JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

An integrated neural decoder of linguistic and experiential meaning

https://doi.org/10.1523/JNEUROSCI.2575-18.2019

Cite as: J. Neurosci 2019; 10.1523/JNEUROSCI.2575-18.2019

Received: 4 October 2018
Revised: 26 August 2019
Accepted: 31 August 2019

An integrated neural decoder of linguistic and experiential meaning

Abbreviated title: A neural decoder of linguistic/experiential meaning

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

[1]Department of Neuroscience, University of Rochester, Rochester, NY 14642, USA

[2]Department of Neurology, Medical College of Wisconsin, Milwaukee, WI, 53226, USA

[3]Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, 14627, USA

[4]School of Nursing, University of Rochester, Rochester, NY 14642, USA

[5]Psychiatry, University of Rochester, Rochester, NY 14642, USA

[6]Department of Biomedical Engineering, University of Rochester, Rochester, NY, 14627, USA

[7]School of Engineering, Trinity Centre for Bioengineering, and Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland

Corresponding author with complete address, including an email address and postal code: Andrew James Anderson, Department of Neuroscience, University of Rochester, Rochester, NY 14627, aander41@ur.rochester.edu

Number of pages: 48

Number of figures: 11

Number of tables: 0

Number of multimedia and 3D models: 0

Number of words for Abstract: 237, Significance Statement 120 words, Introduction: 675, and Discussion: 1502

Conflict of Interest: None

Acknowledgements

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

33 CAREER award 1652127. We thank 3 reviewers for their insightful comments and time and
34 Douwe Kiela and Katrin Erk for models used in supporting analyses.

35

Abstract 237/250 words

The brain is thought to combine linguistic knowledge of words and non-linguistic knowledge of their referents to encode sentence meaning. However, functional neuroimaging studies aiming at decoding language meaning from neural activity have mostly relied on distributional models of word semantics, which are based on patterns of word co-occurrence in text corpora. Here, we present initial evidence that modeling non-linguistic "experiential" knowledge contributes to decoding neural representations of sentence meaning. We model attributes of peoples' sensory, motor, social, emotional and cognitive experiences with words using behavioral ratings. We demonstrate that functional Magnetic Resonance Imaging (fMRI) activation elicited in sentence reading is more accurately decoded when this experiential attribute model is integrated with a text-based model than when either model is applied in isolation (participants were 5 males and 9 females). Our decoding approach exploits a representation-similarity-based framework which benefits from being parameter free, whilst performing at accuracy levels comparable to those from parameter fitting approaches such as ridge-regression. We find that the text-based model contributes particularly to the decoding of sentences containing linguistically oriented "abstract" words and reveal tentative evidence that the experiential model improves decoding of more concrete sentences. Finally, we introduce a cross-participant decoding method to estimate an upper-bound on model-based decoding accuracy. We demonstrate that a substantial fraction of neural signal remains unexplained, and leverage this gap to pinpoint characteristics of weakly decoded sentences and hence identify model weaknesses to guide future model development.

Significance statement 120 words

Language gives humans the unique ability to communicate about historical events, theoretical concepts and fiction. Although words are learnt through language and defined by their relations to other words in dictionaries, our understanding of word meaning presumably draws heavily on our non-linguistic sensory, motor, interoceptive and emotional experiences with words and their referents. Behavioral experiments lend support to the intuition that word meaning integrates aspects of linguistic and non-

linguistic “experiential” knowledge. However behavioral measures do not provide a window on how meaning is represented in the brain and tend to necessitate artificial experimental paradigms. We present a model-based approach that reveals early evidence that experiential and linguistically acquired knowledge can be detected in brain activity elicited in reading natural sentences.

Introduction 675/650 words

Humans’ knowledge of historical events, theoretical concepts and fiction is acquired through language. Whilst a considerable body of linguistic information is stored in text and media repositories, the meaning of language is a biological construct, instantiated in our brains during comprehension. Recent advances in neuroimaging technology, big data and computational modeling, have led to an ability to decode brain activity associated with linguistic meaning. The dominant decoding approach achieves this using only *text-based* information. Word meaning is modeled as a vector of values reflecting how often each word co-occurred with other words across a huge body of text (Landauer and Dumais 1997, Burgess and Lund 1997, Mikolov et al. 2013, Pennington et al. 2014). Despite never having experienced walking or eating, the model “learns” that walk and eat mean different things because they appear in different textual contexts, but that walking relates to legs/shoes and that eating relates to hunger/food because these words frequently co-occur (and end up with similar semantic vectors). This approach supports the construction of conceptual knowledge hierarchies, e.g., a dragonfly is an insect is an animal (Fu et al. 2014), and enables some level of analogical reasoning, e.g., Einstein is to scientist as Picasso is to painter (Mikolov et al. 2013). By registering model and brain activity for corresponding words, and mapping between the two, brain activity for new words, sentences and narratives can be predicted and decoded (Mitchell et al. 2008; Pereira et al. 2013, Wehbe et al. 2014, Huth et al. 2016, de Heer et al. 2017, Pereira et al. 2018).

Behavioral experiments suggest that in addition to linguistic experience, word meaning is shaped by non-linguistic perceptual, motor and interoceptive experiences (Paivio, 1971, Barsalou 1999, Barsalou et al. 2008, Vigliocco et al. 2009, Andrews et al. 2009, Riordan and Jones, 2010, Louwerse and

Jeuniaux 2010, Kousta et al. 2011, Dove 2014, Vigliocco et al. 2014, Andrews et al. 2014, Zwaan 2014, Louwerse, 2018). Indirect evidence that sentence comprehension induces perceptual/motor simulations related to sentence content has been amassed in reaction time studies (Stanfield and Zwaan 2001, Zwaan et al. 2002, Glenberg and Kaschak 2002, Kaschak et al. 2005, Kaschak et al. 2006, Connell 2007, Glenberg et al. 2008, Winter and Bergen 2012, Zwaan and Pecher 2012, Speed and Vigliocco 2014). However, there is little direct *neura*l evidence concerning how linguistic and non-linguistic "experiential" sources of knowledge are encoded in brain activity (though see Anderson et al. 2015 and Wang et al. 2018). The ability to estimate linguistic and non-linguistic contributions to neural representations of meaning is necessary to fully characterize human language and related clinical conditions (Patterson et al. 2007; Fernandino et al. 2013; Lambon Ralph et al. 2017, Anderson and Lin, 2019, Bruffaerts et al. 2019). Critically, neural measures can also help provide a window on natural language comprehension, removing the necessity for artificial behavioral response tasks (which may perturb linguistic systems), and bespoke stimulus materials (that may not reflect natural language). See Hasson et al. (2018) and Hamilton and Huth (2018) for related discussion.

We reveal initial evidence that non-linguistic experiential knowledge can be detected in brain activity elicited in sentence reading by combining an *experiential attribute* model (Binder et al. 2016) with a state-of-the-art text-based semantic model (Pennington et al. 2014) to enhance decoding of a large fMRI dataset. The experiential model is based on peoples' ratings of their sensory/motor/affective/cognitive experiences with words and their referents (building on Cree and McRae 2003, Vinson et al. 2003, Lynott and Connell, 2013). Whilst experiential models have provided a basis for neural decoding (Chang et al. 2010, Fernandino et al. 2015, Fernandino et al. 2016, Anderson et al. 2016a, Yang et al. 2017, Wang et al. 2017, Anderson et al. 2018) it has never been clear whether and how text-based and experiential approaches complement one other. We newly demonstrate that text-based and experiential models differentially contribute to decoding sentences that do/don't contain linguistically oriented "abstract" words. Finally, we introduce a cross-participant decoding method to

estimate the room for improvement in model-based decoding and pinpoint model weaknesses for future development.

Methods

Overview

We reanalyzed an fMRI data set scanned as 14 people read 240 sentences describing everyday situations (Anderson et al. 2016a and summarized below). Sentences were 3 to 9 words long and formed from 242 different content words. 10 participants saw the set of sentences repeated 12 times in total (randomly shuffled each time), and the remaining 4 participants who attended half the number of visits saw the sentences 6 times. Following standard fMRI preprocessing steps, each sentence presentation was represented as a single fMRI volume (there were 12 replicate volumes per sentence for 10 participants and 6 replicates for the remaining 4 participants. Analyses were focused on a "semantic network" of 22 anatomical regions-of-interest (ROIs) that had been detected in previous analysis of the same data (Anderson et al. 2018) testing for regional sensitivity to experiential semantic features associated with words with different grammatical roles. ROIs included left temporal, inferior parietal, inferior/superior frontal cortex as well as some right hemispheric homologs (illustrated later in **Figure 6**). These regions have well established associations with semantic processing (e.g. Binder et al., 2009, Binder and Desai, 2011) and broadly adhere to the "language network" identified by Fedorenko and Thompson-Schill (2014). fMRI activation across the network of brain regions was then decoded using a text-based model, an experiential model, and the two models integrated as detailed below. We refer to the integrated text/experiential approach as "multimodal" to reflect the combination of linguistic information with behavioral ratings. Though to dispel any confusion the experiential model serves as a proxy for knowledge acquired through multiple modalities of experience, just each modality is estimated through the same rating procedure.

Materials

All sentences were pre-selected as experimental materials for the Knowledge Representation in Neural Systems (KRNS) project (Glasgow et al. 2016, www.iarpa.gov/index.php/research-programs/krns), sponsored by the Intelligence Advanced Research Projects Activity (IARPA). The stimuli consisted of 240 written sentences containing 3-9 words and 2-5 (mean±sd = 3.33±.76) content words, formed from different combinations of 141 nouns, 62 verbs, and 39 adjectives (242 words). The sentences are listed in full in the Supplementary Materials of Anderson et al. (2016) and Anderson et al. (2018). 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 (200/240) 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"). The entire list is in Table S1. 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 213/242 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, range 21-55) with no history of neurological 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 days (SD = 3.14). The range of the intervals between first and last visits was 15 - 43 days. 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 approximately 6 minutes. 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. Nouns, verbs, adjectives, and prepositions were presented for 400 ms each, followed by a 200-ms inter-stimulus interval (ISI). Articles ("the") were presented for 150 ms followed by a 50-ms ISI. Mean sentence duration was 2.8 s. Words subtended an average horizontal visual angle of approximately 2.5°. A jittered inter-trial interval, ranging from 400 ms 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 echoplanar images (EPI) were collected with TR = 2000 ms, TE = 24 ms, flip angle = 77º, 41 axial slices, FOV = 192 mm, in-plane matrix = 64 x 64, slice thickness = 3 mm, resulting in 3 x 3 x 3 mm voxels. T1-weighted anatomical images were obtained using a 3D spoiled gradient-echo sequence with voxel dimensions of 1 x 1 x 1 mm. fMRI data were pre-processed using AFNI (Cox, 1996). EPI volumes

were corrected for slice acquisition time and head motion. Functional volumes were aligned to the T1-weighted anatomical volume, transformed into a standardized space (Talairach and Tournoux 1988), and smoothed with a 6mm FWHM Gaussian kernel. The data were analyzed using a general linear model with a duration-modulated HRF, and the model included one regressor for each sentence. Neural activity was modeled as a gamma 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.

