Behavioral/Cognitive

Deep Artificial Neural Networks Reveal a Distributed Cortical Network Encoding Propositional Sentence-Level Meaning

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

[1]Department of Neuroscience, University of Rochester, Rochester, New York 14642, [2]Del Monte Institute for Neuroscience, University of Rochester, Rochester, New York 14642, [3]Facebook AI Research, New York, New York 10003, [4]Department of Neurology, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, [5]Department of Brain and Cognitive Sciences, University of Rochester, Rochester, New York 14627, [6]Department of Linguistics, University of Rochester, Rochester, New York 14627, and [7]Department of Biomedical Engineering, University of Rochester, Rochester, New York 14627

Understanding how and where in the brain sentence-level meaning is constructed from words presents a major scientific challenge. Recent advances have begun to explain brain activation elicited by sentences using vector models of word meaning derived from patterns of word co-occurrence in text corpora. These studies have helped map out semantic representation across a distributed brain network spanning temporal, parietal, and frontal cortex. However, it remains unclear whether activation patterns within regions reflect unified representations of sentence-level meaning, as opposed to superpositions of context-independent component words. This is because models have typically represented sentences as "bags-of-words" that neglect sentence-level structure. To address this issue, we interrogated fMRI activation elicited as 240 sentences were read by 14 participants (9 female, 5 male), using sentences encoded by a recurrent deep artificial neural-network trained on a sentence inference task (InferSent). Recurrent connections and nonlinear filters enable InferSent to transform sequences of word vectors into unified "propositional" sentence representations suitable for evaluating intersentence entailment relations. Using voxelwise encoding modeling, we demonstrate that InferSent predicts elements of fMRI activation that cannot be predicted by bag-of-words models and sentence models using grammatical rules to assemble word vectors. This effect occurs throughout a distributed network, which suggests that propositional sentence-level meaning is represented within and across multiple cortical regions rather than at any single site. In follow-up analyses, we place results in the context of other deep network approaches (ELMo and BERT) and estimate the degree of unpredicted neural signal using an "experiential" semantic model and cross-participant encoding.

Key words: distributional semantics; fMRI; lexical semantics; sentence comprehension; voxelwise encoding; word embedding

Significance Statement

A modern-day scientific challenge is to understand how the human brain transforms word sequences into representations of sentence meaning. A recent approach, emerging from advances in functional neuroimaging, big data, and machine learning, is to computationally model meaning, and use models to predict brain activity. Such models have helped map a cortical semantic information-processing network. However, how unified sentence-level information, as opposed to word-level units, is represented throughout this network remains unclear. This is because models have typically represented sentences as unordered "bags-of-words." Using a deep artificial neural network that recurrently and nonlinearly combines word representations into unified propositional sentence representations, we provide evidence that sentence-level information is encoded throughout a cortical network, rather than in a single region.

Received May 11, 2020; revised Feb. 3, 2021; accepted Feb. 22, 2021.

Author contributions: A.J.A. and E.C.L. designed research; A.J.A., J.R.B., L.F., C.J.H., L.L.C., R.D.S.R., and E.C.L. performed research; A.J.A., L.F., and C.J.H. analyzed data; A.J.A. wrote the first draft of the paper; A.J.A. wrote the paper; D.K. and S.G. contributed unpublished reagents/analytic tools; D.K., L.F., and E.C.L. edited the paper.

A.J.A. and E.C.L. were supported by a University of Rochester Medical Center Schmitt Program on Integrative Neuroscience award. J.R.B., L.F., C.J.H., L.L.C., R.D.S.R., and A.J.A. were supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory under grant FA8650-14-C-7357. R.D.S.R. was supported by NSF CAREER award 1652127.

The authors declare no competing financial interests.

Correspondence should be addressed to Andrew James Anderson at aander41@ur.rochester.edu.

https://doi.org/10.1523/JNEUROSCI.1152-20.2021

Introduction

Sentence comprehension is known to engage a distributed cortical network, spanning temporal, parietal, and inferior/superior temporal cortex (Lau et al., 2008; Binder et al., 2009). However, how sentence-level meaning is constructed from words and represented throughout this "semantic network" remains weakly understood. Pioneering studies linked the semantic composition of words to particular brain regions, such as the anterior temporal lobe (e.g., Baron and Osherson, 2011; Bemis and Pylkkänen, 2011; Westerlund and Pylkkänen, 2014; Zhang and Pylkkänen,

2015) or mid-superior temporal cortex (Frankland and Greene, 2015). However, recent studies suggest a more distributed cortical operation. Word-by-word construction of sentence meaning is marked by electrocorticographic activation across distributed cortical regions (Fedorenko et al., 2016; Nelson et al., 2017), and electro/magnetoencephalographic representations of nouns in-context have been detected in inferior frontal, temporal, and inferior parietal cortex (Lyu et al., 2019). Relatedly, fMRI studies of sentences/narratives have revealed that distributed cortical regions encode similar components of semantic information (Wehbe et al., 2014; Huth et al., 2016; Anderson et al., 2017a; de Heer et al., 2017; Yang et al., 2017; Pereira et al., 2018; Deniz et al., 2019) as stimulated by words regardless of their grammatical role/sentence position (Anderson et al., 2019a).

While fMRI studies strongly suggest that sentence-level semantics are encoded within and across multiple cortical regions, they have largely fallen short of exposing unified sentence-level representations in one critical respect. To identify semantic information in fMRI, studies usually rely on models representing sentences as superpositions of context-independent words. Because such "bag-of-words" (BoW) models ignore context effects, word order, and syntactic structure, it remains unclear whether the fMRI representations they capture reflect sentence-level semantics, as opposed to context invariant word activation elicited in early-stage comprehension (e.g., Swinney, 1979; Tanenhaus et al., 1979; Till et al., 1988).

Computational linguistics research has begun to address the limitations of BoW approaches, using deep artificial neural networks to create sentence representations reflecting within-sentence contexts and word order (e.g., Conneau et al., 2017; Peters et al., 2018; Subramanian et al., 2018; Devlin et al., 2019). These models present a new opportunity to characterize how sentence meaning is encoded across the cortex. We here exploit these models to test for evidence that representations of sentence-level meaning are encoded in multiple regions of the semantic network.

We examined an fMRI dataset scanned as 14 participants read 240 sentences (Anderson et al., 2017a). We used a voxelwise encoding modeling approach to conduct a controlled comparison of how semantic models reflecting different sentence characteristics contributed to predicting fMRI activation throughout the language network (Fedorenko et al., 2010). Models implemented different "composition functions" that combined words into sentences with/without contextual/grammatical/sequential information. Critically, all composition functions operated on the same word-level semantic input, modeled by GloVe (Pennington et al., 2014). Thus, differences in models' ability to explain fMRI activation solely reflected the characteristics of composition. GloVe approximates word-level meaning using numeric vectors of values reflecting how often the modeled word co-occurred with other words across a large corpus of text. Since words with similar meanings tend to appear in similar linguistic contexts, they end up with similar vector representations.

To model sentence-level semantics, we primarily focused on InferSent (Conneau et al., 2017). InferSent is a recurrent nonlinear deep network optimized to produce "propositional" sentence representations for classifying intersentence entailment relationships. InferSent has yielded state-of-the-art performance in several natural language processing tasks and in decoding fMRI activation elicited by sentences, albeit at a whole-brain level (Sun et al., 2019). We focused on InferSent because: (1) sentence entailment has a close tie to the traditional notion of sentences as propositions; and (2) InferSent represents sentences by combining word-level models (e.g., GloVe) that are well established in the neuroimaging literature (since Mitchell et al., 2008; J. Wang et al., 2017; Pereira et al., 2018; Anderson et al., 2019a) with a relatively simple deep network architecture.

We hypothesized that InferSent would predict activation patterns in fMRI data across the semantic network that cannot be accounted for by the baseline models, evidencing that propositional sentence-level meaning is encoded throughout this network rather than being localized to a single site. In follow-up analyses, we tested whether high-performance deep networks with more complex architectures and more parameters afforded stronger fMRI predictions. These networks were "Embeddings from language models" (ELMo) (Peters et al., 2018) and "Bidirectional encoder representations from transformers" (BERT) (Devlin et al., 2019). Finally, we tested for hitherto unpredicted fMRI signal using an "experiential" behavioral semantic model, and cross-participant encoding.

Materials and Methods

Overview

We reanalyzed an fMRI dataset scanned as 14 people read 240 sentences describing everyday situations (Anderson et al., 2017a; and summarized below). Sentences were 3-9 words long and formed from 242 different content words. Ten of the participants saw the set of sentences repeated 12 times in total, and the remaining 4 participants who attended half the number of visits saw the sentences 6 times. Sentence order was randomly shuffled each time. Following standard fMRI preprocessing steps (detailed in later sections), each sentence was represented as a single fMRI volume per participant.

In our primary analyses, we tested how well fMRI sentence activation patterns could be predicted using InferSent-based representations of propositional sentence-level semantics comparative to a series of baseline models (including BoW) that reflected simple rule-based strategies for combining word-level semantic representations into sentences (Fig. 1). We additionally included three other control models in the analyses that captured the grammatical structure and the visual appearance of sentence stimuli. Prediction was implemented using voxelwise encoding modeling with ridge regression (Hoerl and Kennard, 1970) in a leave-one-sentence-out nested cross-validation framework. All semantic sentence models in our primary analyses (InferSent and the baseline/control models) were constructed from the same word-level representations (GloVe) (Pennington et al., 2014). Analyses tested for patterns in fMRI representations of sentences that InferSent could predict but the other models could not. The analysis was initially undertaken on voxels sampled across the whole cortex. To test for evidence that fMRI representations within multiple distributed brain regions reflected propositional meaning, the analysis was repeated within regions of a predefined language network (Fedorenko et al., 2010).

Next, in a follow-up analysis, we placed the results in the context of two other deep networks (ELMo and BERT) that have recently broken various Natural Language Processing (NLP) benchmarks and contributed to high-performance solutions to the Stanford Natural Language Inference (SNLI) sentence entailment task (on which InferSent was trained). ELMo and BERTs' high NLP performance gives good reason to hypothesize they would provide highly accurate fMRI predictions. On the flipside, while BERT is probably the strongest and most complex NLP approach, architecturally it might be the least cognitively plausible for the current serial reading task because BERT processes every word in a sentence in parallel (for a focused investigation of related matters, see also Merkx and Frank, 2020). So, the added complexity may provide no benefit here. To find out, we repeated the cortex-level analysis using ELMo and BERT.

In a final analysis, we estimated the room for improvement in modeling the current fMRI dataset. We first tested for semantic signal that was unpredicted by the deep network models using an "experiential attribute" model (Binder et al., 2016). The experiential model was acquired via

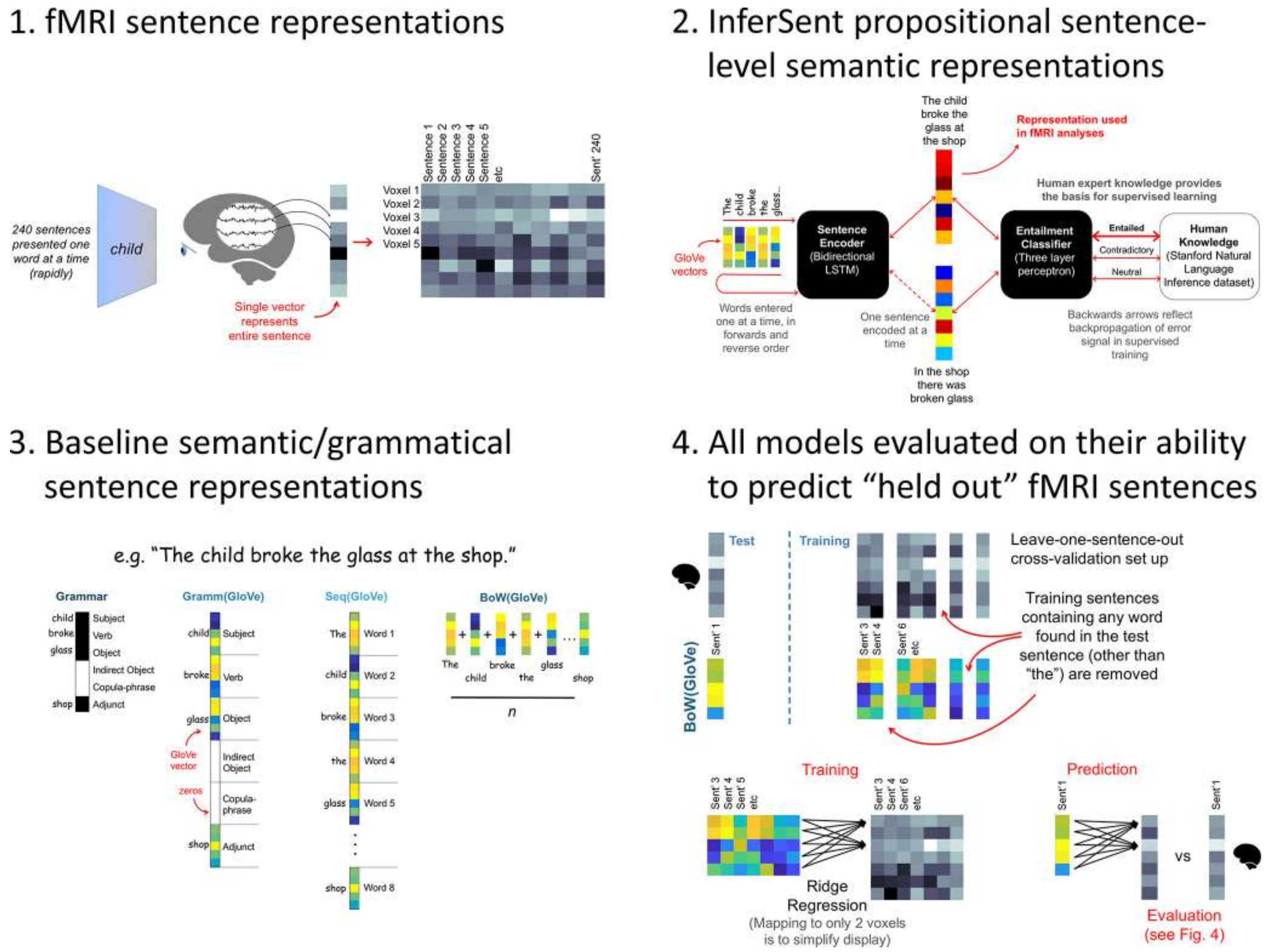

1. fMRI sentence representations
240 sentences presented one word at a time (rapidly)
child
Single vector represents entire sentence
Voxel 1
Voxel 2
Voxel 3
Voxel 4
Voxel 5
Sentence 1
Sentence 2
Sentence 3
Sentence 4
Sentence 5
etc
Sent' 240
2. InferSent propositional sentence-level semantic representations
The child broke the glass at the shop
Representation used in fMRI analyses
Human expert knowledge provides the basis for supervised learning
The child broke the glass...
GloVe vectors
Sentence Encoder (Bidirectional LSTM)
Entailment Classifier (Three layer perceptron)
Entailed
Contradictory
Neutral
Human Knowledge (Stanford Natural Language Inference dataset)
Words entered one at a time, in forwards and reverse order
One sentence encoded at a time
In the shop there was broken glass
Backwards arrows reflect backpropagation of error signal in supervised training
3. Baseline semantic/grammatical sentence representations
e.g. "The child broke the glass at the shop."
Grammar
Gramm(GloVe)
Seq(GloVe)
BoW(GloVe)
child
broke
glass
shop
Subject
Verb
Object
Indirect Object
Copula-phrase
Adjunct
GloVe vector
zeros
The
the
Word 1
Word 2
Word 3
Word 4
Word 5
Word 8
n
4. All models evaluated on their ability to predict "held out" fMRI sentences
Test
Training
Leave-one-sentence-out cross-validation set up
Training sentences containing any word found in the test sentence (other than "the") are removed
BoW(GloVe)
Sent' 1
Sent' 3
Sent' 4
Sent' 5
Sent' 6
etc
Training
Prediction
Ridge Regression
(Mapping to only 2 voxels is to simplify display)
vs
Evaluation (see Fig. 4)

Figure 1. Overview of our primary analysis. Top left, fMRI sentence reading experimental protocol. Top right, A schematic illustrating how InferSent forms sentence-level representations from context-independent word-level semantic representations (GloVe). Bottom left, Baseline models, which formed sentences by combining GloVe vectors according to grammatical rules, concatenation, or averaging. Bottom right, The ridge regression approach used to map the sentence models to predict fMRI activation. In the diagram, this is illustrated for only the BoW(GloVe) vector; however, mappings were estimated in the same way for all models. The procedure used to evaluate model-based predictions is illustrated in Figure 4.

behavioral ratings of sensory, motor, cognitive, interoceptive, and affective attributes of worldly experience, and potentially reflects semantic information that cannot be estimated from text corpora. Finally, to estimate how much sentence processing fMRI signal (semantic or otherwise) that was left unpredicted by all of the models, we applied a cross-participant fMRI encoding analysis (similar to Anderson et al., 2019b).

