Predicting the brain activation pattern associated with the propositional content of a sentence: modeling neural representations of events and states

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

[a] Center for Cognitive Brain Imaging, Psychology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

[1] To whom correspondence should be addressed. email: just@cmu.edu

Key words: fMRI, multi-concept propositions, predictive modeling, neural representations, concept representations

Abstract

Even though much has recently been learned about the neural representation of individual concepts and categories, neuroimaging research is only beginning to reveal how more complex thoughts, such as event and state descriptions, are neurally represented. We present a predictive computational theory of the neural representations of individual events and states as they are described in 240 sentences. Regression models were trained to determine the mapping between 42 neurally plausible semantic features (NPSFs) and thematic roles of the concepts of a proposition and the fMRI activation patterns of various cortical regions that process different types of information. Given a semantic characterization of the content of a sentence that is new to the model, the model can reliably predict the resulting neural signature, or, given an observed neural signature of a new sentence, the model can predict its semantic content. The models were also reliably generalizable across participants. This computational model provides an account of the brain representation of a complex yet fundamental unit of thought, namely the conceptual content of a proposition.

In addition to characterizing a sentence representation at the level of the semantic and thematic features of its component concepts, factor analysis was used to develop a higher level characterization of a sentence, specifying the general type of event representation that the sentence evokes (e.g. a social interaction versus a change of physical state), as well as the voxel locations most strongly associated with each of the factors.

Introduction

Although individual concepts may be the fundamental elements of thought, the generativity and complexity of human thought stem from the ability to combine *multiple* concepts into propositions. Functional neuroimaging approaches have provided remarkable insights concerning the neural representations of individual concepts [Anderson et al., 2014; Bauer and Just, 2015; Coutanche and Thompson-Schill, 2015; Ghio et al., 2016; Huth et al., 2012; Just et al., 2010; Mason and Just, 2016; Mitchell et al., 2008; Peelen and Caramazza, 2012; Pereira et al., 2011; Shinkareva et al., 2011; Wang et al., 2013] and segments of stories [Huth et al., 2016; Wehbe et al., 2014], inter-concept relations [Frankland and Greene, 2015; Wang et al., 2016] or combined concepts [Baron and Osherson, 2011]. But characterizing the neural representations of more complex thoughts, such as event and state descriptions, has remained a considerable challenge.

This paper develops a computational account of the mapping between the concepts of a proposition describing an event or state and the neural representation that it evokes. The major advance of the current study is to characterize the neural representation of events and states as they are described by sentences, by developing a predictive, bi-directional mapping between the conceptual content of a sentence and the corresponding brain activation pattern. In this article, we describe a computational model that can predict the neural representation of 240 different propositions that describe events or states with reasonable accuracy and with reliable capture of the gist of the proposition.

The initial approach taken here is that the neural representation of a sentence is composed of the neural representations of its component concepts and the roles those concepts play in the thematic structure of the sentence. Each word concept in a sentence can be characterized in terms

of a set of neurally plausible semantic features (NPSFs) and in terms of the concepts' thematic role in the sentence or event. The NPSFs are intended to correspond to various brain subsystems, such as the motor system representation of how one interacts with a concrete object, or the perceptual system's representation of the perceptual properties of the object. The main analytic technique that is used here to determine the mapping between semantic features and neural activation patterns is multiple regression modeling. Once the mapping is determined in a large subset of the data, it can then be used to make predictions in the remaining independent subset.

The neural representation of the concepts of a sentence or proposition must be expressed within the brain's organizational system or ontology that categorizes and relates a large number of different concepts. We make an initial attempt at specifying this neurally-based organizational system, by developing 42 NPSFs that provide a starting point for a neurally-motivated feature representational system. These features are considered neurally plausible based on previous findings concerning the neural representations of various types of semantic knowledge. For example, some subsets of these features code perceptual [Kanwisher et al., 1997; Martin, 2007; Oliver and Thompson-Schill, 2003], motor [Hauk et al., 2004; Pulvermüller et al., 2005], and affect-related [Baucom et al., 2012; Chikazoe et al., 2014; Kassam et al., 2013] features of an entity; another subset codes different aspects of human activity in society [Iacoboni et al., 2004; Spitzer et al., 2007]; and another subset codes properties pertaining to time or space [Fernandino et al., 2016; Kranjec and Chatterjee, 2010; Lai and Desai, 2016].

The current study develops a model that relates such semantic features of concepts to the activation level of pre-specified clusters of voxels. The model extends the general approach to all of the concepts in a proposition, and also takes into account the thematic role of each concept. Given the semantic and thematic characterization of the component word concepts of a

proposition, the model can predict the activation pattern the reading of the corresponding sentence will evoke.

The 42 NPSFs were coded in binary form, as applicable or not applicable to each concept. The types of information the NPSFs coded included the perceptual and affective characteristics of an entity (10 NPSFs coded such features, such as man-made, size, color, temperature, positive affective valence, high affective arousal), animate beings (person, human-group, animal), and time and space properties (e.g. unenclosed setting, change of location). For example, a concept such as the noun *judge* was coded with the following NPSFs: person, social norms, knowledge, communication. The general assumption was that individual NPSFs could be related to activation levels in particular clusters of voxels.

In addition to relating individual concepts to the activation patterns they evoked, the model assumes that the neural signatures of the concepts in a sentence are sensitive to the roles they play in the proposition's argument structure. The current model incorporates a thematic role component [Fillmore, 1967] to account for context-dependent, propositional-level semantics. Thus each content word is also coded in terms of its thematic role in the sentence such as agent or patient.

The predictive ability of these NPSFs and thematic features can be compared with other accounts provided by semantic vector representations of concepts [Fellbaum, 1998; Landauer and Dumais, 1997; Niles and Pease, 2001] that don't take neural organization into account. Semantic vector representations are based on a word's distributional properties in a large language corpus, indicating a word's co-occurrence patterns with other words. A word like *judge* might be characterized in terms of its co-occurrence with other items like *jury* or *decide.* Latent Semantic Analysis (LSA) [Landauer and Dumais, 1997] is a prominent example of a semantic

vector representational approach. It is possible to construct a predictive model of the neural representations of concepts and sentences based on the semantic vector representations of the concepts. In the context of the current article, semantic vector representations provide a point of comparison to NPSFs.

In addition to characterizing a sentence representation at the level of the semantic and thematic features of its component concepts, we developed an additional characterization of a sentence at a coarser level of granularity that specifies the general type of thought that the sentence evokes (e.g. the thought of a social interaction versus a physical change of state). The main analytic technique that was used to determine these coarser dimensions of representation was factor analysis applied to the sentence activation data, to identify the emerging meaning dimensions that characterize sentence representations. This facet of the analysis was data-driven (rather than driven by hypothesis); the resulting clustering of sentences with similar activation patterns indicated that they shared coarse underlying semantic dimensions. This coarser level of analysis plays three roles in the current study. It identifies the 4 main types of meaning dimensions that underlie the set of 240 stimulus sentences. It identifies the main regions of brain activation that correspond to the processing of the 240 sentences. And it indicates the correlation between each of the 42 NPSFs and the 4 main meaning dimensions.

One methodological challenge presented by the study of the neural representations of sentences was to overcome fMRI's convolution of the signals from temporally adjacent stimuli, in this case, the successive words of a sentence. Previous studies of the neural processing of a story modeled the text in terms of semantically arbitrary segments determined by image acquisition segments [Huth et al., 2016; Wehbe et al., 2014]. The current approach of modeling a proposition and its component concepts uses the brain image that occurs at the end of the reading

of a sentence. Our preliminary findings showed that this image contained the neural representation of all of the word concepts in the sentence; thus there was no attempt to segment the observed fMRI signal evoked by a sentence into arbitrary time intervals. To obtain an estimate of the neural representation of an individual word concept, the end-of-sentence brain images of the several sentences that contained the target concept were averaged, on the assumption that the images of other words are averaged out. The resulting estimates of the neural representations of concepts provide a robust basis for developing a mapping between semantic representations and brain activation patterns.

The study thus has two main goals. The main goal was to develop a mapping between a semantic characterization of a sentence (based on the individual word concepts in the sentence and their thematic roles) and the resulting brain activation pattern that occurs when the sentence is read. The mapping can then be used to predict the activation pattern of an entirely new sentence containing new words, simply based on its semantic characterization. Furthermore, the mapping that is developed in the model is bi-directional, such that it is also possible to predict the semantic properties of the word concepts in a new sentence, given the activation pattern that the sentence evokes. Closely related to the main goal was an assessment of the similarity across participants of the mapping between semantic features and neural representations. It is possible that every participant has a systematic mapping between semantic features and activation patterns, but this mapping could be similar or dissimilar across participants.

A second main goal of the study was to characterize the broad semantic dimensions that underlie the activation of sentences that describe simple events and states, obtained using factor analysis of the activation patterns that accompany a large number of sentences. The emerging

factors provide a characterization of the neural organization of semantics at a coarser grain size than individual semantic features.

Methods

Participants

Seven healthy, right-handed, native English-speaking adults (mean age = 25, range = 20 - 35, two males) from the Carnegie Mellon community gave written informed consent approved by the Carnegie Mellon Institutional Review Board.

Stimuli and Procedures

The 240 stimulus sentences, generated by [Glasgow et al., [2016] described primarily events but also some states, with each sentence containing a mean of 3.3 content words (range = 2 - 5) and a total of 242 unique words. Examples of stimulus sentences are shown in Table 1 and the full set of stimulus sentences is presented in Table S1. The component concepts of the sentences were chosen to be representative of general human knowledge (objects, human roles, actions, activities, attributes, places, and time). All sentences were literal, used active voice and past tense. A word appeared on average in 3.3 sentences (range = 1 - 7). There were 45 words that appeared only in a single sentence.