Experimental Design and Statistical Analysis

Semantic models

Analyses used the following two semantic models of word meaning. As a proxy for the representational structure that can be acquired the distributional statistics of words in language we used "GloVe" (Pennington et al. 2014). GloVe is a freely downloadable text-based semantic model that represents individual words as 300 dimensional floating point vectors derived by factorizing a word co-occurrence matrix (vocabulary size is 2.2million words and co-occurrences were measured across 840 billion tokens from Common Crawl https://commoncrawl.org). GloVe in particular was used because it yielded state-of-the-art performance decoding fMRI activation associated with sentences in Pereira et al.'s (2018) "universal neural decoder of linguistic meaning", although we found there to be relatively minor differences between using GloVe and other models such as "word2vec" (Mikolov et al. 2013) – see also the final **Results** section.

As a proxy for the representational structure that can be acquired from direct experience with the world we used an experiential attribute model (Binder et al. 2016). This model represents words in terms of human ratings of their degree of association with different attributes of experience (e.g., "On a scale of 0 to 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. As such this 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 same experiential model has previously been used as the basis for predicting and decoding the same fMRI dataset as the current study (Anderson et al. 2016a; Anderson et al. 2018).

Some overlap in the semantic information content of text-based and experiential models is very much expected (see also Riordan and Jones, 2010). Because text describes worldly experiences we expect it to partially capture the structure of experiential knowledge. On the flipside, the experiential attribute model seeks to comprehensively model experiential knowledge, and of course language contributes to our experience. Beyond this, the experiential model was itself built through a linguistically guided rating procedure. However, systematic differences between the models are also expected. This is in part because a lot of experiential information goes unstated in natural verbal communication. For instance, borrowing an example from Bruni and Baroni (2014), it is rarely useful to communicate the color of bananas because it is obvious to all those with experience of bananas. Likewise, it would be unusual to specify that dropping things involves movement. Consequently, whilst an analysis of natural text may indicate that a banana is an edible berry, it may not capture the dominance of color as a perceptual attribute. Therefore, despite being derived via language, attribute ratings can potentially anchor to experiential neural systems and access information that would not otherwise have been reported or experienced in natural verbal communication. Conversely, the experiential attribute model as it stands may be less well suited to capturing the meaning of so-called abstract words which tend to be more amenable to verbal description in terms of their relationships with other words (e.g. "fiction" is an imaginary story), rather than through physical example (e.g. being presented with a cat).

We consider the extent to which experiential information is available in text (and language) to be an empirical question. For instance, representational similarity analysis (Kriegeskorte et al. 2008) can be applied to compare the information structure of the text-based to the experiential models. This yields a statistically significant correlation coefficient of $\rho=0.2$, $p<10^{-6}$ (reflecting Spearman correlation between the below diagonal triangles of inter-word Pearson correlation matrices derived using the text-based and experiential models, n=29161 correlation coefficients per triangle as computed from 242 words). A related regression analysis conducted by Utsumi (2018) demonstrated that text-based models were weakly capable of predicting spatial, temporal and affective attributes of the current experiential model. However, demonstrating that text-based models do not contain aspects of experiential information does not also entail that the missing information is relevant for explaining semantic brain activity. Consequently, we conduct our forthcoming analyses by testing for brain activity that is explainable using the experiential model but not using the text-based model and vice versa (see also Anderson et al. 2015, Popov et al. 2018). This approach takes the *assumption* that the text-based model accurately captures all semantic information that can be extracted from text alone. Whilst we acknowledge that this is a strong assumption that is not likely to have been strictly met here, we believe that current models are sufficiently advanced to begin to segregate experiential from linguistic aspects of semantic representation in brain activity (see also the **Discussion**).

Modeling sentences by summing word-level semantic vectors

To turn word-level semantic vectors into representations of sentences we identified all constituent content words in each sentence, and then pointwise summed together these semantic vectors (**Figure 1**). Although such additive composition is obviously an oversimplification that neglects the effects of word order, syntax and morphology, it has endured as a practically successful technique in both computational linguistics (Mitchell and Lapata 2010, Kiela et al. 2014), and fMRI analyses (Anderson et al. 2016a; Yang et al. 2016; Wang et al. 2017; Pereira et al. 2018; Anderson et al. 2018). Indeed,

attempts to incorporate other linguistic factors such as syntax into models have yet to make appreciable difference to neural decoding performance (Pereira et al. 2018; Anderson et al. 2018).

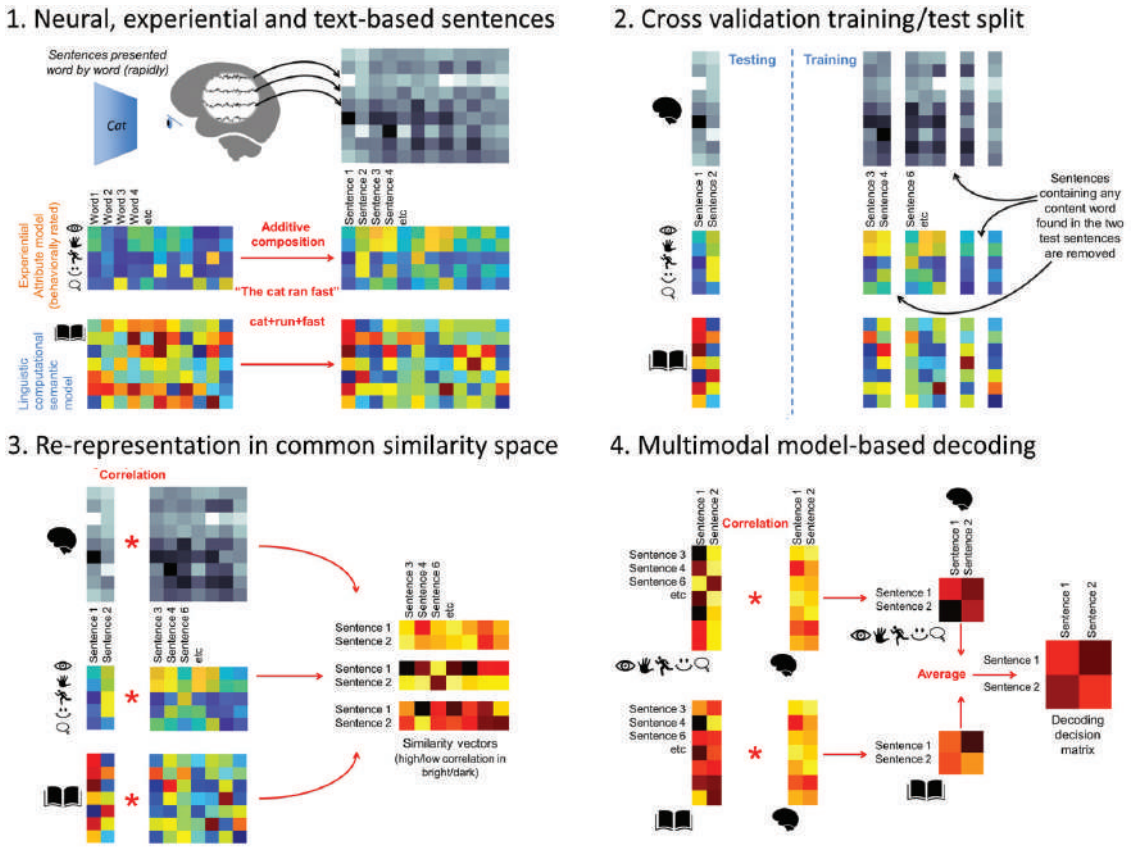


Figure 1. Representational similarity-based decoding algorithm set up to support multiple model-based decoding. Multimodal model combination takes place in stage 4 by averaging 2*2 decoding decision matrices generated by the different models. An alternative approach would have been to pointwise average together the two similarity vectors for the experiential model with those of the text-based model in stage 3. This was disfavored to avoid having to introduce an extra normalization step to deal with correlation coefficients arising from the different models being on different scales (correlation coefficient magnitudes tend to diminish as the number of features correlated becomes large, and here the experiential and text-based models widely differ in the number of features – 65 and 300 respectively). This problem is naturally dealt with in stage 4, because the 2*2 decision matrices are based on correlations between similarity vectors that are all matched in their dimensions.

Representational similarity-based neural decoding set up

To decode fMRI activation, we applied the representational similarity (see also Kriegeskorte et al. 2008) decoding framework introduced in Anderson et al. (2016b) and Anderson et al. (2017) and further extended here to support simultaneous multi-model / multi-ROI / multi-participant decoding of sentences. This was a considered departure from the more commonplace strategy of using multiple regression to map between model and brain (Mitchell et al. 2008, Chang et al. 2010, Sudre et al. 2012, Pereira et al. 2013, Wehbe et al. 2014, Fernandino et al. 2015, Fernandino et al. 2016, Huth et al. 2016, Anderson et al. 2016a, Yang et al. 2017, Wang et al. 2017, Pereira et al. 2018, Anderson et al. 2018). The main reason for choosing the similarity-based approach over (ridge) regression here was for simplicity: to avoid repeating the analyses multiple times over with different regularization penalties and the need to introduce a decision over which penalty to use. For the current analyses, this would complicate the process of integrating information across models, ROIs and individuals (because in each case there would be multiple results associated with the different penalties and multiple decisions to be made). The current focus on the similarity-based approach should not be misconstrued as a claim that similarity-based methods are superior to regression, and we report on some strengths and weaknesses of the two approaches in a supporting analysis using ridge regression (see Results, and **Figure 10**). Other comparative analyses are in Anderson et al. (2016b) and Bulat et al. (2017).

fMRI data was decoded according to a commonly used leave-2-item-out cross validation procedure (Mitchell et al. 2008, Chang et al. 2010, Sudre et al. 2012, Pereira et al. 2013, Wehbe et al. 2014, Anderson et al. 2016a,b, Yang et al. 2017, Wang et al. 2017, Pereira et al. 2018, Anderson et al. 2018). At each cross-validation iteration the 240 sentences were split into a test set of 2 sentences, and a training set of 238 sentences. Then both fMRI and model data for any of the 238 training sentences that contained any content word within the 2 test sentences was deleted from the training data (**Figure 1**

stage 2). This was to enable testing of how well the approach generalized to decoding novel fMRI sentences using sentence vectors built from an entirely novel set of semantic vectors. The mean+/-SD number of sentences in the training set for each iteration was 218+/-5, containing a mean+/-SD of 232+/-2 words). Model and fMRI data corresponding to the training set was featurewise/voxelwise z-scored. Model and fMRI test sentence data were also normalized by subtracting and dividing by the feature/voxelwise mean and SD of the training data.

Because not all fMRI voxels contain informative signal, we estimated which ones were likely to be informative using a commonly used strategy (e.g. Mitchell et al. 2008, Chang et al. 2010, Pereira et al. 2013, Anderson et al. 2015, Anderson et al. 2016b, Anderson et al. 2017, Yang et al. 2017, Wang et al. 2017). For each participant, and separately for each ROI, we took each of the 12 (or 6) fMRI runs through the entire set of sentences, selected only the 218 (or so) training sentences from this and then voxelwise correlated each unique pair of runs together. For the ten participants with 12 runs, this left 66 pairwise correlation coefficients per voxel, and for the four participants with 6 runs this left 15 pairwise correlation coefficients per voxel. A single score was assigned to each voxel by taking the mean of these (66 or 15) correlation coefficients. The 50 voxels with the largest mean value per ROI were selected for analysis. This choice of 50 voxels was ultimately arbitrary though guided by previous work (e.g. Anderson et al. 2013; Anderson et al. 2015, Anderson et al. 2017). In subsequent post hoc analyses (reported later) the interpretation of results was found to be unchanged when using 100 voxels, however in both cases voxel selection systematically improved decoding accuracies over no voxel selection at all.