As additional background, despite their high NLP performance, we had considered ELMo and BERT to be less suitable vehicles for our primary analysis because: (1) ELMo and BERT do not strictly define how sentence representations should be constructed from words (unlike InferSent, which produces a single sentence representation); (2) ELMo and BERT would provide more obscure baseline models (e.g., context-independent BoW) because they operate on character-level/subword units (see Materials and Methods), so a BoW would be a bag-of-word-parts; and (3) ELMo and BERT are architecturally more complex than InferSent, which makes analyses and interpretation more complex.

Materials

All sentences were preselected as experimental materials for the Knowledge Representation in Neural Systems project (Glasgow et al., 2016) (www.iarpa.gov/index.php/research-programs/krns), sponsored by the Intelligence Advanced Research Projects Activity. The stimuli consisted of 240 written sentences containing 3-9 words and 2-5 (mean ± SD = 3.33 ± 0.76) content words, formed from different combinations of 141 nouns, 62 verbs, and 39 adjectives (242 words). The sentences are listed in full by Anderson et al. (2017a, 2019a). Sentences were in active voice and consisted of a noun phrase followed by a verb phrase in past tense, with no relative clauses. Most sentences (200 of 240) contained an action verb and involved interactions between humans, animals, and objects, or described situations involving different entities, events, locations, and affective connotations. The remaining 40 sentences contained only a linking verb ("was"). Each word occurs a mean ± SD (range) of 3.3 ± 1.7 (1-7) times throughout the entire set of sentences and co-occurs with 8.1 ± 4.3 (1-19) other unique words. The same two words rarely co-occur in more than one sentence, and 213 of 242 words never co-occur more than once with any other single word. Forty-two sentences contained instances of words not found in any of the other 239 sentences, and 3 of these sentences contained 2 unique words. There were thus 45 words that occurred in only one sentence, of which 29 were nouns, 7 were verbs, and 9 were adjectives.

Participants

Participants were 14 healthy, native speakers of English (5 males, 9 females; mean age = 32.5 years, range 21-55 years) with no history of neurologic or psychiatric disorders. All were right-handed according to the Edinburgh Handedness Inventory (Oldfield, 1971). Participants received monetary compensation and gave informed consent in conformity with the protocol approved by the Medical College of Wisconsin Institutional Review Board.

Procedure

Participants took part in either 4 or 8 scanning visits. The mean interval between sessions was 3.5 d (SD = 3.14 d). The range of the intervals between first and last visits was 15-43 d. In each visit, the entire list of sentences was presented 1.5 times, resulting in 12 presentations of each sentence over the 8 visits in 10 participants, and 6 presentations over 4 visits in 4 participants. Each visit consisted of 12 scanning runs, each run containing 30 trials (one sentence per trial) and lasting ∼6 min. The presentation order of each set of 240 sentences was randomly shuffled.

The stimuli were back-projected on a screen in white Courier font on a black background. Participants viewed the screen while in the scanner through a mirror attached to the head coil. Sentences were presented word-by-word using a rapid serial visual presentation paradigm (Forster, 1970). Nouns, verbs, adjectives, and prepositions were presented for 400 ms each, followed by a 200 ms interstimulus interval. Articles ("the") were presented for 150 ms followed by a 50 ms interstimulus interval. Mean sentence duration was 2.8 s (range, 1.4-4.2 s). Words subtended an average horizontal visual angle of ∼2.5°. A jittered intertrial interval, ranging from 400 to 6000 ms (mean = 3200 ms), was used to facilitate deconvolution of the BOLD signal. Participants were instructed to read the sentences and think about their overall meaning. They were told that some sentences would be followed by a probe word, and that in those trials they should respond whether the probe word was semantically related to the overall meaning of the sentence by pressing one of two response keys (10% of trials contained a probe). Participants' mean accuracy was 86% correct, with a minimum accuracy of 81%. Participants were given practice with the task outside the scanner with a different set of sentences. Response hand was counterbalanced across scanning visits.

MRI parameters and preprocessing

MRI data were acquired with a whole-body 3T GE 750 scanner at the Center for Imaging Research of the Medical College of Wisconsin using a GE 32-channel head coil. Functional T2*-weighted EPIs were collected with TR = 2000 ms, TE = 24 ms, flip angle = 77°, 41 axial slices, FOV = 192 mm, in-plane matrix = 64 × 64, slice thickness = 3 mm, resulting in 3 × 3 × 3 mm voxels. T1-weighted anatomic images were obtained using a 3D spoiled gradient-echo sequence with voxel dimensions of 1 × 1 × 1 mm. fMRI data were preprocessed using AFNI (Cox, 1996). EPI volumes were corrected for slice acquisition time and head motion. Functional volumes were aligned to the T1-weighted anatomic volume, transformed into a standardized space (Talairach and Tournoux, 1988), and smoothed with a 6 mm FWHM Gaussian kernel. The data were analyzed using a GLM with a duration-modulated HRF, and the model included one regressor for each sentence. fMRI activity was modeled as a γ function convolved with a square wave with the same duration as the presentation of the sentence, as implemented in AFNI's 3dDeconvolve with the option dmBLOCK. Duration was coded separately for each individual sentence. Finally, a single sentence-level fMRI representation was created for each unique sentence by taking the voxelwise mean of all replicates of the sentence.

Cortical language network

To test for evidence that multiple cortical regions encode propositional semantics, we analyzed a "language network" that was predefined by Fedorenko et al. (2010) and can be freely downloaded from https://evlab.mit.edu/funcloc/. The language network specifies cortical regions that were more functionally activated when stimulated with real sentences than by sequences of pseudowords. The language network spans bilateral temporal, inferior parietal, and inferior and mid frontal regions. These regions were initially identified on the left hemisphere and flipped onto the right. We selected the "original" language network parcellation that contains 16 ROIs over other alternatives on the website (e.g., an updated network with 12 ROIs) because the original network covers more cortex and contains more anterior-posterior subdivisions of the temporal cortex (which was convenient for our current goal of testing multiple cortical ROIs). The cortical location of ROIs is illustrated (see Figs. 7, 8), and the breakdown of anatomic regions contributing voxels analyses within each ROI is listed in Table 1.

Experimental design and statistical analysis

Overview of the sentence models

To predict fMRI representations of sentence meaning, we deployed a selection of semantic models of sentence meaning. Our primary analysis used InferSent, which computes a complex recurrent nonlinear composition of words into unified sentence representations. We implemented three baseline models, Bow(GloVe), Seq(GloVe), and Gramm(GloVe), which we used to account for linear/sequential/grammatical rule-based composition of words into sentences. Importantly, all four models, including InferSent, were derived from the same context-independent word-level semantic model, GloVe (Pennington et al., 2014). To serve as a control for syntax, we constructed a model of the grammatical structure of sentences (Grammar). To further control for fMRI activation associated with visual processing of the written sentence stimuli, we constructed two additional models that coded the descriptive statistics of word/sentence lengths (Char stats) and the rudimentary visual appearance of sentences (Word overlay), respectively. These models echo approaches taken by previous studies to control for the visual appearance of words (e.g., Just et al., 2010; Devereux et al., 2013; Wehbe et al., 2014; Fernandino et al., 2016).

In our follow-up analyses, we placed the predictions made by InferSent (and the other models) in the context of two other deep network models, ELMo and BERT, and a behavioral rating-based experiential attribute model. The models are described in the following text in the order of the analyses in which they first appear. We provide a more cursory description of ELMo, BERT, and the attribute model because they are a secondary focus of the current study.

GloVe semantic vectors: word-level units in primary analysis

There has been a long tradition in computational linguistics of leveraging the distributional statistics of word co-occurrences in natural text as a basis for estimating semantic representation. So-called distributional semantic models (also known as word embeddings) approximate word meaning using vectors of values reflecting how often a target word co-occurred with other words across a huge body of text (Lund and Burgess, 1996; Landauer and Dumais, 1997; Mikolov et al., 2013; Pennington et al., 2014). The current article is principally based on one such model, GloVe (Pennington et al., 2014). GloVe represents individual words as 300-dimensional floating point vectors derived by factorizing a word co-occurrence matrix (vocabulary size is 2.2 million words; and co-occurrences were measured across 840 billion tokens from Common Crawl: https://commoncrawl.org). GloVe initially came to prominence in the fMRI literature for yielding state-of-the-art performance decoding fMRI activation associated with sentences in Pereira et al.'s (2018) "universal neural decoder of linguistic meaning." Additionally, GloVe was the basis for the initial implementation of InferSent (Conneau et al., 2017). For these reasons, we used GloVe as the basis word-level unit in our analyses.

InferSent(GloVe): primary analysis

InferSent is a supervised recurrent nonlinear deep learning approach that was presented as an alternative to previous unsupervised corpus-based methods for modeling sentence-level semantics (Conneau et al., 2017). In supervised learning, InferSent leverages human expert knowledge to optimize neural network weights to recurrently combine and refine word-level representations to model sentences. The expert knowledge was provided by the SNLI dataset (Bowman et al., 2015), which contains 570,000 English sentence pairs, with each pair manually categorized according to (1) whether one sentence entailed the other, (2) whether sentence pairs were contradictory, or (3) neutral. Thus, the final sentence representations produced are optimized to support accurate computations of intersentence entailment relations. Successful evaluations on separate natural language inference datasets are presented in Conneau et al. (2017).

We refer to the unified sentence representations produced by InferSent as propositional because classifying entailment relies on an informal comparison of the propositional content of sentences: for example, whether the relationship between entities referenced by a sentence appears to be true or false, and whether a second sentence appears

Table 1. Neuroanatomical regions contributing to the language network ROI analyses[a]

LROI 1	112 ±15.9 voxels	LROI 2	113.2 ±17.5	RROI 1	154.4 ±26.0	RROI 2	116.4 ±18.8
ctx_lh_S_temporal_sup	37.2%	ctx_lh_S_temporal_sup	40.5	ctx_rh_S_temporal_sup	31.4	ctx_rh_S_temporal_sup	25
ctx_lh_G_temporal_middle	18.9%	ctx_lh_G_temporal_middle	14.9	ctx_rh_G_temporal_middle	17.8	ctx_rh_G_occipital_middle	21.1
ctx_lh_G_temp_sup-Lateral	11.4%	ctx_lh_G_occipital_middle	9.6	ctx_rh_G_pariet_inf-Angular	15.1	ctx_rh_G_pariet_inf-Angular	20.1
ctx_lh_G_temp_sup-Plan_tempo	10.8%	ctx_lh_G_temporal_inf	8	ctx_rh_G_pariet_inf-Supramar	9.1	ctx_rh_S_occipital_ant	10
ctx_lh_G_pariet_inf-Supramar	7.8%	ctx_lh_G_pariet_inf-Angular	5.7	ctx_rh_G_temp_sup-Lateral	7.3	ctx_rh_G_and_S_occipital_inf	8.7
		ctx_lh_S_occipital_ant	5			ctx_rh_G_temporal_middle	7.5
LROI 3	111.6 ±13.9	LROI 4	123.9 ±21.5	RROI 3	120.3 ±12.0	RROI 4	141.1 ±24.5
ctx_lh_S_temporal_sup	32.4	ctx_lh_S_front_inf	32.2	ctx_rh_S_temporal_sup	32.2	ctx_rh_S_front_inf	19.8
ctx_lh_G_temporal_middle	25.3	ctx_lh_G_front_inf-Opercular	22.5	ctx_rh_G_temporal_middle	30.6	ctx_rh_S_precentral-inf-part	18
ctx_lh_G_temp_sup-Lateral	14.3	ctx_lh_S_precentral-inf-part	14.2	ctx_rh_G_temporal_inf	11.9	ctx_rh_G_front_inf-Triangul	15.9
ctx_lh_G_temporal_inf	8.7	ctx_lh_G_front_inf-Triangul	11.6	ctx_rh_G_temp_sup-Lateral	10.2	ctx_rh_G_front_inf-Opercular	14.8
ctx_lh_S_temporal_inf	6.8	ctx_lh_G_front_middle	11.5	ctx_rh_S_temporal_inf	5.7	ctx_rh_G_front_middle	14.5
ctx_lh_S_collat_transv_ant	5.4	ctx_lh_G_precentral	5.9			ctx_rh_G_precentral	9.1
LROI 5	111.9 ±18.7	LROI 6	130.6 ±15.6	RROI 5	120.9 ±17.4	RROI 6	138.3 ±19.2
ctx_lh_G_temp_sup-Lateral	30.6	ctx_lh_G_front_inf-Triangul	24.1	ctx_rh_S_temporal_sup	28.3	ctx_rh_G_orbital	19.4
ctx_lh_S_temporal_sup	19.3	ctx_lh_G_orbital	13	ctx_rh_G_temporal_middle	24.7	ctx_rh_G_front_inf-Triangul	16.7
ctx_lh_G_temporal_middle	19.2	ctx_lh_G_front_inf-Orbital	11.8	ctx_rh_G_temp_sup-Lateral	21.7	ctx_rh_G_front_inf-Orbital	12
ctx_lh_S_circular_insula_inf	8.2	ctx_lh_Lat_Fis-ant-Horizont	9.7	ctx_rh_G_temp_sup-Plan_polar	6.2	ctx_rh_G_insular_short	7.7
ctx_lh_G_temp_sup-Plan_polar	5.3	ctx_lh_G_temp_sup-Lateral	6.2			ctx_rh_Lat_Fis-ant-Horizont	6.8
						ctx_rh_G_temp_sup-Lateral	5.3
LROI 7	116.6 ±25.5	LROI 8	117.6 ±25.3	RROI 7	122.0 ±21.3	RROI 8	135.5 ±15.3
ctx_lh_G_pariet_inf-Angular	33.2	ctx_lh_G_precentral	62.2	ctx_rh_G_occipital_middle	32.6	ctx_rh_G_precentral	47
ctx_lh_G_occipital_middle	24.4	ctx_lh_G_front_middle	13.8	ctx_rh_S_oc_sup_and_transversal	20.7	ctx_rh_G_front_middle	15.6
ctx_lh_S_temporal_sup	19.3	ctx_lh_S_precentral-inf-part	7.5	ctx_rh_G_pariet_inf-Angular	17.7	ctx_rh_S_precentral-sup-part	8.7
ctx_lh_S_oc_sup_and_transversal	11	ctx_lh_S_precentral-sup-part	5.4	ctx_rh_G_occipital_sup	10.8	ctx_rh_S_central	7.2
				ctx_rh_S_oc_middle_and_Lunatus	7.1	ctx_rh_G_postcentral	7.2
						ctx_rh_S_precentral-inf-part	6.2

[a]Data are mean ± SD for 14 participants. Each entry displays the no. of stable voxels within each language network ROI that contributed to analyses. For example, within LROI 1, on average 112 voxels contributed to the analysis of each participant. Anatomical labels (Destrieux Atlas) and corresponding percentages indicate the mean percentage of times (across 14 participants) that a selected voxel belonged to that anatomic region. For example, within LROI 1, 37.2% of voxels selected were from left superior temporal sulcus (ctx_lh_S_temporal_sup) when pooling across all cross-validation iterations and participants.

to specify a coherent or contradictory (propositional) interentity relationship. We say "informal" and "appears to be" because strict logical relationships rarely hold in natural language. For instance, although most people would consider that "Socrates got caught out by the rain" probably entails that "Socrates got wet," this is not 100% assured; perhaps Socrates was stuck sheltering from the rain. Likewise, our use of the word "proposition" should be considered to be more graded and probabilistic in nature than the TRUE/FALSE logical propositions typical of propositional calculus.