The sentences were presented 4 times each over 4 days of scans. This slow event-related design allowed acquisition of reliable single-trial neural signatures of individual sentences. The stimulus onset asynchrony between sentences was 12 sec (5 sec of sentence reading + 7 sec fixation). The details of the stimulus presentation timing are described in Supporting Information. This general experimental paradigm and choice of parameters has been applied in many studies

on neural representations of various types of concepts [Just et al., 2010; Mason & Just, 2016; Shinkareva et al., 2011; Wang et al., 2013].

Data were collected using a Siemens Verio 3-T MRI scanner at the Scientific Imaging and Brain Research 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 5mm thick, AC-PC aligned slices were imaged with a gap of 1mm between slices. The acquisition matrix was 64 × 64 with 3.125 × 3.125 × 6 mm^3 voxels.

Data preprocessing

Whole cortex fMRI data were corrected for slice timing, head motion, linear trend, low frequency trends (by applying a high-pass temporal filter at 0.005Hz), and normalized into MNI space using SPM8 (Wellcome Department of Cognitive Neurology, London). Further analyses were performed using in-house scripts written in Matlab7 (Mathworks, MA, USA).

In the processing of the fMRI data for each presentation of a sentence, the percent signal change (PSC) relative to a baseline was computed at each voxel in the brain image. The baseline activation level was measured during and averaged over sixteen 17sec fixation conditions. The baseline measurement started at 4 sec after each fixation presentation onset to account for the hemodynamic response delay. The fMRI data to assess sentence reading consisted of the mean of 5 brain images, collected 1sec apart from each other, the first starting at 7sec from sentence onset. The temporal window for the sentence data analysis was chosen based on pilot investigations. The temporal window within which the neural signatures of *all* the content words in a simple sentence, regardless of their serial position in the sentence, was located in time *after* the entire sentence had been read (as shown in Figure S1). The PSC was then normalized to a mean of 0 and variance of 1 across voxels within each image, to equate the overall intensities across scans.

Neurally Plausible Semantic Features

The mapping between the semantics of a sentence's conceptual content and its neural signature is presumed to be mediated by a layer of semantic features, enabling the prediction of the activation of a new sentence based on its semantic content. A set of 42 neurally plausible semantic features (NPSF) was developed to both characterize a semantic property of some of the 242 unique corpus words and to also correspond to a known or plausible neural processing mechanism. These binary features were defined for each of the 242 word concepts in the corpus, where a coding of 1 indicated that the feature was applicable to the word concept. The coding was performed by two raters with disagreements adjudicated by a third rater. Each word was coded by taking into account the meaning of the word in all of the stimulus sentences in which it appeared, and coding it the same way in the various sentences in which it appeared. The NPSF definitions and sample concepts pertaining to each NPSF are shown in Table 2. Table 3 shows the NPSF coding of several example word concepts.

Thematic role variables that characterize the sentence-level role of concepts

In addition to the NPSF feature set, six additional variables were coded indicating the sentence-level role of each concept: main verb, agent or experiencer, patient or recipient, predicate of a copular sentence (*The window was **dusty***), modifier (*The **angry** activist broke the chair*), and complement in adjunct and propositional phrase, including direction, location, and time (*The restaurant was loud at **night***). These variables covered all the roles that occurred in the stimulus sentences. Thus, each word concept in a sentence was coded with respect to not only the applicability of the 42 NPSFs but also with respect to the 6 thematic roles (indicating which one of the 6 roles is applicable). If a given word played a different thematic role in different sentences then its thematic role would be coded differently. The generative model can thus

predict different neural activity patterns for the same words in different roles. This model has the capability of being expanded to incorporate more thematic roles and different types of sentence roles (such as negation).

Factor analysis

To evaluate the data-driven dimensions of neural sentence representation and identify the brain locations associated with these dimensions, factor analysis (FA) methods were applied.

The main FA was performed on data of the 3 participants with the highest preliminary sentence prediction accuracy when using 300 most stable voxels anywhere in the brain. A set of 600 voxels was selected for each participant, based on the stability of their activation profiles over sentences across 4 presentations. Voxel stability refers to responding similarly to the set of items in the training set across multiple presentations of the set, i.e. displaying a repeatable semantic tuning curve. 50-60 voxels were selected from each of bilateral frontal, temporal, parietal, fusiform, and occipital lobes, and 40 voxels were selected from cingulate cortex. A sentence was then characterized by a mean image across 4 presentations in the selected voxels. A two-level hierarchical exploratory factor analysis was applied to the activation data for the 3 participants. The first-level FA was applied to the activation profiles of 600 voxels over the 240 sentences within participants, resulting in 7 factors underlying the voxel activation profiles per participant. The second-level FA was performed on these 21 factors to further aggregate the neural dimensions over participants. The first 5 factors of the second-level FA explained 37% of the variation. These factors formed the group-level characterization of the main dimensions that underlie the activations.

Each of the second-level factors was associated with a set of voxels, by tracing back each second-level factor to the first-level factors with factor loadings greater than |0.4|. For each of the

identified contributing factors, contributing voxels with a factor loading greater than |0.4| were selected. The resulting clusters with a minimum of 5 contributing voxels were identified as the brain locations associated with the second-level factors. To account for the variability of representations across participants, each cluster was approximated by applying an algorithm that grows a cluster by one voxel in all directions.

Model training and prediction of neural activity (activation-predicting model)

Given the semantic content of a sentence, the goal was to predict the neural activation pattern that the reading of the sentence would evoke. The neural activation was limited to the brain locations discovered using FA. The procedure used a cross-validation protocol that leaves one sentence out as the test data in each fold, so that every sentence is tested once after all the iterations of training-test assignment. The training and test sets are always kept independent during the model training and prediction.

The activation data that characterized each sentence reading was the mean percent signal change over a 5-second interval acquired *after* sentence reading (sentences lasted a mean of 1.4 s), starting at 7 s from sentence onset (see Supporting Information for details of presentation).

The normalized PSC images of 240 sentences with 4 presentations were partitioned into training and test set in each cross-validation fold. The data from one test sentence (including all 4 times it was presented) were left out from the model's training set. Using only the data in the training set, brain images corresponding to each content word were constructed by averaging the PSCs of all the training sentences that contain that content word.

The data that were used to train the regression model were the means of the images corresponding to the sentences in the training set that contained the same content words

manifesting the same roles. The voxels that were used for training and testing were the selected stable voxels within the regions identified by a factor analysis (see Supporting Information for specifics of the voxel selection procedure). It is important to note that the factor locations that were used in the modeling of a given participant *always excluded* all of the data from that participant. Specifically, brain locations identified by the FA based on the 3 participants with the highest prediction accuracies were used to perform the sentence decoding of the 4 remaining participants. For decoding of these 3 high-accuracy participants, three additional FAs were performed following the same procedure except for replacing the participant being classified with the fourth best participant's data.

A kernel ridge regression model [Hastie, T., Tibshirani, R. & Friedman, 2009] was trained to map the relations between the activation patterns from selected representative voxels and intermediate semantic features of words. The learned weights specified the contributions of each of the semantic features to the activation of a voxel. The trained models from the selected voxels were then used to predict the neural signatures of all 351 role-specific words, given the corresponding semantic features.

The predicted brain activity pattern associated with each of the 240 sentences was the mean of the activity associated with its component words at certain thematic roles, without any activation information from the target sentence. In each cross-validation fold, all four presentations of the test sentence were left out of the training set, and the left-out data were used for testing. The four presentations of the test sentence were averaged to minimize noise. Only the voxels selected based on the training set were considered in the analysis. The left-out (observed) image of the test sentence was compared to each of the 240 predicted sentence images, in terms

of cosine similarity. The performance of the prediction was assessed by the normalized rank, or rank accuracy, of the similarity of the predicted test sentence in the list of 240 alternatives.

The statistical significance of the prediction accuracy was determined based on the null distribution generated by a 10^5-iteration random permutation of data, so that the number of cross-validation folds, number of test items per fold, and number of classes were controlled.

The methods used for the meaning-predicting model used the same framework and are reported in Supporting Information.

Cross-participant prediction of neural representation of sentences

Inter-participant commonality of sentence representation was assessed by training a model on the mapping from the NPSF features to brain images using the fMRI data from all but one participant, and then predicting the brain activity of the left-out participant. The specific methods applied were the same as in the within-participant activation-predicting model, except for the method of voxel selection. For the data of the participants in the training set, representative voxels were selected based on their stability of activation at the sentence level. The cross-participant stability was defined as follows: for each of the 6 training participants, the response profiles of word concepts at each voxel were averaged across 4 presentations. The cross-participant stability was the mean pairwise profile similaritiy between participants. Only voxels within the pre-specified factor locations identified using data of the three most accurate participants in the training set were considered. The same set of representative voxels was then used to characterize the data of the test participant.

The model's capture of event semantics: behavioral judgment of event similarity

Human judgments of the similarity of pairs of sentences were crowdsourced using Amazon Mechanical Turk (MT, www.mturk.com) in accordance with the Carnegie Mellon Institutional Review Boards. The sentence stimuli were 195 non-copula sentences that described events, omitting the 45 copula sentences (e.g., *The flower was yellow*), to limit the number of pairs to 18915. These pairs of sentences were organized in 195 groups of 97 pairs of sentences, arranged in different groupings for different MT workers.

Because the vast majority of the sentences were semantically dissimilar, each MT worker was instructed to select only 1 pair of sentences describing the most similar pair of events within each group of 97 pairs, resulting in 195 pairs of sentences judged as similar. Among the sentence pairs judged as most similar, there were 18 pairs of sentences that were selected by three out of the four MT workers.