Decoding of the fMRI data was accomplished by independently re-representing both model and fMRI test data in a common representational similarity space, and then matching model to fMRI sentences in this space. For each fMRI/model dataset in turn, we correlated the two test sentence vectors with each one of the 218 (or so) training sentence vectors. This enabled us to newly represent every single test sentence (whether model or fMRI) as a similarity vector of 218 (or so) correlation coefficients (**Figure 1**

stage 3). We transformed all correlation coefficients in all similarity vectors using Fisher's r-to-z (arctanh) as is a customary treatment when comparing correlation values (though this had only a marginal effect on the current results). Then decoding was achieved by cross correlating the two model similarity-vectors with the two fMRI similarity-vectors (**Figure 1 stage 4**). This resulting 2 * 2 matrix of correlation coefficients was r-to-z transformed (arctanh) and this constituted the "decoding decision". This was evaluated as correct (and scored as a 1 as opposed to zero) only if the sum of z-transformed correlations between the correctly matched model vs fMRI test sentence pair (the matrix diagonal in **Figure 1 stage 4**), exceeded the sum for the incongruent pair (anti-diagonal in **Figure 1 stage 4**). Decoding was repeated for all possible unique pairs of the 240 sentences (28 680 cross validation iterations in total) and the mean score used as the final metric of decoding accuracy. When operating at random (e.g. if the fMRI data contained no semantic signal, or the semantic vectors did not reflect meaning) a mean decoding accuracy of 0.5 is expected. Permutation testing was applied to test whether decoding accuracies were significantly better than chance (by randomly shuffling semantic model sentences and repeating the entire cross-validation analysis 1000 times over, as described in Anderson et al. 2017). Typically, final decoding accuracies above 0.56 were significant at the p=0.05 level (95% of metrics derived in analyzing randomly shuffled sentences were less than or equal to 0.56).

For each participant 2 * 2 decoding decisions were computed in parallel for all 22 ROIs and each of the 2 models. To generate a single decoding decision corresponding to the entire semantic network of all 22 ROIs, we applied an "ensemble averaging" strategy and pointwise averaged together decision matrices across the 22 ROIs (**Figure 2 top row**). We use this network-level decoding estimate as the basis for our main analyses but also report results for individual ROIs.

In a similar vein, we used ensemble averaging as the basis for testing whether models have complementary strengths in decoding. If the models have complementary strengths, then integrating their decisions together as an ensemble will counteract individual model's weaknesses and boost

overall decoding accuracy. Multimodal model integration was achieved by pointwise averaging decision matrices associated with the different models as illustrated in **Figure 1 stage 4**. In all ensemble averaging tests (whether integrating ROIs and/or models) decoding decision matrices were scored as correct precisely as before by testing whether correlations on the diagonal were greater than the antidiagonal. To produce a final metric, scores were then averaged across all 28680 cross validation iterations. Importantly, this ensemble averaging strategy is *not* guaranteed to produce equivalent or better final accuracies than the strongest model of the pair (which would limit its applicability for testing for a multimodal decoding advantage). Specifically, if one model is sufficiently noisy, then the final multimodal decoding accuracy will be lower than the strongest model.

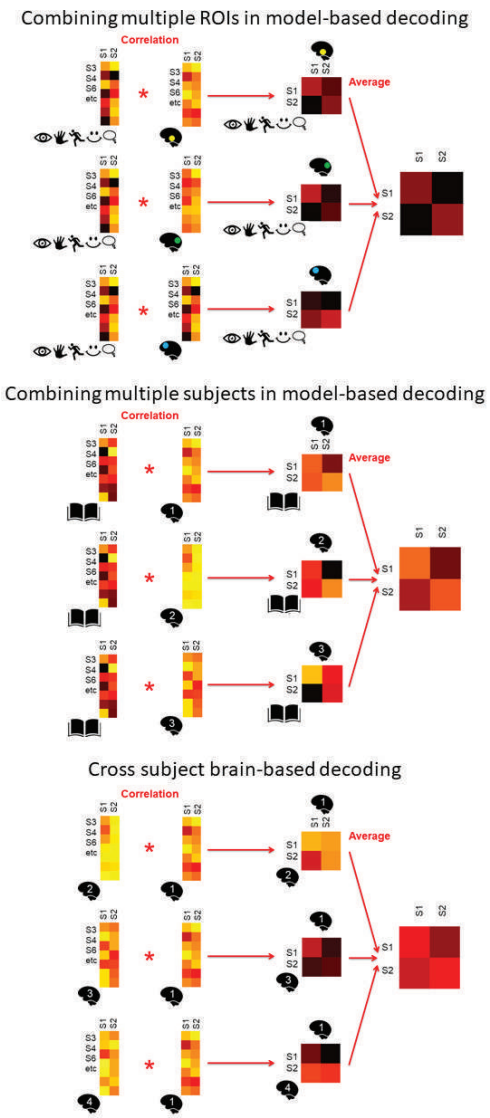


Figure 2. Representational similarity-based algorithm set ups for ensemble decoding. Top, model-based decoding of multiple brain regions in the same participant (see also results in **Figures 4,5,6,8,9,11**). Middle, model-based decoding of multiple participants (see also results in **Figure 5**). Bottom, cross-subject decoding (see also results in **Figure 8**).

The entire cross-validation procedure described above was repeated for each ROI (22) within each participant (14) using both semantic models (2). Ensemble averaging of model decision matrices was

used to derive multimodal decoding accuracies for each participant and ROI. Ensemble averaging of all 22 ROIs' decision matrices was used to generate a single network-level decoding accuracy for each participant and model combination (text, experiential and multimodal). Differences in accuracy between different models were evaluated using t-tests, and p-values associated with multiple ROIs corrected for multiple comparisons according to False Discovery Rate (FDR) (Benjamini and Yekutieli, 2001).

Post hoc analyses partitioning the variance in neural similarity structure explained by each model

ROIs for which decoding accuracy was boosted through model integration were taken to post hoc analyses. We further estimated the unique contribution made by each model to explaining variance in the neural sentence similarity structure and what could be explained equally by both models. This analysis is inspired by de Heer et al. (2017) who partitioned the variance in heard speech fMRI data that was unique to acoustic, articulatory and semantic voxelwise encoding models and shared across them. Besides using different models to de Heer et al. (2017) the forthcoming analysis differs in that it is conducted in representational similarity space. The analysis is illustrated in **Figure 3**.

The analysis was conducted on the entire similarity space defined by all 240 sentences as is a standard approach in Representational Similarity Analysis (RSA) (Kriegeskorte, 2008). Sentence similarity matrices were computed for each participant, for each ROI, by inter-correlating the neural representations of the 240 sentences. In each case this yielded a 240 by 240 correlation matrix. Pearson correlation was used and the correlation coefficients were subsequently r to z transformed (arctanh). Prior to this, voxel selection was conducted to estimate the 50 informative voxels per ROI. Voxel selection used the same correlation-based approach as detailed above for the previous decoding analysis. However, in the case at hand voxel selection was computed only once on all 240 sentences together (because the current RSA did not employ cross validation). To generate a single correlation matrix capturing the 22 ROI ensemble, the 22 r to z transformed correlation matrices corresponding to each ROI were pointwise averaged.

Correlation matrices for the text-based and experiential models were computed in the same fashion by inter-correlating the 240 sentences. Again, Pearson correlation was used and all coefficients were r to z transformed. Next, the unique information contained within every single correlation matrix was extracted by segmenting the below diagonal matrix triangle. Each triangle was then vectorized to create a 28680 element similarity vector. Similarity vectors were subsequently normalized by z-scoring to support the forthcoming regression analysis.

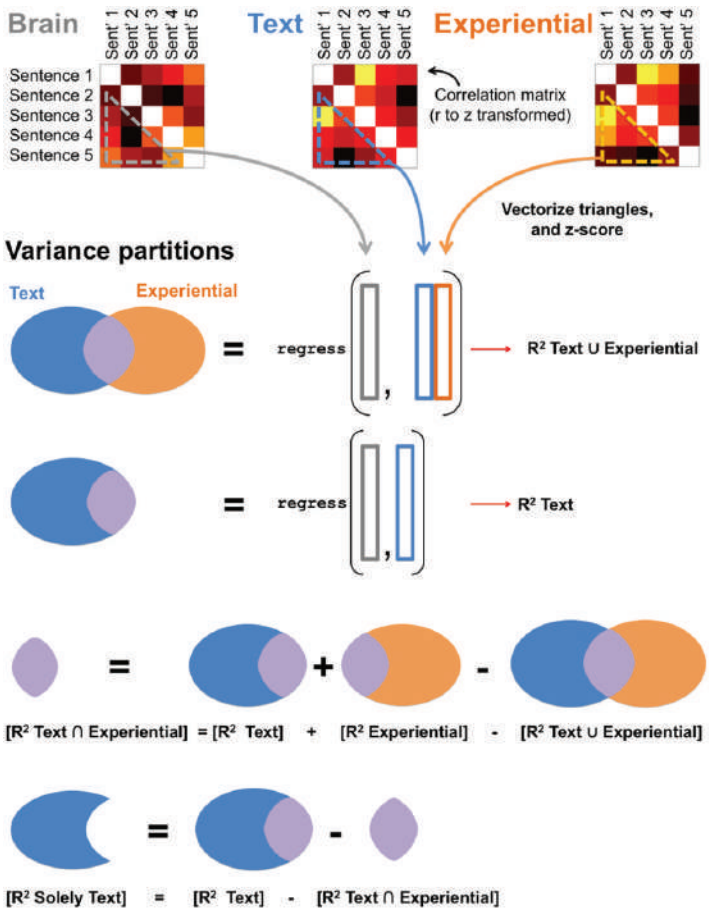


Figure 3 Partitioning the variance in neural similarity structure that is solely accounted for individual models and shared between them.

We applied the set theoretic approach of de Heer et al. (2017) to estimate the variance in neural similarity structure that was explained by the *union* of both models, the *shared* variance that is equally explained by either model, and the variance *solely* accountable to one model (see **Figure 3**). The union [R^2 Text ∪ Experiential] was first estimated using a multiple regression in which the similarity vectors for both models were used as predictors and the neural similarity vector was the response variable. The variance associated with but not necessarily exclusive to the individual models ([R^2 Text] and [R^2 Experiential]) was estimated in 2 separate regression analyses. In each a single similarity vector (associated with one model) was the predictor. To estimate the shared variance [R^2 Text ∩ Experiential] we subtracted away the variance explained by the model union [R^2 Text ∪ Experiential] from the sum of the variance explained by the Text and Experiential models ([R^2 Text] + [R^2 Experiential]). Then, to estimate the variance solely accountable to the text and experiential models ([R^2 Solely Text] and [R^2 Solely Experiential]) we subtracted the shared variance away from the variance explained by the models (e.g. [R^2 Text] - R^2 Text ∪ Experiential]). To produce positive correlation coefficients from these measures, we took the square root of R^2 values as undertaken by de Heer et al. (2017). However, because variance/positive correlation measures do not facilitate testing whether individual models made significantly greater than zero contribution to explaining neural data (because they are always greater than or equal to zero) we also undertook a partial correlation analysis (see also Anderson et al. 2015, Wang et al. 2018). Here we computed the correlation between neural similarity structure and one model, whilst controlling for the other model. To test the generality of partial correlation coefficients across participants, we compared them to zero using one sample t-tests (n=14).

