental Health (Grant MH029617).

Author contributions

M.A.J., V.C. and J.W. designed the research and performed the experiment. J.W. and V.C. analyzed the data. J.W. and M.A.J. wrote the manuscript.

Acknowledgments

We are grateful to N. Diana for outstanding technical assistance.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.neuroimage.2017.06.033.

References

Anderson, A.J., Bruni, E., Lopopolo, A., Poesio, M., Baroni, M., 2015. Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text. NeuroImage 120, 309–322. http://dx.doi.org/10.1016/j.neuroimage.2015.06.093.

Anderson, A.J., Murphy, B., Poesio, M., 2014. Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. J. Cogn. Neurosci. 26, 658–681. http://dx.doi.org/10.1162/jocn_a_00508.

Baroni, M., Zamparelli, R., 2010. Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space, pp. 1183–1193.

Bauer, A.J., Just, M.A., 2015. Monitoring the growth of the neural representations of new animal concepts. Hum. Brain Mapp. 36, 3213–3226. http://dx.doi.org/10.1002/hbm.22842.

Buchweitz, A., Shinkareva, S.V., Mason, R.A., Mitchell, T.M., Just, M.A., 2012. Identifying bilingual semantic neural representations across languages. Brain Lang. 120, 282–289. http://dx.doi.org/10.1016/j.bandl.2011.09.003.

Capitani, E., Laiacona, M., Mahon, B., Caramazza, A., 2003. What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. Cogn. Neuropsychol. 20, 213–261. http://dx.doi.org/10.1080/02643290244000266.

Carlson, T.A., Schrater, P., He, S., 2003. Patterns of activity in the categorical representations of objects. J. Cogn. Neurosci. 15, 704–717. http://dx.doi.org/10.1162/jocn.2003.15.5.704.

Coecke, B., Sadrzadeh, M., Clark, S., 2010. Mathematical foundations for a compositional distributional model of meaning. arXiv:1003.4394 [cs.CL].

Connolly, A.C., Guntupalli, J.S., Gors, J., Hanke, M., Halchenko, Y.O., Wu, Y.-C., Abdi, H., Haxby, J.V., 2012. The representation of biological classes in the human brain. J. Neurosci. 32, 2608–2618. http://dx.doi.org/10.1523/JNEUROSCI.5547-11.2012.

Cowen, A.S., Chun, M.M., Kuhl, B.A., 2014. Neural portraits of perception: reconstructing face images from evoked brain activity. NeuroImage 94, 12–22. http://dx.doi.org/10.1016/j.neuroimage.2014.03.018.

Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19, 261–270. http://dx.doi.org/10.1016/S1053-8119(03)00049-1.

Damarla, S.R., Just, M.A., 2013. Decoding the representation of numerical values from brain activation patterns. Hum. Brain Mapp. 34, 2624–2634. http://dx.doi.org/10.1002/hbm.22087.

Eger, E., Ashburner, J., Haynes, J.-D., Dolan, R.J., Rees, G., 2008. fMRI activity patterns in human LOC carry information about object exemplars within category. J. Cogn. Neurosci. 20, 356–370. http://dx.doi.org/10.1162/jocn.2008.20019.

Frankland, S.M., Greene, J.D., 2015. An architecture for encoding sentence meaning in left mid-superior temporal cortex. PNAS 112, 11732–11737. http://dx.doi.org/10.1073/pnas.1421236112.

Ghio, M., Vaghi, M.M.S., Perani, D., Tettamanti, M., 2016. Decoding the neural representation of fine-grained conceptual categories. NeuroImage 132, 93–103. http://dx.doi.org/10.1016/j.neuroimage.2016.02.009.

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, 156–166. http://dx.doi.org/10.1016/j.neuroimage.2004.05.020.

Hauk, O., Johnsrude, I., Pulvermu, F., 2004. Somatotopic representation of action words in human motor and premotor cortex. Neuron 41, 301–307.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430. http://dx.doi.org/10.1126/science.1063736.

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458.

Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76, 1210–1224. http://dx.doi.org/10.1016/j.neuron.2012.10.014.

Ishai, A., Ungerleider, L.G., Martin, A., Schouten, J.L., Haxby, J.V., 1999. Distributed representation of objects in the human ventral visual pathway. Proc. Natl. Acad. Sci. USA 96, 9379–9384. http://dx.doi.org/10.1073/pnas.96.16.9379.

Just, M.A., Cherkassky, V.L., Aryal, S., Mitchell, T.M., 2010. A neurosemantic theory of concrete noun representation based on the underlying brain codes. PLoS One 5. http://dx.doi.org/10.1371/journal.pone.0008622.