Alternative semantic representations

Linguists and psychologists have proposed several different types of semantic representations of individual concepts that unlike NPSFs, were not intended to be related to brain function. The ability of four vector-space characterizations of the 242 word concepts was compared to the NPSFs' account of the fMRI data. These vector-space representations characterize a word's meaning based on its co-occurrence with other words in large corpus. These four alternatives included GloVe vectors for word representation (300 dimensions) [Pennington et al., 2014], doc VSM and dep VSM (a documents-based and a dependencies-based vector space model respectively, 1000 dimensions each) [Fyshe et al., 2013], and Latent Semantic Analysis (LSA, 308 dimensions) [Landauer and Dumais, 1997]. The vector used to represent a sentence was the mean of the vectors representing the individual word concepts. The

analyses reported below assessed the accuracy of each of these representational approaches in predicting brain activation patterns associated with comprehending a sentence, in comparison with the accuracy of the NPSFs' account.

Results

Predicting neural signatures of new events from their semantic content

Activation prediction. The neural representation evoked by the reading of an individual sentence was reliably predicted based on its semantic and thematic content. Given the word-level semantic (NPSF) and thematic features of the content words of a left-out sentence, the model can predict the activation pattern of a sentence that is new to the model with a mean rank accuracy of 0.86 among 240 alternatives, where chance is 0.50. The analyses associated with this finding are described in more detail below.

The semantic-feature-to-fMRI mapping was learned by training on the data from 239 sentences using a kernel ridge regression model [Hastie, T., Tibshirani, R. & Friedman, 2009], while the image of the 240th sentence was left out for testing the accuracy of the prediction in each cross-validation fold. A mapping from the 42 neurally plausible semantic features and 6 thematic role variables to the activations in selected voxels in pre-identified brain regions was learned using all the concepts in the 239 training sentences. The *a priori* regions were clusters with high factor loadings on any of the main dimensions revealed by factor analysis of the activation data of other participants (as described below). The neural signatures of all of the component content words of the test sentence were predicted using their NPSF values and thematic roles and applying the learned mapping. The predicted neural signature of a sentence was the mean of the predicted signatures of its component words. To assess the model's

prediction accuracy for a given test sentence, the observed activation pattern of the test sentence was compared for similarity to the predicted activation patterns of all 240 sentences (using the cosine measure of similarity). The accuracy measure was the normalized rank of the similarity of the test sentence's predicted activation pattern, among the 240 predictions, to its actual activation pattern. This procedure iterated through 240 cross-validation folds, testing each sentence once. At the group level (where the fMRI images of all 7 participants were averaged), the mean rank accuracy for the 240 sentence predictions was 0.86, as shown in Figure 1, significantly different from chance ($p < 10^{-5}$).

While the rank accuracy provides a measure of how distinct the predicted neural signature of a sentence is from those of other sentences, it is indifferent to the *absolute* similarity between the observed and the model-predicted neural signatures. To assess the absolute similarity, cosine similarity between the actual and predicted signatures was computed over the voxels being modeled. The mean observation-prediction similarity over sentences was 0.42, significantly different from chance ($p < 10^{-5}$). To estimate the ceiling that this similarity measure can reach given the noise in the current data, the similarity between the images of the same sentences as observed in different presentations was computed. The mean image of two presentations was compared with the mean image of the other two presentations by exhausting all possible combinations of presentation pairs when all the other parameters were controlled. This measure of observation-observation similarity results in a mean of 0.58. Thus the mean absolute similarity of 0.42 between the observed and the model-predicted neural signatures of the 240 sentences captures more than two thirds of the maximum possible similarity of 0.58.

When the model was implemented at the level of each individual participant, the mean rank accuracy of the predictions across participants was 0.82 (range = .79 - .84), as shown in Table

S2). Aggregating the fMRI data over participants apparently reduces the noise and provides a more accurate mapping (mean rank accuracy of .86, described above). The remaining analyses were performed at the individual level unless otherwise specified.

Although the model above was trained without any direct information about the activation associated with the test sentence, a word in the test sentence could have occurred in other sentences in the training set. To test the model's predictive ability in the absence of any such information, a second classifier was trained only on sentences that did not contain any of the words in the test sentence (resulting in a meaning training set size of 229 sentences), thus excluding all activation levels associated with any of the word concepts in the test sentence. The resulting accuracy of the sentence prediction decreased only marginally, remaining at a mean rank accuracy of 0.81 over participants (range = 0.78 - 0.83). Thus, the sentence prediction accuracies are primarily based on the mapping between brain activation patterns and semantic and thematic features, rather than on the mapping to observed activation patterns of particular word concepts. In addition, this analysis establishes the model's ability to predict activation for the previously unseen words.

Meaning prediction. The mapping in the other direction, decoding semantic content from activation patterns, yielded a similar outcome. Given an unlabeled fMRI image evoked by the reading of a test sentence, the activation-to-features mapping model was trained on the data from the other 239 sentences and used to predict a vector of the NPSF semantic features of the test sentence. For example, given only the activation associated with the sentence *The witness shouted during the trial*, the model predicts that the sentence contained the following non-zero NPSFs, ordered in terms of their relative weights: negative affective valence, human group, communication, high affective arousal, person, social action, mental action, social norms,

perceptual. The accuracy measure assessed the similarity of the actual (coded) vector of NPSFs to the predicted vector, relative to its similarity to the predicted vectors of all 240 sentences. The group-level and individual-level models produced a rank accuracy of 0.87 and mean rank accuracy of 0.83 (range .80 - .87) respectively, very similar to the activation-predicting model (Table S2). For brevity, further analyses were performed using the activation-predicting model at the individual participant level, unless otherwise specified.

The model's capture of event semantics

The model's confusion errors and the decoding outcome of the meaning-predicting model indicated that it was capturing the semantics of the events described by the sentences. The activation predictions of sentences describing similar types of events, such as *The flood damaged the hospital* and *The storm destroyed the theater,* were similar to each other, as Figure 2 illustrates. The model characterizes similar events similarly even when the specific words and concepts of the sentences are dissimilar.

To systematically assess how well the model captured such event similarities, the model's measure of event similarity was compared to independent human judgments of sentence meaning similarity for the 18915 pairs of sentences formed by only the 195 out of the 240 sentences that described events. The model-based inter-sentence similarity was measured by the mean pairwise similarity of the predicted activation patterns across individual participants. The 18 pairs of sentences that were judged to be most similar by the human raters had a mean rank accuracy of 0.83 in the list of sentence pairs ranked by model-predicted similarity, as shown in Table 4. This high consistency between the human judgment and the model-based neural similarity measures demonstrates that the model captures the core meaning of events at a level of abstraction higher than the meanings of the individual words.

Comparing NPSFs with other semantic representations

The prediction performance of the models above was based on the mapping between the neural signatures of concepts and a postulated layer of meaning elements, namely the neurally plausible semantic features (NPSFs). Four alternative semantic characterizations of the sentence content were compared to the NPSFs' account. These alternatives, all based on semantic vector space models, use features derived from word co-occurrence norms in text corpora: GloVe vectors for word representation [Pennington et al., 2014], a documents-based and a dependencies-based vector space model (doc VSM and dep VSM) [Fyshe et al., 2013], and *LSA* [Landauer and Dumais, 1997]. The NPSF features performed reliably better than any of the other feature sets in this comparison, as shown in Figure 3. Dimensionality as an intrinsic attribute of each vector space model might be responsible for the accuracy differences. Nevertheless, the decoding performance provides a pragmatic test of the information content that can be explained by these features. The corpus-based, automatically-generated vector representations also produced accounts that were very reliably above chance level, demonstrating the robustness of the general approach using alternative semantic characterizations.

Commonality of neural representations of sentences across individuals

To determine whether the brain activation pattern for a given sentence was similar across people, a model that predicted brain activation patterns from semantic features was trained on the data of all but one participant, and tested on the data of the left-out participant. The mean accuracy of the activation-predicting model in the left-out participant's brain based on the neural representations of other 6 people was 0.77 (range = 0.72 - 0.79). This finding indicates a considerable degree of cross-individual commonality in the neural representation of sentences, and in the pattern of assocations between the semantic and neural elements.

Thematic role representation: distinguishing agent and patient roles in reversible sentences

The concepts that filled the agent and patient roles in many of the stimulus sentences made the thematic role assignment unambiguous, such as *The happy child found the dime*. When the thematic role variables were omitted from the model, the accuracy of sentence prediction decreased slightly and significantly from a mean of 0.82 to 0.80 (paired-sample $t_6 = 8.4$, $p < 0.001$). An ancillary study was run using sentences whose agent and patient concepts were reversible (e.g. *The rabbit punches the monkey*), to determine how accurately the thematic roles of each concept could be decoded and whether any additional locations of brain activation came to light. (The methods are described in Supporting Information). The main finding was that the agent and patient roles of the two animal concepts could be reliably decoded from the neural representation with a mean accuracy of 0.80 across five participants, demonstrating the decodability of this level of information from the structure of a sentence, independently of the conceptual content of the role fillers [Frankland and Greene, 2015]. Furthermore, the voxels used in this decoding included 17 clusters, 13 of which were within or adjacent to the regions identified in the main study (Figure S2). The remaining 4 clusters were in spatially distinct regions in the precuneus, right amygdala, left anterior middle temporal lobe and the right precentral gyrus (circled in Figure S2). Without these 4 additional clusters, the accuracy of identifying the thematic roles in the ancillary study decreased to 0.75, but remained reliably above chance ($p < 0.05$).

The contribution of the ancillary study is to identify additional regions for representing the thematic organization of a sentence when this organization has to be inferred from the sentence structure. Activation in the amygdala has been shown to be mediated by superior temporal cortex; the activation patterns in this region differentiate pairs of reversible sentences [Frankland &

Greene, 2015]. The precuneus cluster has been found to be sensitive to violation of social expectations [Berthoz et al., 2006; Petrini et al., 2014], suggesting its role in interpreting the inter-concept relations from a perspective of social norms. A related account of the regions involved in thematic role processing emerges from studies of aphasia [Thompson, C. K., Meltzer-Asscher, 2014].