Code accessibility

MATLAB similarity-based decoding code is available on request from the corresponding author.

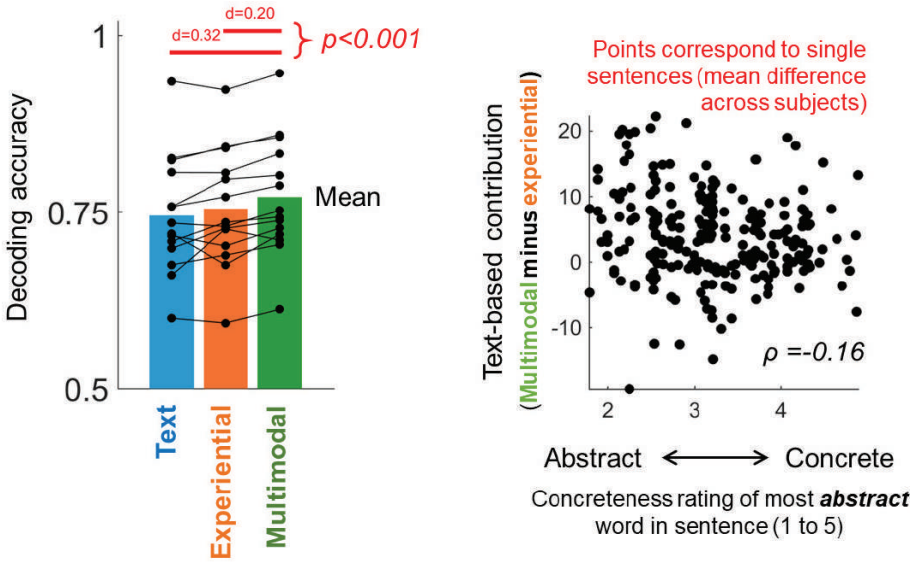


Figure 4. Integrating text-based and experiential models produces stronger decoding. Individual-level accuracies arising from decoding the 22 ROI ensemble (see **Figure 2** top row). The contribution of the text-based model to multimodal decoding was particularly pronounced for sentences containing abstract words (right). Effect sizes (d) were estimated according to Dunlap et al. (1996) as d=t*(2*(1-r)/n), where t is the t-statistic arising from the corresponding paired t-test, r is Pearson correlation, and n is the number of participants (14).

Results

Integrating text-based and experiential semantic models produces stronger decoding than either alone

To test for evidence that both linguistic and non-linguistic experiential aspects of meaning were present in neural activation, we tested whether combining the two models together improved decoding accuracies above using either model in isolation. Decoding accuracies for both models and their combination are illustrated in **Figure 4**. T-tests revealed that whilst there were no differences in decoding accuracy between the text-based and experiential models, when the models were combined accuracies were significantly greater than for either model in isolation. (Difference between multimodal and text, t=6.9, p<0.0001; difference between multimodal and experiential, t=4.8, p=0.0004). This key

result provides direct neural evidence that linguistic and experiential semantic information were present in brain activation elicited in sentence reading. Otherwise, decoding accuracies were significant for all participants and all models.

Text-based model enhances discrimination of sentences containing abstract words

We next examined the nature of the contribution made by the text-based and experiential model in more detail. An obvious area to anticipate divergence between models is for more linguistically oriented "abstract" words. These words do not directly correspond to "concrete" entities in the world that can be directly sensed. As such an abstract word's meaning is language dependent and most amenable to description in terms of other words which could contribute different parts of that meaning (Brysbaert et al. 2014). Examples of words that are relatively abstract within the current dataset include: 'negotiate', 'agreement', 'wealthy', 'famous' and 'clever'. We hypothesized that the text-based model (which is built from word co-occurrence statistics) would make a particular contribution to multimodal decoding of sentences containing abstract words.

To test for an abstractness advantage associated with integrating the text-based model in decoding, we looked up concreteness ratings (Brysbaert et al. 2014) for each of the 242 content words in the sentence set. Each of the 240 sentences was then scored according to (1) the concreteness of the least concrete (most abstract) word in the sentence; (2) the concreteness of the most concrete word in the sentence; (3) the mean concreteness of all words in the sentence. Because the experimental sentences also varied across many other factors which through coincidence might be confounded with concreteness measures, we attempted to take these into account. Specifically, we additionally scored each sentence according to the least frequent, most frequent and mean frequency of content words in the sentence (derived from log2 transformed SUBTLEX-US counts of Brysbaert and New, 2009), the number of words in each sentence, and the minimum, maximum and mean word length (number of characters). After this each sentence was represented with 10 measures (including the 3 concreteness measures).

We then identified how well each individual sentence was decoded using the different models. For each participant (n=14), this yielded a vector of 240 sentence decoding scores for the text-based model, the experiential model and the multimodal combination. The maximum score attainable (and maximum value in the vector) was 239, which could have been achieved if the corresponding sentence was successfully discriminated from all 239 other sentences during the entire leave-2-out cross validation analysis.

To estimate the independent contribution made by the text-based model to multimodal decoding, we pointwise subtracted the experiential model-based decoding scores for individual sentences away from corresponding multimodal model scores (and repeated for each participant). Positive scores arising from this subtraction indicate sentences that were better discriminated by integrating the text-based model, which in turn we hypothesized relate to a measure of sentence abstractness. To test this, we correlated (Spearman) each participant's vector of "boost" values with the three different sentence concreteness measures (leaving 14 * 3 correlation coefficients). We repeated these correlations for the other seven sentence measures (re: word frequencies and lengths of constituent words and the number of words in the sentence).

To test for the generality of positive correlations across participants, participants' correlation coefficients were r-to-z transformed (arctanh) and then compared to zero using a one sample t-test (n=14). T-tests were repeated on correlation coefficients associated with each of the 10 sentence measures. The resultant 10 p-values were FDR corrected. Only one of the ten t-tests yielded a statistically significant result. This was the test based on correlations between the concreteness rating of the most abstract word in the sentence and the decoding boost ($t=-3.7$, $p=0.04$, FDR corrected). The mean correlation coefficient across participants was -0.08. This provided evidence that the decoding advantage brought by integrating the text-based model was related to better discrimination of sentences containing abstract words. This relationship is illustrated in **Figure 4** (note that the figure illustrates the mean

accuracy boost per sentence across all 14 participants with associated correlation coefficient of -0.16, whilst the above t-test was based on 14 individual-level correlation coefficients, with a mean value of -0.08).

In an attempt to gain an intuition of whether there was a common theme to the sentences containing abstract words that had received a decoding boost we listed them. However, we were unable to confidently pinpoint any systematic pattern. The ten sentences that received the greatest decoding boost are as follows with the most abstract word in each sentence in italics. The abstract word's concreteness rating (C) and the mean decoding boost (B) associated with the sentence is in brackets after the sentence. "The family was *happy*." (C=2.6, B=22); "The team *celebrated*." (C=2.9, B=21); "The *patient* survived." (C=2.5, B=21); "The family *survived* the powerful hurricane." (C=2.6, B=20); "The pilot was *friendly*." (C=2.3, B=20); "The flood was *dangerous*." (C=2.1, B=20); "The man *lost* the ticket to soccer." (C=2.3, B=20); "The jury watched the *witness*." (C=4.1, B=19); "The council read the *agreement*." (C=2.2, B=18); "The artist *shouted* in the hotel." (C=4.2 B=18). On face value a commonality would appear to be that many of the sentences have affective connotations. However, such affective connotations were not exclusive to the boosted sentences. Indeed, three of the four sentences that were most disadvantaged by integrating the text-based model were also valenced. These four sentences were "The team *lost* the football in the forest." (C=2.3 B=-19), "The teacher broke the *small* camera." (C=3.2, B=-15), "The *aggressive* team took the baseball." (C=2.5, B=-12.4), "The actor *gave* the football to the team." (C=2.8, B=-12.5). We therefore presume simply that the text-based model helped explain neural signal reflecting the additional linguistic information exposed in accessing abstract words and integrating their meaning into sentences. However, in the future it will also be valuable to consider stimuli that are more controlled in their content and less dominated by concrete sentences than the current dataset.

We next ran an analogous analysis attempting to understand the contribution of the experiential model which we anticipated might be associated with concreteness. Sentence-wise decoding score vectors

(length 240, maximum value 239) for the text-based model were subtracted from the multimodal model to generate an "accuracy boost" vector for each participant. Correlations between boost vectors and all ten sentence measures were computed. Whilst a significant positive correlation was detected with the concreteness of the most abstract word in the sentence – i.e. the contribution of the experiential model was especially associated with concrete sentences without any abstract words (average correlation across participants=-0.045, p=0.02, disappointingly this result did not survive FDR correction for multiple comparisons and should be treated tentatively for the time being.

Model-based neural decoding at group-level leverages cross-participant regularities

A fundamental difference between either of the semantic models and individuals' fMRI data was that models were generic representations of concepts built from group-level information (either from texts written by many authors, or ratings given by many people), whereas fMRI activation captured snapshots of an individual's interpretation of a sentence at the particular time of scanning, and in the broader context of their own personal experiences. We therefore reasoned that combining individuals' fMRI data together as a group should expose regularities in semantic representations across individuals and lead to a stronger pattern match to the group-level models (**Figure 2 middle**). Aside from this the group-level combination should also iron out noise (whether this arises due to technological reasons or participants attention levels and/or compliance with the task). Either way, this should lead towards a less noisy comparison between model and fMRI data, albeit at the expense that the results may not generalize to individuals (though this has already been tested in **Figure 4**).

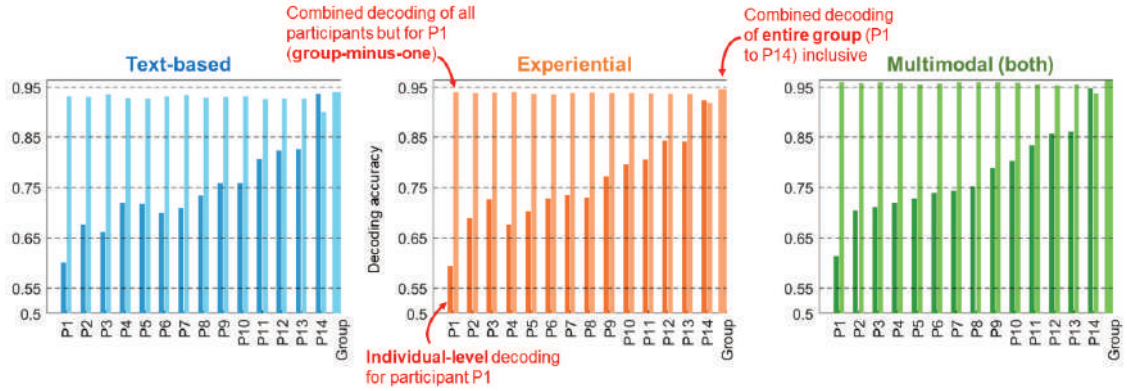


Figure 5. Decoding neural data at group-level exploits cross participant regularities. Model-based decoding accuracies at group-level (**Figure 2 middle**) and for each individual corresponding to all 22 ROIs decoded as an ensemble. Individual participants' decoding accuracies (dark) are plotted beside group-minus-one (light) decoding accuracies derived using all other participants. "Group" is all 14 participants combined.

