Cerebral Cortex, 2018; 1–16

doi: 10.1093/cercor/bhy110
Original Article

ORIGINAL ARTICLE

Multiple Regions of a Cortical Network Commonly Encode the Meaning of Words in Multiple Grammatical Positions of Read Sentences

Andrew James Anderson[1], Edmund C. Lalor[1,2,3], Feng Lin[4,5], Jeffrey R. Binder[6], Leonardo Fernandino[6], Colin J. Humphries[6], Lisa L. Conant[6], Rajeev D. S. Raizada[7], Scott Grimm[8] and Xixi Wang[1]

[1]Department of Biomedical Engineering, University of Rochester, Rochester, NY 14627, USA, [2]Department of Neuroscience, University of Rochester, Rochester, NY 14642, USA, [3]School of Engineering, Trinity Centre for Bioengineering, and Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland, [4]School of Nursing, University of Rochester, Rochester, NY 14642, USA, [5]Psychiatry, University of Rochester, Rochester, NY 14642, USA, [6]Department of Neurology, Medical College of Wisconsin, Milwaukee, WI 53226, USA, [7]Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA and [8]Department of Linguistics, University of Rochester, Rochester, NY 14627, USA

Address correspondence to Andrew James Anderson, Department of Biomedical Engineering, 201 Robert B. Goergen Hall, University of Rochester, Rochester, NY 14627, USA. Email: aander41@ur.rochester.edu.

Abstract

Deciphering how sentence meaning is represented in the brain remains a major challenge to science. Semantically related neural activity has recently been shown to arise concurrently in distributed brain regions as successive words in a sentence are read. However, what semantic content is represented by different regions, what is common across them, and how this relates to words in different grammatical positions of sentences is weakly understood. To address these questions, we apply a semantic model of word meaning to interpret brain activation patterns elicited in sentence reading. The model is based on human ratings of 65 sensory/motor/emotional and cognitive features of experience with words (and their referents). Through a process of mapping functional Magnetic Resonance Imaging activation back into model space we test: which brain regions semantically encode content words in different grammatical positions (e.g., subject/verb/object); and what semantic features are encoded by different regions. In left temporal, inferior parietal, and inferior/superior frontal regions we detect the semantic encoding of words in all grammatical positions tested and reveal multiple common components of semantic representation. This suggests that sentence comprehension involves a common core representation of multiple words' meaning being encoded in a network of regions distributed across the brain.

Key words: concepts, fMRI, lexical semantics, semantic model, sentence processing

Introduction

Whilst it is now established that sentence comprehension engages a widely distributed network of sensory, motor, linguistic, cognitive and affective neural systems (Lau et al. 2008; Binder et al. 2009; Lambon Ralph et al. 2017), how meaning is represented, extracted and processed by this system is weakly understood. Beyond the academic value of finding answers to this problem, scientific progress in this area may enhance diagnosis and treatment of language disorders and provide guidance for artificial intelligence research.

Two very recent studies revealed that the word by word construction of sentence meaning in the brain is marked by a rise in electrocorticography (ECoG) electrode activity that shows a similar profile in distributed brain regions (Fedorenko et al. 2016; Nelson et al. 2017). These studies suggest that semantic constructs of entire sentences may be represented in distributed brain regions, which has been contrasted (Fedorenko et al. 2016) with proposals that link semantic composition to a particular region (e.g., Baron and Osherson 2011; Westerlund and Pylkkänen 2014; Zhang and Pylkkänen 2015). It remains unclear what semantic content is represented in different sentence processing regions. For instance, is semantic activation associated with a similar representation of meaning accumulating in many different regions, or do different regions tend to represent different semantic features (or even features of words in select grammatical positions of sentences, e.g., subject/verb/object, etc)?

The intracranial electrodes analyzed by Fedorenko et al. (2016) and Nelson et al. (2017) may be too few and positioned too sparsely to adequately test questions of detailed semantic representation. We turn to functional Magnetic Resonance Imaging (fMRI) which enables activation across the entire brain to be scanned with relatively high spatial resolution, but relatively slow sample rate. The goals of the current article are to identify from fMRI data those brain regions that are engaged in processing sentence meaning, then: test whether they semantically encode content words in multiple grammatical positions; filter out and interpret the components of semantic information processed by those regions; and identify commonalities in semantic components represented across regions. To do this we build on very recent advances in semantic modeling that have made addressing this type of question possible.

The last decade has seen semantic models becoming an increasingly popular tool with which to predict and decipher scans of brain activity as people perform conceptual tasks (e.g., Mitchell et al. 2008; Chang et al. 2010; Devereux et al. 2010; Sudre et al. 2012; Pereira et al. 2013; Wehbe et al. 2014; Fernandino et al. 2015b; Anderson et al. 2016; Fernandino et al. 2016; Huth et al. 2016; Yang et al. 2017). These models typically represent meaning as a vector of weights on a set of discrete semantic features. For instance, the concept "football" would weigh heavily on "*shape*" and "*lower-limb*" related features and weakly on "*taste*" and "*temperature*". In contrast "beer" would show a reversed pattern, and both "beer" and "football" would weigh heavily on "*socialization*". Such a semantic model can be applied to factor brain activation patterns associated with different concepts (e.g., "football") into a set of component brain activation maps, each one associated with a model feature. Thus, brain maps for *taste*, *temperature*, *lower-limb*…, can be weighted and summed together to predict neural activation for "football". Recent modeling work has transitioned from the prediction of neural activation patterns associated with isolated words (Mitchell et al. 2008; Chang et al. 2010; Devereux et al. 2010; Sudre et al. 2012; Pereira et al. 2013; Fernandino et al. 2015b, 2016), to more ambitious questions concerning the prediction of neural activation associated with sentences (Anderson et al. 2016; Pereira et al. 2016; Yang et al. 2017) and narratives (Wehbe et al. 2014; Huth et al. 2016).

We here reanalyze fMRI data from one of these studies that scanned 14 participants whilst they read 240 sentences describing simple everyday situations (Anderson et al. 2016). Anderson et al. (2016) used a semantic model (Binder et al. 2016) to predict fMRI representations of words contained in sentences, then superposed these to predict fMRI representations of the meanings of new sentences. Whilst this study showed that neural activation associated with sentence meaning is predictable, it left many questions open surrounding how sentence meaning is distributed across the cortex, and the entirety (or otherwise) to which sentence meaning is locally represented within different brain regions. In particular, it did not reveal how regional activation relates to words in different grammatical positions of sentences, which semantic features capture neural activation in different regions, and commonalities in both of the previous respects across different regions of the semantic network.

This article newly reveals how semantic features underpinning sentence meaning are commonly encoded in multiple distributed brain regions. To show this we first develop computational methods to test which regions encode semantic content associated with words in different grammatical positions (e.g., in the sentence "the footballer bought the beer", different semantic features are triggered by the sentence's subject and verb and object). We reason that if a region is sensitive to semantic features supplied by all elements of grammatical structure (subject/verb/object inclusive in the previous example) then it is a candidate for representing the semantic constructs of sentences observed by Fedorenko et al. (2016). Secondly, we test which semantic features can be reconstructed from fMRI activation in different regions, before finally estimating semantic structure that is common across regions. We discover that semantic features associated with words in each grammatical position tested can commonly be reconstructed from multiple brain regions. Results are consistent with comprehension involving a common multiword core of semantic information being encoded in each of those regions.

Materials and Methods

Fourteen people were scanned as they read 240 simple 3 to 9 word long sentences describing everyday situations (sentences are detailed in Anderson et al. 2016; Glasgow et al. 2016 and listed in Supplementary materials Table S1). Following standard preprocessing (documented in detail in Anderson et al. 2016 and summarized in Supplementary materials) a single fMRI volume was created for each sentence per participant. An "*experiential attribute*" semantic model was built that represented each of the 242 content words that appeared in the 240 sentences in terms of human ratings of the degree of association between each lexical concept and a particular type of experience (e.g., "On a scale of 0 to 6, to what degree do you think of a football as having a characteristic or defining color?"). Ratings were collected via Amazon Mechanical Turk for a total of 65 features (attributes) of experience (listed in Supplementary materials Table S3), that ranged over sensory, motor and affective processes, systems processing spatial, temporal, and causal information, and social cognition and abstract cognitive operations (Binder et al. 2016). Ratings were averaged across subjects to give a single 65 dimensional "semantic feature vector" for each word.

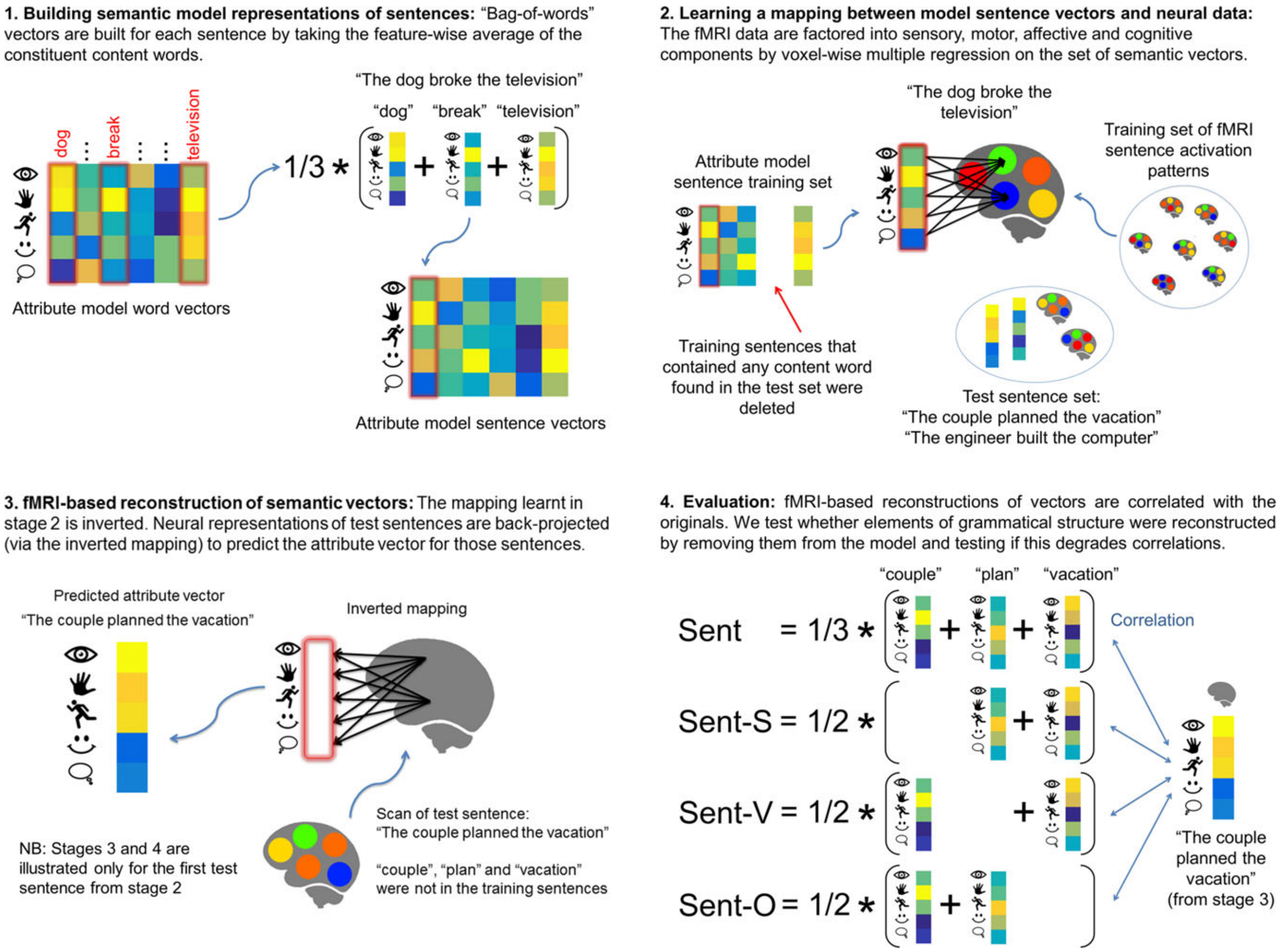
Analytic Procedure: Mapping the Semantic Model to fMRI Sentences and Back

All subsequent analyses were based on the following procedure, which was later used to test which brain regions were sensitive to semantic information associated with content words in different grammatical positions, and which semantic features could be extracted from fMRI activation within different brain regions.