Just, M.A., Cherkassky, V.L., Buchweitz, A., Keller, T.A., Mitchell, T.M., 2014. Identifying autism from neural representations of social interactions: neurocognitive markers of autism. PLoS One 9, e113879. http://dx.doi.org/10.1371/journal.pone.0113879.

Kemmerer, D., Castillo, J.G., Talavage, T., Patterson, S., Wiley, C., 2008. Neuroanatomical distribution of five semantic components of verbs: evidence from fMRI. Brain Lang. 107, 16–43. http://dx.doi.org/10.1016/j.bandl.2007.09.003.

Kintsch, W., 2001. Predication. Cogn. Sci. 25, 173–202.

Kotz, S.A., Cappa, S.F., von Cramon, D.Y., Friederici, A.D., 2002. Modulation of the lexical–semantic network by auditory semantic priming: an event-related functional MRI study. NeuroImage 17, 1761–1772. http://dx.doi.org/10.1006/nimg.2002.1316.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P.A., 2008. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60, 1126–1141. http://dx.doi.org/10.1016/j.neuron.2008.10.043.

Mason, R.A., Just, M.A., 2016. Neural representations of physics concepts. Psychol. Sci. 27, 904–913. http://dx.doi.org/10.1177/0956797616641941.

Mitchell, J., Lapata, M., 2010. Composition in distributional models of semantics. Cogn. Sci. 34, 1388–1429. http://dx.doi.org/10.1111/j.1551-6709.2010.01106.x.

Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M., Newman, S., 2004. Learning to decode cognitive states from brain images. Mach. Learn. 57, 145–175. http://dx.doi.org/10.1023/B:MACH.0000035475.85309.1b.

Mitchell, T.M., 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. http://dx.doi.org/10.1126/science.1152876.

Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., Gallant, J.L., 2009. Bayesian reconstruction of natural images from human brain activity. Neuron 63, 902–915. http://dx.doi.org/10.1016/j.neuron.2009.09.006.

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L., 2011. Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. 21, 1641–1646. http://dx.doi.org/10.1016/j.cub.2011.08.031.

O'Toole, A.J., Jiang, F., Abdi, H., Haxby, J.V., 2005. Partially distributed representations

of objects and faces in ventral temporal cortex. J. Cogn. Neurosci. 17, 580–590. http://dx.doi.org/10.1162/0898929053467550.
Peelen, M.V., Caramazza, A., 2012. Conceptual object representations in human anterior temporal cortex. J. Neurosci. 32, 15728–15736. http://dx.doi.org/10.1523/JNEUROSCI.1953-12.2012.
Pereira, F., Detre, G., Botvinick, M., 2011. Generating text from functional brain images. Front. Hum. Neurosci. 5, 72. http://dx.doi.org/10.3389/fnhum.2011.00072.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45, S199–S209. http://dx.doi.org/10.1016/j.neuroimage.2008.11.007.
Polyn, S.M., Natu, V.S., Cohen, J.D., Norman, K.A., 2005. Category-specific cortical activity precedes retrieval during memory search. Science 310, 1963–1966. http://dx.doi.org/10.1126/science.1117645.
Pulvermüller, F., Hauk, O., Nikulin, V.V., Ilmoniemi, R.J., 2005. Functional links between motor and language systems. Eur. J. Neurosci. 21, 793–797. http://dx.doi.org/10.1111/j.1460-9568.2005.03900.x.
Reddy, L., Tsuchiya, N., Serre, T., 2010. Reading the mind's eye: Decoding category information during mental imagery. NeuroImage 50, 818–825. http://dx.doi.org/10.1016/j.neuroimage.2009.11.084.
Rissman, J., Eliassen, J.C., Blumstein, S.E., 2003. An event-related fMRI investigation of implicit semantic priming. J. Cogn. Neurosci. 15, 1160–1175. http://dx.doi.org/10.1162/089892903322598120.
Rossell, S.L., Price, C.J., Nobre, A.C., 2003. The anatomy and time course of semantic priming investigated by fMRI and ERPs. Neuropsychologia 41, 550–564. http://dx.doi.org/10.1016/S0028-3932(02)00181-1.
Sachs, O., Weis, S., Zellagui, N., Sass, K., Huber, W., Zvyagintsev, M., Mathiak, K., Kircher, T., 2011. How different types of conceptual relations modulate brain activation during semantic priming. J. Cogn. Neurosci. 23, 1263–1273. http://dx.doi.org/10.1162/jocn.2010.21483.
Shinkareva, S.V., Malave, V.L., Mason, R.A., Mitchell, T.M., Just, M.A., 2011. Commonality of neural representations of words and pictures. NeuroImage 54, 2418–2425. http://dx.doi.org/10.1016/j.neuroimage.2010.10.042.
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, e1394. http://dx.doi.org/10.1371/journal.pone.0001394.
Smolensky, P., 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. Artif. Intell. 46, 159–216.
Socher, R., Huval, B., Manning, C.D., Ng, A.Y., 2012. Semantic compositionality through recursive matrix-vector spaces, pp. 1201–1211.
Vuilleumier, P., Henson, R.N., Driver, J., Dolan, R.J., 2002. Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. Nat. Neurosci. 5, 491–499. http://dx.doi.org/10.1038/nn839.
Wang, J., Cherkassky, V.L., Just, M.A., 2017. Characterizing the neural content of complex thoughts: Computational modeling of brain representations of events and states, Human Brain Mapping (in press).
Wang, J., Baucom, L.B., Shinkareva, S.V., 2013. Decoding abstract and concrete concept representations based on single-trial fMRI data. Hum. Brain Mapp. 34, 1133–1147. http://dx.doi.org/10.1002/hbm.21498.
Wang, J., Cherkassky, V.L., Yang, Y., Chang, K.K., Vargas, R., Diana, N., Just, M.A., 2016. Identifying thematic roles from neural representations measured by functional magnetic resonance imaging. Cogn. Neuropsychol., 1–8. http://dx.doi.org/10.1080/02643294.2016.1182480.
Wittgenstein, L., 1953. Philosophical Investigations. Blackwell, Oxford.