Combining fMRI activation across participants is in general complicated by both anatomical and functional differences between individuals. Whilst sophisticated "hyper-alignment' methods for combining group fMRI data exist (Haxby et al. 2011, Guntupalli et al. 2016), in the current case it is also possible to combine individuals together as a group, by averaging together individual's decoding decision matrices (see **Figure 2 middle**) in precisely the same way as we have combined models and ROIs.

Group-level decoding accuracies, illustrated in **Figure 5**, were unanimously greater than average individual-level decoding accuracies. For example, where the mean individual-level decoding accuracy was 0.77 the group-level result was 0.97 (as a side note, although this may seem surprisingly high accuracy, this score is consistent with previous word-level decoding studies (Anderson et al. 2017, Anderson et al. 2016b) and also reflects the high sensitivity of the leave-2-out test). To confirm the statistical significance of this effect we used on sample t-tests to test the individual-level decoding

results against the single group-level score. The outcome was significant ($p<0.05$) in every test following FDR correction (for all ROIs and for both models and their multimodal combination).

In conducting group-level analyses it is possible that individual participants play a dominant role in results (an individual may have just happened to elicit semantic representations that match the models well, or have been particularly attentive to the task). To examine the influence of individual participants we re-ran the group-level analysis holding out each individual from the group in turn. Group minus participant decoding accuracies are plotted next to the held out participant in **Figure 5**. It became clear that one participant (P14) indeed played a dominant role relative to the other participants (the decoding accuracy for P14 was slightly greater than group-level decoding based on P1 to P13). However, when P14 was excluded from the group, decoding accuracies remained high (0.94) and this accuracy was substantially greater than individual results for P1 to P13. Therefore, although P14 played a dominant role, the group-level advantage persisted when this participant was excluded. Interestingly the decoding accuracy for the entire group (all 14 participants) always exceeded every single individual-level decoding accuracy, and every group-minus-one decoding accuracy, which indicates that every single participant, including the poorest decoded (P1) beneficially contributed to the group-level decision.

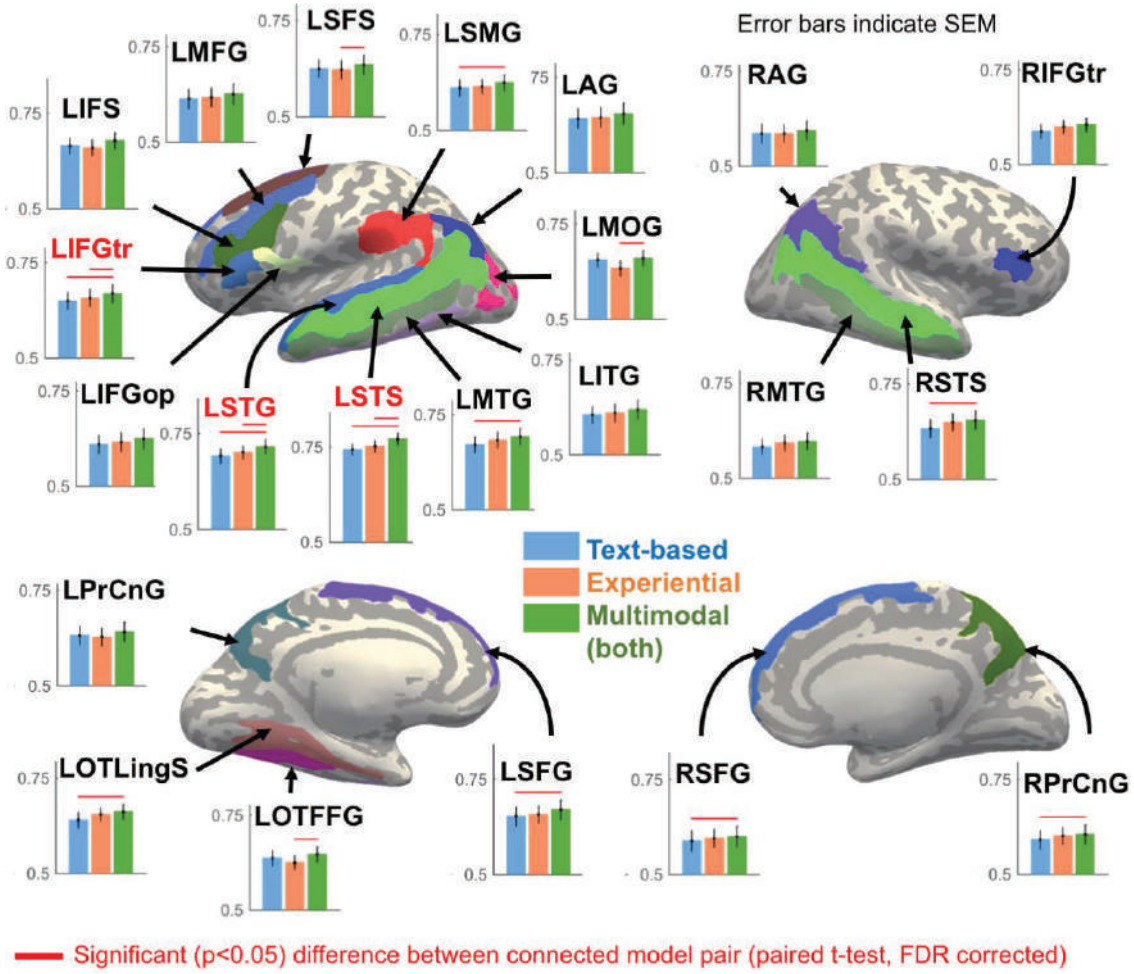


Figure 6. Multimodal model integration improves decoding of superior temporal and inferior frontal regions. Mean±SEM decoding accuracies across 14 participants derived using the text-based and experiential model independently, and then when combined together (i.e. multimodal decoding, see **Figure 1 stage 4**).

Model integration improves decoding of superior temporal and inferior frontal brain regions

Thus far analyses have been based on decoding the semantic network of all 22 ROIs as an ensemble. Mean±SEM decoding accuracies across participants for individual ROIs are in **Figure 6**. As displayed on **Figure 6** decoding accuracies arising from the multimodal approach were significantly greater than

those for either model in isolation in the left superior temporal sulcus (LSTS), left superior temporal gyrus (LSTG) and the triangular part of the left inferior frontal gyrus (LIFGtr), and not reduced for any ROI. LSTS and LSTG yielded the highest and second highest decoding accuracies across all ROIs. All ROIs were on average decoded better than chance by both models. However, no statistically significant differences in overall decoding accuracy were detected between models for any ROI. Otherwise decoding accuracies varied across ROIs in a similar fashion to that observed in previous results (Anderson et al. 2016a, Anderson et al. 2018) and were strongest in LSTS with a mean accuracy of 0.75 (permutation tests revealed all individuals were decoded at accuracies significantly above chance level ($p=0.05$). The lowest average decoding accuracy was for the right angular gyrus (RAG), mean=0.59, with 7/14 participants returning results that were significantly above chance ($p=0.05$).

Post hoc tests partitioning the variance in the neural similarity structure explained by each model

The fraction of variance in the overall representational similarity structure (derived from all 240 sentences) of LSTS, LSTG, LIFGtr and the 22 ROI ensemble that was uniquely explained by each model or commonly explained by both models is in the Venn diagrams of **Figure 7**. In each of the four tests, approximately 50% of the captured variance in neural similarity structure could be explained equally by either model (within participant percentage averaged across participants). Around 40% of the remaining variance was associated with the experiential model and the other 10% associated with the text-based model. Also displayed in **Figure 7 (right)** are partial correlation coefficients, reflecting the correlation between the neural sentence-level similarity structure, and the text-based model, whilst controlling for the experiential model (and vice versa). For all ROIs, partial correlation coefficients for both the text-based and experiential models were found to be significantly greater than zero (across participants, test statistics are in **Figure 7)**. This provides further evidence that both models independently contributed to explaining the neural data.

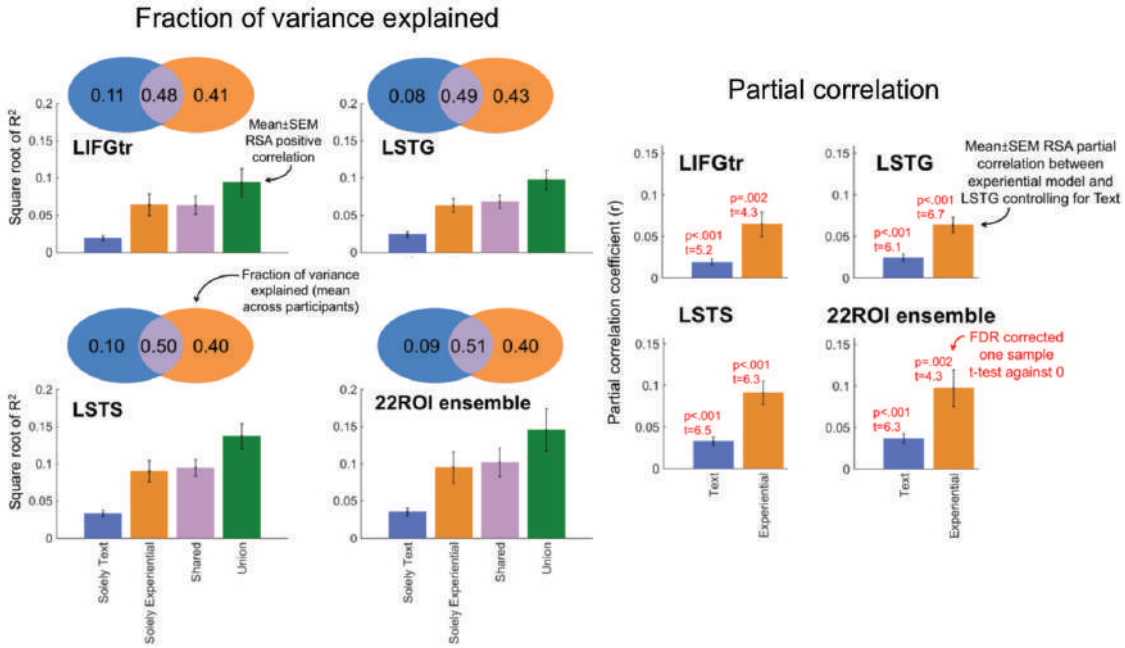


Figure 7. Partitioning the contribution made by the text-based and experiential models to explaining neural similarity structure across the entire set of 240 sentences. The Venn diagrams (left) illustrate the mean (across participants) fraction of variance that is solely accounted for by the individual models and shared between them in the RSA analyses (see **Figure 3**). The bar plots (left) display the associated mean±SEM positive correlations (square root of R^2). Deserving of additional explanation, in LIFGtr the mean experiential coefficient is marginally greater than the shared coefficient whereas the mean fraction of variance explained by the experiential model is less than the shared component. This occurred because the experiential model tended to uniquely explain more variance in participants with large Union R^2 values (relative to shared variance) and vice versa. The averages of raw coefficients (in the bar plots) reflect this trend, but the Venn diagrams do not, because the trend was removed by computing fractions within each participant, prior to averaging across participants. Mean±SEM partial correlation coefficients for the two models in the same RSA analyses are displayed in the four bar plots to the right (and tested against zero).