Firstly, to enable sentence-level fMRI data to be matched to the word-level semantic model, a "bag-of-words" semantic model representation for each of the 240 sentences was built by featurewise averaging all constituent content words in the sentence (Fig. 1 stage 1). Although this approach to combining words is simplistic because it ignores word order and syntactic structure, similar additive phrase construction methods have proved practically effective in both predicting brain activation (Anderson et al. 2016) and in computational linguistics (Mitchell and Lapata 2010.

Next, to estimate how well semantic vectors could be reconstructed from new fMRI data an iterative leave-2-sentence out cross-validation procedure was run for each participant. At each cross-validation iteration, the 240 sentences were split into a test set of 2 sentences and a training set of 238 sentences. Then both fMRI and model data for any of the 238 training sentences that contained any content word within the 2 test sentences was deleted. This was done to enable testing of how well the predictive model could generalize to decode the semantic content of sentences constructed from untrained words. The mean ± SD number of sentences in the training set for each iteration was 218 ± 5, containing a mean±SD of 232 ± 2 words. Model and fMRI data for the remaining training sentences were then feature/voxelwise z-scored. Model and fMRI test sentence data were also normalized by subtracting and dividing by the feature/voxelwise mean and SD of the training data.

At each iteration, semantic vectors were reconstructed from the test fMRI data using a two-stage regression analysis (Pereira et al. 2016, see Fig. 2). First a "forward model" that mapped the 65 semantic model features to activation in each individual voxel was learnt from the training sentences using an independent multiple regression for each voxel. Regression was implemented using the Moore–Penrose pseudoinverse (function pinv in MATLAB, with the default tolerance of 1e–6) to invert the semantic model matrix (see top of Fig. 2). The resultant matrix of beta coefficients (with number-of-features rows and number-of-voxels columns) constitutes the forward model and can be thought of in terms of a set of brain maps, one map for each semantic feature, of the degree of modulation in activity at each voxel by that feature. Whilst weighting these brain maps with corresponding semantic feature values for a test sentence (achieved by matrix multiplication of a test semantic vector with the forward model) would have predicted voxel activation associated with the sentence (as undertaken in



1. Building semantic model representations of sentences: "Bag-of-words" vectors are built for each sentence by taking the feature-wise average of the constituent content words.

dog
break
television
Attribute model word vectors
"The dog broke the television"
"dog" "break" "television"
1/3 *
+
+
Attribute model sentence vectors

2. Learning a mapping between model sentence vectors and neural data: The fMRI data are factored into sensory, motor, affective and cognitive components by voxel-wise multiple regression on the set of semantic vectors.

Attribute model sentence training set
"The dog broke the television"
Training set of fMRI sentence activation patterns
Training sentences that contained any content word found in the test set were deleted
Test sentence set:
"The couple planned the vacation"
"The engineer built the computer"

3. fMRI-based reconstruction of semantic vectors: The mapping learnt in stage 2 is inverted. Neural representations of test sentences are back-projected (via the inverted mapping) to predict the attribute vector for those sentences.

Predicted attribute vector
"The couple planned the vacation"
Inverted mapping
NB: Stages 3 and 4 are illustrated only for the first test sentence from stage 2
Scan of test sentence:
"The couple planned the vacation"
"couple", "plan" and "vacation" were not in the training sentences

4. Evaluation: fMRI-based reconstructions of vectors are correlated with the originals. We test whether elements of grammatical structure were reconstructed by removing them from the model and testing if this degrades correlations.

"couple" "plan" "vacation"
Sent = 1/3 *
Sent-S = 1/2 *
Sent-V = 1/2 *
Sent-O = 1/2 *
+
Correlation
"The couple planned the vacation" (from stage 3)

Figure 1. Model to brain to model algorithm for reconstructing the semantic content of sentences. Stage 2 and 3 are illustrated in detail in Figure 2. In stage 4 "Sent" is short for sentence, and S, V, and O are abbreviations for Subject, Verb and Object. Exclusion of only S, V and O are illustrated to save space, however analyses involved withholding all 6 grammatical elements in turn (also Indirect Object, Copula-phrase and Adjunct).

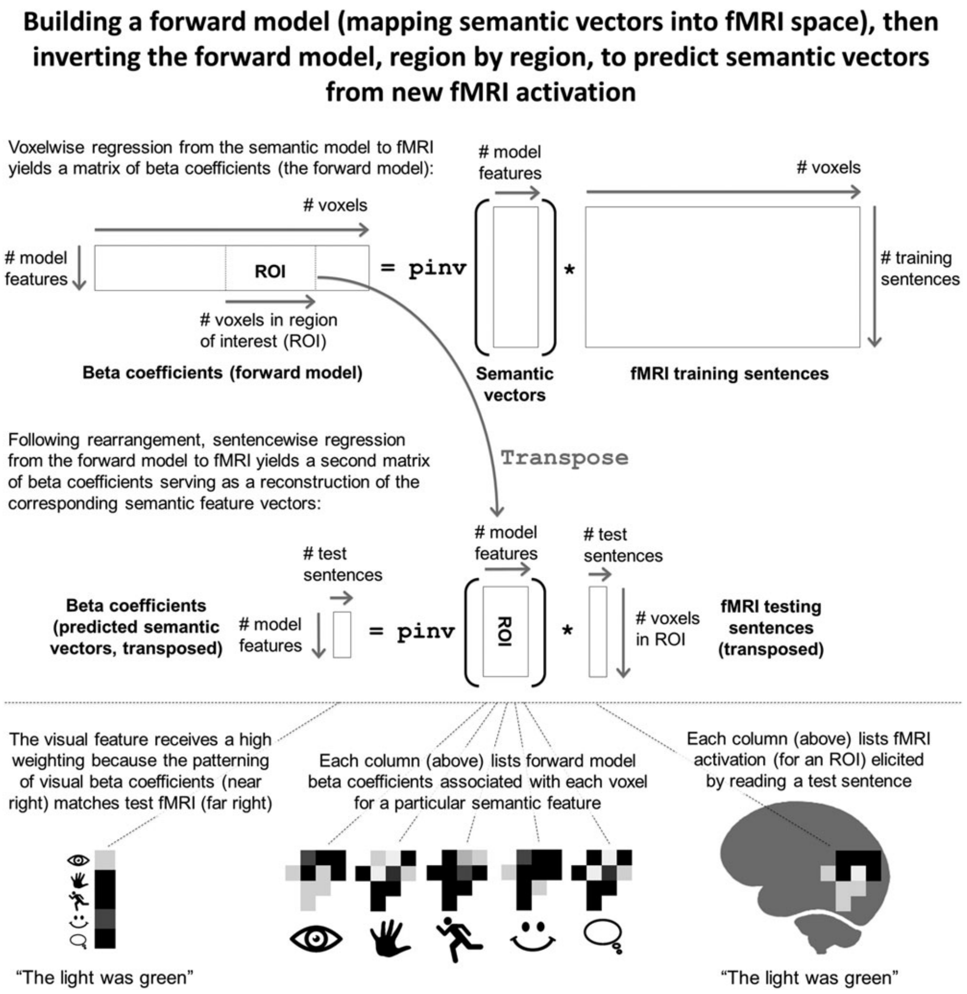

Building a forward model (mapping semantic vectors into fMRI space), then inverting the forward model, region by region, to predict semantic vectors from new fMRI activation
Voxelwise regression from the semantic model to fMRI yields a matrix of beta coefficients (the forward model):
# voxels
# model features
ROI
= pinv
# voxels in region of interest (ROI)
Beta coefficients (forward model)
# model features
*
# voxels
# training sentences
Semantic vectors
fMRI training sentences
Transpose
Following rearrangement, sentencewise regression from the forward model to fMRI yields a second matrix of beta coefficients serving as a reconstruction of the corresponding semantic feature vectors:
# test sentences
Beta coefficients (predicted semantic vectors, transposed)
# model features
= pinv
# model features
ROI
*
# test sentences
# voxels in ROI
fMRI testing sentences (transposed)
The visual feature receives a high weighting because the patterning of visual beta coefficients (near right) matches test fMRI (far right)
Each column (above) lists forward model beta coefficients associated with each voxel for a particular semantic feature
Each column (above) lists fMRI activation (for an ROI) elicited by reading a test sentence
"The light was green"
"The light was green"

Figure 2. Detailed illustration of the computational stages involved in reconstructing semantic features from fMRI activation. The top row corresponds to Figure 1 stage 2, and the middle/bottom row Figure 1 stage 3. pinv corresponds to the Moore–Penrose pseudoinverse as implemented for example by Matlab function pinv.

Anderson et al. 2016), our central interest was instead in the semantic information that could be reconstructed from fMRI activation patterns local to different anatomical regions of interest (ROI).

To this end, we identified voxels within each ROI of the Destrieux atlas (Fischl et al. 2004), segmented the columns corresponding to those voxels from the forward model (beta coefficient) matrix, and then inverted the segmented beta coefficients (again using the Moore–Penrose pseudoinverse). This built a mapping projecting from ROI space back into semantic model space. Matrix multiplication of test sentence fMRI data, with the inverse of the forward model enables semantic vectors to be reconstructed from the test sentences. This process was repeated for each of the 22 ROIs.

This second stage of inverting the forward model can be understood in terms of a second set of multiple regression analyses, each conducted independently for each test sentence (Pereira et al. 2016) as illustrated in Figure 2 (middle and bottom). Here regression is from the forward model (transposed to have number of voxels rows, and number of features columns) to the vector of voxel activation values for a single test sentence (number of voxels rows). The vector of beta coefficients (number of features rows) arising from each second stage regression analysis serves as the reconstruction of the semantic vector for that test sentence: reconstructed semantic features will receive high values if the brain map of forward model beta coefficients for that feature matches the actual activation profile of the test fMRI data, and low values otherwise (see Fig. 2, bottom).

To evaluate the quality of the reconstructed model data for each ROI, we applied a similar procedure to that introduced by Mitchell et al. (2008) which has been in common usage since. For each pair of test sentences the two ROI-based reconstructed semantic model vectors were cross-correlated with the 2 original model vectors using Pearson correlation. The 4 resulting coefficients (2 for matching sentences and 2 for non-matching sentences) were transformed using Fisher's r to z transform (arctanh), as is customary when comparing correlation values. If the sum of values corresponding to the correctly matched reconstructed versus original pairs was greater than the sum for the non-matched pairs, decoding was scored as a success (1), otherwise a failure (0).

This process was repeated, leaving out each possible pair of held out test sentences in turn (28 680 iterations in total). For each participant, a single metric of success was computed for each ROI by taking the mean of the 28 680 test scores. If there is no correspondence between model and fMRI data, a mean accuracy of 0.5 is expected. Statistical significance of decoding performance was estimated for each individual participant using permutation testing. Specifically, prior to running the analysis above, the word-level semantic vectors were shuffled relative to the word-labels (where the label is the written word). For each ROI, sentence-level semantic vectors were constructed according to the original word labels, meaning that sentences were now built from random word-level semantic content. From then on, the leave-2-out analysis proceeded as normal. This was repeated 1000 times over, using different random shuffles each time, to give a null distribution of 1000 mean decoding accuracies. *P*-values associated with (non-randomized) decoding accuracies were calculated as the proportion of chance accuracies greater than or equal to the observed decoding accuracy.