InferSent is constructed from two modules (Figs. 1, 2): (1) a sentence encoder, which is a recurrent artificial neural network that iteratively combines an input sequence of word vectors (currently GloVe) presented one at a time, into a sentence-level vector output; and (2) an entailment classifier, which takes two sentence-level vectors as input (two outputs from the sentence encoder) and estimates whether the pair are entailed, contradictory, or neutral. The entailment classifier plays a critical role in the training procedure. Specifically, the difference between the classifier estimate and the correct classification (from SNLI) provides an error signal that can be propagated back through the sentence encoder to optimize network weights. After the sentence encoder weights have been optimized via the entailment classifier, the encoder module can be used in isolation to generate new representations of novel sentences (as was the case for encoding the 240 sentences tested in this article). The architecture of the two modules is outlined in more detail as follows.

Sentence encoder. The basic unit of the sentence encoder is the Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) recurrent neural network. Recurrent networks are broadly characterized by having feedback loops that enable previous network outputs to inform processing of and to be integrated with new inputs. This enables new words to be interpreted in fed back context, which is critical for correct interpretation of polysemous words and homonyms. For instance, "bat" has multiple meanings (e.g., flying mammal or sports tool) that can be selected by context (e.g., fruit bat vs baseball bat). LSTMs were introduced to overcome difficulties faced by standard recurrent approaches in retaining critical information over long time intervals. For instance, to correctly interpret "...the bat flew through the air," given the preceding context "The batsman lost his grip and...", it is necessary to relate bat to batsman at the start of the sentence. LSTMs accommodate long-term dependencies through a network architecture that uses a memory cell to remember information across consecutive network cycles, and a series of three nonlinear information gates that manage how information flows through the network. Gates are as follows: (1) a forget gate, which deletes irrelevant information from the cell; (2) an input gate, which selects what new information should be added to the cell; and (3) an output gate, which retrieves information from the cell to form network output. All three gates select information to delete/store/retrieve based on the new input (e.g., a word) and previous output (e.g., the sentence so far). Thus, on each iteration, a new word is input, the cell memory is updated, and a filtered version of the cell memory is output from the network, and fed back for the next iteration.

The sentence encoder uses a bidirectional LSTM (Graves and Schmidhuber, 2005). As the name implies, bidirectional LSTMs combine two LSTMs, which, respectively, cycle forwards and in reverse order

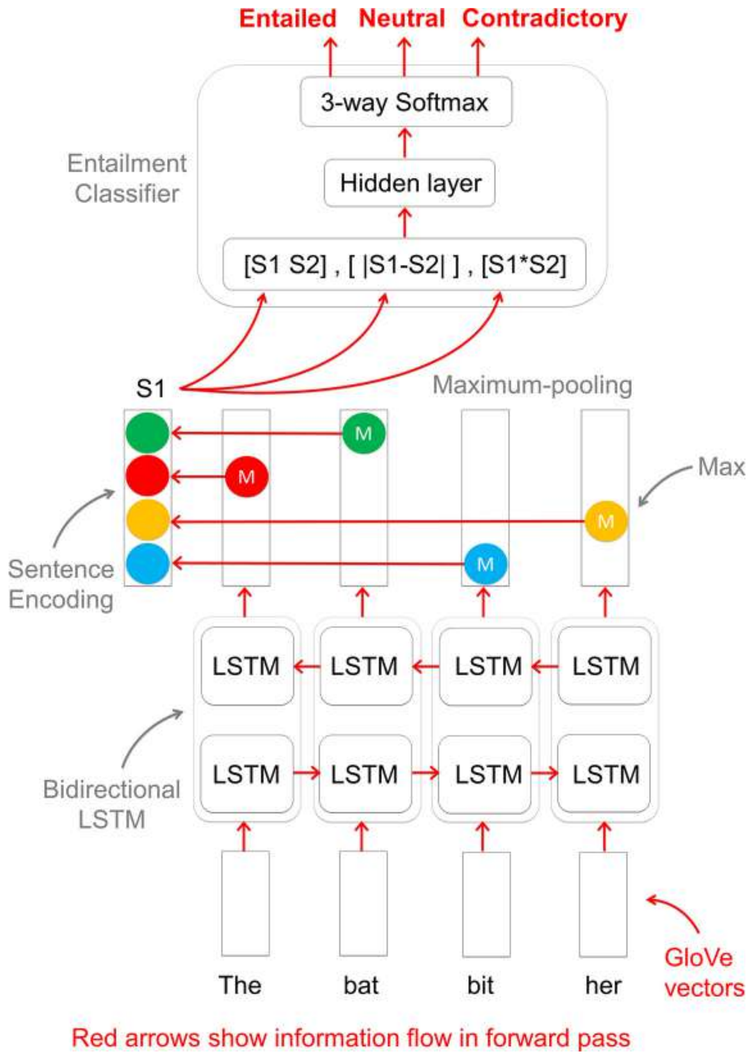


Figure 2. InferSent algorithm encoding unified propositional sentence representations. While the Entailment classifier operates on two sentences (S1 and S2), the encoding of only a single sentence is illustrated to simplify display. In practice, the second sentence would be encoded separately, in precisely the same way as illustrated for the first sentence. Once the two sentences have been encoded, they are combined and integrated, and the composite representation is evaluated to estimate whether the sentences entailed one another as opposed to being contradictory or neutral.

through input (here words). Thus, for any word in a sentence, the bidirectional LSTM simultaneously supplies information on historic and future context, which was collated by concatenating outputs from the two LSTMs. Intuition of how modeling future context could be advantageous for language comprehension is easy to supply; it is not until the end of the sentence, "The bat flew through the air, and bit her on the neck," that it is clear that the bat is a vampire. Indeed, encoding of past and future context often affords practical advantages over unidirectional LSTMs (e.g., Graves and Schmidhuber, 2005), as also was the case for InferSent in tests on a battery of standard NLP tasks (Conneau et al., 2017).

However, because bidirectional operation produces an output vector for each word (reflecting forwards/reverse information across the entire sentence), one must be chosen to reflect a sentence, or they must be combined. To form a single sentence representation, InferSent selects the maximum pointwise activation value across all output vectors. In tests on NLP tasks, this "max-pooling" representation was found to have practical advantages over the obvious alternative of mean-pooling (Conneau et al., 2017). The max-pooled sentence vector was the representation used in all analyses in this article. Sentence vector length was 4096.

Entailment classifier. To classify whether sentence pairs are entailed, contradictory, or neutral, a fully connected feedforward neural network, with a single hidden layer and a 3 class Softmax output layer, is used. The Softmax function normalizes the three network outputs to sum to 1, to estimate the probability of each entailment class being correct. Classifier input was supplied by the sentence encoder, which was run twice per classification to generate two sentence representations for testing. The classifier input representation combined information within and across the two sentences, as was computed in three ways: (1) concatenating the sentence vectors, (2) taking the absolute difference between sentence vectors, and (3) taking the pointwise multiplication of the two vectors. These three representations were concatenated for entry into the classifier.

InferSent Training. As mentioned above, training was performed with sentences from the SNLI dataset (Bowman et al., 2015). Optimization of weights was achieved by stochastic gradient descent, run on minibatches of 64 sentence pairs. The error signal computed across the 64 sentence classifications was used as a basis for adapting weights across both the classifier and the sentence encoder. Learning rate and weight decay parameters were set at 0.1 and 0.99, respectively. Learning weight was divided by 5 if classification accuracy decreased. Training was terminated when the learning rate fell below a threshold of 10^{-5}. Source code is freely available (see Code availability).

Following training, the encoder module was used in isolation to generate vector representations of the 240 sentences included in the fMRI study. For all of the analyses in this article, we used the pretrained version of InferSent constructed by the original authors (Conneau et al., 2017).

BoW(GloVe): primary analysis baseline

BoW models represent sentences as an unordered linear assembly of constituent words (e.g., by computing the pointwise average of word vectors), and have endured as a practically successful technique in both computational linguistics (Mitchell and Lapata, 2010; Kiela and Clark, 2014), and fMRI analyses (Anderson et al., 2017a, 2019a,b; J. Wang et al., 2017; Yang et al., 2017; Pereira et al., 2018) despite their obvious shortcomings in neglecting word order, syntax, and morphology. We implemented the BoW model by taking the pointwise mean of GloVe vectors for constituent words in sentences (Fig. 1). In previous work (e.g., Anderson et al., 2019b), we had excluded function words (e.g., "The," "in") from our analyses; however, we included them here to serve as a control for InferSent, which does operate using function words. Relatedly, in pilot tests, we observed that including function words led to slightly greater prediction scores. Also, because InferSent uses a max-pooling stage (described in the previous section), we experimented with computing the maximum featurewise value across word vectors within sentences, rather than averaging them. This brought no performance benefit, and these results are not discussed further.

Gramm(GloVe): primary analysis baseline

To capture the grammatical structure of sentences in a semantic model, we created a canonical sentence representation that contained slots for semantic vectors of words with different grammatical roles. Nine different grammatical structures were identified in the 240 experimental sentences (Anderson et al., 2019a). These nine structures were constructed from combinations of the following elements: Subject (240), Verb (196), Direct Object (128), Indirect Object (27), Copula-Phrase (44), and

Adjunct (74). The number in parentheses indicates the number of sentences containing the grammatical element. The nine different grammatical structures were as follows, with the number of sentences following that structure listed in parentheses:

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][Adjunct: *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] [Adjunct: *at the* party]

Subject, Copula-Phrase (39): [S: *The* family] [Copula-Phrase: *was* happy]

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

The canonical sentence model (illustrated in Fig. 1) was a long vector containing 6 slots for GloVe vectors corresponding to each of the grammatical elements (hence 1800 dimensions in total, given each GloVe vector is 300 dimensions). The vector was first initialized with zeros. Then, GloVe vectors for content words were slotted into the vector, according to their grammatical role. In cases where a grammatical element is modified by an adjective (e.g., "The wealthy family" in the above example), then semantic vectors for the adjective were pointwise averaged with the noun. If sentences did not contain a particular grammatical entry, their value remained as zero. Function words (illustrated in italics) in the above example were not included in the model. This was because we had attempted to implement the role of function words by segregating content words within the grammatical frame.

Seq(GloVe): primary analysis baseline

To model the order that constituent content words in sentences in which appeared, we simply concatenated GloVe vectors for words end on end in order of presentation (Fig. 1). Specifically, each sentence was modeled as a long vector containing 9 slots, to accommodate 9 or fewer words (the longest sentence contained 9 function and content words). Thus, the vector had 2700 dimensions. If the sentence contained only 3 words, GloVe vectors would be slotted in to the first, second, and third slots (the first 900 dimensions) and the remainder of the vector would be padded with zeros.

Grammar: primary analysis baseline

To capture only the grammatical structure of sentences, we created a binary model encoding the presence or absence of grammatical elements in sentences. In line with the grammatically structured semantic model described previously (Gramm(GloVe)), the current model was a vector containing 6 entries for Subject, Verb, Direct Object, Indirect Object, Copula-Phrase, and Adjunct. If a sentence contained a grammatical element, the corresponding vector entry was assigned the value 1. Otherwise, it was assigned a 0.

Visual appearance of textual sentence stimuli: primary analysis controls

To model fMRI activation associated with the visual processing of the written sentence stimuli, we constructed two models. "Char stats" coded descriptive statistics of the number of characters and words in sentences. Each sentence was represented as a 6 element vector coding the number of words in the sentence, the mean, SD, maximum and minimum number of characters per word, and the number of characters in the entire sentence. "Word overlay" coded a coarse representation of the visual appearance of the sentence stimuli, which had been presented word by word. Each sentence was represented as a vector of 11 elements. Eleven reflected the number of characters comprising the longest words within the set of experimental sentences. To construct the sentence representation, each word was first also modeled as a binary vector of eleven elements. The presence/absence of a character in respective positions of the word was modeled as a 1/0, respectively. Then all word vectors in the sentence were pointwise summed. As a dummy example, illustrated with a 6 element vector, "A cat jumped" would be represented as: [3 2 2 1 1 1] = [1 0 0 0 0 0] + [1 1 1 0 0 0] + [1 1 1 1 1 1].

ELMo: follow-up analyses

ELMo is a recurrent deep network approach that was introduced to generate contextualized vector representations of words that capture elements of both syntax and semantics and accommodate polysemy (Peters et al., 2018). Architecturally, ELMo incorporates three layers of subnetworks that process word sequences. Each layer outputs word vectors with different degrees of contextualization. The three vectors are typically integrated via a weighted average that may be fine-tuned according to the specific task.

ELMo's first layer (L1) is a convolutional neural network that generates context-independent word vectors from the series of characters forming each word. Operating on characters enables ELMo to leverage morphologic similarities between words to generate semantic representations for new untrained words. For example, "computer," "computed," and "computerized" are all semantically related and all share the same character stem ("compute"), which would suggest that a new word "computerization" would have a related meaning. The convolutional network approach to transforming character sequences into word vectors is described in detail by Kim et al. (2016).

The next two layers of ELMo are bidirectional LSTMs (L2 and L3) that memorize features of past/future words to produce increasingly contextualized word representations (for a more detailed description of LSTMs, see InferSent). To encode bidirectional context, the output vectors extracted from forwards and backwards LSTMs are concatenated. The output of the second bidirectional LSTM (L3) feeds into a final Softmax layer that predicts the identity of a forthcoming or preceding word based on past or future words, respectively, depending on the direction of operation. Thus, unlike InferSent, ELMo's training is entirely based on the statistics of word use in text corpora (rather than human judgments of entailment relations). Weights across all layers of ELMo are optimized to improve the word prediction accuracy.