In both of the post hoc RSA analyses the contribution of the experiential model is on face value stronger and the text-based model weaker than would have been anticipated from the previous decoding analyses in which accuracies arising from each individual model were fairly well balanced

(**Figure 6**). We are currently unsure of the precise reason for this. The decoding analysis differs from the current RSA in two key respects. First it repeatedly tests different sub-samples of the representational similarity matrix that correspond to particular sentence pairs. Specifically, each cross-validation iteration is based on a test of 238 * 2 correlation coefficients rather than the global set of all 28680 coefficients as tested by the RSA. Second the decoding analysis incorporates a decision function to best match up model sentences with fMRI sentences. Thus, it seems that one, other or both of these differences render the decoding analysis less sensitive to emphasizing the contribution of the experiential model. We leave detailed investigation of these differences over to future work. However, in the meantime the current post hoc analyses underlines the value of the experiential model in explaining the neural data.

Cross-participant neural decoding estimates an upper bound on decoding accuracy

Having demonstrated the benefits of combining models, we questioned how much room there is left over for improvement in decoding. We asserted that in the general case the best decoder of an individual's fMRI activation will be other people's fMRI activation (at least in the absence of a personalized semantic model). To this end we decoded each individual's fMRI data using all other individuals in turn and then combined the collective decoding decisions together as an estimate for the upper bound on decoding accuracy. In advance, such cross-participant decoding is liable to be advantaged over the semantic models by additionally decoding non-semantic information – e.g. activation reflecting orthographic or syntactic processing. Nevertheless, this information is still useful to identify ROIs for which there is room for improvement in decoding.

As illustrated in **Figure 2 bottom**, we first used each participant to decode each other participant (repeated for each ROI, and the combination of 22 ROIs). For each individual, this left 13 decision matrices (2*2) at each cross-validation iteration (for each of the 13 other participants). These 13 matrices were pointwise averaged, scored as previously by comparing the sums on the matrix diagonal and antidiagonal, and then scores were averaged across all cross-validation iterations to give a final

cross-participant decoding accuracy. We considered this final accuracy as an estimate on the upper bound decoding achievable for that individual (in absence of a personalized semantic model).

Mean+/-SEM cross participant decoding accuracies for each ROI are compared to the multimodal decoding accuracies in **Figure 8**. Only multimodal decoding accuracies (i.e. the best decoder so far) are displayed as a comparison to avoid visual clutter (the isolated model results are in **Figure 6**). FDR corrected paired t-tests between cross participant and multimodal model-based accuracies revealed significantly stronger decoding for the cross-participant approach in LSTS, LSTG, left occipitotemporal fusiform gyrus (LOTFFG) and left mid occipital gyrus (LMOG) and for the set of 22 ROIs decoded as an ensemble (all $p<0.05$). This indicated that there was sentence-related signal present in the fMRI data within these regions that had not been decoded by the models. This was particularly the case for LMOG and LOTFFG ($d=1.67$ and 1.24 respectively) however improvements for LSTS and LSTG were notable ($d=0.81$ and 0.56 respectively).

Also noteworthy was the gap separating cross participant decoding accuracies for the 22ROI ensemble and individual ROIs. For instance, the difference between cross participant decoding of the highest accuracy ROI (LSTS) and the 22ROI ensemble was significant and sizeable ($t=7.7$, $p<0.0001$, $d=1.47$). This provides evidence that complementary sentence-related information (which could be semantic or orthographic or syntactic) was distributed across the different ROIs. In contrast for the semantic model-based analysis most of the decodable information appears to be in LSTS (see also **Figures 9-11**).

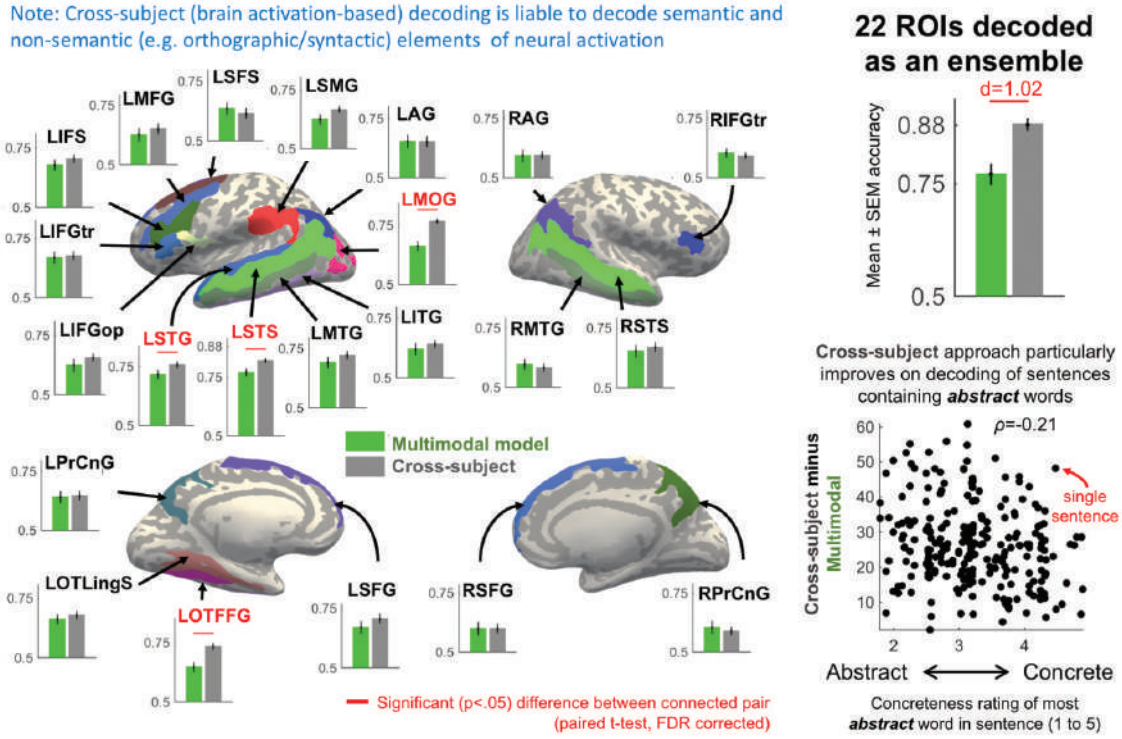


Figure 8. Estimating the room for improvement: how cross-participant decoding improves on the multimodal model-based approach. Mean±SEM cross participant (brain-based) decoding accuracies (**Figure 2 bottom**) across all 14 participants beside comparative results for the multimodal model (also shown in **Figure 4**). Detailed results arising from decoding using the combination of all 22 ROIs (**Figure 2 top**) are to the right. The scatter plots indicate characteristics of sentences for which cross-participant (brain-based) decoding was advantaged over the multimodal model-based approach. The effect size (d) was estimated as described in **Figure 4**.

Sentences decoded weakly by the models tended to contain abstract words

To attempt to get a handle on what semantic information could have been left undecoded by the models, and in so doing identify model weaknesses we finally tested whether particular types of sentences were better decoded by the cross-participant approach. We concentrated analyses on decoding all 22 ROIs as an ensemble. Then for each participant (14) we extracted sentence-wise decoding scores (240) for both the cross-participant and multimodal approaches. The vector of 240

multimodal model-based decoding scores was pointwise subtracted from the cross-participant decoding scores to generate an “accuracy boost vector” indicating which sentences were better decoded by cross participant decoding. For each participant we correlated (Spearman) this accuracy boost vector with min/mean/max concreteness of constituent content words in sentences, the min/mean/max (log2 transformed) word frequency per sentence, the minimum/mean/maximum word length per sentence and the number of words per sentence. The resulting correlation coefficients were Fisher’s r to z transformed (arctanh). For each set of 10 tests, transformed coefficients for all 14 participants were compared to zero using a t-test. P-values were then FDR corrected. Four tests yielded significant results. These were the concreteness rating of the most abstract word in the sentence (mean r=-0.08, t=-4.1, p=0.01 FDR corrected, see also **Figure 8**), mean word frequency (mean r=-0.06, t=3.6, p=0.03 FDR corrected), minimum word length (mean r=-0.05, t=-3.5 , p=0.03 FDR corrected) and the number of words in the sentence (mean r=0.07, t=4.5, p=0.01, FDR corrected). To counteract possible effects of spurious intercorrelations between these four sentence measures, a second round of partial correlation analyses was run. Each participant’s accuracy boost vectors were partially correlated (Spearman) with vectors associated with each of the four measures in turn whilst controlling for the other three. Of the four measures, only correlations with the concreteness rating of the most abstract word in the sentence were found to be significantly lower than zero (mean partial r=-0.06, t=-3.6, p=.003). This provided evidence that cross participant decoding gained a particular advantage over the multimodal model approach for sentences containing abstract words. The relationship between the decoding advantage and the number of words per sentence, constituent word frequencies and word lengths is unclear due to their intercorrelation.

Secondary supporting analyses

The following two sections present the outcome of a series of secondary supporting analyses that are possibly best suited to the dedicated reader.

Supporting analysis: multimodal decoding advantage is preserved in analyses of different cortical networks

Our main analysis was focused on a semantic network of 22 regions that had been identified in Anderson et al. (2018) which differently used a two-stage regression-based analysis and the experiential model alone. Reasons for this were to maintain continuity, so as to enable current results to be referenced back to the results of Anderson et al. (2018), and also for simplicity. However, because the network of 22 regions was derived using the experiential model only it is possible that the current analysis could have been biased towards the experiential model.

To confirm that results were not specific to the 22ROI network, we repeated our initial decoding analysis using different network configurations. We first derived decoding accuracies for all 150 ROIs in the Destrieux atlas using the text-based and experiential models independently. At this stage the analysis was conducted without voxel selection to cut down on computational overheads. We then identified a set of "high accuracy" brain regions that were decoded significantly at the p=0.01 level (FDR corrected) by either model. We then reperformed the decoding analysis, this time with voxel selection (50 voxels per ROI). We firstly decoded the *union* of high accuracy ROIs associated with either or both of the models. Secondly, we decoded the *intersection* of high accuracy ROIs associated with both models.

The "intersection" network contained 12 ROIs. Of these, the following 9/12 ROIs also appeared in the 22 ROI ensemble: LSTS, LSTG, LMTG, LIFGtr, LIFS, LAG, LPrCnG, LOTLingS, RSTS. Destrieux atlas names for the 3 ROIs that were not in the 22 were: ctx_lh_G_cingul-Post-dorsal, ctx_lh_S_oc-temp_lat, ctx_lh_S_subparietal. The "union" network contained 15 ROIs. These included the 12 listed above as well as LOTFFG and LMOG which were in the original 22 ROI ensemble and ctx_lh_S_precentral-inf-part which was not. The 4 new ROIs that had not been in the original 22 were all relatively low scoring (ranked 11th highest or below out of the union of 15 ROIs). Mean±SEM decoding accuracies for the text-based, experiential and multimodal for the 4 new ROIs were respectively: ctx_lh_G_cingul-Post-dorsal 0.63±0.02, 63±0.02, 64±0.02; ctx_lh_S_oc-temp_lat 0.63±0.02, 63±0.01, 65±0.02;

ctx_lh_S_subparietal: 0.62±0.02, 63±0.02, 64±0.02 and ctx_lh_S_precentral-inf-part 0.62±0.02, 61±0.02, 63±0.02. T-tests revealed no significant differences in decoding accuracy between the text-based and experiential models for these 4 ROIs.