Analysis Overview

We initially undertook a preliminary analysis to identify a network of brain regions processing semantics on which to focus our 4 main analyses. The first of the main analyses tests which elements of sentences' grammatical structure are semantically encoded in different brain regions. We anticipate that some regions will semantically encode all elements of grammatical structure based on the results of Fedorenko et al. (2016) and Nelson et al. (2017). Secondly, we test which sentences are discriminable in different brain regions to confirm that the same sentences were encoded in different regions. Thirdly we test which experiential semantic features can be reconstructed from fMRI activity in different regions. Finally, we identify commonalities in the experiential semantic features that can be reconstructed from different regions. In advance, some degree of correspondence between the results of these analyses is to be anticipated: if different brain regions correlate on the semantic components they reconstruct, then those regions will also correlate on the sentences they can discriminate. However, the strength of each correlation would be difficult to ascertain without explicit tests of each factor.

Results

Identification of a Network of Brain Regions Processing Semantic Information

In a preliminary analysis we identified a network of brain regions engaged in the semantic processing of sentences, based on the ROIs returning the highest mean sentence decoding accuracies across participants. We used permutation testing to identify an individual-level significance threshold associated with $P = 0.01$ and selected all ROIs that had a mean decoding accuracy across participants that was greater than this threshold, i.e., For the selected ROIs, on average, each participant's decoding results were significant at the $P < 0.01$ level.

This yielded 22 regions of interest (ROIs), which it turned out were widely distributed across the cortex. Decoding accuracies and selected ROIs are illustrated in Figure 3. Sixteen ROIs were in the left hemisphere and 6 in the right hemisphere. The set of ROIs encompassed regions associated with processing sentences identified by Pallier et al. (2011), Fedorenko et al. (2016) and Nelson et al. (2017), regions that show a similar profile as speakers of different languages listen to narrative in their native language (Honey et al. 2012), and adhere closely to the spread of the semantic processing network as defined in recent literature (e.g., Binder et al. 2009; Binder and Desai 2011), as well as variants of the "language network" as identified by Fedorenko and Thompson-Schill (2014). The highest scoring ROI was the left superior temporal sulcus, for which decoding accuracies were significant ($P < 0.01$) for all 14 participants. The weakest scoring of the ROIs was the right precuneus, for which 7/14 participants yielded significant results at the $P = 0.01$ level (the cumulative binomial probability of achieving ≥7 results significant at $P = 0.01$ is $P < 0.0001$). Destrieux atlas labels of ROIs, followed in parentheses by the abbreviation used in the subsequent text of this article and the mean ± SD number of voxels across participants in each are as follows:

ctx_lh_S_temporal_sup (**LSTS**, 533±66), ctx_lh_G_temp_sup-Lateral (**LSTG**, 435±49), ctx_lh_G_temporal_middle (**LMTG**, 588 ±62), ctx_lh_G_front_sup (**LSFG**, 1269±111), ctx_rh_S_temporal_sup (**RSTS**, 567±68), ctx_lh_G_pariet_inf-Angular (**LAG**, 441±51), ctx_lh_S_front_inf (**LIFS**, 217±29), ctx_lh_G_precuneus (**LPrCnG**, 397±65), ctx_lh_G_front_inf-Triangul (**LIFGtr**, 197±44), ctx_lh_G_front_middle (**LMFG**, 856±84), ctx_lh_S_front_sup (**LSFS**, 313 ±38), ctx_lh_G_pariet_inf-Supramar (**LSMG**, 515±65), ctx_lh_G_occipital_middle (**LMOG**, 352±36), ctx_lh_S_oc-temp_med_and_Lingual (**LOTLingS**, 187±19), ctx_lh_G_front_inf-Opercular (**LIFGop**, 266±40), ctx_rh_G_front_sup (**RSFG**, 1200±111), ctx_lh_G_temporal_inf (**LITG**, 534±78), ctx_rh_G_temporal_middle (**RMTG**, 650±75), ctx_lh_G_oc-temp_lat-fusifor (**LOTFFG**, 309±42), ctx_rh_G_pariet_inf-Angular (**RAG**, 563±64), ctx_rh_G_front_inf-Triangul (**RIFGtr**, 223±53), ctx_rh_G_precuneus (**RPrCnG**, 413±72).

Elements of Grammatical Structure that are Semantically Encoded in Different Brain Regions

Nine different grammatical structures were identified within the 240 experimental sentences. These were constructed from combinations of the following grammatical elements (where the number in parentheses indicates the number of sentences containing each element): {**S**ubject (240), **V**erb (196), direct **O**bject, (128) Indirect **O**bject (27), **Cop**ula-phrase (44), **Adju**nct (74)}. The 9 different grammatical structures, with examples, were as follows (NB sentences containing a copula-phrase, in which the linking verb was exclusively "was" were not regarded as containing a verb conveying any intrinsic meaning):

Subject, Verb (3): [S: The patient][V: survived]

Subject, Verb, Object (98): [S: The family][V: survived][O: the powerful hurricane]

Subject, Verb, Object, Indirect Object (7): [S: The child][V: gave] [O: the flower][IO: to the artist]

Subject, Verb, Object, Adjunct (23): [S: The child][V: broke] [O: the glass][Adju: at the restaurant]

Subject, Verb, Indirect Object (19): [S: The parent][V: shouted] [IO: at the child]

Subject, Verb, Indirect Object, Adjunct (1) [S: The judge] [V: stayed] [IO: at the hotel][Adju: during the vacation]

Subject, Verb, Adjunct (45): [S: The wealthy family][V: celebrated][Adju: at the party]

Subject, Copula-phrase (39): [S: The family] [Cop: was happy]

Subject, Copula-phrase, Adjunct (5): [S: The school] [Cop: was empty][Adju: during summer]

A full grammatical breakdown of the test sentences is included in Supplementary Materials Table S1, and Table S2

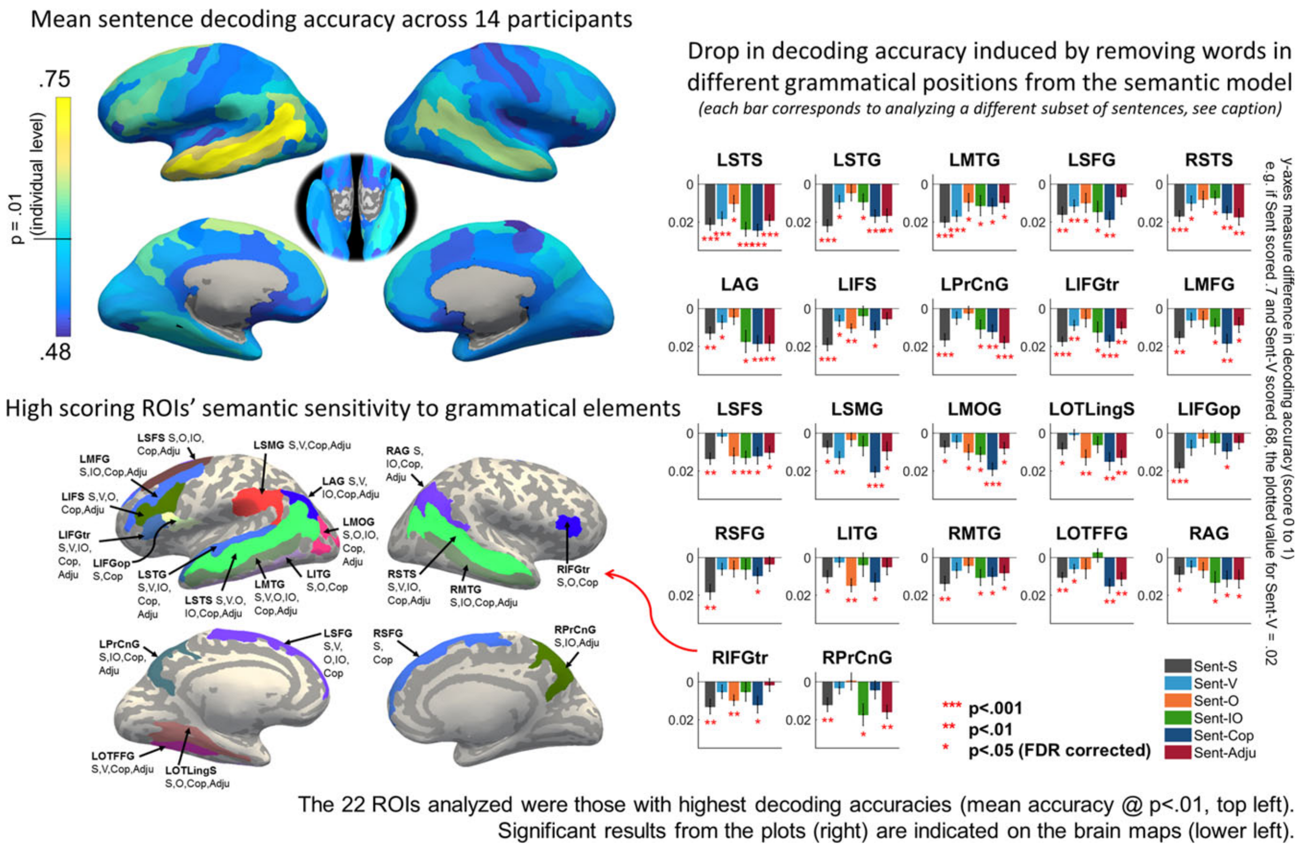


Figure 3. Mean sentence decoding accuracy across participants (top left). Mean ± SEM drop in decoding accuracy induced by withholding semantic vectors associated with each grammatical element from the sentence decoding model for the highest scoring ROIs (right). ROIs are arranged in order of decreasing decoding accuracy (LSTS at the top left gives the strongest result). Significant results from each ROI from the grammatically reduced model analysis (right) are identified on a brain map (bottom left). ROI colors for this plot (bottom left) are arbitrary and used to delineate ROIs. Supplementary Figure 4 illustrates a companion post hoc analysis undertaken on anterior, mid and posterior subregions of LSTS, LSTG, and LMTG. Because different grammatical elements occurred with different frequencies across the sentences (and withholding elements consequently affected different subsets of sentences) different bars correspond to tests on different subsets of the experimental sentences. See main text for further details. The number of test pairs contributing to each comparison were: Sent-S: 28667 tests, Sent-V: 27733 tests, Sent-O: 22464 tests, Sent-IO: 6102 tests, Sent-Cop: 9568 tests, Sent-Adju: 14985 tests.

identifies the frequency that each content word appeared in each grammatical position.

We then tested which regions encode semantic information associated with different elements of sentences' grammatical structure, and in particular regions that represent all elements inclusive (which might represent the sentence constructs observed by Fedorenko et al. 2016). To estimate which grammatical elements were semantically encoded in a brain region, we tested whether excluding the semantic vector associated with each element {S, V, O, IO, Cop, Adju} from the semantic sentence model decreased decoding accuracy. We reasoned that if an element of grammatical structure is not semantically encoded in an ROI, excluding that element from the sentence model should either have no effect on decoding accuracy or even improve it (because semantic information associated with other grammatical elements is irrelevant to decoding that region). We also reasoned that a statistically significant drop in decoding accuracy induced by excluding all 6 grammatical elements in turn would reveal that any of the 6 grammatical elements can semantically activate the region in question. We consider that such a region can represent semantic information provided by a content word in any grammatical position of a sentence. In addition, if decoding is still significantly better than chance when using the reduced sentence model that is missing a grammatical element, this is evidence that semantic information associated with (some of) the remaining grammatical elements of the sentence is also encoded.