Underlying large-scale dimensions of neural representation of sentences and their brain locations

To determine the nature of the meaning dimensions of the neural representation of the stimulus sentences at a more molar level than NPSFs, a two-level hierarchical factor analysis [Just et al., 2014] was applied to the activation data. There were four main factors or semantic dimensions that emerged, that characterized the neural organization of the sentence representation. These factors can be interpreted as 1. *people and social interactions, 2. spatial and temporal settings, 3. actions (and affected objects),* and *4. feelings.* These interpretations are based in large part on the NPSF semantic features and sentences that were most associated with each factor (Supporting Information). For example, the factor scores of the first factor correlated with the semantic features of *Person* (0.63), *Communication* (0.42), *Intellectual* (0.37), etc. Sentences such as *The young author spoke to the editor,* and *The judge met the mayor* had high scores on this factor. Thus this factor was interpreted as representing people and social interactions. The factor analysis also identifies the voxels that have the highest factor loadings for each factor (Table S3).

A condensed account of the relation between brain locations, large-scale semantic dimensions, and concept-level NPSFs is depicted in Figure 4. Each main dimension is described by a cloud of semantic features (Figure 4B), which are color-mapped to their corresponding

brain locations (Figure 4A). Each NPSF tends to be correlated well with only one of the dimensions (Figure 4C).

This factor analysis interpretably organizes the main meaning dimensions that characterize the 240 sentences in this stimulus set, relates the emerging factors to the NPSFs, and identifies the brain locations (factor-related voxel clusters) that can be used to instantiate the mapping between the semantic features (NPSFs) of concepts and the resulting activation patterns.

Discussion

The main contribution of this paper is the integrated, computational account of the relation between the semantic content of a sentence and the brain activation pattern evoked by the reading of the sentence. The integration applies to several facets of the account. First, the model integrates over the multiple concepts in a sentence to predict the activation of the sentence as a whole. Second, the model also takes into consideration the thematic structure that integrates the roles that the concepts play in the sentence. Third, the model integrates over a wide range of 42 neurally plausible semantic features of different types, from concrete perceptual and motor features of an object (that can be thought of as embodied) to social and abstract features of actions that have much less connection to concrete properties. Many previous studies have reported the correspondence between only a handful of features and brain activation patterns (e.g. Fernandino et al., 2016; Just et al., 2010), but the current study relates activation to many different types of semantic features. Fourth, the model integrates the account of the activation attributable to the 42 concept-level features to the larger-scale dimensions of meaning that emerged from the factor analysis of the activation patterns of the sentences. In all four senses, the current model provides a more integrated, larger scale account of the brain activation associated with concept and sentence meaning. Moreover, it provides a point of departure for

future studies of the neural representation of complex thoughts composed of multiple components.

Although sentences are not merely the sum of their parts, this study shows the extent to which a linear combination of thematically-coded concept representations is able to characterize the neural representation of simple sentences. We note moreover, that the quantities being added are estimates of word concept representations *as they occur in context*, rather than in isolation. The addends that were used to generate sentence representations may be sentence-context-sensitive. Elsewhere, we demonstrate systematic interactions or context effects among the NPSFs of the concepts of a proposition [Just, M. A., Wang, J., Cherkassky, 2017, under review].

The NPSFs provide a basis of prediction of sentence activation in terms of the co-activation of various brain subsystems that have characteristic processing specializations. The NPSFs were developed with these specializations in mind, and they result in more accurate predictions of sentence activation patterns than do feature sets derived from language-based corpora. Moreover, the NPSF-based characterization can be expanded to include additional types of concepts and, if needed, additional neural systems. That is, this general approach has the potential of extensibility to any number of concepts and simple sentences constructed from the concepts.

The theory-driven NPSFs and the data-driven neural dimensions (factors) derived from the factor analysis of the sentence activations provide accounts at two levels of granularity. The factors provide an account at a coarser, higher level that specifies the general type of thought associated with a simple sentence (e.g. the thought of an action and its consequences). The NPSFs provide an account at a finer level that specifies the properties of each of the concepts in the proposition (e.g. a change of location of an object). Together these two levels of account

suggest that the generativity of conceptual and propositional representation is enabled by the possible combinatorial space of these fundamental neural components and their values.

The initial success of the modeling using neurally plausible features suggests that the building blocks for constructing complex thoughts are shaped by neural systems rather than by lexicographic considerations. This approach predicts that the neural dimensions of concept representation might be universal across languages, as studies are beginning to suggest [Yang et al., 2017]. In this perspective, the concepts in each language would be underpinned by some subset of a universal set of NPSFs. The predictive modeling of the neural signatures of new concepts in a variety of dissimilar languages is a possible way to test the hypotheses reflected by these neurally plausible semantic features, in contrast to hypotheses based on models that are blind to neural capabilities.

The current study demonstrates the possibility of modeling the neural representation of semantic information beyond the single word level by taking into consideration the role of a concept in a proposition. Although only six roles present in the current stimulus set were implemented in the model, it could be expanded to incorporate more elaborated thematic roles. Furthermore, computationally, the model has the potential to identify neural signatures of other aspects of sentence meaning, such as negation, tense, syntactic roles of concepts, etc. Future studies may use this approach to test more complex compositional semantic representations.

Direction of prediction

The regression modeling enables the prediction of brain activity of a sentence as well as the comparably accurate prediction of the sentence semantics. Given the main scientific goal of understanding the process and representation that the brain generates during sentence comprehension, the corresponding direction is predicting the activation from knowledge of the

sentence, as was implemented in previous studies [Huth et al., 2016; Mitchell et al., 2008; Wehbe et al., 2014]. However, the other direction of decoding semantics from activations, provides a more intuitive assessment of the model's success, constituting a "mindreading" capability. There is an immediate, human-readable assessment of how unfaithful or faithful the prediction is to the actual sentence. The two directions of mapping serve different purposes, and their comparable accuracies speak to the robustness of the general approach.

In summary, the new findings and the accompanying theoretical account provide an initial explanation of how complex meaning is neurally represented. It presents an initial mapping between brain activity and that part of the mind that represents events and states as described by sentences.

Large-scale semantic dimensions and brain locations emerging from factor analysis of sentence activation patterns

The activation profiles of many regions identified in the factor analysis (Figure 4A and Table S3) are consistent with previous findings of the role of these regions in semantic knowledge representation, such as right anterior temporal lobe for semantic knowledge of people [Gesierich et al., 2012], fusiform gyrus for representing objects [Martin, 2007], parahippocampal areas for representing places [Epstein and Kanwisher, 1998], etc. Moreover, the response profiles of several other regions suggest that reading of simple sentences that describe events and states also involve various non-language-specific neural systems associated with the processing of social, affective, motor and visual properties of events, as discussed below.

People. One of the large scale dimensions emerging from the factor analysis assigned high scores to sentences describing people and social interactions. The location of the largest factor-related cluster (posterior cingulate cortex and the adjacent precuneus region) is known for its role

in episodic memory and being a signature of the default mode network [Cavanna and Trimble, 2006; Gusnard and Raichle, 2001], and it has also been found to activate during social sharing of emotions vs. processing emotion alone [Wagner et al., 2015], and during thinking about intentional vs. physical causality [Ouden et al., 2005]. Similarly, the medial prefrontal cortex has been associated with social and moral reasoning [Moll et al., 2005; Van Overwalle, 2011]. The left superior frontal gyrus has been associated with reasoning about social contracts [Fiddick et al., 2005], social comparison [Dvash et al., 2010], and found to be part of a theory of mind network [Wolf et al., 2010]. In other words, processing social content during the reading of sentences also involves a neural network for social processing. This factor is correlated with the NPSFs of person, communication, social norms, and social interaction, as shown in Figure 4C.

Places (spatial and temporal settings). The sentences with high scores on this factor included most of the sentences that described a scene rather than an event (e.g. *The restaurant was loud at night*). The brain regions associated with this factor include the parahippocampal area, and the posterior cingulate cortices, which have been linked to viewing scenes and representing semantic knowledge of shelters [Henderson et al., 2008; Just et al., 2010]. The cluster in the right angular gyrus has been associated with accurate temporal memory [Jenkins and Ranganath, 2010] and retrieval of learned ordinal movement sequences [Bengtsson et al., 2004], suggesting its role in the representation of time-related concepts. This factor is correlated with the NPSFs of setting, openness (nonenclosure) and shelter (enclosure), as shown in Figure 4C.

Actions (and affected objects). Sentences that included main verbs such as *break* or *kick* had high scores on this factor. The cluster associated with this factor in the middle frontal gyrus has been associated with motor imagery [Szameitat et al., 2007] and retrieval of visually or

haptically encoded objects [Stock et al., 2008]. This factor was correlated with the NPSFs of physical impact and change of location, as shown in Figure 4C.

Feelings. The right temporoparietal junction, which is known for its role in representing belief of others [Saxe and Powell, 2006], has been found to activate for processing fear-inducing pictures [Stark et al., 2007] and fearful body expressions [Grèzes et al., 2007]. Moreover, this region has been found to represent various emotions by the appraisal-related features, such as expectedness, fairness, etc. of affective events [Skerry and Saxe, 2015]. The right inferior frontal gyrus has been associated with recognizing visually presented objects with negative valence [Kensinger and Schacter, 2007]. This factor was correlated with the NPSFs of high affective arousal and negative affective valence, as shown in Figure 4C.