Decoding accuracies arising from the intersection and union networks are displayed in **Figure 9**. For comparison we display results of the original 22 ROI ensemble analysis conducted on either 50 or 100 voxels selected per region. Additionally, we display results for LSTS (the highest scoring ROI) and also results when the analysis was undertaken on the entire cortex without any voxel selection. Also displayed are correlations relating the contribution of the text-based model to multimodal decoding to ratings of sentence abstractness echoing **Figure 3**.

Figure 9 reveals that the decoding advantage brought through model integration is preserved across all tests irrespective of the network configuration. Also preserved is the "abstractness" advantage brought by the text-based model to decoding sentences containing abstract words. It is visually apparent that that decoding accuracy was modulated by the network configuration tested. Because it is not a focus of the current article we leave in depth treatment of these differences and how to select the optimal network over to future work. Suffice it to say that activity in LSTS appears to have been the linchpin of network-based decoding and there was no dominant network configuration that yielded significantly greater accuracy than all others.

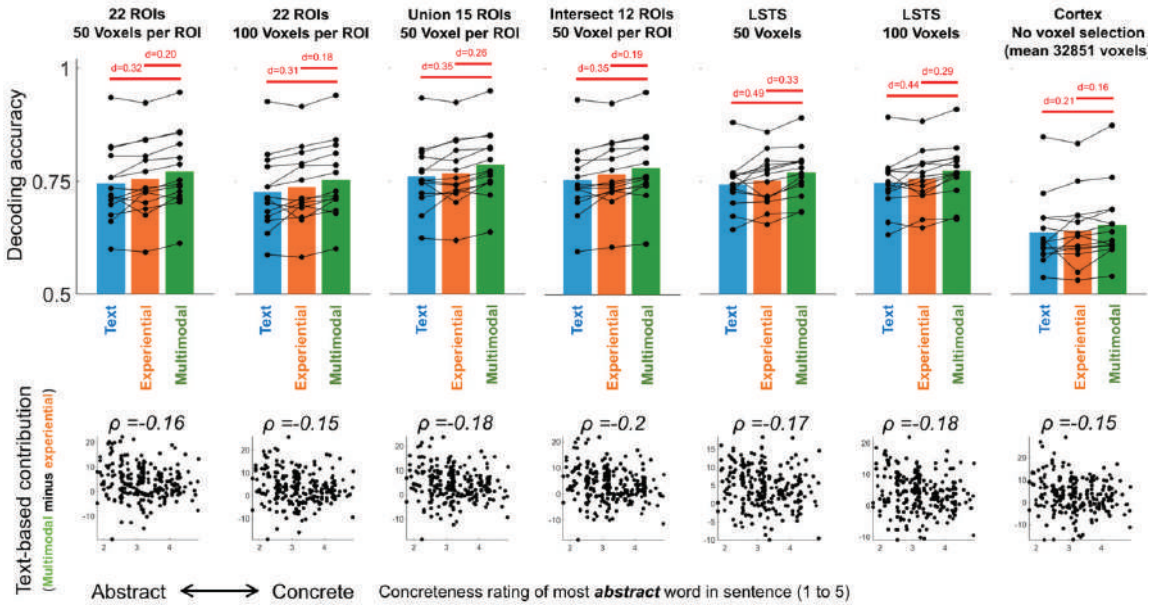


Figure 9. Decoding accuracies arising from decoding different networks of ROIs using different model combinations. This Figure is companion to **Figure 4** which describes how effect sizes (d) were estimated.

Supporting analysis: comparison and integration of ridge regression with similarity-based decoding

Prompted by a concern that the current RSA approach might inherently bias results in favor of one or other model we used ridge regression (Hoerl and Kennard, 1970) to reanalyze multimodal ROIs: LSTS, LSTG, LIFGtr and the 22ROI ensemble. Ridge regression was selected because it has been popular in recent fMRI studies using text-based models (e.g. Huth et al. 2016; de Heer et al. 2017; Pereira et al. 2018). Importantly, this reanalysis provides a quantitative estimate of the decoding advantages brought by using experiential attributes and similarity-based methods comparative to a state-of-the-art text-based/ridge regression approach.

For each participant and each ROI we reran precisely the same leave-2-sentence-out cross validation procedure with the same training/testing data splits and same voxel selection (50 voxels per ROI) as

our main similarity-based analysis. At each of the 28 680 cross validation iterations, we fit a separate ridge regression for the text-based model and then for the experiential model to predict activation in each individual voxel (i.e. forward encoding). We repeated regression fitting using each of the following 9 regularization penalties (λ) [1 10 100 10^3 10^4 10^5 10^6 10^7 10^8]. As before, at each cross validation iteration, we computed a 2*2 decoding decision matrix for both the text-based and experiential models. This was repeated for each λ by correlating predicted and actual fMRI activation patterns and r-to-z transforming correlation coefficients (9 decoding decision matrices per model). To create a multimodal 2*2 decoding decision matrix we pointwise averaged together decision matrices arising from each of the two models precisely as described in **Figure 1 stage 4**. We repeated this for every combination of λ pairs leaving 9*9=81 multimodal decision matrices per iteration per ROI. We created a 22 ROI ensemble decoding matrix for each model and λ combination by pointwise averaging each of the 9 + 9 + 81 decision matrices across ROIs. Note that for the multimodal approach there were 81^{22} possible ways that λ could have been combined across ROIs and therefore the pointwise averaging approach we have taken could have missed out on the ideal combination. Alternative approaches could have included selecting an optimal λ for each ROI using nested cross-validation, or building a single decision matrix from all voxels. Because regression is not our prime focus we leave detailed investigation of this to future work. As for all our other analyses, at each iteration decoding decision matrices were evaluated as correct (1) if the sum of coefficients on the diagonal exceeded the sum on the antidiagonal, otherwise they were incorrect (0). A final decoding accuracy score was assigned as the mean correctness across trials for each model and λ combination (yielding 9 + 9 + 81 accuracies per ROI per participant).

Decoding accuracies were in general qualitatively similar to those observed for the similarity-based analysis and are illustrated in **Figure 10**. Decoding accuracies for both models tended to reach a maxima and flatten off at λ values greater than $\lambda=10^3$ or 10^4. For “flattened” λ values, in tests on LSTS and LIFGtr the multimodal approach was significantly more accurate than either of the text-based and experiential models alone. For LSTG and the 22ROI ensemble, whilst the multimodal approach yielded

significantly greater decoding accuracies than the text-based model (**Figure 10 bottom left**), accuracies were not significantly greater than the experiential model (at least for "flattened" λ values **Figure 10 bottom right**). This weak multimodal improvement reflects both the comparatively weak decoding accuracies achieved for the text-based model using ridge regression, and the relatively strong accuracies for the experiential model (see caption of **Figure 10** for statistical test results). Ridge regression's weak performance with the text-based model could reflect a combination of difficulties surrounding: scaling up to the 300 text-based features (in comparison to 65 experiential attributes, with 240 sentences); and/or the distributional properties of the text-based data; and/or our current selection of λ missing out on the optimal value. Ridge regression's stronger performance with the experiential model evidences that parameter fitting can lead to decoding improvements with the current data. There was little to separate multimodal decoding accuracies for the two decoders, although for a particular λ value (10^3) ridge regression yielded stronger performance in LSTS alone.

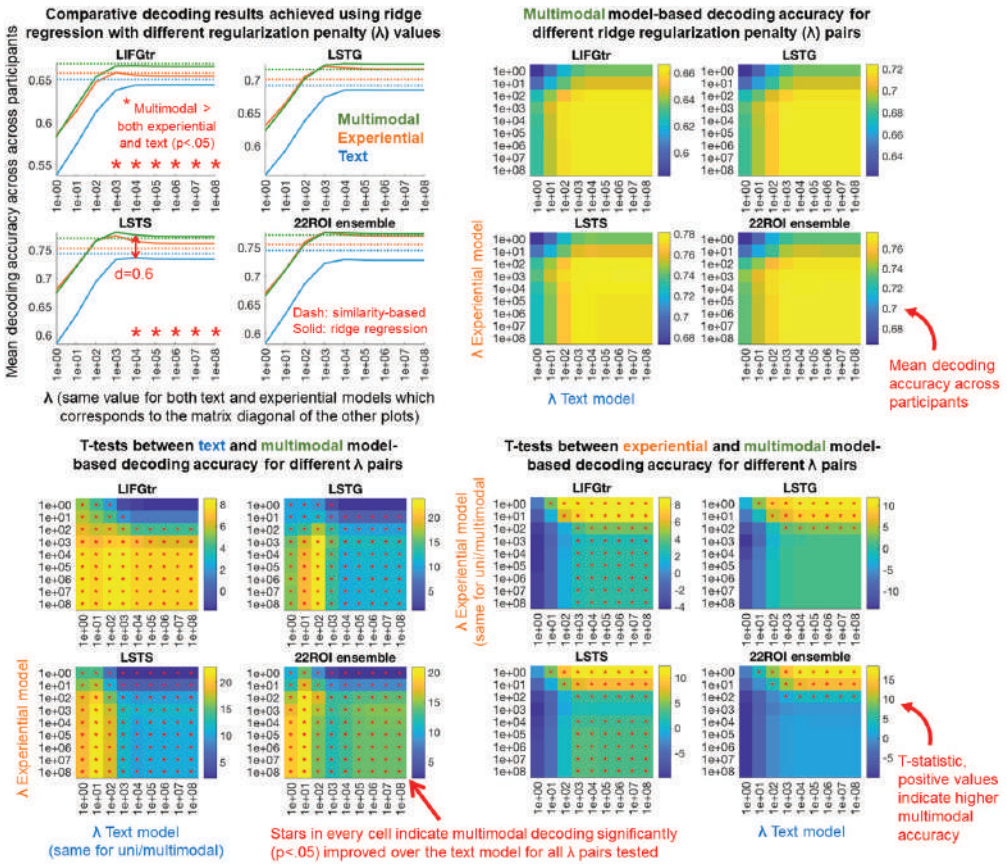


Figure 10. **Comparative text-based, experiential and multimodal decoding accuracies acquired using ridge regression** *Multimodal advantages* for a particular selection of λ values are in the top left. Multimodal decoding accuracies for all λ configurations are in the top right. Results of paired t-tests comparing multimodal decoding accuracies to text-based decoding accuracies and experiential decoding accuracies for each λ configuration are in the bottom left and right plots. All tests were one-tailed, in anticipation of the multimodal advantage observed in our initial analyses. The illustrated effect size (d) provides a conservative estimate of the benefit of integrating experiential features into a conventional text-based ridge regression approach. d was computed as described in the caption of **Figure 4**. *Differences between ridge regression and similarity-based decoding:* The top left plot illustrates both similarity-based (dashed lines) and regression-based (solid lines) results. Decoding accuracies using ridge regression with the text model and the top performing λ (always $\lambda=10^4$) were unanimously significantly lower than for the similarity-based approach (LSTS: t=5.8, $p=6*10^{-5}$, df=13; LSTG: t=5.3, $p=1.4*10^{-4}$, df=13; LIFGtr: t=5.2, $p=1.8*10^{-4}$, df=13; 22ROI: t=8, $p=2*10^{-6}$, df=13; all 2-tailed paired t-tests, df=13). Conversely, in 75% of tests using the experiential model with the top

scoring λ (always $\lambda=10^3$), ridge regression yielded significantly stronger decoding accuracies than the similarity-based analysis (LSTS: t=5.6, $p=8.5*10^{-5}$; LSTG: t=4, p=0.001; LIFGtr: t=0.1, p=0.9; 22ROI: t=7, $p=9.8*10^{-6}$; all 2-tailed paired t-tests, df=13). For multimodal decoding, ridge regression yielded stronger decoding in LSTS with the top scoring $\lambda=10^3$ (t=3.3, p=0.006) but not other λ values, and otherwise there were no significant differences for the other ROIs (LSTG: t=1.9, p=0.07; LIFGtr: t=-.88, p=0.4; 22ROI: t=1.4, p=0.2; all 2-tailed paired t-tests, df=13). All p-values in plots and captions are not corrected for multiple comparisons.