In interpreting this analysis, it is first important to point out that a regional drop in decoding accuracy induced by removing a grammatical element from the sentence model does not entail that grammatical information (as opposed to semantics) is explicitly encoded in brain activation. However, one strategy the brain might use to parse sentences into different elements of grammatical structure is to spatially partition semantic information associated with different elements of grammatical structure in the cortex (as there is some evidence for e.g., Frankland and Greene 2015). In this case, a region exclusively dedicated to representing a specific grammatical element would be revealed by a selective decoding deficit induced by removing that element (and possibly a boost in decoding accuracy induced by removing other grammatical elements containing irrelevant information). However, to foreshadow the forthcoming results, the current analysis is consistent with most regions we test representing semantics supplied by words in most grammatical positions.

We: (1) Built model sentence vectors with {S, V, O, IO, Cop, Adju} excluded by taking the featurewise mean of all other content words in each sentence (Fig. 1 stage 4). If the grammatical element contained an adjective (e.g., "The wealthy patient…"), the adjective was also excluded. These vectors are denoted by {Sent-S, Sent-V etc} where Sent refers to the full sentence model. (2) Used each of {Sent-S, Sent-V etc} in place of Sent, to decode the ROI-based reconstructed model data derived in the initial analysis (when fMRI data was regressed on the full sentence model). This was repeated for each ROI and participant. (3) Contrasted per-participant mean decoding accuracies for Sent and each of Sent-{ S, V, O, IndO, Cop, Adju } using paired *t*-tests (df = 13,1-tail, with the prediction that accuracy for the full sentence is greatest). This was repeated for each ROI, and *P*-values were corrected for multiple comparisons by False Discovery Rate (FDR) (Benjamini and Hochberg 1995).

The comparisons between Sent and each experimental condition {Sent-S, Sent-V etc} were restricted to only those held out sentence pairs for which either one of or both sentences in the pair contained the grammatical element (we also present companion results in Supplementary materials for the subset of these tests for which both sentences contained the dropped grammatical element). Because different grammatical elements occurred with different frequencies across the sentences, this meant that {Sent-S, Sent-V etc} were based on pairwise decoding comparisons of different subsets of the 240 sentences (with each subset having a different size). In addition, because removing elements of grammatical structure resulted in occasional cases of semantic sentence vectors for different sentences being made identical (consider removing the subject from “the criminal broke the television” and “the dog broke the television”), cross-validation iterations involving pairs of identical sentence vectors were also excluded from the analysis to avoid biasing results. In practice, there were few such instances, and the most severely affected condition was Sent-S, where 13/28 680 of the pairwise comparisons were deleted from both Sent and Sent-S. Finally the number of pairwise decoding comparisons undertaken on tests of each experimental condition were as follows: Sent-S: 28667 tests, Sent-V: 27733 tests, Sent-O: 22464 tests, Sent-IO: 6102 tests, Sent-Cop: 9568 tests, Sent-Adju: 14985 tests. To recap, each test was restricted to sentence pairs where either one or both sentences originally contained the grammatical element (e.g., for Sent-V 19110/27733 tests involved both sentences containing V, and the remaining 8623/27733 test pairs involved only a single sentence containing V, a single test was deleted due to the creation of identical sentence vectors).

The drops in decoding accuracy incurred by withholding grammatical elements from the sentence model {Sent-S, Sent-V etc} are shown in Figure 3. LSTS and LMTG showed a significant FDR corrected ($P < 0.05$) drop in decoding accuracy incurred by removing each of the 6 grammatical elements, LSTG, LSFG, RSTS, LAG, LIFGtr, LSFS, LMOG showed a significant drop associated with 5 elements, LIFS, LPrCnG, LMFG, LSMG, LOTLingS, RMTG, LOTFFG, RAG showed a significant drop for 4 elements, LITG, RIFGtr and RPrCnG showed a drop for 3 elements, and LIFGop and RSFG showed a drop for 2 elements. Companion results for the subset of tests when both sentences contained the dropped grammatical element are in Supplementary Figure 1. The pattern of results is broadly similar in the key respect that each ROI is sensitive to the removal of a similar selection of grammatical elements, though the magnitude of differences varies for some of the tests.

Aggregating results across contiguous ROIs (Fig. 3), it can be seen that the removal of all 6 elements of grammatical structure incurred a significant drop in decoding accuracy in left temporal, inferior parietal, inferior frontal, and superior frontal brain regions (e.g., LSTS, LMTG, [LAG & LMOG], [LIFG & LIFS], [LSFG & LSFS]). This is evidence that all of these (collated) regions were activated by semantic information exclusive to each grammatical element. Furthermore, following the removal of each grammatical element, discrimination accuracies for each ROI were all still significantly greater than chance-level (0.5), (one sample *t*-tests, df = 13, all $P < 0.001$, FDR corrected). This is evidence that each ROI contained semantic information associated with both the removed element, together with one or more other grammatical elements of the sentences.

An additional analysis (Supplementary Fig. 2) took this one step further to test for semantic information associated with the presence of two or more grammatical elements within activation patterns. Decoding accuracies using the grammatically reduced models that were already missing an element were compared to those achieved using only a single grammatical element—the sentence subject (e.g., Sent-V was contrasted with “S Only” with the expectation that “S Only” would have lower accuracy). The sentence subject was chosen as the single element for experimental convenience because it was the only element to appear in all 240 sentences (and therefore Sent-V, Sent-O, Sent-IO, Sent-Cop and Sent-Adju could all be compared to “S Only”). There was a significant drop in decoding accuracy between each of Sent-V, Sent-O, Sent-IO, Sent-Cop and Sent-Adju when compared to “S Only” in each of LSTS, LSTG, LMTG, LSFG, LSFS, LAG, LIFGtr and LOTLingS. Taken alongside the previous evidence that left temporal, inferior parietal, inferior frontal and superior frontal brain regions encode semantic information associated with any of the 6 grammatical elements, this provides evidence that activation patterns contain semantic information associated with more than two grammatical elements. We therefore infer that left temporal, inferior parietal, inferior frontal and superior frontal brain regions represent semantic information associated with content words in multiple grammatical positions within sentences. These results echo the ECoG-based findings of Fedorenko et al. (2016).

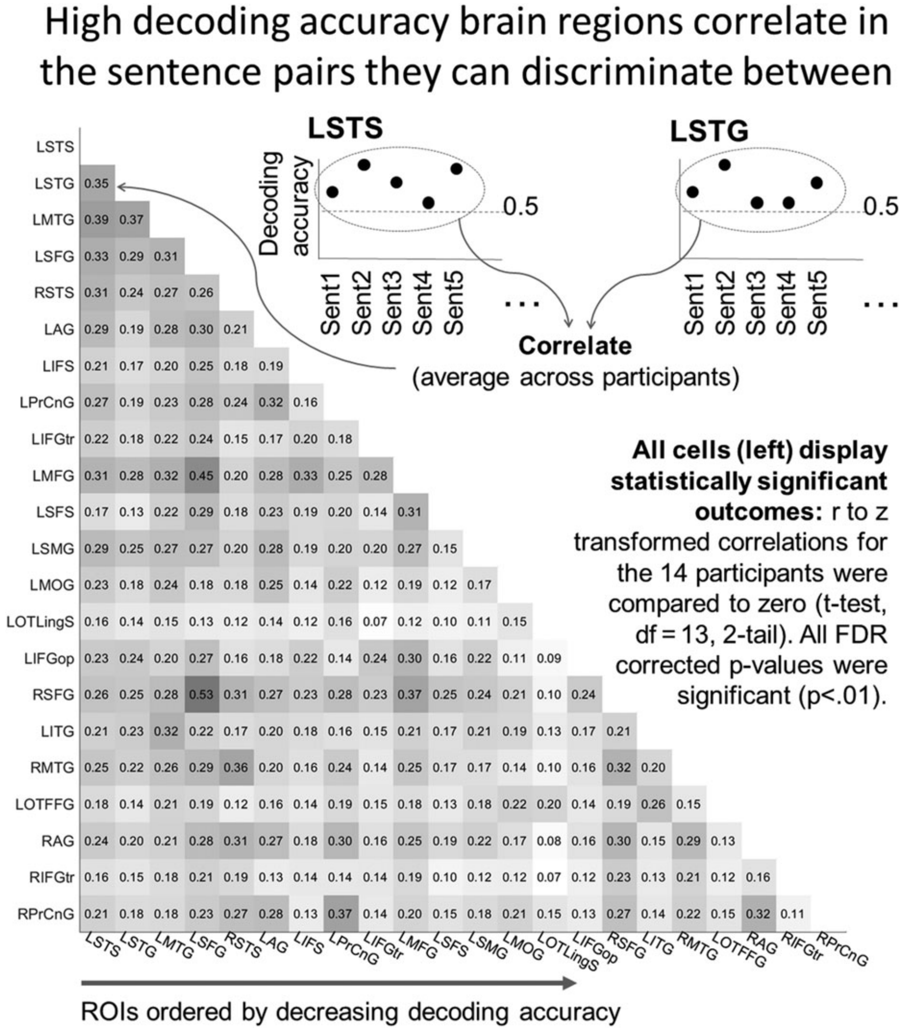
A qualitative observation is that different ROIs show different sensitivities to withholding different grammatical elements, e.g., LMTG and LSTS are sensitive to withholding verbs, and LITG and LOTLingS are especially sensitive to withholding objects (but not verbs). We do not assign importance to differences in the magnitude of drop for different grammatical elements *within* each ROI because these were not experimentally controlled. Different grammatical elements appeared with different frequency across the 240 sentences (e.g., S was in all sentences and IO was in 27) and were also not controlled according to semantic content (e.g., sentences’ subjects were often social roles and rarely places, whereas sentences’ adjuncts tended to be places). This also entails that there is an experimental confound between grammar and semantics in the experimental sentences, which had been initially preselected as experimental materials for the Knowledge Representation in Neural Systems (KRNS) project (Glasgow et al. 2016, www.iarpa.gov/index.php/researchprograms/krns), sponsored by the Intelligence Advanced Research Projects Activity (IARPA). Indeed, to some degree this confound is embedded in the English language; e.g., nouns tend to represent objects and verbs tend to represent actions.

To provide some resolve to this semantics/grammar confound, we ran a parallel analysis that decoded the subset of the sentences that all had a similar grammatical structure (according to the criteria used in this article). These were the 98/240 Subject-Verb-Object sentences. These SVO sentences were decoded using the following sets of sentence vectors: Sent-S, Sent-V, Sent-O, and S only, V only and O only. The results of this analysis are in Supplementary Figure 3. Results across ROIs show a similar general trend to Figure 3. Even when using semantic vectors for just the subject, verb or object of the sentence to decode ROI-based reconstructions of full sentences, discrimination accuracy was significantly greater than 0.5 for all ROIs (one sample *t*-tests, df = 13, all $P < 0.05$, FDR corrected). This strongly suggests that the current decoding results are predominantly accountable to semantic content.