In summary, the factor analysis yields broad semantic dimensions that characterize the events and states in the 240 stimulus sentences. The brain locations associated with the each of the factors suggest that the brain structures that are activated in the processing of various aspects of everyday life events also encode the corresponding semantic information. The factors can be construed as higher-order dimensions that subsume relevant NPSFs.

Generalizability across individuals

The generalizability of the findings is first demonstrated when individual participants' data are modeled independently. The ability to predict and decode a new sentence was present for every one of the seven participants with high accuracies. When individual participant's data are modeled, the reliability of the results is assessed not in terms of the number of participants, but in terms of the large number of stimulus items (240 sentences). The model was further tested for its generality across people: a model trained on the data of all but one participant is able to reliably predict the neural signature of the left-out participant. This result indicates the generalizability of

the neural representations across individuals. The fact that a small number of participants are sufficient to build a cross-participant model suggests the consistency of the effect.

Study limitations and implications

One limitation is that despite the large number of stimulus sentences examined, the admixture of narrative and descriptive sentences is limited in its structure and content. Sentences can provide far wider types of information and they can have far wider syntactic forms than those examined here. The set of NPSFs would certainly have to be expanded to code all possible sentences, but perhaps their number would asymptote at a few hundred. More complex syntactic processing would evoke more activation, but the neural representations of the concepts involved may not be affected by the syntactic complexity of the sentence. Although the specific account provided by the model is thus limited to the types of sentences examined in the study, the general principles may well be extensible to the neural representation of concepts in any sentence.

Another limitation is that the factor analysis produced dimensions with reasonable interpretations, but the interpretations were not put to a rigorous test. It should be possible to independently assess the large-scale underlying dimensions of a sentence and thus predict its factor scores on the 4 factors. Furthermore, the outcome of the factor analysis here was limited to the sample of 240 stimulus sentences. It is likely that a much larger sample of sentence types would yield additional factors, which could also be independently assessed and tested for predictive ability.

The study was also limited to the processing of visually presented sentences, and the neural signature at the end of the reading of a sentence contained the representations of all of the component concepts in the sentence. If the sentences were presented in the auditory modality, it is possible the neural signature at the end of the listening to a sentence might not be the optimal

decoding window for all of the component concepts in the sentence.

Despite its limitations, the study opens new avenues of inquiry concerning the neural representation of complex inputs. For example, the model described here can predict the activation patterns evoked by the reading of translation-equivalent sentences in another language [Yang et al., 2017]. Not only can the model be extended to different languages, but it may also extend to different media of communication: the neural representation of a short video depicting a simple event might also be characterized by using the methods developed here. A broader avenue of extension might be the characterization of the full neurally-based ontology that organizes information in a human brain, regardless of the form of the input. The full potential of fMRI has yet to be realized as it is applied to the representation of increasingly complex information.

Summary

This study leads to an initial theoretical and computational account of the neural representation of the propositional content of event-describing and state-describing sentences. The main contribution is the predictive bidirectional mapping between the neurosemantic properties of concepts and the neural signatures that characterize how the brain represents events and states described by simple sentences.

The novelty of the approach mainly lies in the mapping that incorporates neurally driven properties of concepts and sentences, the roles concepts play in a proposition, and identification of an end-of-sentence neural signature of the semantics of all the component concepts of the proposition. The findings indicate that (1) The neural representation of an event or state-describing proposition entails brain subsystems specialized in representing particular semantic information that can be characterized by a set of neurally plausible semantic features; (2) It is

possible to reliably predict sentence-level brain activity from this set of specialized neural bases and the knowledge of the semantic properties of the component words of a sentence and their inter-relations. (3) It is also possible to decode the semantic properties of the concepts in a sentence from the observed activation patterns; and (4) The neural representation of the meaning of events and states is largely common across individuals.

Acknowledgements

We thank Nick Diana for outstanding technical assistance. This work 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).

The authors declare no conflict of interests.

References

Anderson AJ, Murphy B, Poesio M (2014): Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. J Cogn Neurosci 26:658–681.

Baron SG, Osherson D (2011): Evidence for conceptual combination in the left anterior temporal lobe. Neuroimage 55:1847–1852.

Baucom LB, Wedell DH, Wang J, Blitzer DN, Shinkareva S V (2012): Decoding the neural representation of affective states. Neuroimage 59:718–27.

Bauer AJ, Just MA (2015): Monitoring the growth of the neural representations of new animal concepts. Hum Brain Mapp 36:3213–3226.

Bengtsson SL, Ehrsson HH, Forssberg H, Ullen F (2004): Dissociating brain regions controlling the temporal and ordinal structure of learned movement sequences. Eur J Neurosci 19:2591–2602.

Berthoz S, Grèzes J, Armony JL, Passingham RE, Dolan RJ (2006): Affective response to one's own moral violations. Neuroimage 31:945–950.

Cavanna AE, Trimble MR (2006): The precuneus: a review of its functional anatomy and behavioural correlates. Brain 129:564–583.

Chikazoe J, Lee DH, Kriegeskorte N, Anderson AK (2014): Population coding of affect across stimuli, modalities and individuals. Nat Neurosci 17:1114–1122.

Coutanche MN, Thompson-Schill SL (2015): Creating concepts from converging features in human cortex. Cereb Cortex 25:2584–2593.

Dvash J, Gilam G, Ben-Ze'ev A, Hendler T, Shamay-Tsoory SG (2010): The envious brain: The neural basis of social comparison. Hum Brain Mapp 31:1741–1750.

Epstein R, Kanwisher N (1998): A cortical representation of the local visual environment. Nature 392:598–601.

Fellbaum C ed. (1998): WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Fernandino L, Binder JR, Desai RH, Pendl SL, Humphries CJ, Gross WL, Conant LL, Seidenberg MS (2016): Concept representation reflects multimodal abstraction: a framework for embodied semantics. Cereb Cortex 26:2018–2034.

Fiddick L, Spampinato MV, Grafman J (2005): Social contracts and precautions activate different neurological systems: An fMRI investigation of deontic reasoning. Neuroimage 28:778–786.

Fillmore CJ (1967): The case for case. In: Bach, E, Harms, R, editors. Proceedings of the Texas Symposium on Language Universals.

Frankland SM, Greene JD (2015): An architecture for encoding sentence meaning in left mid-superior temporal cortex. PNAS 112:11732–11737.

Fyshe A, Talukdar P, Murphy B, Mitchell T (2013): Documents and dependencies: an exploration of vector space models for semantic composition. In: . International Conference on Computational Natural Language Learning (CoNLL 2013). Sofia, Bulgaria.

Gesierich B, Jovicich J, Riello M, Adriani M, Monti A, Brentari V, Robinson SD, Wilson SM, Fairhall SL, Gorno-Tempini ML (2012): Distinct neural substrates for semantic knowledge and naming in the temporoparietal network. Cereb Cortex 22:2217–26.

Ghio M, Vaghi MMS, Perani D, Tettamanti M (2016): Decoding the neural representation of fine-grained conceptual categories. Neuroimage 132:93–103.

Glasgow K, Roos M, Haufler A, Chevillet M, Wolmetz M (2016): Evaluating semantic models with word-sentence relatedness. arXiv: 160307253 [csCL].

Grèzes J, Pichon S, de Gelder B (2007): Perceiving fear in dynamic body expressions. Neuroimage 35:959–967.

Gusnard DA, Raichle ME (2001): Searching for a baseline: Functional imaging and the resting human brain. Nat Rev Neurosci 2:685–694.

Hastie, T., Tibshirani, R. & Friedman J (2009): The Elements of statistical learning: Data mining,

inference, and prediction. Springer.

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

Henderson JM, Larson CL, Zhu DC (2008): Full scenes produce more activation than close-up scenes and scene-diagnostic objects in parahippocampal and retrosplenial cortex: an fMRI study. Brain Cogn 66:40–49.

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016): Natural speech reveals the semantic maps that tile human cerebral cortex. JOUR. Nature 532:453–458.

Huth AG, Nishimoto S, Vu AT, Gallant JL (2012): A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76:1210–1224.

Iacoboni M, Lieberman MD, Knowlton BJ, Molnar-Szakacs I, Moritz M, Throop CJ, Fiske AP (2004): Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. Neuroimage 21:1167–1173.

Jenkins LJ, Ranganath C (2010): Prefrontal and medial temporal lobe activity at encoding predicts temporal context memory. J Neurosci 30:15558–15565.

Just, M. A., Wang, J., Cherkassky VL (2017): Neural representations of the semantic content of individual simple sentences: Concept combinatorics and sentence context effects. Neuroimage (Under review).

Just MA, Cherkassky VL, Buchweitz A, Keller TA, Mitchell TM (2014): Identifying autism from neural representations of social interactions: Neurocognitive markers of autism. PLoS One 9:e113879.

Just MA, Cherkassky VL, Aryal S, Mitchell TM (2010): A neurosemantic theory of concrete noun representation based on the underlying brain codes. Article. PLoS One 5:e8622.

Kanwisher N, Woods RP, Iacoboni M, Mazziotta JC (1997): A locus in human extrastriate cortex for visual shape analysis. J Cogn Neurosci 9:133–42.

Kassam KS, Markey AR, Cherkassky VL, Loewenstein G, Just MA (2013): Identifying emotions on the basis of neural activation. PLoS One 8:e66032.

Kensinger EA, Schacter DL (2007): Remembering the specific visual details of presented objects: Neuroimaging evidence for effects of emotion. Neuropsychologia 45:2951–2962.

Kranjec A, Chatterjee A (2010): Are temporal concepts embodied? A challenge for cognitive neuroscience. Front Psychol 1:240.

Lai VT, Desai RH (2016): The grounding of temporal metaphors. Cortex 76:43–50.

Landauer TK, Dumais ST (1997): A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol Rev 104:211–240.

Martin A (2007): The representation of object concepts in the brain. Annu Rev Psychol 58:25–45.