We downloaded a pretrained implementation of ELMo constructed by the original authors (Peters et al., 2018). In this implementation, all word and sentence vectors have 1024 entries. Because ELMo does not strictly commit to a specific way to transform word-level vectors into a sentence representation, we pointwise averaged vectors for constituent words in sentences to produce sentence vectors. We repeated this for each layer to produce one context-independent and two contextualized vectors for each sentence (ELMo L1, ELMo L2, and ELMo L3, respectively). To combine all layers together, we concatenated them (ELMo L1, L2,L3).

BERT: follow-up analyses

Transformer encoders, such as BERT (Devlin et al., 2019), were introduced in part to address the practical limitations of recurrent networks (including LSTMs) in capturing long-term dependencies between distant words, and in part to alleviate their lengthy serial processing times. BERT saves on time by parallelized processing of all word/subword vectors within a sentence/paragraph (for ease of explanation, we shall refer to both words and subwords as words). BERT captures long-term dependencies via a "self-attention" computation that places each word into the context of all other word inputs. In particular, having simultaneous access to both past and future words enables BERT to form richer representations of bidirectional context than bidirectional LSTMs (which process past and future separately and concatenate the respective outputs). However, such parallelization necessitates additional measures to retain word order (which recurrent networks encode implicitly). To resolve this, BERT explicitly integrates a code reflecting both the absolute and relative sentential position of each word into that word's input vector.

Architecturally, BERT is formed from multiple layers of identical so-called transformer blocks, with each block formed from multiple layers of nodes. Each subsequent block produces an increasingly contextualized

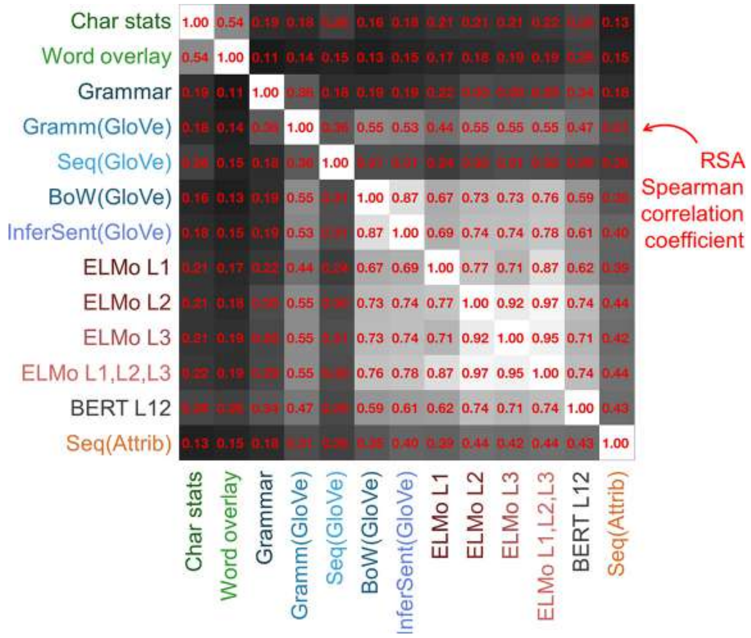


Figure 3. RSA revealed the interrelationship between the different sentence models tested. Each entry in the matrix corresponds to an RSA between two models. To compute RSA, intersentence Pearson correlation matrices were constructed for each model. To compare models, the below diagonal correlation matrix triangles were extracted from each matrix and vectorized. Spearman correlation was then computed between vectorized triangles corresponding to each model pair. Statistical significance was evaluated by permutation testing. The order of the sentences for one model was randomly shuffled, and the rows and columns of the respective correlation matrix were rearranged according to the shuffled order. The correlation between the vectorized matrix triangle of the shuffled matrix and the triangle of the other unshuffled model matrix was then computed. The 1001 correlation coefficients (associated with shuffled and unshuffled matrix comparisons) were ranked in descending order, and a *p*-value was computed as the rank associated with the unshuffled coefficient divided by 1001. All correlations displayed were highly significant ($p = 0.001$). BERT L12 was selected for display *post hoc* because it yielded strong predictions in later analyses (Fig. 9).

representation of each word. Contextualization is achieved via self-attention, which quantifies how relevant each of the other input words is to the current word (the self). In practice, multiple-self attention estimates are computed (Multi-Headed Attention), which potentially capture different types of contextual relationships. The outputs across all attention heads within a block are integrated, and passed on through a fully connected feedforward network to form the input to the next block. The last block feeds into two separate Softmax layers that either serve to predict the identity of a masked input word or classify whether pairs of input sentences were consecutive. All weights throughout BERT were optimized according to these two tasks. The optimized architecture constitutes a "pretrained" BERT model that can later be fine-tuned for a particular task of interest (e.g., question answering) through additional training.

In the current article, we apply the pretrained "BERT-large-uncased" model, downloaded from Hugging Face (Wolf et al., 2019). This implementation has 24 layers of blocks, and 16 attention heads per block. The layer outputs provide 24 candidate representations of words. Each word vector has 1024 entries. We experimented with two ways to form sentences. The first was to use word combinations assembled by BERT in response to a key token "[CLS]." [CLS] is included in input sequences to assist in the classification of consecutive sentences. The second was to take the pointwise average across word representations within each layer. Both approaches generated 24 sentence representations, one per layer. In pilot analyses, we observed the [CLS] sentences to yield slightly weaker fMRI predictions across layers, and we therefore focused on the averaging approach.

Seq(experiential attribute): follow-up analyses

The experiential attribute model (Binder et al., 2016) presented an alternative to modeling complex meaning using only word usage statistics, and seeks to account for knowledge acquired from sensory, motor, cognitive, interoceptive, and affective experience interacting with the world, not just language and text. BoW based experiential attribute models have been extensively tested on the current fMRI dataset (Anderson et al., 2017a, 2019a) and complement BoW GloVe models in explaining representational structure (Anderson et al., 2019b).

The experiential attribute model broadly aligns with "embodiment" theories that posit representations of word meaning reflect a summarization of the brain states involved in experiencing that word, partially reenacted across sensory/motor/affective/cognitive subsystems (e.g., Barsalou et al., 2008; Glenberg, 2010; Pulvermüller, 2013; Binder et al., 2016). This model represents words in terms of human ratings of how strongly people associate words with different attributes of experience (e.g., "On a scale of 0-6, to what degree do you think of a banana as having a characteristic or defining color?"). Ratings were collected via Amazon Mechanical Turk for a total of 65 attributes spanning sensory, motor, affective, spatial, temporal, causal, social, and abstract cognitive experiences. Ratings for each attribute were averaged across workers to derive a single 65 dimensional vector for each word.

To model sentences, attribute vectors for constituent content words in sentences were concatenated in order of their presentation. This was in precisely the same fashion as Seq(GloVe), but excluding function words (which had not been behaviorally rated). The Seq(Attribute) model was implemented in light of the results of our primary analysis (see Fig. 6) when Seq(GloVe) yielded higher prediction scores than the BoW(Glove) or Gramm(GloVe). Gramm(Attribute) and BoW(Attribute) had weaker explanatory power than the sequential model and are not discussed further.

Predictive mapping using ridge regression and leave-one-sentence-out nested cross-validation

Because interpreting analyses of fMRI activation elicited by "uncontrolled" sentence/natural language stimuli can be complicated by spurious correlations between representations arising at different processing levels (from vision to semantics), we first estimated the extent to which this affected the current data by computing representational similarity analyses (RSA) (Kriegeskorte et al., 2008) between all model pairs. These analyses demonstrated that, despite similarities within broad model classes (semantic models correlated more strongly with one another than they did with models of visual processing), every model pair was significantly related (Fig. 3; all $p = 0.001$). In particular, BERT correlated relatively strongly with the visual appearance models, which might reflect BERT's explicit encoding of word positions (which reflects sentence length).

We therefore configured both our model fitting approach and evaluation procedure to reduce the impact of such confounds between models. As we were primarily interested in evaluating how accurately the semantic models would generalize to predict neural representations of entirely new sentences, we undertook our analyses within a cross-validation framework where sentences used to test predictions minimally overlapped in their constituent words with the sentences used to train the predictive mapping between model and fMRI data.

To map between each model and the fMRI data, we applied ridge regression (Hoerl and Kennard, 1970) in a leave-one-sentence-out cross validation framework. This was repeated for each model and participant. Cross-validation iterated across the 240 sentences as follows. At each cross-validation iteration, the data were split into the following: (1) A

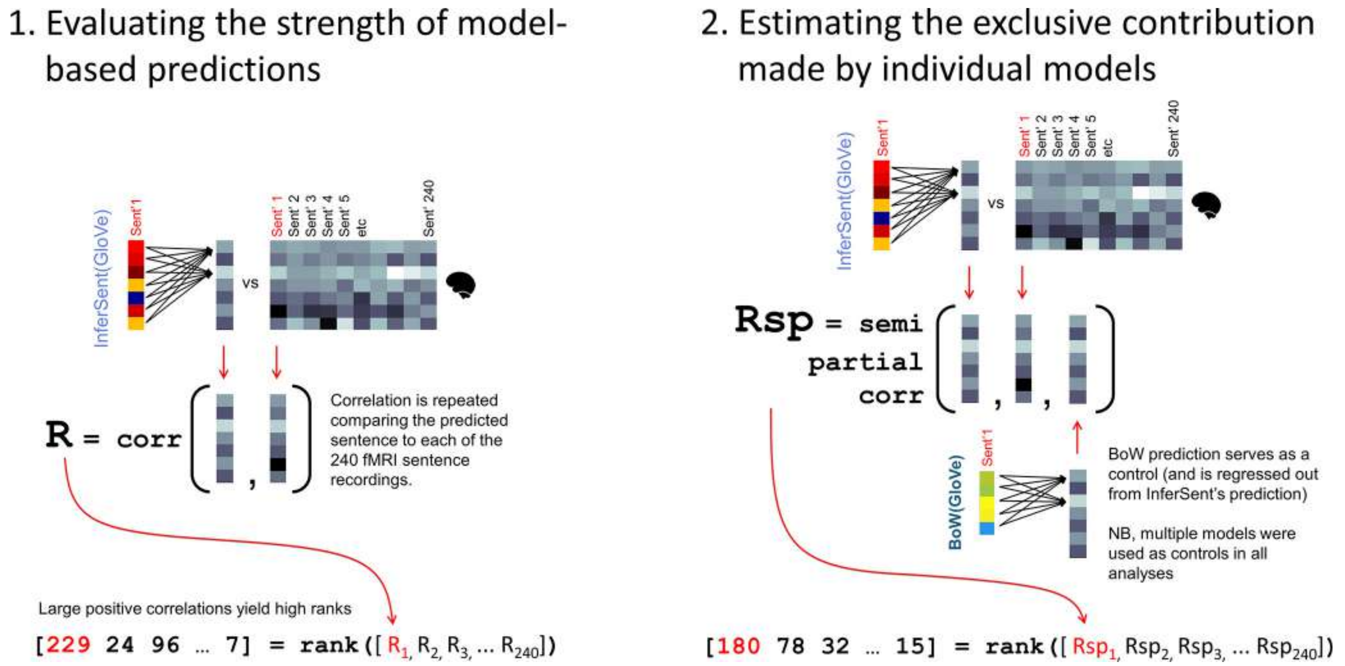

1. Evaluating the strength of model-based predictions
InferSent(GloVe)
Sent'1
vs
Sent' 1
Sent' 2
Sent' 3
Sent' 4
Sent' 5
etc
Sent' 240
R = corr
Correlation is repeated comparing the predicted sentence to each of the 240 fMRI sentence recordings.
Large positive correlations yield high ranks
[229 24 96 ... 7] = rank([R1, R2, R3, ... R240])
2. Estimating the exclusive contribution made by individual models
InferSent(GloVe)
Sent'1
vs
Sent' 1
Sent' 2
Sent' 3
Sent' 4
Sent' 5
etc
Sent' 240
Rsp = semi partial corr
BoW(GloVe)
Sent'1
BoW prediction serves as a control (and is regressed out from InferSent's prediction)
NB, multiple models were used as controls in all analyses
[180 78 32 ... 15] = rank([Rsp1, Rsp2, Rsp3, ... Rsp240])

Figure 4. Evaluation of predicted fMRI sentences. Predictions were evaluated using two metrics. Left, The rank metric provided a measure of how accurately the predicted fMRI representation reflected the original sentence recording, comparative to recordings of all of the other 239 mismatched sentences. Right, The semipartial rank metric controlled for the predictions made by other models in evaluation. Specifically, before evaluating the prediction made by one model (InferSent above), the predictions made by other model(s) are regressed out from InferSent's prediction (in the above illustration BoW(GloVe) is regressed out). The residual is then compared with the original fMRI activation, and ranked relative to the other sentences. Both rank and semipartial rank metrics were normalized to the range 0-1. For instance, if the score was 229, the normalized score was 0.95, computed as (229 − 1)/239.

test subset, formed from a single sentence (changing each iteration), which was used to evaluate the predictive power of the model. (2) A training subset, which was used to fit the predictive mapping between each model and the fMRI data. The training subset was formed from all sentences that did not contain any word in the test sentence other than "the." As a result of this stipulation, the number of sentences in the training set varied with each iteration (mean ± SD, 214 ± 17 sentences). (3) A tuning subset. The tuning subset was used to estimate the "ideal" ridge regression regularization parameter λ at each cross-validation iteration (as is described later in this section). The mapping associated with the best λ was used to predict the test sentence from 1. The tuning sentences were the subset that remained after the training and testing sentences had been selected (mean ± SD, 25 ± 17 sentences), and they could overlap in constituent words with both the training and test sentence. It would have been preferable if there were no such overlap, but this would have resulted in small training data splits for the current dataset.

At each cross-validation iteration, ridge regression was applied to the training sentences to fit a many-to-one predictive mapping from the multidimensional model to each single voxel. Before fitting the regression, both voxel activation and model feature values were normalized by subtracting the mean across the training sentences, and dividing by the SD (i.e., *z* scoring). Model and fMRI data for the tuning and test sentences were likewise normalized according to same means and SDs that had been computed on the training sentences. A separate ridge regression was fit for each voxel and each of a series of regularization parameters (λ) used to counteract overfitting.

To estimate the ideal λ at each cross-validation iteration, voxel activation for each of the tuning sentences was predicted using the regression mapping derived from each λ. The mean prediction accuracy across the tuning sentences was evaluated according to the rank scoring procedure detailed in the following section (Fig. 4, left). The regression mapping associated with the highest scoring λ was selected to predict the test sentence. The prediction accuracy for the test sentence was evaluated via the same rank scoring procedure.

For the cortex-level analysis, we set candidate λs to range across the following: [1 1e1 1e2 1e3 1e4 1e5]. Because the "peak" selected λs were on average (across iterations and participants) within the range 1e1 to 1e4 for each model, we believe a satisfactory fit was found in each case. For the ROI analysis, we extended the range of candidate λs to the following: [1 1e1 1e2 1e3 1e4 1e5 1e6 1e7 1e8] following pilot tests. The average "peak" λ value (across iterations and participants) selected for the different models across all ROIs was always within the range 1e1 and 1e6. This again suggested that a satisfactory fit was achieved for each model.

Rank scoring prediction accuracy

To evaluate the predictions made by each model, we first computed Pearson correlation between each predicted fMRI sentence (a vector of activation values across voxels) and the original fMRI recording. Next, we evaluated how the correlation coefficient ranked comparative to coefficients computed between the predicted sentence and the original fMRI recordings of each of the other 239 mismatched sentences (Fig. 4, left) (Pereira et al., 2018). Under this setup, the ideal prediction would be marked by a high positive correlation coefficient yielding a rank of 240. The rank associated with the correctly matched sentence was subsequently normalized to the range 0-1, by subtracting 1 and dividing by 239. A final normalized rank score was assigned to each model as the mean normalized rank across all 240 sentence predictions (one per each cross-validation iteration). If there were no sentence-related signal in either the fMRI data or the model, the expected score would be 0.5.