Finally, motivated by theories that the anterior temporal lobe plays a central role in conceptual representation (e.g., Patterson et al. 2007) and additionally by theories that it is central to conceptual combination (Bemis and Pylkkänen 2012; Brennan and Pylkkänen 2012; Westerlund and Pylkkänen 2014; Zhang and Pylkkänen 2015) a post hoc analysis was run that repeated the analysis withholding semantic vectors associated with all 6 grammatical elements, this time on anterior, mid and posterior subregions of LSTS, LSTG, and LMTG. The results of these tests are in Supplementary Figure 4. In summary, no subregion showed statistical sensitivity to the removal of all 6 grammatical elements (unlike LSTS and LMTG when treated as a whole). All posterior subregions and mid LSTS were statistically sensitive to the removal of either 4 or 5 grammatical elements, each including verbs. Collectively the 3 anterior subregions were sensitive to the removal of 5 of the 6 grammatical elements, but not verbs. Beyond this insensitivity to verb removal there were no visually obvious differences distinguishing anterior subregions.

Different Brain Regions Reconstruct the Same Sentences

Although unlikely, it was still possible that different experimental sentences were semantically processed in different brain regions (this would have been more likely if the experimental sentences had been designed to group into distinctive categories such as mathematical concepts and food). To verify correspondence in the sentences processed by different brain regions, we identified which sentence pairs were discriminable in each region and correlated results across regions. More specifically, each individual test sentence was assigned a discriminability score, given by the number of times that sentence was successfully discriminated from another sentence in the 28 680 pairwise sentence comparisons. The maximum discriminability score attainable for each sentence is therefore one less than the number of test sentences (240 − 1 = 239). This process was repeated for each ROI and for each participant, yielding a total of 22 (ROIs) ∗ 14 (participants) discriminability scores for each of the 240 sentences. To test whether a pair of ROIs decoded similar sentence pairs, the set of 240 discriminability scores for each ROI were intercorrelated using Pearson's correlation.



	LSTS	LSTG	LMTG	LSFG	RSTS	LAG	LIFS	LPrCnG	LIFGtr	LMFG	LSFS	LSMG	LMOG	LOTLingS	LIFGop	RSFG	LITG	RMTG	LOTFFG	RAG	RIFGtr	RPrCnG
LSTS																						
LSTG	0.35																					
LMTG	0.39	0.37																				
LSFG	0.33	0.29	0.31																			
RSTS	0.31	0.24	0.27	0.26																		
LAG	0.29	0.19	0.28	0.30	0.21																	
LIFS	0.21	0.17	0.20	0.25	0.18	0.19																
LPrCnG	0.27	0.19	0.23	0.28	0.24	0.32	0.16															
LIFGtr	0.22	0.18	0.22	0.24	0.15	0.17	0.20	0.18														
LMFG	0.31	0.28	0.32	0.45	0.20	0.28	0.33	0.25	0.28													
LSFS	0.17	0.13	0.22	0.29	0.18	0.23	0.19	0.20	0.14	0.31												
LSMG	0.29	0.25	0.27	0.27	0.20	0.28	0.19	0.20	0.20	0.27	0.15											
LMOG	0.23	0.18	0.24	0.18	0.18	0.25	0.14	0.22	0.12	0.19	0.12	0.17										
LOTLingS	0.16	0.14	0.15	0.13	0.12	0.14	0.12	0.16	0.07	0.12	0.10	0.11	0.15									
LIFGop	0.23	0.24	0.20	0.27	0.16	0.18	0.22	0.14	0.24	0.30	0.16	0.22	0.11	0.09								
RSFG	0.26	0.25	0.28	0.53	0.31	0.27	0.23	0.28	0.23	0.37	0.25	0.24	0.21	0.10	0.24							
LITG	0.21	0.23	0.32	0.22	0.17	0.20	0.18	0.16	0.15	0.21	0.17	0.21	0.19	0.13	0.17	0.21						
RMTG	0.25	0.22	0.26	0.29	0.36	0.20	0.16	0.24	0.14	0.25	0.17	0.17	0.14	0.10	0.16	0.32	0.20					
LOTFFG	0.18	0.14	0.21	0.19	0.12	0.16	0.14	0.19	0.15	0.18	0.13	0.18	0.22	0.20	0.14	0.19	0.26	0.15				
RAG	0.24	0.20	0.21	0.28	0.31	0.27	0.18	0.30	0.16	0.25	0.19	0.22	0.17	0.08	0.16	0.30	0.15	0.29	0.13			
RIFGtr	0.16	0.15	0.18	0.21	0.19	0.13	0.14	0.14	0.14	0.19	0.10	0.12	0.12	0.07	0.12	0.23	0.13	0.21	0.12	0.16		
RPrCnG	0.21	0.18	0.18	0.23	0.27	0.28	0.13	0.37	0.14	0.20	0.15	0.18	0.21	0.15	0.13	0.27	0.14	0.22	0.15	0.32	0.11	

Figure 4. Correlation between the 22 ROIs (selected in Fig. 3) in the sentence pairs that could be discriminated.

Repeating for each participant, this yielded 14 correlation coefficients (one per participant) for each pair of ROIs. To evaluate the statistical significance of these 14 coefficients, the coefficients were r-to-z transformed and the resulting values were compared to zero using one sample *t*-tests (df = 13,2-tail). The set of *P*-values across ROI pairs were corrected using False Discovery Rate (Benjamini and Hochberg 1995). Mean±SD correlations of discrimination profiles between ROIs are shown in Figure 4. All comparisons yielded statistically significant results.

Multiple Semantic Features can be Reconstructed from Activation in Multiple Brain Regions

To estimate which semantic features could be reconstructed from different brain regions' activation during sentence comprehension, we correlated model features that were reconstructed from test fMRI data (held out from the training phase) with the features of the original sentence model. Reconstructions of all 65 features for all 240 sentences, for each of the 22 ROIs were accumulated over the leave-2-out cross-validation procedure. As each sentence was held out a total of 239 times over the 28 680 cross-validation iterations (per participant), it was also reconstructed 239 times, and the featurewise mean of these 239 reconstructions was taken to analysis. NB the current analysis could have been approached with less computational effort using a leave-one-out procedure, however as we already had the data available from the leave-2-out analysis, we elected to use this for all analyses.

To evaluate whether a semantic feature was successfully reconstructed the row corresponding to that feature's values across all 240 sentences was extracted from both the reconstructed matrix and the original model matrix. The two row vectors were correlated (Pearson's correlation). This process was repeated for all semantic features, for all ROIs, for all participants (giving 65 correlations, for each ROI for each participant). Results are in Supplementary Figure 5. To test statistical significances of the correlations, all coefficients were r-to-z transformed, and for each ROI, for each feature, the set of 14 r-to-z transformed correlations were compared to zero using one sample t-tests (df = 13, 2-tail). *P*-values were corrected according to False Discovery Rate (Benjamini and Hochberg 1995).

18/65 semantic features were reconstructed significantly by all 22 of the ROIs, and all 22 ROIs significantly reconstructed more than half of the semantic features. The number of semantic features that were significantly reconstructed can be seen to visibly decline (Supplementary Fig. 5) with ROIs' mean decoding accuracies. Temporal ROIs (LSTS, LSTG, and LMTG) significantly reconstructed the majority of features (the maximum was LSTS, at 61/65). Inferior frontal, superior frontal and inferior parietal ROIs significantly reconstructed ≥70% of features.

In general ROIs reconstructed a diverse array of features associated with different modalities of experience (Supplementary Fig. 5), which we explore in more detail in the next section. Features that were most strongly reconstructed apparently related to human traits and socialization (e.g., human, speech, social). The only feature that was not significantly reconstructed by any ROI was *music*. This may be because music was of generally low relevance to the majority of sentences. Whilst people may infer the likelihood of *music* from experimental sentences referencing e.g., "theater", "party", "television", and "cellphone", none of the sentences explicitly described a musical scenario.

Taken together with the results of the previous two analyses, these results suggest that multiple grammatical elements of sentences are semantically encoded within a network of brain regions, that each reconstruct multiple dimensions of experience with words and their referents. Given the limited number of experimental sentences analyzed (and subject matter), it is natural to expect strong components of covariation across semantic features (e.g., *color*, *shape*, and *pattern* might be modulated by the same underlying "visual" component). We next test for underlying semantic dimensions and commonalities in representation across different brain regions.

Underlying Semantic Components and their Commonality Across Brain Regions

Finally, we present the results of factor analyses seeking to (1) uncover the underlying dimensionality of semantic features in the original semantic model; (2) interpret the underlying dimensionality of the 22 ROI-based reconstructions by referencing them back to the latent dimensions of the original model identified in (1); (3) identify commonalities in the underlying dimensions of the reconstructed data across brain regions.

Analyses focused on only the semantic features that could be reconstructed from fMRI data (as identified by a single significant correlation for any of the 22 ROIs in the previous section). This step was originally implemented with the intention of removing noise in the data (i.e., removing semantic features that were not neurally supported), however as this was only *music* this turned out to be a minor adjustment.

PCA was applied to the 64 features (minus *music*) *240 sentences model matrix, after each feature had been z-scored. The first 10 principal components were retained for further investigation and varimax rotated (Fig. 5). These 10 components all had eigenvalues greater than one {16.1, 9, 6.6, 5.1, 3.4, 3.4, 2.7, 2, 1.7, 1.2}, all had apparently straightforward interpretations and they collectively explained 81% of variance in the data. Components 11 and 12 also had eigenvalues greater than one (1.09 and 1.04 respectively), however had less obvious interpretations and consequently were discarded to simplify the upcoming discussion. If they had been included components 11 and 12 would have explained an extra 3% of variance.

We offer the following interpretation of each component (features that receive heaviest loadings for that component are identified in italics): **pc1 "human traits"** (*human, speech, body, face, biomotion*); **pc2 "negative valence"** (*fearful, disgusted, sad, unpleasant, harm, pain*); **pc3 "locomotion"** (*path, lower-limb, fast, motion, toward, away*); **pc4 "place"** (scene, *landmark, long, large*); **pc5 "handling"** (upper-limb, practice, near, needs); **pc6 "hearing"** (*sound, high, low, loud, audition*); **pc7 "eating/drinking"** (*taste, smell, temperature*). **pc8 "timing"** (*time, duration, short, caused, number*). **pc9 "positive valence"** (*happy, pleasant, arousal, attention*). **pc10 "object properties"**, (*vision, shape, small, weight*). These components were subsequently used as references to interpret the underlying dimensions of the ROI-based reconstructed data.