Supplementary Materials

Optimal temporal window for decoding component word concepts that are part of a sentence

Many previous studies (e.g. Just et al., 2010) have found that the optimal temporal window in the fMRI signal for decoding words presented in isolation is 2-5 s after the word's presentation *offset*. To find the optimal temporal window for analysis of the main current study in which the three words in a sentence are presented in quick succession, a pilot study was run and several temporal windows were explored. The accuracy of decoding the concepts in each serial position (agents, verbs, and objects) was computed using a narrow temporal window containing only 2 s of images, but starting at various times relative to the stimulus onset. The results showed that agents, verbs, and objects were all most accurately predicted when the time window consisted of 5 images starting at 6 s from the sentence onset. This outcome determined the choice of temporal window used in analyzing the fMRI data associated with the words in the stimulus sentences.

Table S1. Behavioral ratings for the 27 content words on three semantic dimensions

Word type	Primary Dimension	Word	Behavioral ratings		
			Shelter	Manipulation	Eating
Agent	Shelter	explorer	3.94	4.06	2.71
		hiker	4.06	3.94	2.82
		tourist	3.76	2.29	5.06
	Manipulation	carpenter	5.24	6.65	1.76
		mechanic	2.65	6.65	1.94
		plumber	3.06	6.35	2.41
	Eating	diner	3.00	3.35	6.88
		glutton	1.53	2.12	6.47
		picnicker	1.94	2.65	6.65
Verb	Shelter	enters	5.65	3.06	1.47
		exits	5.29	3.00	1.24
		repairs	4.18	6.65	1.18
	Manipulation	drops	1.53	4.76	2.24
		grabs	1.59	6.18	3.53
		lifts	1.76	6.53	3.06
	Eating	bites	1.18	3.35	6.65
		chews	1.41	3.65	6.82
		tastes	1.24	2.00	7.00
Object	Shelter	car	5.29	4.47	2.06
		church	6.53	1.88	2.06
		house	7.00	2.76	3.29
	Manipulation	hammer	2.71	6.47	1.24
		knife	1.41	6.35	5.65
		pliers	1.65	6.53	1.18
	Eating	carrot	1.12	2.88	7.00
		celery	1.12	3.24	7.00
		tomato	1.12	2.29	7.00

Behavioral ratings (on a scale of 1-7) are the averaged ratings of the salience of the content words to three semantic dimensions, obtained from an independent group of 17 participants.

Table S2. Locations (MNI centroid coordinates and radii) of the spheres associated with the three semantic factors

Factor	**Sphere Location**	**x**	**y**	**z**	**Radius (mm)**
Shelter	L Fusiform / Parahippocampal Gyrus (PPA)	-32	-42	-18	6
	R Fusiform / Parahippocampal Gyrus (PPA)	26	-38	-20	4
	L Precuneus	-12	-60	16	8
	R Precuneus	16	-54	14	8
	L Inf Temporal Gyrus	-56	-56	-8	4
Manipulation	L Supramarginal Gyrus	-60	-30	34	10
	L Postcentral/Supramarginal Gyrus	-38	-40	48	12
	L Precentral Gyrus	-54	4	10	6
	L Inf Temporal Gyrus	-46	-70	-4	8
Eating	L Inf Frontal Gyrus	-54	10	18	8
	L Mid/Inf Frontal Gyri	-48	28	18	6
	L Inf Temporal Gyrus	-52	-62	-14	4

Table S3. Rank accuracy of decoding sentences when different numbers of voxels were selected. The main analysis reported in the main text selected at least 3 voxels per region and 120 voxels in total.