To further estimate the exclusive contribution that each model made to predicting fMRI activation, we computed a "semipartial rank score" as is illustrated in Figure 4 (right). This semipartial score complemented the rank score, which risked obscuring whether or not different models were predicting complementary information. For instance, two models may obtain the same rank score by predicting different components of the fMRI signal.

The semipartial rank score was estimated by computing the semipartial correlation between the original fMRI recording of a sentence S, and the representation of S predicted by one model (e.g., A) while controlling for predictions of S made by the combination of other models (e.g., B, C, D, and so on). Practically, the prediction of S made by A was regressed on the prediction made by B, C, and D (using multiple regression). Then the residual was computed, which reflected the pattern exclusively predicted by A, but not the other models. To quantify the exclusive prediction accuracy, Pearson's correlation between the residual and the original fMRI sentence representation was computed. Then the residual was compared with the original fMRI recordings of each of the other 239 mismatched sentences using Pearson correlation. The correlation coefficients were then ranked, and the rank associated with sentence S was normalized to the range 0-1. A final normalized semipartial rank score was assigned as the mean normalized rank across all 240 sentence predictions. If there were no signal exclusive to the test model (here A), then the expected semipartial score would be 0.5.

We note that there could have been other ways to approach evaluating the unique contribution of individual models to predicting variance (in particular, see the variance partitioning approach of de Heer et al., 2017). For instance, sentence vectors associated with different models could have been stacked together to form one large composite predictor in each multiple regression. For the case at hand, the computational overheads associated with fitting the regression to a composite of multiple models with thousands of features and potentially repeating this to find λ combinations for the constituent models made this approach prohibitive.

Voxel selection and averaging sentence replicates

Because not all fMRI voxels contain informative signal, we estimated which ones were likely to be informative using a commonly used "stability" strategy (e.g., Mitchell et al., 2008; Chang et al., 2011; Pereira et al., 2013; Anderson et al., 2015, 2017b, 2019b; J. Wang et al., 2017; Yang et al., 2017). At each cross-validation iteration, we took the subset of training sentences and estimated which voxels were informative by taking each of the 12 (or 6) fMRI runs through training sentences and voxelwise correlating each unique pair of runs together. For the 10 participants with 12 runs, this left 66 pairwise correlation coefficients per voxel; and for the 4 participants with 6 runs, this left 15 pairwise correlation coefficients per voxel. A single stability score was assigned to each voxel by taking the mean of these (66 or 15) correlation coefficients. We selected and then segmented the *n* voxels with the highest correlation coefficients to enter the voxelwise encoding modeling analysis, while discarding the other voxels.

The value for *n* differed for our initial analysis that sampled voxels across the entire cortex where *n* was 500 to echo the number chosen in Mitchell et al.'s (2008) seminal analysis, and for our ROI analysis (where *n* was 50 within each ROI, maintaining consistency with a previous study: Anderson et al., 2019b). For the cortex-level analysis, we excluded the occipital pole and calcarine sulcus ROIs (ctx_lh_Pole_occipital, ctx_rh_Pole_occipital, ctx_lh_S_calcarine, ctx_rh_S_calcarine) from voxel selection to cut down on signal associated with early visual processing of the sentence stimuli (which we were not interested in testing).

Both numbers of voxels selected (500 and 50) are ultimately arbitrary; however, we are confident that this particular parameterization has little bearing on the pattern of critical results we present here. For instance, in the current article, we often obtain similar patterns of results across models whether looking at 50 voxels within an individual ROIs or 500 voxels across the brain. Likewise, in previous work, we have observed model-based results to be robust when different numbers of voxels are explicitly tested (Anderson et al., 2019b).

Following voxel selection, and before regression, fMRI activation patterns for the 12 (or 6) replicates of the same sentence were voxelwise averaged. This produced a single fMRI representation for each of the 240 sentences for each of the 14 participants.

Cross-participant similarity-based encoding

To provide an estimate of how much signal in each participant's fMRI data had not been predicted by the models, we computed a cross-participant analysis. This analysis used fMRI data from other participants as a basis for predicting sentence representations in the test participant. This followed the reasoning that, in the general case, commonalities in representation estimated across a group will provide the strongest estimate of a separate individual, assuming a sufficiently sized group, and the lack of personalized models (see also Anderson et al., 2020).

To perform cross-participant encoding, we applied a representational similarity-based approach to sidestep problems associated with structural and functional misalignments across participants' fMRI data (that persist despite anatomic normalization). This is a simpler and more approximate alternative to hyperalignment (Haxby et al., 2011) or stacking fMRI data across multiple participants to use as predictors in a large multiple regression (which rapidly becomes computationally prohibitive). Nonetheless, we have found it to be practically effective in related work (Anderson et al., 2016, 2019b) as we do again here.

Cross-participant similarity-based encoding (Fig. 5) uses the array of similarities computed between a test sentence, and a selection of training sentences from one participant's fMRI data as the basis for predicting the representation of the test sentence in a different participant (from their corresponding training sentences). Similarities were β coefficients arising from the regression of the test sentence fMRI vector on each of the training sentences. Thus, if there are 230 training sentences, there were 230 β coefficients. The fMRI representation of the test sentence in a new participant was predicted via a weighted average of the second participant's 230 training sentences, where the weights were the original participants' β coefficients. Group-level commonalities were estimated by taking the mean of β coefficients across multiple participants. For instance, the pointwise average of the 230 β coefficients across Participants 2-14 was applied to encode the test sentence in Participant 1.

Tests of statistical significance

The accuracy that each model could predict fMRI representations of sentences (whether at cortex or ROI level) was scored at an individual participant level, using the normalized rank score and normalized semipartial rank score metrics described in Rank scoring prediction accuracy and illustrated in Figure 4. Both rank scores ranged from 0 to 1, with 0.5 representing the theoretical chance level (i.e., the result to be expected if the analysis were repeated with models and fMRI data that contained no correlated signal).

To provide an individual participant-level estimate of the statistical significance of the rank scores generated by each model, we ran permutation tests. Model sentence vectors were randomly shuffled so that they were no longer aligned with their sentence labels (e.g., such that the representation of "meaning" was misaligned with the text label). The fMRI data remained untouched. Normalized rank scores were computed with the shuffled sentences. This process was repeated 100 times with different random shuffles to supply a null distribution of scores. The 101 scores (associated with shuffled and unshuffled models) were ranked in descending order, and a *p*-value was computed as the rank associated with the unshuffled score divided by 101.

To test for the generality of results across participants, nonparametric signed ranks tests were applied. To test for differences in prediction accuracy between models, signed rank tests were applied to the 14 participants' normalized rank scores arising from each model. To test whether individual models could exclusively predict fMRI signal components that were not predicted by other models, normalized semipartial rank scores for the 14 participants were compared with the chance level of 0.5 using one-sample signed rank tests.

Multiple comparisons were corrected for as indicated in the Results using false discovery rate (FDR) (Benjamini and Hochberg, 1995).

Data and code availability

Preprocessed fMRI data and MATLAB version 2020a code to compute all analyses are available at https://osf.io/7uvmg/.

Language network ROIs were downloaded from https://evlab.mit.edu/funcloc/, selecting the download option for "A subset of the original parcels from Fedorenko et al. (2010) which include the 8 parcels in the frontal and temporal lobe" and ["... flipped onto the RH"].

Pretrained GloVe was downloaded from https://nlp.stanford.edu/projects/glove.

Pretrained InferSent was downloaded from https://github.com/facebookresearch/InferSent.

Pretrained ELMo was downloaded from AllenNLP: https://allennlp.org/elmo.

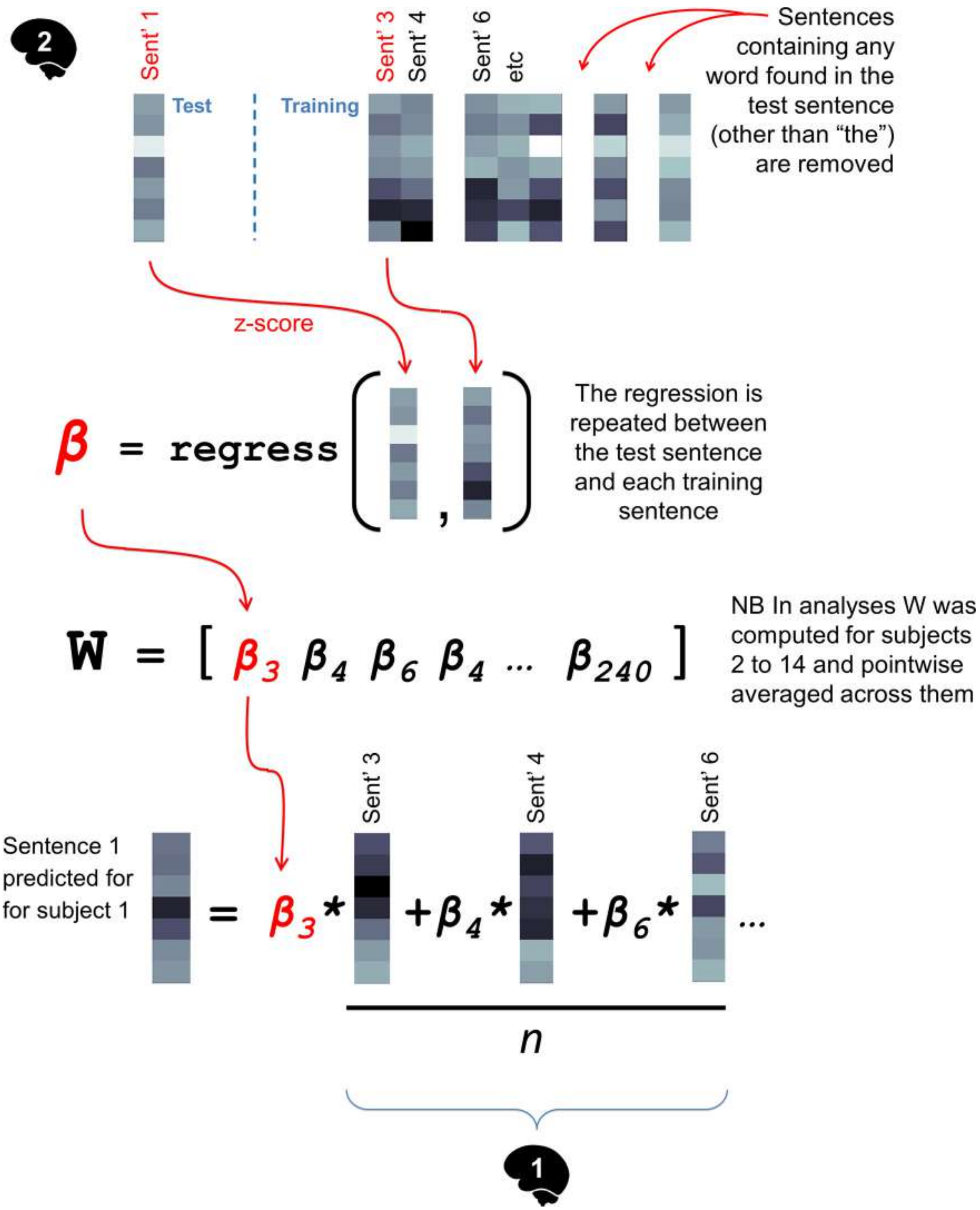

2
Sent' 1
Test
Training
Sent' 3
Sent' 4
Sent' 6
etc
Sentences containing any word found in the test sentence (other than "the") are removed
z-score
β = regress(,)
The regression is repeated between the test sentence and each training sentence
W = [β3 β4 β6 β4 ... β240]
NB In analyses W was computed for subjects 2 to 14 and pointwise averaged across them
Sentence 1 predicted for for subject 1
Sent' 3
Sent' 4
Sent' 6
= β3* +β4* +β6* ...
n
1

Figure 5. Using cross-participant, similarity-based encoding to predict the fMRI representation of a sentence for Participant 1, from Participant 2's data. This approach was used to estimate an upper bound on prediction accuracy and reveal signal that had not been predicted by the models (when the weight vectors W were averaged across participants).

Pretrained BERT was downloaded from https://huggingface.co/transformers/pretrained_models.html. We selected the "bert-large-uncased" implementation, run under PyTorch 1.6.0.

Results

fMRI data reflect propositional sentence-level semantics encoded by InferSent

To first establish how accurately fMRI activation could be predicted by InferSent, the three GloVe-baseline models, and the grammatical structure model, we computed normalized rank scores (Fig. 4, left) for each model and participant in a cortex-level analysis on 500 stable voxels (for a neuroanatomical illustration of the voxels, see Fig. 6, right). All models predicted fMRI sentence representations at accuracies that were significantly greater than chance level (permutation-based p values were 0.01 for each model and participant, with 100 permutations). Comparative results for the different models are illustrated in Figure 6. From qualitative visual inspection of Figure 6 (top left), it is clear that InferSent yielded the strongest rank scores, followed by Seq(GloVe), Gramm(GloVe), BoW(GloVe), and Grammar. This was supported by the results of two-tailed signed rank tests that compared scores between model pairs (results shown in Fig. 6, top left).

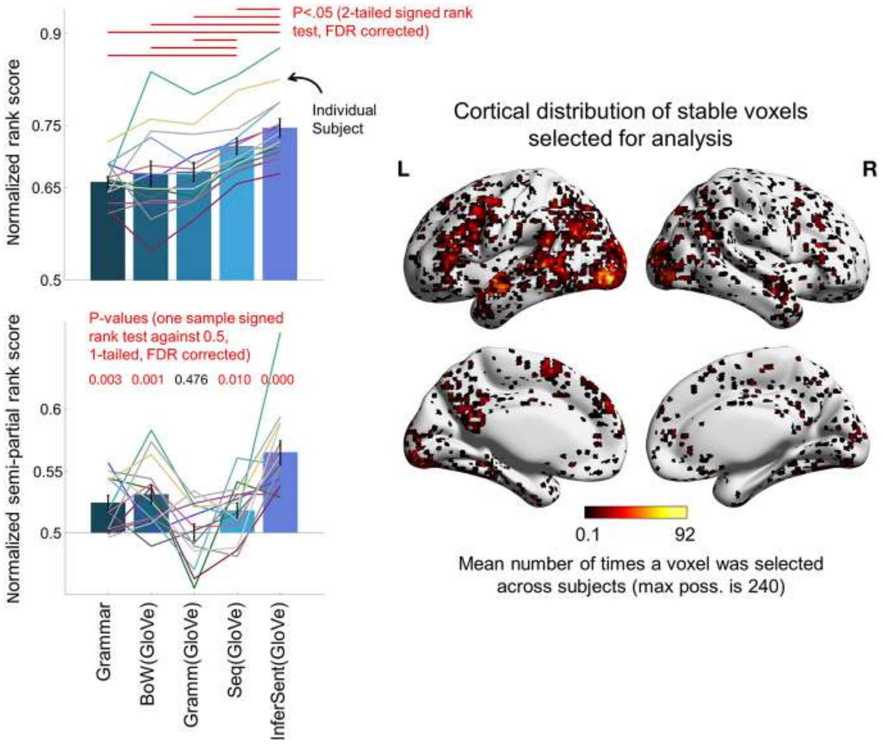


Figure 6. fMRI activation patterns reflect propositional sentence-level semantics encoded by InferSent. Results correspond to the 500 most stable voxels per participant, reselected across the cortex at each cross-validation iteration (right). Left column displays the critical results. Top left bar plot represents that normalized rank scores associated with InferSent were greater than each of the other models. Top right bar plot, InferSent predicted components of fMRI activation that could not be predicted by the other models. Bar heights are mean scores across the 14 participants. Error bars indicate SEM. The Char stats and Word overlay models were included as controls in all semipartial rank score analyses but are not illustrated to simplify the display and are later included in Figure 9. A complete listing of statistical test values for the two-tailed signed rank tests (top left) and for one-sample signed rank tests (bottom left) is in Extended Data Figures 6-1 and 6-2, respectively.