PCA was then undertaken on the reconstructed data for each ROI. As the ROI-based semantic reconstruction converts fMRI into the same representational format across participants (a number of features ∗ number of sentences matrix), we took the opportunity to aggregate the reconstructed data into group-level representations by pointwise averaging ROI-based reconstructions across participants to yield 22 group-level matrices, one for each ROI (and each matrix was featurewise z-scored prior to PCA). This step would have been more difficult to approach on the fMRI data itself, where anatomical/functional differences between different brains may result in relatively similar activity patterns being spatially mismatched even in

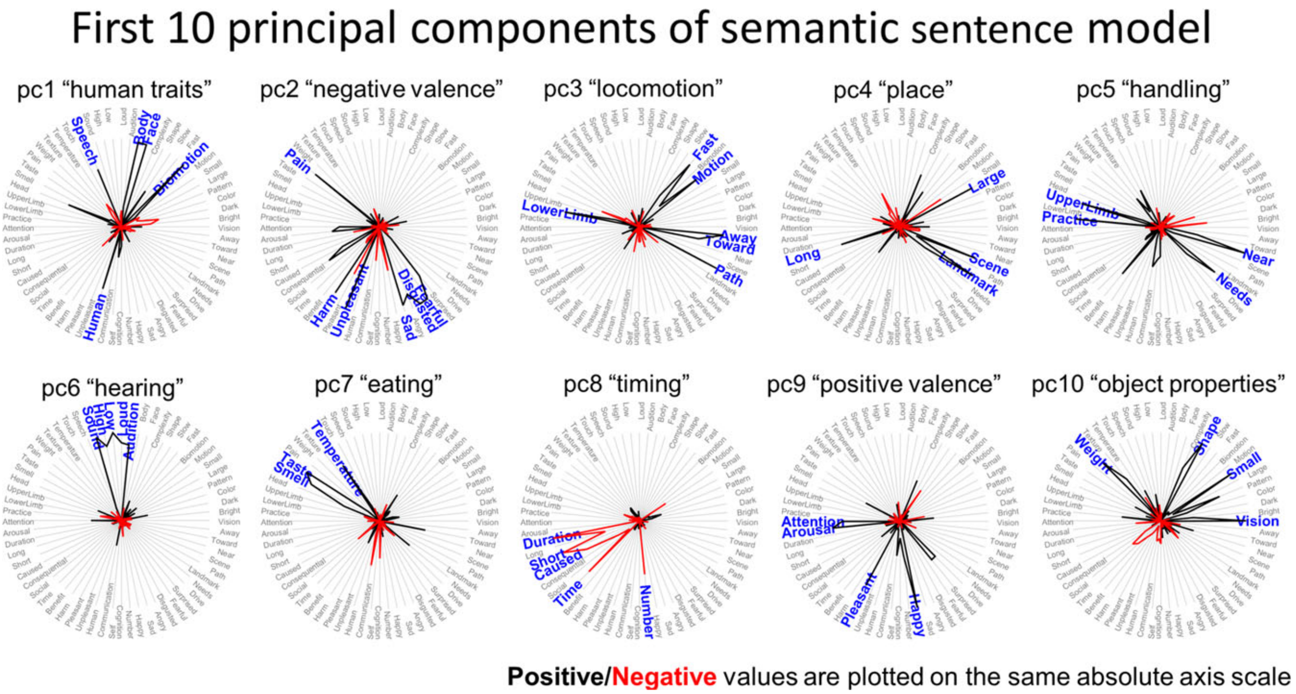


Figure 5. First 10 principal components of the original semantic sentence model varimax rotated. The 10 components collectively explain 81% of variance in the model data. Interpretations of each component are in quotes above each plot. Highlighted in bold blue font are features that had absolute loadings greater than one standard deviation above the mean (absolute) loading. That pc8 has negative loadings is an incidental byproduct of PCA: positive expression of negatively loaded semantic features is induced by negative principal component scores.

the spatially normalized fMRI space. The motivation behind building group-level representations was to expose group-level regularities in semantic representation (that previous work leads us to expect, e.g., Shinkareva et al. 2011; Anderson et al. 2015; Fernandino et al. 2015b; Huth et al. 2016; Anderson et al. 2017) and consequently to reduce noise in the data. We also might expect group-level reconstructions to match better to the semantic model itself, because this too was built from group-level averages of individuals' behavioral ratings. The downside of building group-level representations is, however, that inferences are not guaranteed to generalize to individuals in the group (a related discussion is in Anderson et al. 2015). Nevertheless, comparative results derived at individual-level are included in Supplementary materials that reflect broadly similar patterns, with weaker correlation strengths.

For each ROI, the first 10 principal component scores of the reconstructed data were retained. The 10 components were related back to the previously computed 10 model reference components (“human traits”, “negative valence”, etc.) by cross-correlating each row of each ROI's 10 component *240 sentences matrix of principal component scores with rows of the 10∗240 model reference matrix using Pearson correlation. This resulted in a 10∗10 correlation matrix of r values (and associated P-values) for each of the 22 ROIs. The set of 22∗10∗10 P-values were collectively corrected for multiple comparisons according to False Discovery Rate (Benjamini and Hochberg 1995). Correlation matrices for a selection of 9/22 ROIs are illustrated in Figure 6. The choice to display only 9 ROIs was made to avoid visually cluttering diagrams, and complete results are included in Supplementary Figure 6 for the group-level analysis and Supplementary Figure 7 for the individual-level analysis.

Absolute correlation values were displayed in Figure 6 to ease visualization by remedying cases when otherwise matching model and ROI-based principal components turned out to be flipped in polarity relative to one another (e.g., see Fig. 5 pc8), and model and ROI-based component scores were consequently negatively correlated. We verified that cases of significant negative correlations were a byproduct of flipped eigenvectors by ensuring that negatively correlated component scores coincided with negative correlations in eigenvector loadings. If both component scores and eigenvector loadings are negatively correlated between model and ROI, the matrix multiplication of scores and target reconstructs semantic features that are positively correlated between model and ROI. Consequently, the negative correlation in principal component scores does not signify a true negative relationship between the model and ROI-based reconstruction. Of the 52 statistically significant negative correlations in principal component scores (Supplementary Fig. 6), 51 indeed coincided with negatively correlated eigenvectors. The single exception was the correlation between model pc8 “timing” and LPrCnG pc10.

Common to the majority of ROIs (18/22 ROIs at group-level and 12/22 ROIs at individual-level, see Supplementary Figs 6 and 7) was a significant correlation between the first 4 principal component score row vectors and the first 4 model reference components (“human traits”, “negative valence”, “locomotion”, “place”). This is visible as the diagonal of red stars (significant correlations) in the upper left quadrant of most of the ROI's correlation matrices. This suggests that a common semantic information core, broadly related to these 4 components (for this set of 240 sentences), was regionally encoded across the semantic network (and in particular left lateral temporal, inferior frontal, superior frontal, and inferior parietal regions, respectively).

Beyond these 4 common components, significant correlations in the group-level data were observed between component score row vectors and “handling” in 11/22 ROIs, “hearing”

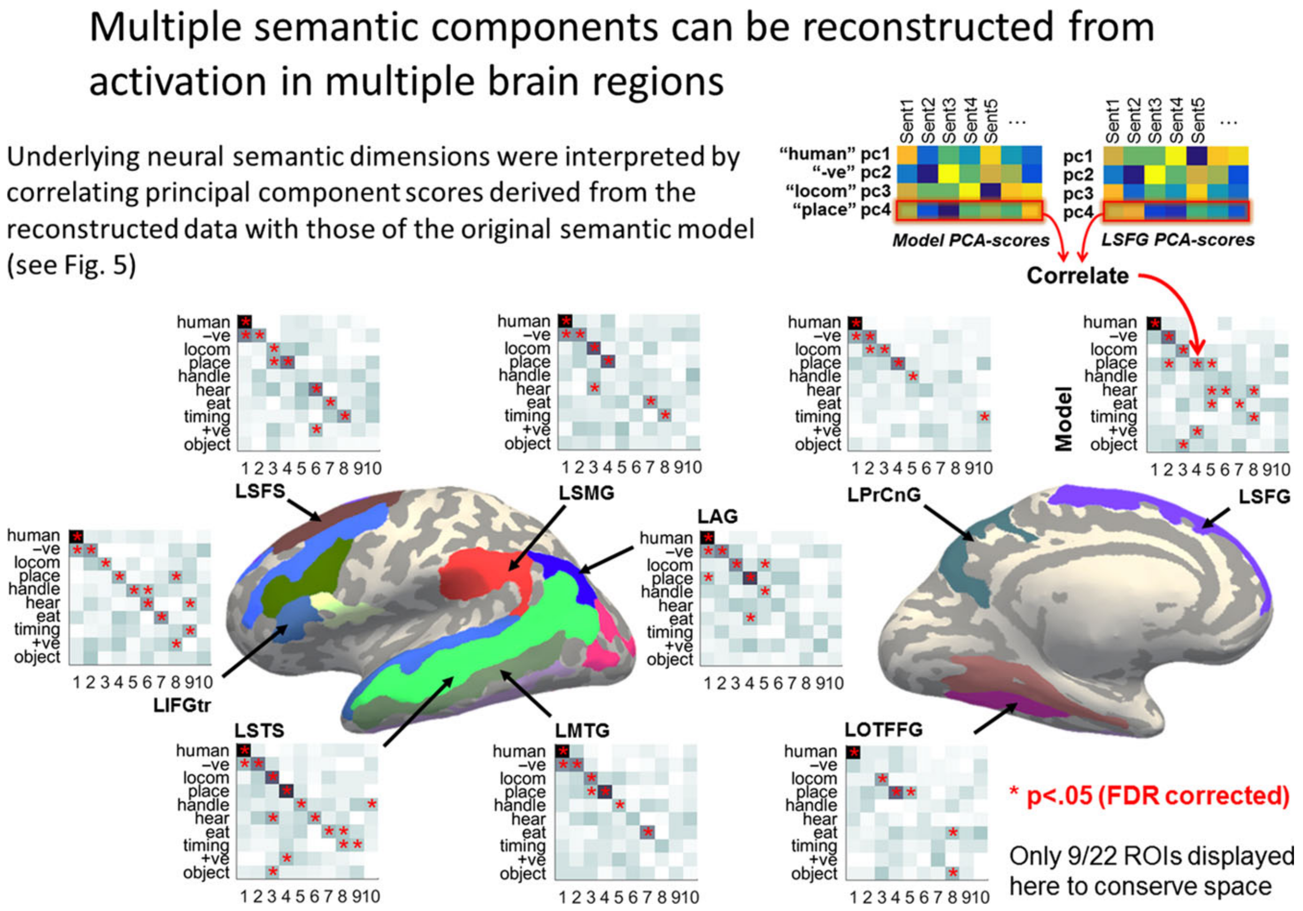


Figure 6. Correlation between principal component scores derived from the original semantic sentence model, and those derived from group-level ROI-based reconstructions of the semantic model. Absolute correlation coefficients are displayed in the matrices to ease visualization. This remedies cases when otherwise matched model and ROI-based principal components turned out to be flipped in polarity (e.g., pc8 in Fig. 5) causing negative correlation coefficients (see main text for further details and tests of this). Dark indicates a high correlation coefficient and light a low coefficient. Correlation coefficients in each correlation matrix were scaled differently to the greyscale range to optimize visual contrast for each plot. Actual ranges are in Supplementary Figure 6, alongside correlation matrices for the remaining 13 ROIs, and a post hoc companion analysis focusing on anterior, mid and posterior regions of LSTS, LSTG, and LMTG.

in 16/22, "eating" in 14/22, "positive valence" in 5/22, and "object properties" in 5/22. As would be expected, ROIs supporting high decoding accuracy (Fig. 3) tended to show greater numbers of significant correlations. The highest scoring ROI (LSTS) had principal component score vectors that correlated with all 10 model reference components, however there was not a one-to-one match between all LSTS component score vectors and those of the model (4 of the LSTS component score vectors each correlated with 2 or 3 model component score vectors). A similar pattern of some one-to-one matches and some one-to-many matches was observed for all ROIs. On average, across all 22 ROIs, 6 component score vectors per ROI correlated with an average of 7 model reference components (i.e., on average across the 22 correlation matrices, 6 columns had one or more stars and 7 rows had one or more stars). Collating across ROIs, 79/220 component scores did not correlate with any model reference component score, 94/220 correlated with a single model reference score, 42/220 correlated with 2, 4/220 correlated with 3, and 1 with 4.

Following up on the previous post hoc analyses that were run using the grammatically reduced models to decode anterior, mid, and posterior subregions of LSTS, LSTG, and LMTG, the current analysis was repeated on the same subregions. The results of this analysis are illustrated in Supplementary Figure 6. In brief, component scores reconstructed from all subregions except for mid MTG significantly correlated with the 4 core model components: "human traits", "negative valence", "locomotion", and "place". Mid MTG reconstructed only "human traits" and "eating". All 3 posterior subregions, and mid LSTS reconstructed around 7 components in total, whereas anterior regions reconstructed 4 or 5 components.