		Mandatory minimal number of voxels selected per sphere					
		0	3	6	10	12	20
Total number of voxels being selected	50	0.88	0.90	NA	NA	NA	NA
	100	0.90	0.91	0.90	NA	NA	NA
	120	0.91	0.91*	0.90	0.90	NA	NA
	150	0.90	0.91	0.91	0.90	0.90	NA
	200	0.91	0.91	0.91	0.91	0.91	NA
	500	0.90	0.90	0.90	0.90	0.90	0.90

* Main analysis

Table S4. Frequency of a voxel being selected in multiple participants when only voxels *outside* the factor-related spheres were considered. The largest commonality was in one voxel that was selected in 6 participants. The voxel selection procedure was the same as in the decoding analysis, except that no cross-validation was implemented.

Frequency	Number of voxels
1	1385
2	179
3	53
4	13
5	3
6	1

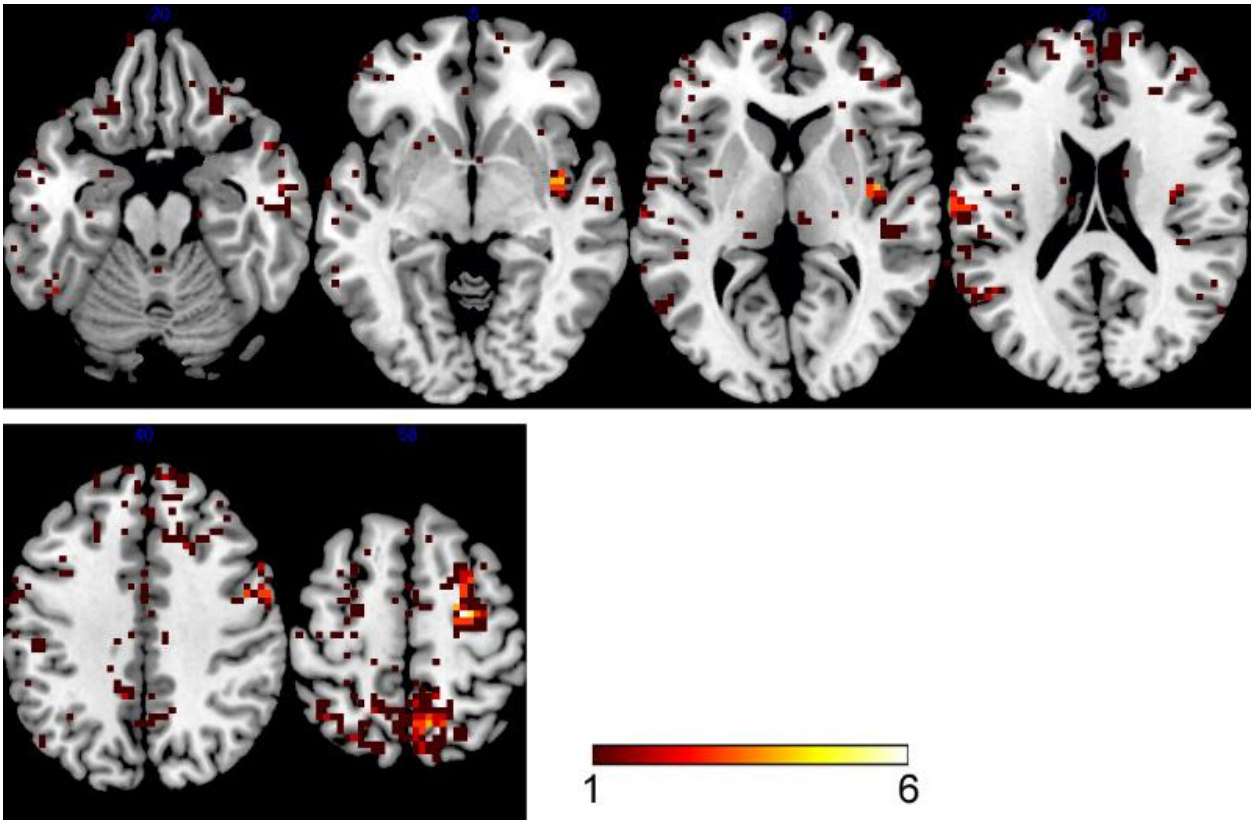

Figure S1. Locations of selected voxels when only locations outside the factor-related spheres were considered. Red, yellow, and white voxels were selected respectively for one, two, and three participants. Color map indicates frequencies of a voxel being selected over 10 participants.