To estimate whether InferSent could exclusively predict components of sentence activation in the fMRI data that had not been predicted by the combination of Seq(GloVe), Gramm(GloVe), BoW(GloVe), Grammar, and the visual appearance models (Char stats and Word overlay), we computed normalized semipartial rank scores (Fig. 4, right). This analysis was undertaken on the same 500 stable voxels as the previous rank score analysis (Fig. 6, right). Results for the different models are illustrated in Figure 6 (bottom left) and revealed that InferSent could predict components of activation in sentence-level fMRI data that could not be predicted by any of the other models (mean ± SD, semipartial rank scores were 0.065 ± 0.035). Statistical significance was evaluated using one-sample signed rank tests to evaluate whether normalized semipartial rank scores were greater than the chance level of 0.5. For InferSent, this was found to be significant (W = 105, p = 0.0003, one-tailed, FDR-corrected across the six models). We also detected evidence that Seq(GloVe), Bow (GloVe), and Grammar, but not Gramm(GloVe), could exclusively predict components of fMRI activation, albeit with lower semipartial rank scores (Fig. 6, bottom left).

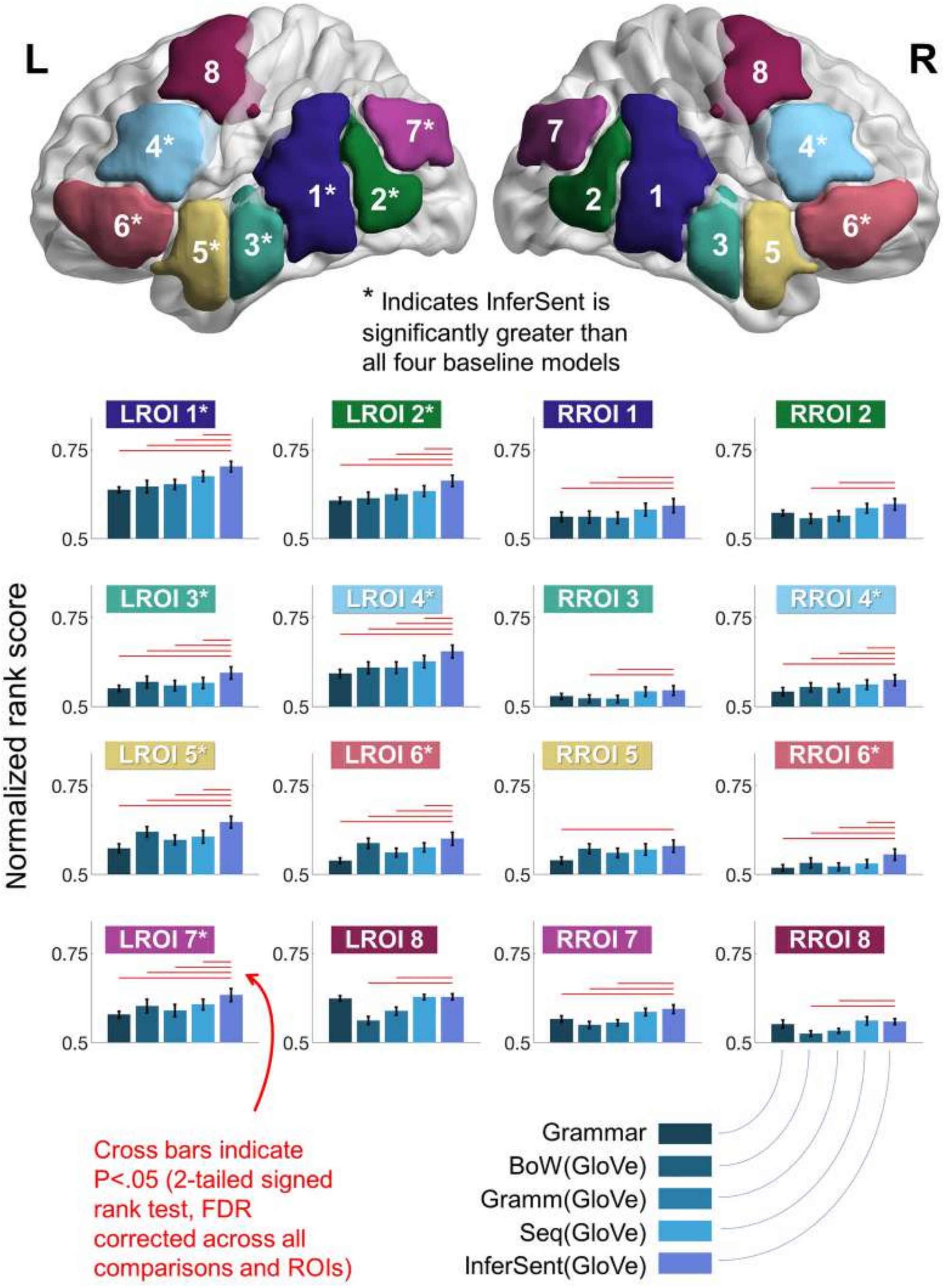
These two analyses provided evidence that propositional sentence-level semantic representations can be detected in fMRI data, but did not indicate whether sentence-level representations were anatomically localized to a particular subset of the 500 voxels analyzed, as opposed to being distributed across multiple cortical sites.

Multiple cortical regions encode propositional sentence-level semantics

To estimate whether different regions of a distributed cortical language network locally encoded propositional sentence-level semantics, we evaluated voxelwise encoding models on regional fMRI activation. We repeated the analyses of the previous section and computed normalized rank scores and semipartial normalized rank scores (Fig. 4) for each participant and model on each of the 16 language network ROIs. Results are displayed in Figures 7 and 8, respectively, and ROI-level results broadly echo the outcomes of the cortex-level analysis (Fig. 6). To simplify display, we did not illustrate rank scores for the visual appearance models in Figure 7, but we list them in Table 2.

As illustrated in Figure 7, InferSent yielded significantly stronger rank scores than all four of the baseline models in 7 of 8 left hemispheric ROIs. These included anterior, mid, and posterior regions of the temporal lobe, inferior parietal cortex, and inferior frontal gyrus, but excluded the mid-frontal ROI (LROI 8). The same effect was observed in 2 of 8 right hemispheric ROIs (both right inferior frontal: RROI 4 and 6). We note here that the rank scores for right hemispheric ROIs tended to be visibly weaker than their left hemispheric counterparts, suggesting lower signal and potentially less power to discern differences between models. Statistical significance was evaluated with two-tailed signed rank tests that compared InferSent with each of the four baseline models. These four tests were repeated for each ROI, and the entire set of p values were FDR-corrected. All test statistics are listed in full in Extended Data Figure 7-1.

As illustrated in Figure 8, InferSent exclusively predicted components of fMRI activation in multiple regions of the language network that were not predicted by: Seq(GloVe), Gramm(GloVe), BoW(GloVe), Grammar, Char stats, and Word overlay when combined together. Specifically, in the left hemisphere, semipartial rank scores yielded by InferSent were significantly greater than chance level in 7 of 8 temporal, inferior parietal, and inferior temporal ROIs (one-sample signed rank tests against 0.5, one-tailed, FDR-corrected). These were the same left hemispheric ROIs for which InferSent previously yielded significantly stronger rank scores than the four baseline models (Fig. 7). In the right hemisphere, the same effect was observed in 5 of 8 ROIs, which included inferior frontal cortex (RROI 4 and 6 echoing Fig. 7) and additionally regions of anterior and posterior temporal cortex. Other noteworthy contributions to prediction came from the Grammar baseline model in mid frontal ROI (LROI 8). Also, Grammar, Bow(GloVe) and Seq(GloVe) contributed to predicting LROI 1 (posterior temporal lobe). As we considered these baseline model results to be peripheral to our main aims, they are not discussed further. p values



* Indicates InferSent is significantly greater than all four baseline models

Cross bars indicate P<.05 (2-tailed signed rank test, FDR corrected across all comparisons and ROIs)

Figure 7. Regions of temporal, inferior parietal cortex, and inferior frontal cortex were more accurately predicted by InferSent than the four baseline models. Bar heights illustrate the mean normalized rank scores across the 14 participants. Error bars indicate SEM. The anatomic location of the ROIs is illustrated at the top: 50 voxels were selected within each ROI for analysis at each cross-validation iteration (see Table 1). InferSent yielded significantly greater normalized rank scores than all four of the baseline models in 9 of 16 ROIs (*). A complete listing of statistical test values for the two-tailed signed rank tests used to compare InferSent to rank scores derived from the baseline models is in Extended Data Figure 7-1.

were FDR-corrected across all models and ROIs. All test statistics are listed in full in Extended Data Figure 8-1.

Collectively, these results provided evidence that propositional sentence-level representations are encoded within and across multiple sites spanning temporal, parietal, and frontal cortex, and not localized to any particular brain region.

Different deep artificial neural network sentence model predicted fMRI data with similar accuracy

To place the current results acquired using InferSent into a broader landscape of other deep network approaches, we explored how cortex-level predictions made by InferSent compared with ELMo and BERT. We were interested to see whether ELMo and BERT's strong NLP performance and added complexity carried over to strong fMRI prediction accuracy.

To recap, ELMo has three layers: L1 is the representation derived from a character-based convolutional network, and L2 and L3 are contextualized bidirectional LSTM-based representations. We analyzed each layer separately, and all layers concatenated together (L1,L2,L3). BERT has 24 layers reflecting contextualized representations generated by self-attention-based encoder blocks. We

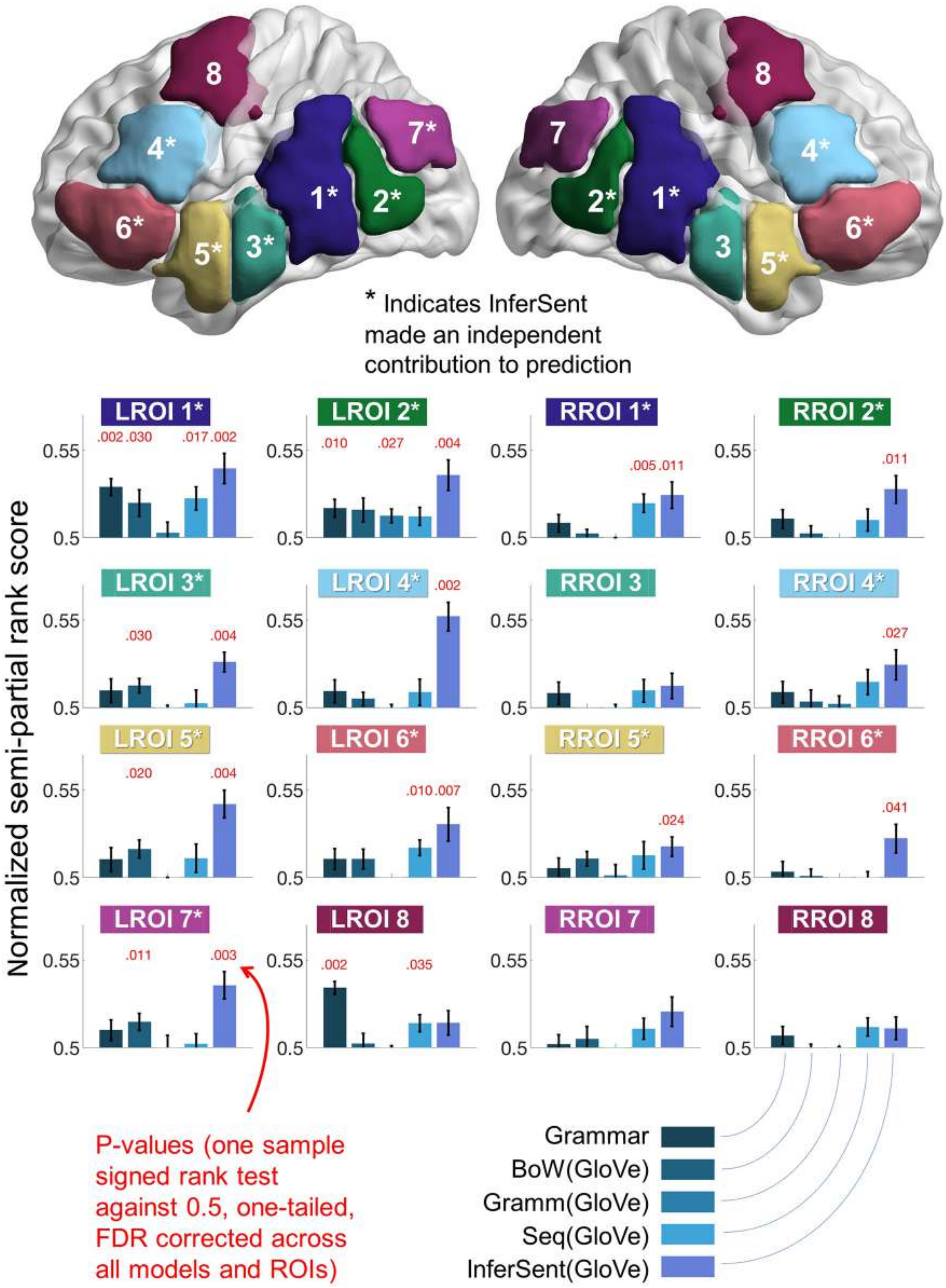


Figure 8. InferSent made independent contributions to predicting fMRI activation within temporal, inferior parietal, and inferior frontal cortex. Bar heights illustrate mean semipartial rank scores across the 14 participants. Error bars indicate SEM. The anatomic location of the ROIs is illustrated at the top: 50 voxels were selected within each ROI for analysis at each cross-validation iteration (see Table 1). InferSent predicted components of fMRI sentence representations that were not predicted by the other models in 12 of 16 ROIs (*). The Char stats and Word overlay models were included as controls in all of the semipartial analyses, but the corresponding rank scores are not illustrated here to simplify display. A complete listing of statistical test values for the one-tailed sample signed rank tests used to test semipartial rank scores against chance level is in Extended Data Figure 8-1.

analyzed each of BERT's 24 layers separately because regression on all 24 layers concatenated together or on combinations of fewer layers would have been computationally prohibitive at this stage of investigation.