To directly test for commonalities in reconstructed data across ROIs, each row vector of each ROI's 10∗240 matrix of principal-component scores was cross-correlated with the equivalent row vectors for each other ROI, yielding a 10∗10 correlation matrix for each ROI pair. The entire set of correlations were collated together and re-arranged by sorting according to principal component number to facilitate visual assimilation. Correlation coefficients for the 9/22 ROIs illustrated in Figure 6 are in Figure 7. Matrices for all 22 ROIs are in Supplementary Figure 8 (for both group and individual-level data). As for Figure 6 absolute correlation values are displayed to ease visualization (in cases of negative correlation arising from flipped components, as discussed earlier in the description of Fig. 6). Similar to Figure 6, we checked that negative correlations between ROI-based principal component scores coincided with negatively correlated eigenvector loadings, finding this to be the case in 98% of the 2128 instances of significant negative correlation displayed in the complete results shown in Supplementary Fig. 8.

The blocked pattern visible around the correlation matrix diagonal (the actual diagonal is self-correlations) reflects similarities in matched component scores across ROIs, e.g., the upper left dark block indicates scores for the first principal component were highly correlated across ROIs, and likewise for the other blocks. A broadly similar pattern is visible in the individual-level data, albeit with visibly weaker correspondence across ROIs beyond principal component 4 (Supplementary Fig. 8). This is direct evidence for commonalities in multiple latent semantic dimensions across widely distributed regions of the brain's semantic network.

In sum, this section has revealed underlying semantic dimensions in both the model data and ROI-based reconstructed data, referenced the underlying neural dimensions back to those of the model, and demonstrated how multiple latent semantic dimensions are commonly encoded across left temporal, inferior frontal, superior frontal and inferior parietal regions, as well as some right hemispheric homologs. For the current selection of 240 sentences, at least 4 common latent dimensions appear to be interpretable. These relate to "human traits", "negative valence", "locomotion" and "place". However, in interpreting this result it is important to remember that the expression and relative importance of these particular components is liable to differ for a different set of sentences emphasizing different semantic features (e.g., if sentences never mention or imply places, the place component is unlikely to be modulated).

Discussion

The prime contribution of this article is to uncover evidence of a new characteristic of sentence processing: that multiple regions of a cortical network commonly encode the meaning of content words in multiple grammatical positions of sentences. More specifically, that when sentence text is converted to meaning in the brain, multiple dimensions of meaning, first decoded by the brain from words in multiple grammatical positions are represented within left temporal, inferior frontal, superior frontal and inferior parietal cortex (as well as some right hemispheric homologs).

This follows up the recent ECoG results of Fedorenko et al. (2016) and Nelson et al. (2017) that revealed neural signal associated with the construction of sentence meaning across different brain regions. However, unlike the current article, they did not estimate what semantic content is represented within different regions, what semantic content is shared across regions, and where different elements of grammatical structure are semantically encoded. Due to the slow sample rate of fMRI, the current results do not reveal the temporal dynamics of how semantic information emerges across the brain. For instance, similar semantic representations may have been generated in parallel in distributed neural regions, and/or modally- or linguistically-specialized neural modules may have independently generated information that was subsequently channeled throughout the brain. More detailed analyses of neural data recorded using a technique with a higher sampling frequency than fMRI will be necessary to resolve this question.

That meaning is processed across a number of "semantic hubs" that integrate information across different modalities (e.g., posterior temporal, anterior temporal, inferior temporal, inferior frontal and inferior parietal regions) is a common component of most contemporary proposals of semantic processing (e.g., Binder et al. 2009; Binder and Desai 2011; Pulvermüller 2013; Lambon Ralph et al. 2017). However, what aspects of meaning are processed by different hubs and what information is common across them is poorly understood. Previous work

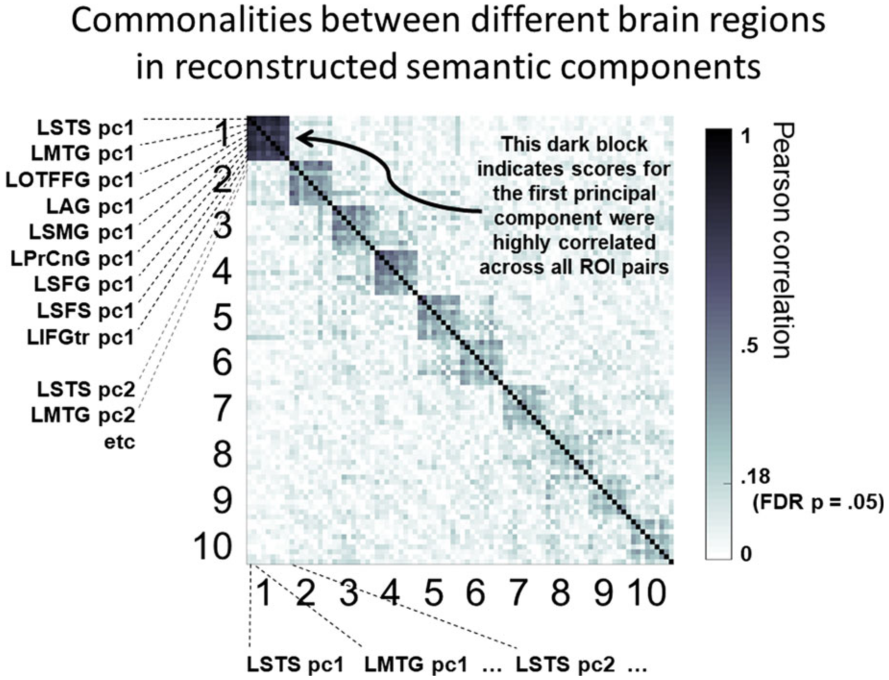


Figure 7. Correlations between the first 10 principal component score vectors derived from the group-level semantic representations reconstructed from each of the 9 ROIs illustrated in Figure 6. Complete data for all 22 ROIs is in Supplementary Figure 8. Absolute correlation coefficients are displayed in the matrix to simplify visualization. This remedies cases when otherwise matched ROI-based principal components turned out to be flipped in polarity (e.g., pc8 in Fig. 5) causing negative correlation coefficients (see main text for further details and tests of this).

targeting cross/multimodal conceptual representation has predominantly focused on fMRI activation elicited by isolated concrete nouns, and either detected the whereabouts of common neural responses to stimuli presented in different modalities, e.g., names and pictures of the same objects (e.g., Akama et al. 2012; Devereux et al. 2013; Fairhall and Caramazza 2013; Shinkareva et al. 2011; Simanova et al. 2014), or analyzed concrete noun activation using models with more restricted coverage (e.g., 5 sensorimotor attributes in Fernandino et al. 2015a, 2015b, 2016, or visual and textual distributional semantic models in Anderson et al. 2015). Where Anderson et al. (2016) demonstrated how a broader set of 65 experiential attributes could be used to predict fMRI activation elicited in sentence comprehension, the current article substantially advances beyond this by linking semantic fMRI activation to elements of grammatical structure, and in so doing newly uncovering evidence that content words in multiple grammatical positions of sentences are semantically encoded in multiple distributed brain regions; newly identifying which experiential attributes support decoding (and can be reconstructed from different regions); and newly revealing commonalities in the semantic representation of sentences across multiple distributed brain regions.

The breadth and distribution of common semantic content observed in this study is perhaps greater than would be expected from prior work that placed emphasis on identifying distinct roles of particular anatomical regions. With specific respect to sentence comprehension, Frankland and Greene (2015) have recently pinpointed a region in the left midsuperior temporal cortex where activation associated with the agent and patient of a small set of sentences is spatially distinct. These authors also showed that sentence manipulations that modulate affective connotations (“the grandfather kicked the baby” versus “the baby kicked the grandfather”) modulate neural activation in the amygdala. Whilst our results are compatible with this architecture, and indeed in Figure 3 we observe highest decoding accuracy in superior temporal regions, they would additionally predict that semantic information associated with sentences’ subject/object and valence is also locally available in many other regions in the brain. Indeed, this prediction receives some external support from Wang et al.’s (2016) analysis of fMRI cued by a small set of videos, where agent and patient were switched and distinct agent/patient activation was revealed in different brain regions. The broader spread of semantic information decoded in the current study compared to Frankland and Greene (2015) is likely to stem from both the greater power afforded to the current analysis by analyzing 240 comparatively diverse sentences (rather than 6), and our analysis of larger regions of interest (compared to their searchlight analysis).

More generally, an extensive body of literature has documented functional specificities associated with different cortical regions/networks in tasks related to semantic processing/language. These include brain regions/networks that are selective for semantic categories (such as animals/tools e.g., Martin et al. 1996), object features (such as shape, color e.g., Martin 2016), body parts (e.g., Hauk et al. 2004), actions (e.g., Desai et al. 2009), valence (e.g., Vigliocco et al. 2014), grammatical classes (such as nouns/verbs e.g., Caramazza and Hillis 1991, also interpreted as object/actions e.g., Vigliocco et al. 2011). Our identification of commonalities in semantic content across regions should not be taken to dispute results demonstrating functional specificities, or to be claiming that the regions exhibiting commonalities in this article do not have their own functional specializations (indeed the 22 ROIs that were the focus of this article can be qualitatively observed to vary in the semantic features/grammatical elements they are activated by). However, irrespective of the functional specificity of different regions, or which regions were initially responsible for turning text into semantic information, the current results do provide evidence that a common core of semantic information is broadly available across the semantic network during sentence comprehension (in particular left temporal, inferior and superior frontal and inferior parietal cortex as seen in Figs 3, 6 and 7). As much of the previous literature has been based on studies of isolated concepts (with no specified context) it will be interesting to see whether the spread of common information across the brain observed in the current sentence decoding article is related to the additional task demands associated with processing multiple words in sentences and/or integrating semantic content according to grammatical structure, and/or the result of analyzing a comparatively large dataset (of 240 sentences constructed from 242 different content words).

The detection of latent semantic features spanning different brain regions observed in this article is not without precedent in the literature. In a factor analysis of neural activation patterns associated with 60 isolated nouns, Just et al. (2010) identified semantic components that they related to *manipulation*, *shelter*, and *eating*, located in 3–4 lobes, and more recently used the same components as the basis for decoding brain activation elicited by reading short sentences themed around manipulation, shelter or eating (Just et al. 2017). Yang et al. (2017) used principal components analysis of fMRI data to select voxel clusters to use as the basis for predicting fMRI activation associated with 60 sentences across languages. Similar to the current study, each component was linked to voxels in distributed brain regions, and components were interpreted as relating to *people*, *actions*, *feelings*, and *places*. More generally, analyses that have applied semantic models to factor fMRI into brain maps associated with different semantic features have consistently detected patterns of high feature weightings that are distributed across the cortex (Mitchell et al. 2008; Fernandino et al. 2015b; Anderson et al. 2016; Huth et al. 2016). The current article has gone beyond this previous work by showing how multiple semantic features associated with words in multiple grammatical positions of sentences can be reconstructed from activation in multiple brain regions (Figs 3, 6 and 7), and identifying commonalities in sentence discriminability across these regions (Fig. 4). Though we acknowledge that the respective findings of the current analyses are inter-related: if two brain regions correlate on semantic feature reconstructions then they are going to correlate in the sentence pairs they can decode.

The current modeling approach is limited in the following respects. Firstly, whilst the “bag-of-words” word combination procedure benefits from its simplicity in interpretation, it is an oversimplification of semantic composition. The effects of word order, syntax and morphology are ignored (see Anderson et al. 2016 for a related discussion of the deficiencies of the bag-of-words approach). For instance, demonstrating that removing semantic information associated with all grammatical elements from the model impairs the decoding of a region, does not entail that the region also explicitly encodes grammar. In preliminary analyses, various attempts were made to grammatically constrain how word-level semantic vectors were combined together into sentence representations. Disappointingly, decoding results were worse than the current bag-of-words approach. These attempts may have been compromised by having insufficient/inadequate training data, or possibly the slow sample rate of fMRI. Related to this, to better understand the interaction between semantics and grammar it would be desirable in the future to test sets of sentences that better balance the frequency with which different grammatical elements appear across sentences, and the frequency that different semantic categories appear in different grammatical positions.