Mason RA, Just MA (2016): Neural representations of physics concepts. Psychol Sci 27:904–913.

Mitchell TM, Shinkareva S V, Carlson A, Chang K-M, Malave VL, Mason RA, Just MA (2008): Predicting human brain activity associated with the meanings of nouns. Science 320:1191–5.

Moll J, Zahn R, de Oliveira-Souza R, Krueger F, Grafman J (2005): The neural basis of human moral cognition. Nat Rev Neurosci 6:799–809.

Niles I, Pease A (2001): Towards a standard upper ontology. In: . Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001 (FOIS '01). New York, NY, USA: ACM. pp 2–9.

Oliver RT, Thompson-Schill SL (2003): Dorsal stream activation during retrieval of object size and shape. Cogn Affect Behav Neurosci 3:309–322.

Ouden HEM Den, Frith U, Frith C, Blakemore S (2005): Thinking about intentions 28:787–796.

Van Overwalle F (2011): A dissociation between social mentalizing and general reasoning. Neuroimage 54:1589–1599.

Peelen M V, Caramazza A (2012): Conceptual object representations in human anterior temporal cortex. J Neurosci 32:15728–36.

Pennington J, Socher R, Manning CD (2014): Glove: Global Vectors for Word Representation. In: . EMNLP Vol. 14, pp 1532–1543.

Pereira F, Detre G, Botvinick M (2011): Generating text from functional brain images. Front Hum Neurosci 5:72.

Petrini K, Piwek L, Crabbe F, Pollick FE, Garrod S (2014): Look at those two!: The precuneus role in unattended third-person perspective of social interactions. Hum Brain Mapp 35:5190–5203.

Pulvermüller F, Hauk O, Nikulin V V., Ilmoniemi RJ (2005): Functional links between motor and language systems. Eur J Neurosci 21:793–797.

Saxe R, Powell LJ (2006): It's the thought that counts: specific brain regions for one component of theory of mind. Psychol Sci 17:692–699.

Shinkareva S V, Malave VL, Mason R a, Mitchell TM, Just MA (2011): Commonality of neural representations of words and pictures. Neuroimage 54:2418–25.

Skerry AE, Saxe R (2015): Neural representations of emotion are organized around abstract event features. Curr Biol 25:1945–1954.

Spitzer M, Fischbacher U, Herrnberger B, Grön G, Fehr E (2007): The neural signature of social norm compliance. Neuron 56:185–96.

Stark R, Zimmermann M, Kagerer S, Schienle A, Walter B, Weygandt M, Vaitl D (2007): Hemodynamic brain correlates of disgust and fear ratings. Neuroimage 37:663–673.

Stock O, Röder B, Burke M, Bien S, Rösler F (2008): Cortical activation patterns during long-term memory retrieval of visually or haptically encoded objects and locations. J Cogn Neurosci 21:58–82.

Szameitat AJ, Shen S, Sterr A (2007): Effector-dependent activity in the left dorsal premotor cortex in motor imagery. Eur J Neurosci 26:3303–3308.

Thompson, C. K., Meltzer-Asscher A (2014): Neurocognitive mechanisms of verb argument structure processing. In: Bachrach, A., Roy, I., Stockall, L, editor. Structuring the Argument: Multidisciplinary research on verb argument structure. Amsterdam: John Benjamins Publishing. pp 141–168.

Wagner U, Galli L, Schott BH, Wold A, van der Schalk J, Manstead ASR, Scherer K, Walter H (2015): Beautiful friendship: Social sharing of emotions improves subjective feelings and activates the neural reward circuitry. Soc Cogn Affect Neurosci 10:801–808.

Wang J, Baucom LB, Shinkareva S V. (2013): Decoding abstract and concrete concept representations based on single-trial fMRI data. Hum Brain Mapp 34:1133–1147.

Wang J, Cherkassky VL, Yang Y, Chang KK, Vargas R, Diana N, Just MA (2016): Identifying thematic roles from neural representations measured by functional magnetic resonance imaging. Cogn Neuropsychol:1–8.

Wehbe L, Murphy B, Talukdar P, Fyshe A, Ramdas A, Mitchell T (2014): Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. PLoS One 9:e112575.

Wolf I, Dziobek I, Heekeren HR (2010): Neural correlates of social cognition in naturalistic settings: A model-free analysis approach. Neuroimage 49:894–904.

Yang Y, Wang J, Bailer C, Cherkassky V, Just MA (2017): Commonality of neural representations of sentences across languages: Predicting brain activation during Portuguese sentence comprehension using an English-based model of brain function. Neuroimage 146:658–666.

Table 1. Examples of Stimulus Sentences

The journalist interviewed the judge.
The angry activist broke the chair.
The flood damaged the hospital.
The happy child found the dime.
The witness shouted during the trial.
The jury listened to the famous businessman.
The young author spoke to the editor.
The judge met the mayor.
The restaurant was loud at night.
The window was dusty.

Table 2. The 42 Neurally Plausible Semantic Features (NPSFs)

Category *	Feature	Definition	Example stimuli
Perceptual and Affective Characteristics of an Entity	Man-made	objects or settings made by humans	bicycle, desk, newspaper, church
	Natural	objects or activities occurring in nature	flower, flood, island
	Inanimate	non-living object	ball, coffee, window
	Visual perception	visual perceptual properties	big, blue, empty, new, shiny
	Size	Physical volume or size	big, heavy, long, small
	Color	self-explanatory	black, blue, green, red, white
	Temperature	related to temperature	sun, summer, winter, cold, hot
	Positive affective valence	self-explanatory	celebrate, laugh, vacation, happy
	Negative affective valence	self-explanatory	destroy, fear, terrorist, dangerous, sick
	High affective arousal	self-explanatory	celebrate, shout, hurricane, angry
Animate Beings	Person	a human being	boy, doctor, farmer, pilot, voter
	Animal	an animal or anatomy of animals	bird, dog, duck, feather, horse
	Human-group	more than one human being	team, couple, family, mob, council
Time and Space	Settings	place or temporal settings	lake, church, park, night, hotel
	Unenclosed	an environment without shelter or enclosure	beach, lake, field, island, street
	Location	actions or events that imply spatial settings	meeting, visit, stay, live
	Shelter	being enclosed, indoors is a salient feature; opposite of unenclosed	car, hotel, school, hospital, store
	Change of location	self-explanatory	approach, hike, throw, car, run
	Event	self-explanatory	dinner, protest, trial, vacation
	Time-related	related to a time period or timing	morning, night, spring, summer, end

Table 2 continued

Category *	Feature	Definition	Example stimuli
Human Activity Type	Violence/conflict	involving aggression and those who commit it	army, guard, soldier, terrorist
	Health	related to improving or threatening health	medicine, doctor, patient, victim, hospital
	Eating/drinking	self-explanatory	drink, eat, dinner, corn, restaurant
	Communication	medium of communication	listen, speak, newspaper, author, reporter
	Sports	related to recreation or competitive physical activities	play, soccer, baseball, bicycle, team
	Technology	related to technology or technical skills	computer, television, engineer, scientist
	Money	related to financial activities or economics	buy, cash, banker, expensive, wealthy
	Arts and literature	objects, actions or professions related to humanities, arts, literature	actor, author, artist, theatre, draw
	Social norms	related to laws and authority structure	trial, criminal, lawyer, court, prison
	Governance	related to civics, politics, dominance	debate, protest, army, mayor, embassy
	Intellectual	requiring, gaining, or providing knowledge or expertise	plan, read, computer, engineer, school
Social Action or State	Transfer of possession	transaction (giving/receiving); change of ownership	give, steal, take, buy
	Social interaction	interaction between two or more subjects	interview, negotiate, party, lonely
	Social support	providing social support is a salient feature	help, family, minister, parent
Physical Action or State	Physical action	self-explanatory	kick, throw, play, walk, march
	Change of physical state	self-explanatory	destroy, fix, grow, break
	Physical impact	two subjects or objects coming in contact with each other	break, destroy, drop, kick
Mental Action or State	Mental action	requiring cognitive processes; occurring internally	liked, plan, want, teacher, clever
	Perceptual action	self-explanatory	listen, watch, read, witness
	Emotion	Emotional state or action	fear, laugh, like, happy
Abstractness	Abstract	detached from sensory or motor properties; low imagability	plan, want, clever
Part of Speech	Attribute	adjectives	aggressive, blue, shiny, sick

* The grouping into categories is included here to facilitate description but was not used in the modeling.

Table 3. Examples of NPSF coding of content word concepts

Word	NPSF Features
interview	Social, Mental Action, Knowledge, Communication, Abstraction
walk	Physical Action, Change of Location
hurricane	Event, Change of Physical State, Health, Natural, Negative Affective Valence, High Affective Arousal
cellphone	Social Action, Communication, Man-Made, Inanimate
judge	Social norms, Knowledge, Communication, Person
clever	Attribute, Mental Action, Knowledge, Positive Affective Valence, Abstraction

Table 4. Normalized rank of the 18 behaviorally-judged most similar pairs of sentences with respect to the model's similarity measure. Each row indicates a pair. The pairs are ordered by their rank in the model's similarity prediction.