Normalized rank scores for ELMo and BERT are presented alongside all other approaches tested in this article in Figure 9. InferSent, ELMo (L2 or L3 or L1,L2,L3), and BERT's best performing layers all yielded broadly similar normalized rank

Table 2. Normalized rank score listings for the visual appearance models, to support Figure 7[a]

ROI	Char stats	Word overlay	InferSent(GloVe)
LROI 1	0.67 ± 0.03	0.66 ± 0.03	0.70 ± 0.06
LROI 2	0.63 ± 0.03	0.63 ± 0.03	0.66 ± 0.06
LROI 3	0.58 ± 0.03	0.58 ± 0.03	0.59 ± 0.06
LROI 4	0.63 ± 0.04	0.62 ± 0.03	0.65 ± 0.07
LROI 5	0.60 ± 0.03	0.59 ± 0.03	0.65 ± 0.06
LROI 6	0.57 ± 0.03	0.56 ± 0.03	0.60 ± 0.07
LROI 7	0.63 ± 0.02	0.62 ± 0.02	0.63 ± 0.07
LROI 8	0.66 ± 0.03	0.64 ± 0.03	0.63 ± 0.03
RROI 1	0.59 ± 0.05	0.58 ± 0.04	0.59 ± 0.07
RROI 2	0.61 ± 0.03	0.60 ± 0.03	0.60 ± 0.06
RROI 3	0.55 ± 0.04	0.55 ± 0.04	0.55 ± 0.05
RROI 4	0.57 ± 0.04	0.57 ± 0.03	0.57 ± 0.06
RROI 5	0.56 ± 0.05	0.56 ± 0.04	0.58 ± 0.06
RROI 6	0.54 ± 0.03	0.53 ± 0.03	0.56 ± 0.06
RROI 7	0.63 ± 0.05	0.63 ± 0.04	0.59 ± 0.05
RROI 8	0.58 ± 0.04	0.58 ± 0.03	0.56 ± 0.03

[a] Data are mean ± SD. Scores for InferSent(GloVe) are illustrated in Figure 7 and are provided again here as a reference point. Scores for Char stats and Word overlay are quite high throughout the language network. This presumably reflects orthographic information that is distributed across the cortex as well as spurious correlations with the semantic models observed in Figure 3.

scores. Mean ± SD scores across participants were as follows: InferSent, 0.747 ± 0.056; ELMo L1, 0.713 ± 0.060; ELMo L2, 0.751 ± 0.058; ELMo L3, 0.748 ± 0.056; ELMo L1,L2,L3, 0.754 ± 0.059. BERT's strongest mean rank scores were derived from mid-level layers as follows: L10, 0.758 ± 0.052; L11, 0.755 ± 0.053; L12, 0.761 ± 0.053; L13, 0.758 ± 0.053; L14, 0.757 ± 0.055. BERT's earlier layers yielded scores of ~0.74, while scores for later layers tailed off to 0.70.

Signed ranks tests detected no significant differences between InferSent and ELMo L1,L2,L3 (W = 79, $p = 0.104$, uncorrected). Signed ranks tests between InferSent and each layer of BERT revealed a significant difference at L12 only when results were not corrected for multiple comparisons (BERT > InferSent, W = 88, $p = 0.025$ uncorrected) and at L18-L24 (InferSent > BERT all $W \geq 99$ and all $p < 0.002$ uncorrected). Signed ranks tests between ELMo and BERT revealed significant differences at L1 to L3 and L17 to L24 (ELMo > BERT, all $W \geq 96$ and all $p \leq 0.004$, uncorrected).

This section has revealed that, for the current fMRI dataset, there were relatively mild differences in prediction accuracy between InferSent, ELMo, and BERT. BERT L12 yielded the highest prediction accuracy of all, but this was not significantly more accurate than ELMo or InferSent. We did not explicitly test whether InferSent, ELMo, and BERT made independent contributions to predicting fMRI data because these could be because of a mixture of uncontrolled differences in network architecture, training paradigms, training data, word/subword inputs, and so on. However, as a byproduct of an analysis performed in the next section, we did uncover evidence that each deep network approach could predict different (uninterpretable) components of fMRI signal (Fig. 9, top right bar plot).

The experiential attribute model revealed semantic signal that was not predicted by deep network sentence models

To estimate how much room there is for improvement in computationally modeling the current fMRI data, we ran two additional analyses. The first leveraged an experiential attribute model (Binder et al., 2016), which was designed to approximate semantic knowledge acquired from sensory, motor, cognitive, interoceptive, and affective experience, and thus potentially capture information that is not available from the text corpora used in deep network training.

Rank scores for Seq(Attribute), comparative to all other models are displayed in Figure 9 (left). The mean ± SD score for Seq (Attribute) was 0.722 ± 0.045, which was similar in magnitude to Seq(GloVe) (0.717 ± 0.0741) and significantly lower than the deep network approaches (see signed rank tests in Fig. 9). To estimate whether Seq(Attribute) predicted fMRI signal components that were not captured by the deep networks, we estimated semipartial rank scores for Seq(Attribute) when controlling for InferSent, ELMo L1,L2,L3, BERT L12, Grammar, Char stats, and Word overlay. The results are illustrated in Figure 9 (top right bar plot). Mean ± SD semipartial rank scores for Seq(Attribute) were 0.054 ± 0.026, which was significantly greater than the chance level of 0.5 (W = 105, $p = 0.0001$, one-tailed, FDR-corrected across all models).

These results provided evidence that there was semantic signal in the fMRI data that had not been predicted by the deep network models. This may reflect knowledge that was acquired through nonlinguistic experience of the word, as opposed to from text and language (Anderson et al., 2019b).

Cross-participant encoding revealed fMRI signal that was not predicted by any model

To estimate whether any fMRI signal was left unpredicted by the models regardless of its nature (semantic or otherwise), we applied cross-participant encoding. To recap, cross-participant encoding assumes that, in the absence of a personalized model (e.g., Anderson et al., 2020) and when armed with a sufficiently composed participant group, the best predictor of fMRI activation for an unseen sentence in an individual will be derived from fMRI activation for that sentence in a group of other individuals.

We first computed rank scores using cross-participant similarity-based encoding at the cortex level, which yielded a mean ± SD normalized rank score of 0.793 ± 0.030, which was significantly greater than all other models (Fig. 9, bottom left; for signed rank test outcomes, see Extended Data Fig. 9-1). To reveal signal that was exclusively predicted by the cross-participant approach, we computed semipartial rank scores in an analysis that included InferSent, ELMo L1, ELMo L1,L2,L3, Seq (Attribute), Grammar, Char stats, and Word overlay. The results are illustrated in Figure 9 (bottom right bar plot). Cross-participant semipartial rank scores were comparatively high (0.153 ± 0.030). Semipartial rank scores for the other models were low (all <0.033), but some were still >0, which suggests that the current cross-participant approach was an underestimate of the true upper bound prediction accuracy.

These results provided evidence that there was substantial fMRI signal that had not been predicted by the models and, more broadly, that there is room for improvement in building sentence models to predict the current fMRI data. In addition, this section has suggested that the current cross-participant predictions have some room for improvement which might come from including more participants or applying more sophisticated strategies to integrate data across individuals (e.g., Haxby et al., 2011).

Discussion

The current study has revealed evidence that propositional sentence-level meaning is encoded in fMRI activation within and across regions of a previously identified cortical language network (Fedorenko et al., 2010). To model sentence-level meaning,

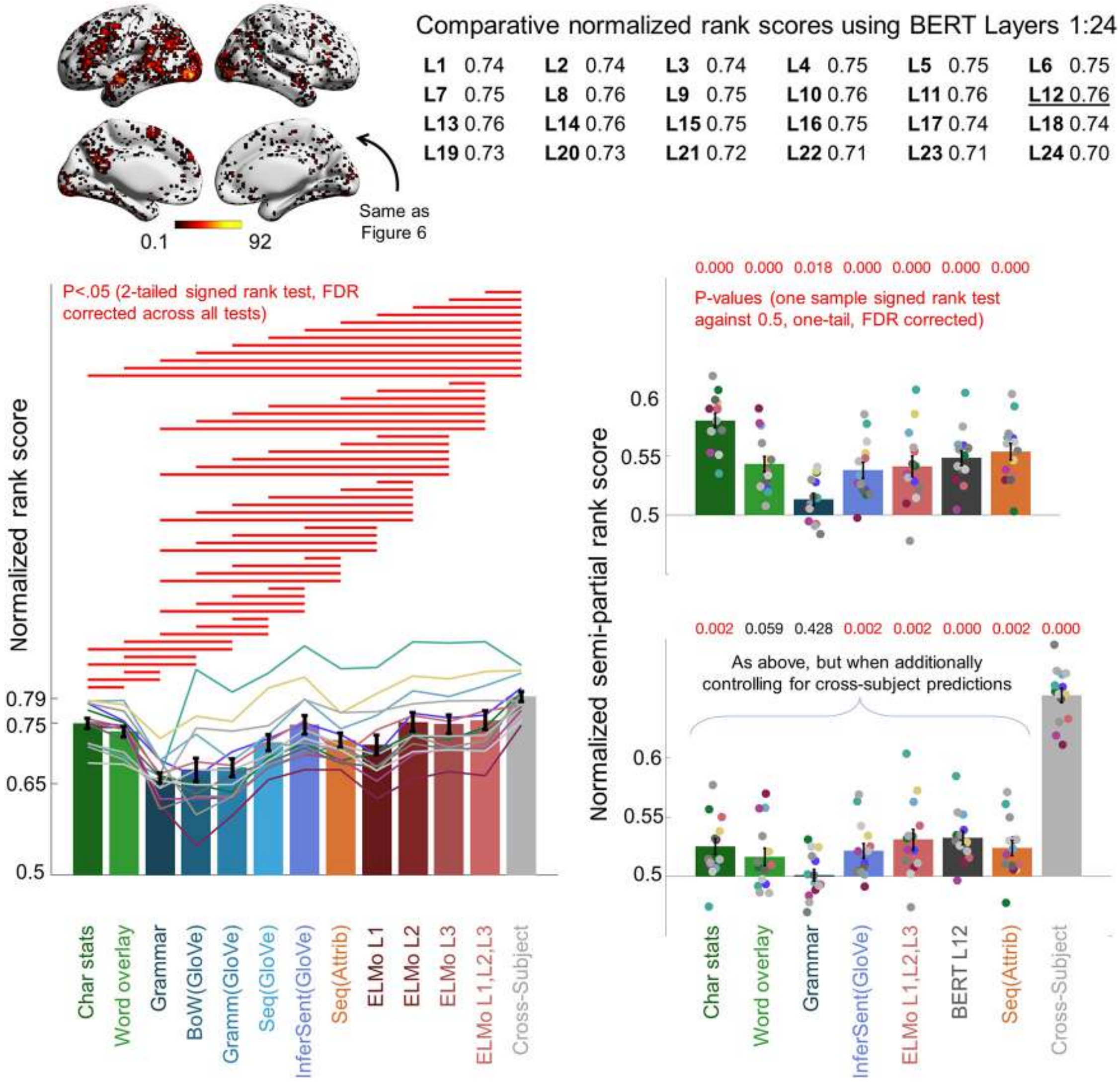


Figure 9. InferSent, ELMo, and BERT yielded similar accuracy predictions, and the experiential attribute model and cross-participant encoding revealed unpredicted fMRI signal. Results correspond to the 500 most stable voxels per participant (top left). Bottom left, Normalized rank scores obtained for the full complement of predictive approaches excluding BERT. Mean rank scores for BERT are listed at the top right to simplify display (because BERT has 24 layers). Right, Semipartial rank scores for a selection of models. Top right, Semipartial scores when the cross-participant encoding approach was excluded. Bottom right, Comparative semipartial scores when cross-participant encoding was included and semipartial scores for the models were expected to dwindle toward 0.5 (because the cross-participant approach in principle accounts for group-level commonalities in neural signal). Bar heights and error bars represent mean and SEM, respectively. Circles represent individual participants. A complete listing of statistical test values for the two-tailed signed rank tests (left bar plot) is in Extended Data Figure 9-1. A complete listing of test values for the one-sample signed rank tests in the top and bottom right bar plots is in Extended Data Figures 9-2 and 9-3, respectively.

we used InferSent, a recurrent nonlinear deep neural network trained to combine sequences of word-level semantic vectors into unified propositional sentence-level representations. Our results showed that InferSent exclusively predicted components of activation in fMRI data that was unaccounted for by baseline models that superposed and/or segregated word-level semantic vectors. We first discuss what aspects of sentence-level semantics that InferSent may have captured that the other models did not, and how this information is distributed across the cortex.

We use the term "propositional" here to refer to sentence-level information that is required to compute entailment relationships. In this sense, the propositional meaning of a sentence combines the meanings of its component words according to their functional roles in the sentence structure. That is, syntactical information must be inferred and integrated with lexical-semantic information, at least to some extent. The precise nature of the propositional information encoded by InferSent is however challenging to pin down, and indeed explaining deep network representations is notoriously difficult (for an examination of the information captured by InferSent and related approaches, see also Conneau et al., 2018). In part, this is because InferSent had opportunities to exploit several interacting factors to build propositional sentence representations. We outline these, but also stress that future work will be necessary to establish the importance of each.

Word sense selection

Behavioral experiments provide evidence that sentence comprehension is characterized by the initial activation of context-independent word representations, such that multiple senses of words, such as "bat," are jointly activated, followed later by sense selection when the appropriate meaning is specified by contextual information (e.g., Swinney, 1979; Tanenhaus et al., 1979; Till et al., 1988). All of the current baseline models had no capacity to select/deselect appropriate/inappropriate word senses. However, InferSent's LSTM architecture does afford this opportunity. Specifically, at each network cycle, the input gate nonlinearly filters the incoming word (GloVe) for features to incorporate into the forthcoming sentence representation, guided by the previous "sentence-so-far" output (that is recurrently fed back). Likewise, "forgettable" features are nonlinearly filtered away from the cell memory that stores an internal sentence representation. Thus, both gating operations could deselect semantic features associated with inappropriate word senses and perform sense selection.

Emphasizing words dominating sentence meaning

Electrophysiological studies have established that neural responses to different words in sentences vary considerably, with the strength of response roughly proportional to how unexpected a word is in the context (Kutas and Federmeier, 2011). Thus, "The dentist brushed the boy's tree" would elicit a greater electrophysiological response (the so-called N400) than if the ending had been "teeth." It is reasonable to consider that fMRI activation associated with semantic features of unexpected words may likewise be exaggerated. It is also reasonable to consider that function words emphasize particular content words ("the bat" rather than "bat" or "a bat"). The current baseline models had no capacity to distinguish and weight unexpected words, nor leverage function words to adapt content word meaning. However, such weightings could have been enabled by InferSent's LSTM. Specifically, the input gate performs a weighted integration of the new GloVe input and the previous sentence-so-far output (recurrently fed-back). Thus, a combination of inhibitory weights on features of the sentence-so-far, and excitatory weights on the new GloVe input would zero out recurring features across words and emphasize novel unexpected information in the new input.

Word order and thematic role assignment

The subject-verb-object word order of the English language is critical for understanding thematic role assignments (e.g., agent and patient) and thus sentence comprehension. Recent studies have provided evidence that agent and patient can be spatially distinguished in fMRI activation (Frankland and Greene, 2015; J. Wang et al., 2016). InferSent likewise has some capacity to map words to different segments of sentence-level vectors. Indeed, relative word order is encoded on each iteration when the new GloVe input word is concatenated with the previous sentence-so-far output (recurrently looped back).