Secondly, despite benefitting from its neurobiological interpretability (see Anderson et al. 2016; Binder et al. 2016), the experiential attribute model is likely to prove limited in its ability to capture some abstract linguistic concepts. Compared to computational models based on the distributional statistics of words in huge text corpora (which have been in frequent use in the neuroimaging literature since Mitchell et al. 2008), the experiential attribute model is advantaged in that it will reliably capture fundamental components of meaning that are so obvious that people are disinclined to report them in writing. For instance, it is rarely useful to point out the *color* and *shape* of a banana in natural text, and consequently this information is liable to be underrepresented in text-based distributional models. However, the experiential attribute model is disadvantaged when it comes to semantic structure that is only available through language (e.g., the phylogeny and natural history of the banana). Consequently, there is likely to be benefit in the future to combining both experiential and linguistic information in modeling. Andrews et al. (2009) demonstrated that the different data sources capture complementary information in describing behavioral data. In the context of fMRI data cued by written concrete nouns, Anderson et al. (2013) and Anderson et al. (2015) have identified benefits to combining visual and text-based distributional semantic models. Additionally, text-based models have formed the only modeling basis thus far to discriminate neural activation patterns elicited by a selection of abstract nouns (Anderson et al. 2017). Indeed, work in progress on the current sentence dataset suggests that experiential and text-based distributional semantic information are complementary.

In conclusion, the current article has revealed new commonalities in the neural encoding of the semantic components of sentences across widely distributed brain regions, specifically, how multiple dimensions of peoples' experience with words (and their referents), arising from words in multiple grammatical positions, can be reconstructed from fMRI activation in multiple brain regions. In so doing, this has provided a methodological foundation for both quantifying the importance of different semantic features and of words from different grammatical positions to the brain wide representation of sentence concepts. This also provides further validation for the experiential attribute model, with many questions left open for future investigation.

Supplementary Material

Supplementary material is available at *Cerebral Cortex* online.

Funding

This work was supported in part by a Schmitt Program on Integrative Neuroscience (University of Rochester Medical Center) award, a University of Rochester Center for Language Sciences Fellowship, and the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory under Grant FA8650-14-C-7357.

Notes

We thank two anonymous reviewers for their insightful comments and suggestions. *Conflict of Interest*: None declared.

References

Akama H, Murphy B, Na L, Shimizu Y, Poesio M. 2012. Decoding semantics across fMRI sessions with different stimulus modalities: a practical MVPA study. Front Neuroinform. 6: 24.

Anderson AJ, Binder JR, Fernandino L, Humphries CJ, Conant LL, Aguilar M, Wang X, Doko D, Raizada RDS. 2016. Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. Cereb Cortex. doi:10.1093/cercor/bhw240.

Anderson AJ, Bruni E, Bordignon U, Poesio M, Baroni M. 2013. Of words, eyes and brains: correlating image-based distributional semantic models with neural representations of concepts. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013); Seattle, WA: Association for Computational Linguistics. pp. 1960–1970.

Anderson AJ, 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.

Anderson AJ, Kiela D, Clark S, Poesio M. 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. Trans Assoc Comput Linguist. 5:17–30.

Andrews M, Vigliocco G, Vinson D. 2009. Integrating experiential and distributional data to learn semantic representations. Psychol Rev. 116(3):463–498.

Baron SG, Osherson D. 2011. Evidence for conceptual combination in the left anterior temporal lobe. Neuroimage. 55: 1847–1852. doi:10.1016/j.neuroimage.2011.01.066.

Bemis DK, Pylkkänen L. 2012. Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. Cereb Cortex. 23(8): 1859–1873. doi:10.1093/cercor/bhs170.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 57(1):289–300.

Binder JR, Conant LL, Humphries CJ, Fernandino L, Simons S, Aguilar M, Desai R. 2016. Toward a brain-based componential semantic representation. Cogn Neuropsychol. 33(3–4): 130–174.

Binder JR, Desai RH. 2011. The neurobiology of semantic memory. Trends Cogn Sci. 15(11):527–5.

Binder JR, Desai RH, Graves WW, Conant LL. 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex. 19:2767–2796.

Brennan J, Pylkkänen L. 2012. The time-course and spatial distribution of brain activity associated with sentence processing. Neuroimage. 60(2):1139–1148. doi:10.1016/j.neuroimage.2012.01.030.

Caramazza A, Hillis A. 1991. Lexical organization of nouns and verbs in the brain. Nature. 349(6312):788–90.

Chang KM, Mitchell TM, Just MA. 2010. Quantitative modeling of the neural representations of objects: how semantic feature norms can account for fMRI activation. Neuroimage. 56: 716–727.

Desai R, Binder JR, Conant LL, Seidenberg MS. 2009. Activation of sensory-motor and visual areas by sentence comprehension. Cereb Cortex. 20:468–478.

Devereux B, Kelly C, Korhonen A. 2010. Using fMRI activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In: Murphy B, Chang KK, Korhonen A, editors. Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics. Los Angeles, USA: Association for Computational Linguistics. p. 70–78.

Devereux BJ, Clarke A, Marouchos A, Tyler LK. 2013. Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. J Neurosci. 33(48):18906–18916.

Fairhall SL, Caramazza A. 2013. Brain regions that represent amodal conceptual knowledge. J Neurosci. 33:10552–10558.

Fedorenko E, Scott TL, Brunner P, Coon WG, Pritchett B, Schalk G, Kanwisher N. 2016. Neural correlate of the construction of sentence meaning. Proc Nat Acad Sci USA. 113(41): E6256–E6262.

Fedorenko E, Thompson-Schill SL. 2014. Reworking the language network. Trends Cogn Sci. 18(3):120–126.

Fernandino L, Binder JR, Desai RH, Pendl SL, Humphries CJ, Gross WL, Conant LL, Seidenberg MS. 2015a. Concept representation reflects multimodal abstraction: a framework for embodied semantics. Cereb Cortex. doi:10.1093/cercor/bhv02.

Fernandino L, Humphries CJ, Conant LL, Seidenberg MS, Binder JR. 2016. Heteromodal cortical areas encode sensory-motor features of word meaning. J Neurosci. 36(38):9763–9769.

Fernandino L, Humphries CJ, Seidenberg MS, Gross WL, Conant LL, Binder JR. 2015b. Predicting brain activation patterns associated with individual lexical concepts based on five sensory-motor attributes. Neuropsychologia. doi:10.1016/j.neuropsychologia.2015.04.009.

Fischl B, van der Kouwe A, Destrieux C, Halgren E, Segonne F, Salat DH, Busa E, Seidman LJ, Goldstein J, Kennedy D, et al. 2004. Automatically parcellating the human cerebral cortex. Cereb Cortex. 14:11–22.

Frankland SM, Greene JD. 2015. An architecture for encoding sentence meaning in left mid-superior temporal cortex. Proc Natl Acad Sci USA. 112(37):11732–11737. doi:10.1073/pnas.1421236112.

Glasgow K, Roos M, Haufler A, Chevillet M, Wolmetz M. 2016. Evaluating semantic models with word-sentence relatedness. arXiv:1603.07253.

Hauk O, Johnsrude I, Pulvermüller F. 2004. Somatotopic representation of action words in human motor and premotor-cortex. Neuron. 41(2):301–7.

Honey CJ, Thompson CR, Lerner Y, Hasson U. 2012. Not lost in translation: neural responses shared across languages. J Neurosci. 32(44):15277–15283.

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

Just MA, Cherkassky VL, Aryal S, Mitchell TM. 2010. A neurosemantic theory of concrete noun representation based on the underlying brain codes. PLoS One. 5(1):e8622.

Just MA, Wang J, Cherkassky VL. 2017. Neural representations of the concepts in simple sentences: concept activation prediction and context effect. Neuroimage. 157:511–520.

Lambon Ralph MA, Jefferies E, Patterson K, Rogers TT. 2017. The neural and computational bases of semantic cognition. Nat Rev Neurosci. 18:42–55.

Lau EF, Phillips C, Poeppel D. 2008. A cortical network for semantics: (de)constructing the N400. Nat Rev Neurosci. 9: 920–933.

Martin A. 2016. GRAPES—Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. Psychon Bull Rev. 23:979–990.

Martin A, Wiggs CL, Ungerleider LG, Haxby JV. 1996. Neural correlates of category-specific knowledge. Nature. 379(6566):649.

Mitchell J, Lapata M. 2010. Composition in distributional models of semantics. Cogn Sci. 34(8):1388–1439.

Mitchell TM, Shinkareva SV, Carlson A, Chang K-M, Malave VL, Mason RA, Just MA. 2008. Predicting human brain activity associated with the meaning of nouns. Science. 320: 1191–1195.

Nelson MJ, Karoui IE, Giber K, Yang X, Cohen L, Koopman H, Cash SS, Lionel Naccache L, Hale JT, Pallier C, et al. 2017. Neurophysiological dynamics of phrase-structure building during sentence processing. Proc Natl Acad Sci USA. 114: E3669–E3678. doi:10.1073/pnas.1701590114.

Pallier C, Devauchelle A-D, Dehaene S. 2011. Cortical representation of the constituent structure of sentences. Proc Natl Acad Sci USA. 108(6):2522–2527.

Patterson K, Nestor PJ, Rogers TT. 2007. Where do you know what you know? The representation of semantic knowledge in the human brain. Nat Rev Neurosci. 8:976–987.

Pereira F, Botvinick M, Detre G. 2013. Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. Artif Intell. 194:240–252.

Pereira F, Lou B, Pritchett B, Kanwisher N, Botvinick M, Fedorenko E. 2016. Decoding of generic mental representations from fMRI data using word embeddings. bioRXiv. doi:10.1101/057216.

Pulvermüller F. 2013. How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. Trends Cogn Sci. 17(9):458–470.

Simanova I, Hagoort P, Oostenveld R, Van Gerven MAJ. 2014. Modality-independent decoding of semantic information from the human brain. Cereb Cortex. 24:426–434.

Shinkareva SV, Malave VL, Mason RA, Mitchell TM, Just MA. 2011. Commonality of neural representations of words and pictures. Neuroimage. 54:2418–2425.

Sudre G, Pomerleau D, Palatucci M, Wehbe L, Fyshe A, Salmelin R, Mitchell T. 2012. Tracking neural coding of perceptual and semantic features of concrete nouns. Neuroimage. 62: 451–463.

Vigliocco G, Kousta ST, Della Rosa PA, Vinson DP, Tettamanti M, Devlin JT, Cappa SF. 2014. The neural representation of abstract words: the role of emotion. Cereb Cortex. 24(7):1767–1777.

Vigliocco G, Vinson DP, Druks J, Barber H, Cappa SF. 2011. Nouns and verbs in the brain: a review of behavioural, electrophysiological, neuropsychological and imaging studies. Neurosci Biobehav Rev. 35(3):407–26.

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. 33(3–4):257–264.

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(11):e11257.

Westerlund M, Pylkkänen L. 2014. The role of the left anterior temporal lobe in semantic composition vs. semantic memory.

Neuropsychologia. 57:59–70. doi:10.1016/j.neuropsychologia.2014.03.001.

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.

Zhang L, Pylkkänen L. 2015. The interplay of composition and concept specificity in the left anterior temporal lobe: an MEG study. Neuroimage. 111:228–240. doi:10.1016/j.neuroimage.2015.02.028.