Sentence 1	Sentence 2	Rank
The editor drank tea at dinner.	The lawyer drank coffee.	1.00
The artist hiked along the mountain.	The tourist hiked through the forest.	0.99
The activist listened to the tired victim.	The policeman interviewed the young victim.	0.99
The fish lived in the river.	The duck lived at the lake.	0.99
The council read the agreement.	The policeman read the newspaper.	0.98
The dangerous criminal stole the television.	The terrorist stole the car.	0.97
The witness shouted during the trial.	The jury listened to the famous businessman.	0.97
The angry lawyer left the office.	The tired jury left the court.	0.94
The witness went to the trial.	The witness spoke to the lawyer.	0.93
The wealthy politician liked coffee.	The lawyer drank coffee.	0.92
The accident damaged the yellow car.	The accident destroyed the empty lab.	0.91
The flood damaged the hospital.	The soldier delivered the medicine during the flood.	0.89
The boy threw the baseball over the fence.	The boy kicked the stone along the street.	0.89
The young girl played soccer.	The happy girl played in the forest.	0.87
The yellow bird flew over the field.	The girl saw the small bird.	0.78
The tourist found a bird in the theater.	The man saw the fish in the river.	0.41
The author kicked the desk.	The horse kicked the fence.	0.33
The couple read on the beach.	The cloud blocked the sun.	0.21
	Mean	**0.83**

Figure Captions

Figure 1. Rank accuracy of identifying the 240 sentences using the activation-predicting model or the meaning-predicting model. Error bars indicate standard deviations. The rank accuracy for each participant individually was significantly above chance level (the critical value of the rank accuracy being significantly different from chance level at $p = 10^{-5}$ is 0.58 based on random permutation tests).

Figure 2. Predicted and observed activation patterns and semantic features (NPSFs) for two pairs of sentences (sentences A and B in the first panel, sentences C and D in the second panel), displaying the similarity between the model-predicted (upper row) and observed (lower row) activation patterns for each of the sentences. See text for quantitative measures of similarity. Also note the similarities between the two sentences within each pair (columns A and B and columns C and D) in terms of their predicted activation, observed activation, and semantic features (NPSFs). The geometric symbols (hexagon, diamond, rectangle, and ellipse) pointing to voxel clusters and adjoining the test sentence semantic features (NPSFs) correspond to locations of voxels clusters with high factor loadings on the large-scale semantic factors of *people, places, actions (and their consequences),* and *feelings,* respectively (see Figure 4C). The font size of the NPSFs indicates their mean values averaged across the content words of the sentence.

Figure 3. Comparison of sentence activation prediction accuracies using different types of predicting semantic features. Error bars indicate standard errors. The *p* values indicate the comparison between the mean rank accuracy using each of the four alternatives versus the

NPSF-based model in paired-sample t-tests over 7 participants. ** two-tailed $p < 0.01$; * two-tailed $p < 0.05$.

Figure 4. Relations between brain locations, large-scale semantic factors underlying sentence representations, and neurally plausible semantic features (NPSFs) of concepts.
(A) Brain regions associated with the four large-scale semantic factors: people (yellow), places (red), actions and their consequences (blue) and feelings (green). Details of the region locations are shown in Table S3. (B) Word clouds associated with each large-scale semantic factor. The clouds are formed using the 7 neurally plausible semantic features most associated with each factor to illustrate some of the main meaning components of each factor.
(C) NPSFs that correlate with at least one factor with $r > 0.3$ ($p < 0.0001$). The pairwise correlations are computed between each NPSF and the factor scores over sentences. The NPSF for a sentence is computed as the mean of the NPSFs of all the concepts in the sentence.

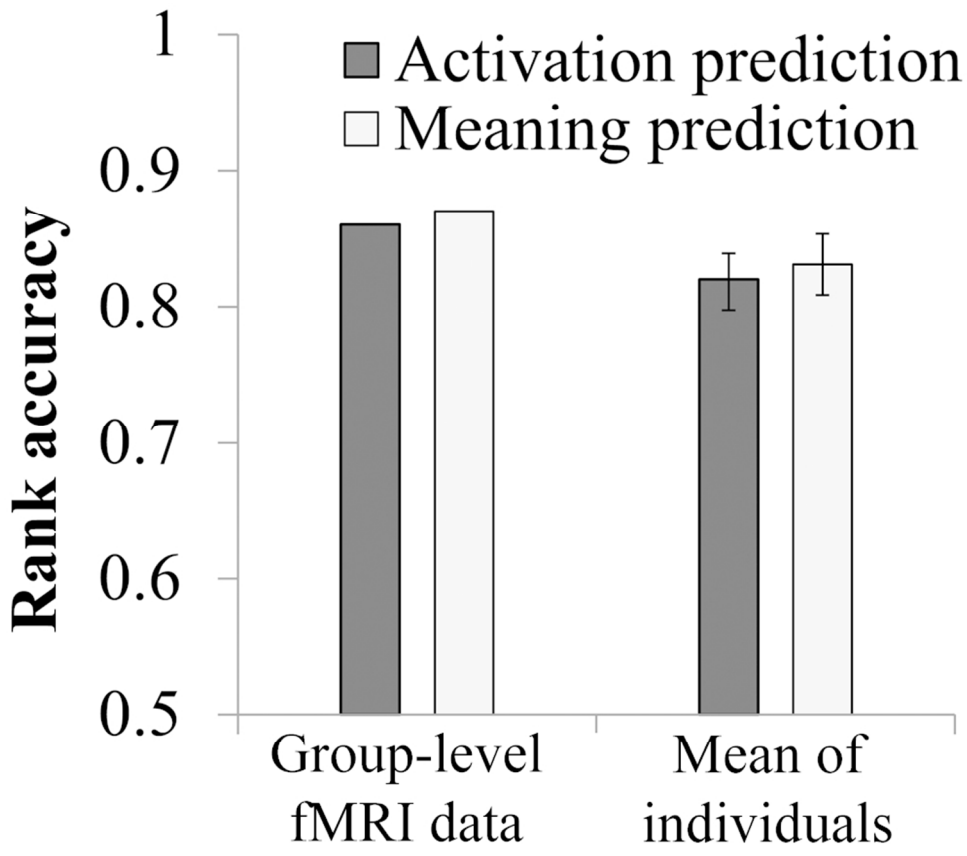

Figure 1. Rank accuracy of identifying the 240 sentences using the activation-predicting model or the meaning-predicting model. Error bars indicate standard deviations. The rank accuracy for each participant individually was significantly above chance level (the critical value of the rank accuracy being significantly different from chance level at p = 10-5 is 0.58 based on random permutation tests).

87x73mm (300 x 300 DPI)

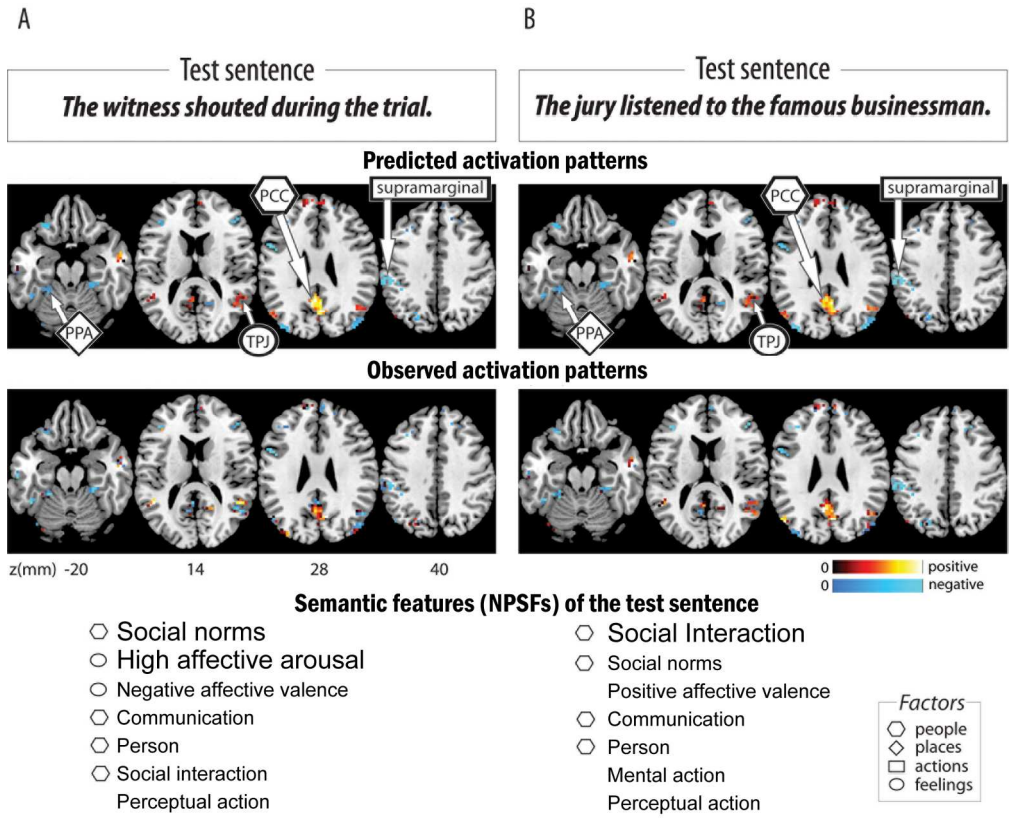

Figure 2. Predicted and observed activation patterns and semantic features (NPSFs) for two pairs of sentences (sentences A and B in the first panel, sentences C and D in the second panel), displaying the similarity between the model-predicted (upper row) and observed (lower row) activation patterns for each of the sentences. See text for quantitative measures of similarity. Also note the similarities between the two sentences within each pair (columns A and B and columns C and D) in terms of their predicted activation, observed activation, and semantic features (NPSFs). The geometric symbols (hexagon, diamond, rectangle, and ellipse) pointing to voxel clusters and adjoining the test sentence semantic features (NPSFs) correspond to locations of voxels clusters with high factor loadings on the large-scale semantic factors of people, places, actions (and their consequences), and feelings, respectively (see Figure 4C). The font size of the NPSFs indicates their mean values averaged across the content words of the sentence.