How the results relate to other brain imaging studies

The current results have provided evidence that propositional sentence-level semantic representations are distributed throughout a cortical language network, rather than being localized within any particular brain region (e.g., Baron and Osherson, 2011; Bemis and Pylkkänen, 2011; Westerlund and Pylkkänen, 2014; Frankland and Greene, 2015; Zhang and Pylkkänen, 2015). This echoes Fedorenko et al. (2016) and Nelson et al. (2017) who revealed electrocorticographic signals associated with sentence construction arising concurrently in distributed brain regions, and Lyu et al. (2019) who detected contextualized noun representations in EEG/MEG recordings. Results extend previous fMRI studies that have revealed distributed neural correlates of semantic features (Huth et al., 2016; Anderson et al., 2017a, 2019b; Yang et al., 2017; Pereira et al., 2018), syntax (Blank et al., 2016; Fedorenko et al., 2020); and words with different grammatical roles (Anderson et al., 2019a); but not whether activation reflected sentence-level meaning. Finally, our results complement work that has used cross-participant comparisons to reveal distributed neural correlates of narrative comprehension (Honey et al., 2012) and how temporal context is encoded (e.g., Chien and Honey, 2020); and also other studies that have begun to use deep network-based models to predict contextualized semantic responses in neural data (Jain and Huth, 2018; Gauthier and Levy, 2019; Sun et al., 2019; Toneva and Wehbe, 2019; Goldstein et al., 2020; Heilbron et al., 2020; Schrimpf et al., 2020), but not revealed the distributed cortical encoding of propositional sentence-level meaning.

Together with the above studies, the current results suggest that lateral temporal, inferior parietal, and inferior frontal cortices comprehensively encode propositional sentence representations that integrate across multiple words and semantic features. However, because of the slow sample rate of fMRI, the time course over which semantic representations arose in different regions remains unclear. Thus, we are unable to estimate whether semantic representations were locally assembled in a particular brain region, such as the anterior temporal lobe and subsequently channeled to different cortical regions; or alternatively, whether sentence-level representations were constructed in parallel across multiple interacting network hubs. Future work leveraging electrocorticography and EEG/MEG will be necessary to expose the spatiotemporal time course of semantic composition (e.g., Fedorenko et al., 2016; Fyshe et al., 2019; Lyu et al., 2019; Toneva and Wehbe, 2019; Caucheteux and King, 2020; Goldstein et al., 2020; Heilbron et al., 2020; Lopopolo et al., 2020).

Similarities and differences between the deep artificial neural network approaches

In follow-up analyses, we placed InferSent's results into the context of two newer deep network approaches: ELMo and BERT. This was to find out how high performance in applied NLP translates to fMRI prediction, and whether the added complexity of the newer deep networks conferred a particular advantage here. The results were equivocal. There were slight differences in fMRI prediction accuracy between the three deep networks, but they all yielded stronger results than the baselines. This might partially reflect representational similarities that the different approaches converged on. Such putative similarities could reflect priors that are incidentally encoded by different deep network architectures regardless of training (Conneau et al., 2018; Wieting and Kiela, 2019) and/or that different semi-supervised/supervised training paradigms produce similar representations. Irrespective, we consider each approach tested here to provide the basis for approximating propositional meaning because ELMo and BERT have both contributed to models breaking SNLI benchmarks (https://paperswithcode.com/sota/natural-language-inference-on-snli) and InferSent's LSTM architecture and training were selected/optimized for SNLI.

We are also careful to point out that results for different deep networks, and their biological plausibility may vary with the brain-imaging task scanned and with fine-tuned training on particular NLP tasks. For instance, architecturally, transformers might better capture knowledge assimilation associated with reading a book when one can flip pages (and one's attention) back and forth to integrate information across long-term dependencies. Conversely, recurrent networks could be more biologically plausible models of speech comprehension, where words are delivered in serial order (see also Merkx and Frank, 2020). Training on next-word prediction could be particularly important for building biologically plausible models of both reading and speech comprehension (Goldstein et al., 2020; Heilbron et al., 2020; Schrimpf et al., 2020), given the extensive evidence that prediction underpins biological language comprehension (Kuperberg and Jaeger, 2016).

Utility of incorporating supra-textual information into semantic models

The study has also emphasized the utility of incorporating supra-textual knowledge into semantic models to explain brain activation. This is in two respects: (1) in corroborating previous findings that modeling nonlinguistic "experiential knowledge" (Anderson et al., 2013, 2015, 2017b, 2019a; Bulat et al., 2017; Abnar et al., 2018; X. Wang et al., 2018; Djokic et al., 2020) can help predict brain activation that cannot fully be explained by text-based language models; and (2) by incorporating human "expert" knowledge on natural language inferences (Bowman et al., 2015) to supervise the InferSent training procedure.

Conclusion

In conclusion, the current study has provided evidence that a distributed cortical network encodes propositional representations of sentence-level meaning. This suggests that unified, integrated representations of sentence meaning are locally encoded in fMRI within multiple brain regions rather than being confined to a particular site. The study has also demonstrated the utility of deep network approaches to capture components of semantic information that have yet to be explained by any other method. However, despite combining state-of-the-art computational and behavioral modeling methods, we have revealed that a substantial fraction of fMRI signal remains unexplained. In going forward, we contend that artificial deep neural network approaches that integrate multimodal information will play a vital role to play in explaining this biological signal.

References

Abnar S, Ahmed R, Mijnheer M, Zuidema W (2018) Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. In: Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL), pp 57–66. Salt Lake City: Association for Computational Linguistics.

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. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1960–1970. Seattle: Association for Computational Linguistics.

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, Zinszer BD, Raizada RD (2016) Representational similarity encoding for fMRI: pattern-based synthesis to predict brain activity using stimulus-model-similarities. Neuroimage 128:44– 53.

Anderson AJ, Binder JR, Fernandino L, Humphries CJ, Conant LL, Aguilar M, Wang X, Doko D, Raizada RD (2017a) Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. Cereb Cortex 27:4379–4395.

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

Anderson AJ, Lalor EC, Lin F, Binder JR, Fernandino L, Humphries CJ, Conant LL, Raizada RD, Grimm S, Wang X (2019a) Multiple regions of a cortical network commonly encode the meaning of words in multiple grammatical positions of read sentences. Cereb Cortex 29:2396–2411.

Anderson AJ, Binder JR, Fernandino L, Humphries CJ, Conant LL, Raizada RD, Lin F, Lalor EC (2019b) An integrated neural decoder of linguistic and experiential meaning. J Neurosci 39:8969–8987.

Anderson AJ, McDermott K, Rooks B, Heffner KL, Dodell-Feder D, Lin FV (2020) Decoding individual identity from brain activity elicited in imagining common experiences. Nat Commun 11:5916.

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

Barsalou LW, Santos A, Simmons WK, Wilson CD (2008) Language and simulation in conceptual processing. In: Symbols, embodiment, and meaning (De Vega M, Glenberg AM, Graesser AC, eds), pp 245–283. Oxford: Oxford UP.

Bemis DK, Pylkkänen L (2011) Simple composition: a magnetoencephalography investigation into the comprehension of minimal linguistic phrases. J Neurosci 31:2801–2814.

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.

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

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

Blank I, Balewski Z, Mahowald K, Fedorenko E (2016) Syntactic processing is distributed across the language system. Neuroimage 215:307–323.

Bowman SR, Angeli G, Potts C, Manning CD (2015) A large annotated corpus for learning natural language inference. arXiv 1508.05326.

Bulat L, Clark S, Shutova E (2017) Speaking, seeing, understanding: correlating semantic models with conceptual representation in the brain. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1092–1102. Copenhagen: Association for Computational Linguistics.

Caucheteux C, King JR (2020) Language processing in brains and deep neural networks: computational convergence and its limits. bioRxiv. doi: 10.1101/2020.07.03.186288.

Chang KM, Mitchell T, Just MA (2011) Quantitative modeling of the neural representation of objects: how semantic feature norms can account for fMRI activation. Neuroimage 56:716–727.

Chien HY, Honey CJ (2020) Constructing and forgetting temporal context in the human cerebral cortex. Neuron 106:675–686.e11.

Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A (2017) Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp 670–680. Copenhagen: Association for Computational Linguistics.

Conneau A, Kruszewski G, Lample G, Barrault L, Baroni M (2018) What you can cram into a single vector: probing sentence embeddings for linguistic properties. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp 2126–2136. Melbourne: Association for Computational Linguistics.

Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162–173.

de Heer WA, Huth AG, Griffiths TL, Gallant JL, Theunissen FE (2017) The hierarchical cortical organization of human speech processing. J Neurosci 37:6539–6557.

Deniz F, Nunez-Elizalde AO, Huth AG, Gallant JL (2019) The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. J Neurosci 39:7722–7736.

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 33:18906–18916.

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol 1, pp 4171–4186. Minneapolis, Minnesota, USA.

Djokic VG, Maillard J, Bulat L, Shutova E (2020) Decoding brain activity associated with literal and metaphoric sentence comprehension using distributional semantic models. Trans Assoc Comput Linguistics 8:231–246.

Fedorenko E, Hsieh PJ, Nieto-Castañón A, Whitfield-Gabrieli S, Kanwisher N (2010) New method for fMRI investigations of language: defining ROIs functionally in individual subjects. J Neurophysiol 104:1177–1194.

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

Fedorenko E, Blank I, Siegelman M, Mineroff Z (2020) Lack of selectivity for syntax relative to word meanings throughout the language network. bioRxiv. doi: 10.1101/477851.

Fernandino L, Humphries CJ, Conant LL, Seidenberg MS, Binder JR (2016) Heteromodal cortical areas encode sensory-motor features of word meaning. J Neurosci 21:9763–9769.

Forster KI (1970) Visual perception of rapidly presented word sequences of varying complexity. Percept Psychophys 8:215–221.

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

Fyshe A, Sudre G, Wehbe L, Rafidi N, Mitchell TM (2019) The lexical semantics of adjective–noun phrases in the human brain. Hum Brain Mapp 40:4457–4469.

Gauthier J, Levy RP (2019) Linking artificial and human neural representations of language. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp 529–539. Hong Kong: Association for Computational Linguistics.

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

Glenberg A (2010) Embodiment as a unifying perspective for psychology. Wiley Interdiscip Rev Cogn Sci 1:586–596.

Goldstein A, Zada Z, Buchnik E, Schain M, Price A, Aubrey B, Nastase SA, Feder A, Emanuel D, Cohen A, Jansen A (2020) Thinking ahead: prediction in context as a keystone of language in humans and machines. bioRxiv. doi: 10.1101/2020.12.02.403477.

Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 18:602–610.

Haxby JV, Guntupalli JS, Connolly AC, Halchenko YO, Conroy BR, Gobbini MI, Hanke M, Ramadge PJ (2011) A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72:404– 416.

Heilbron M, Armeni K, Schoffelen JM, Hagoort P, de Lange FP (2020) A hierarchy of linguistic predictions during natural language comprehension. bioRxiv. doi: 10.1101/2020.12.03.410399.

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780. 15

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67.

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

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.

Jain S, Huth AG (2018) Incorporating context into language encoding models for fMRI. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp 6629–6638. Montreal: Curran.

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:e8622.

Kiela D, Clark S (2014) A systematic study of semantic vector space model parameters. In: Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL, pp 21–30. Gothenburg, Sweden.

Kim Y, Jernite Y, Sontag D, Rush AM (2016) Character-aware neural language models. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp 2741–2749. Phoenix.

Landauer T, Dumais S (1997) A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol Rev 104:211–240.

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

Lopopolo A, Schoffelen JM, van den Bosch A Willems RM (2020) Words in context: tracking context-processing during language comprehension using computational language models and MEG. bioRxiv. doi: 10.1101/2020.06.19.161190.

Lund K, Burgess C (1996) Producing high-dimensional semantic spaces from lexical co-occurrence. Behav Res Methods Instrum Comput 28:203–208.

Lyu B, Choi HS, Marslen-Wilson WD, Clarke A, Randall B, Tyler LK (2019) Neural dynamics of semantic composition. Proc Natl Acad Sci USA 116:21318–21327.

Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis: connecting the branches of systems neuroscience. Front Syst Neurosci 2:4.

Kuperberg GR, Jaeger TF (2016) What do we mean by prediction in language comprehension? Lang Cogn Neurosci 31:32–59.

Kutas M, Federmeier KD (2011) Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu Rev Psychol 62:621–647.

Merkx D, Frank SL (2020) Comparing transformers and RNNs on predicting human sentence processing data. arXiv 2005.09471.

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

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

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: International Conference on Learning Representations (ICLR) Workshop. Scottsdale, AZ.

Nelson MJ, Karoui IE, Giber K, Yang X, Cohen L, Koopman H, Cash SS, L, Naccache L, Hale JT, Pallier C, Dehaene S (2017) Neurophysiological dynamics of phrase-structure building during sentence processing. Proc Natl Acad Sci USA 114:E3669–E3678.

Oldfield RC (1971) The assessment and analysis of handedness: the Edinburgh Inventory. Neuropsychologia 9:97–113.

Pennington J, Socher R, Manning CD (2014) GloVe: Global Vectors for Word Representation. In: Proc Conf Empirical Methods in Natural Language Processing (EMNLP 2014); Doha, pp 1532–1543. Qatar: Association for Computational Linguistics.

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, Ritter S, Gershman S, Kanwisher N, Botvinick M, Fedorenko E (2018) Toward a universal decoder of linguistic meaning from brain activation. Nat Commun 9:963.

Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Deep ZL (2018) Contextualized word representations. Proc 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, pp 2227–2237.

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

Schrimpf M, Blank I, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, Tenenbaum J, Fedorenko E (2020) Artificial neural networks accurately predict language processing in the brain. bioRxiv. doi: 10.1101/2020.06.26.174482.

Subramanian S, Trischler A, Bengio Y, Pal CJ (2018) Learning general purpose distributed sentence representations via large scale multi-task learning. In: Proceedings of the 2018 International Conference on Learning Representations (ICLR). Vancouver, Canada.

Sun J, Wang S, Zhang J, Zong C (2019) Towards sentence-level brain decoding with distributed representations. Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence 2019 Jul 17, Vol 33, pp 7047–7054. Honolulu, Hawaii, USA.

Swinney DA (1979) Lexical access during sentence comprehension: (re)consideration of context effects. J Verb Learn Verb Behav 18:645–659.

Talairach J, Tournoux P (1988) Co-planar stereotaxic atlas of the human brain. In: Three-dimensional proportional system: an approach to cerebral imaging. New York: Thieme.

Tanenhaus MK, Leiman JM, Seidenberg MS (1979) Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. J Verb Lear Verb Behav 18:427–440.

Till RE, Mross EF, Kintsch W (1988) Time course of priming for associate and inference words in a discourse context. Mem Cognit 16:283–298.

Toneva M, Wehbe L (2019) Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In: Advances in Neural Information Processing Systems, pp 14928–14938. Vancouver: Curran.

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: 257–264.

Wang J, Cherkassky VL, Just MA (2017) Predicting the brain activation pattern associated with the propositional content of a sentence: modeling neural representations of events and states. Hum Brain Mapp 38:4865–4881.

Wang X, Wu W, Ling Z, Xu Y, Fang Y, Wang X, Binder JR, Men W, Gao JH, Bi Y (2018) Organizational principles of abstract words in the human brain. Cereb Cortex 28:4305–4318.

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:e11257.

Wieting J, Kiela D (2019) No training required: exploring random encoders for sentence classification. In: Proceedings of International Conference on Learning Representations (ICLR). New Orleans.

Westerlund M, Pylkkänen L (2014) The role of the left anterior temporal lobe in semantic composition vs. semantic memory. Neuropsychologia 57:59–70.

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Le ST, Gugger S, Drame M, et al. (2019) Hugging Face's transformers: state-of-the-art natural language processing. arXiv 1910.03771.

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.