177x146mm (300 x 300 DPI)

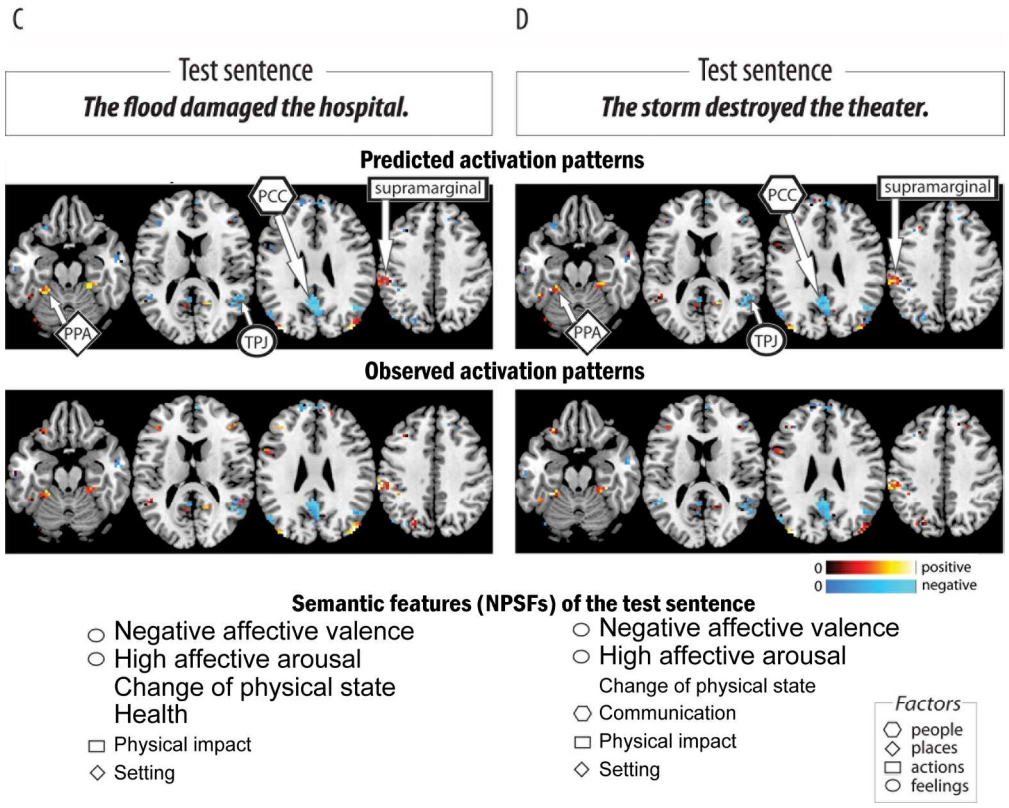

Figure 2. Predicted and observed activation patterns and semantic features (NPSFs) for two pairs of sentences (sentences A and B in the first panel, sentences C and D in the second panel), displaying the similarity between the model-predicted (upper row) and observed (lower row) activation patterns for each of the sentences. See text for quantitative measures of similarity. Also note the similarities between the two sentences within each pair (columns A and B and columns C and D) in terms of their predicted activation, observed activation, and semantic features (NPSFs). The geometric symbols (hexagon, diamond, rectangle, and ellipse) pointing to voxel clusters and adjoining the test sentence semantic features (NPSFs) correspond to locations of voxels clusters with high factor loadings on the large-scale semantic factors of people, places, actions (and their consequences), and feelings, respectively (see Figure 4C). The font size of the NPSFs indicates their mean values averaged across the content words of the sentence.

177x142mm (300 x 300 DPI)

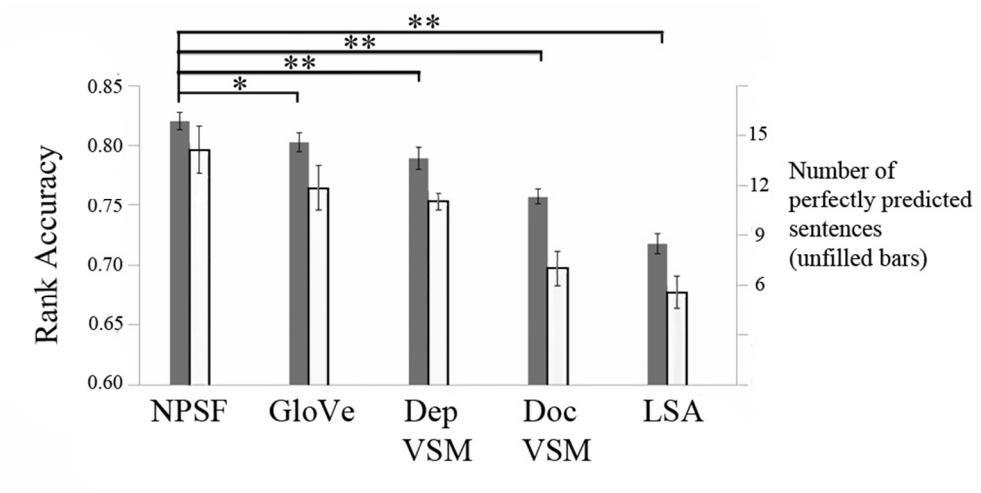

Figure 3. Comparison of sentence activation prediction accuracies using different types of predicting semantic features. Error bars indicate standard errors. The p values indicate the comparison between the mean rank accuracy using each of the four alternatives versus the NPSF-based model in paired-sample t-tests over 7 participants. ** two-tailed $p < 0.01$; * two-tailed $p < 0.05$.

87x42mm (300 x 300 DPI)

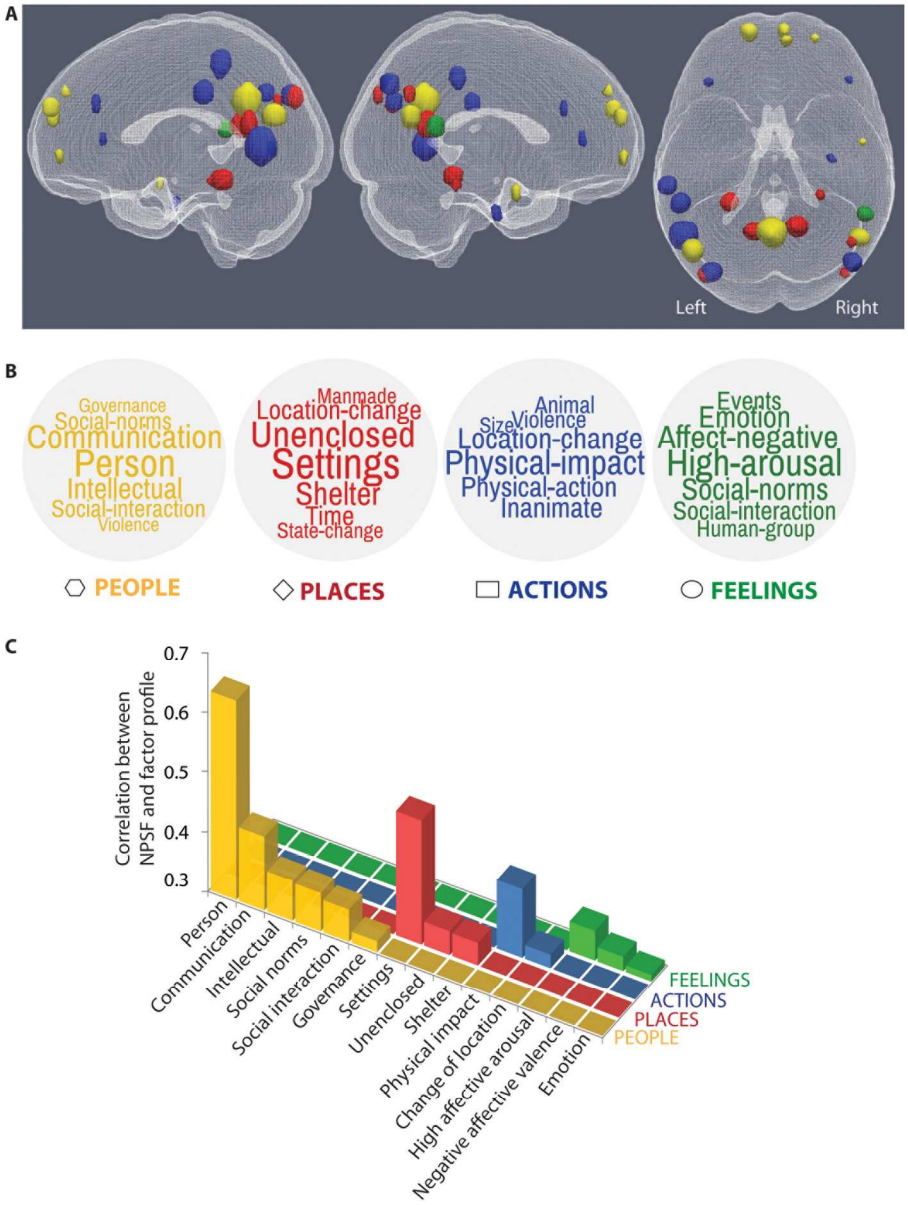

Figure 4. Relations between brain locations, large-scale semantic factors underlying sentence representations, and neurally plausible semantic features (NPSFs) of concepts.
(A) Brain regions associated with the four large-scale semantic factors: people (yellow), places (red), actions and their consequences (blue) and feelings (green). Details of the region locations are shown in Table S3. (B) Word clouds associated with each large-scale semantic factor. The clouds are formed using the 7 neurally plausible semantic features most associated with each factor to illustrate some of the main meaning components of each factor.
(C) NPSFs that correlate with at least one factor with $r > 0.3$ ($p < 0.0001$). The pairwise correlations are computed between each NPSF and the factor scores over sentences. The NPSF for a sentence is computed as the mean of the NPSFs of all the concepts in the sentence.

177x237mm (300 x 300 DPI)