The comparatively weak decoding accuracies obtained with ridge regression for the text-based model begged the question of whether the ridge regression multimodal decoder was failing to capitalize on neural information that was decoded successfully by the similarity-based approach. Relatedly, whether decoding of LSTG and the 22ROI ensemble would be advantaged by a "best of both worlds" multimodal approach that jointly leverages similarity-based decoding and ridge regression. To answer this question, we reran the cross-validation analysis using the similarity-based approach with the text-based model in parallel with ridge regression on the experiential model (using top scoring $\lambda=10^3$). At each cross validation iteration we integrated the respective decoding decisions made by the different models/decoders by pointwise summing respective 2*2 decoding decision matrices (**as in Figure 1 stage 4**). The multimodal joint similarity/regression approach indeed yielded significantly greater decoding accuracies than both the regression-based experiential decoder and the text similarity-based decoder in LSTG and the 22ROI ensemble as well as LSTS and LIFGtr (see **Figure 11**, statistical test results are in the caption).

In sum this section has provided further evidence that experiential semantic features explain variance in sentence-level fMRI data that cannot be accounted for by state-of-the-art text-based regression approaches, and further support for the claim that multimodal approaches provide the most accurate models of fMRI to date.

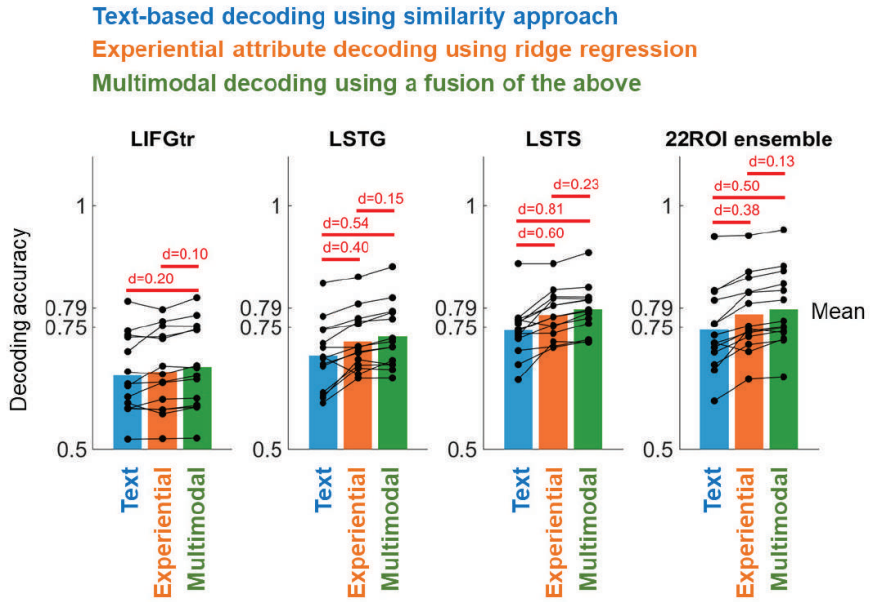


Figure 11. Multimodal decoder integrating the best text-based decoder (similarity) with the best experiential decoder (ridge regression, $\lambda=10^3$) Effect sizes (d) are displayed in cases of statistically significant differences (paired t-test, all p ≤.01, FDR corrected). d was computed as described in the caption of **Figure 4**. Paired t-test results were: for contrasts between multimodal decoding and experiential regression-based decoding: LSTS: t=4.8, p=0.002; LSTG: t=3.6, p=0.01; LIFGtr: t=4.5, p=0.003; 22ROI: t=5, p=0.002; for contrasts between multimodal decoding and text similarity-based decoding: LSTS: t=8.3, $p<10^{-4}$; LSTG: t=8.0, $p<10^{-4}$; LIFGtr: t=4.0, p=0.006; 22ROI: t=5, $p<10^{-4}$; for contrasts between experiential regression-based decoding and text similarity-based decoding: LSTS: t=4.2, p<0.005; LSTG: t=3.8, p<0.009; LIFGtr: t=1.39, p=0.58; 22ROI: t=4.8, p=0.002. All p-values FDR corrected.

Supporting analysis: persistence of multimodal advantage using different text-based models

The claim that the experiential model enhanced decoding by capturing non-linguistic experiential knowledge rests on the assumption that the current text-based model captured all of the experiential structure that is possible to obtain from word use statistics (see **Methods**). As both the text-based and

experiential modeling approaches are in an ongoing state of development it cannot be concluded that the current results will be the same for all future models and/or neural data sets. Additionally, it is possible that idiosyncrasies surrounding how the experiential model was constructed and its statistical properties (e.g. representational sparseness) could have contributed to the decoding advantage it conferred. Vice versa for the text-based model. Consequently, we consider that the current results provide "early evidence" that linguistic and non-linguistically acquired knowledge is represented in fMRI activation elicited in sentence comprehension. However, our core finding that an integrated experiential and text-based decoding approach yields significantly higher accuracy than either model alone has held true for all text-based models we have tested thus far, which have been built using different algorithms from different text corpora. We are not aware of another experiential model suitable for the current analysis.

In preliminary investigations we had tested word co-occurrence models (e.g. Roller et al. 2014), word2vec (Mikolov et al. 2013, Baroni et al. 2014), that yielded similar decoding accuracy levels (e.g. LSTS: mean ≤0.75) and a similar core result that integrating experiential and text-based models yielded significantly greater decoding accuracies (e.g. LSTS: Multimodal > cooccurrence $t=10.1$, $p<10^{-5}$, $d=0.78$; Multimodal>word2vec $t=6.9$, $p<10^{-5}$, $d=.44$, both 2-tailed paired t-tests). We focused on GloVe (Pennington et al. 2014) principally because it was the basis for Pereira et al.'s (2018) fMRI sentence decoding study.

In the interim a number of new computational text-based approaches have emerged (e.g. Conneau et al. 2017, Peters et al. 2018, Subramanian et al. 2018, Devlin et al. 2019). These typically leverage deep artificial neural networks to derive sentence representations that reflect word order and within-sentence contexts. A thorough comparison of deep network text-based approaches and the experiential model is beyond the scope of the current work, and perhaps would be best undertaken using an experiential model that also accommodates word order and context effects (which could be achieved by rating entire sentences, words in context, or entering experiential word vectors as deep network input).

Nevertheless, to date we have tested one deep model, InferSent (Conneau et al. 2017), which is notable in having recently yielded state-of-the-art decoding of Pereira et al.'s (2018) sentence-level fMRI dataset (Sun et al. 2019). We hope to present a full treatment of results in future work, however to foreshadow those, the current core finding still holds: Integrating the current sentence-level experiential model with InferSent yields significantly greater decoding accuracy than either model in isolation (e.g. LSTS: Multimodal > InferSent $t=7.2$, $p<10^{-5}$, $d=0.34$, 2-tailed paired t-test).

Discussion 1502/1500 words

This study has revealed early evidence that modeling both linguistic knowledge of word usage and experiential knowledge of words' referents enhances decoding of brain activation patterns associated with sentence meaning. This suggests that non-linguistic experiential knowledge is represented in sentence-level fMRI activation. Importantly, because this result is based on direct measures of brain activation elicited during the comprehension of natural sentences, it is an advance on previous behavioral evidence that has been indirectly inferred from experimental responses (e.g. Paivio, 1971, Stanfield and Zwaan, 2001 Zwaan et al. 2002; Andrews et al. 2009, Louwerse and Jeuniaux, 2010, Kousta et al. 2011). More generally this is evidence that it is now possible to use brain data to quantitatively estimate the contribution that linguistic and non-linguistically acquired knowledge make to representing the meaning of natural language. This is especially relevant to theories that conceptual representations are acquired through and partially embodied within experiential neural systems (Barsalou et al. 2008, Glenberg, 2010, Pulvermüller 2013, Binder et al. 2016). Relatedly, it suggests we can begin to estimate how close "ungrounded" semantic models (e.g. text-based) can get to representing human conceptual knowledge (see Harnad, 1990 for discussion of the "symbol grounding problem"). With respect to questions of grounding and embodiment we should be clear that the current analyses provide no guarantee that brain activation that was selectively decoded by the experiential model was actually represented within primary perceptual/modal processing systems, or was critical to comprehension rather than epiphenomenal (Mahon and Caramazza 2009, Mahon 2015).

At a more practical level, the current study advances on previous state-of-the-art text-based neural encoders/decoders (Huth et al. 2016, Huth et al. 2016b, Pereira et al. 2018) because multimodal integration boosts decoding performance. This provides further evidence that combining multiple modalities of information in semantic models leads to more human-like representations of meaning (Andrews et al. 2009, Bruni et al. 2014, Anderson et al. 2015).

We have also demonstrated that the text-based model contributed particularly to decoding sentences containing abstract words. Although this was hypothesized (Anderson et al. 2016a) and text-based models have previously helped explain abstract concept fMRI (Anderson et al. 2017, Wang et al. 2018, Pereira et al. 2018), it was not a foregone conclusion. This was because abstract concepts are thought to be grounded relatively strongly upon affective experiences (Kousta et al. 2011, Vigliocco et al. 2014) and contemporary text-based models generate relatively weak predictions of affective experiential attributes (Utsumi, 2018). As it turned out many sentences that the text-based model helped explain contained words with affective connotations (e.g. “happy”, “celebrated”, “survived”). It seems likely that the advantage conferred in these cases (and others) was down to the extra linguistic/contextual information in the text-based model. Otherwise, Utsumi (2018) found text-based models to be disadvantaged in predicting spatial/temporal attributes. Testing how spatial/temporal attributes contribute to semantic representations in the brain may thus provide an interesting avenue for future investigation. Whilst we here detected tentative evidence that the experiential model contributed to decoding concrete sentences (without any abstract words), this result did not survive correction for multiple comparisons.

The demonstration that sentence-level neural activation is best decoded using a multimodal approach is not without foreshadow. Anderson et al. (2015) found that a visually grounded semantic model (derived from natural images) and a text-based model differentially correlated with fMRI activation in brain regions with known visual/linguistic processing roles. However, because participants were tasked to read concrete nouns and actively contemplate their semantic properties (Just et al. 2010) it is not

clear whether the results reflect active visual imagery as opposed to more passive language comprehension (see also Willems et al. 2010). In other work, Abnar et al. (2018) used a joint text-based/experiential approach to better predict fMRI elicited by drawings of nouns alongside their names, and Wang et al. (2018) revealed partial correlations between fMRI elicited by abstract Chinese words, a Chinese text-based model and a model built from 12 behavioral ratings that interestingly included valence, space and time. In both cases it is not clear how results would extend to decoding read sentences (in English).

The current study extended a representational similarity-based decoding method (Anderson et al. 2016b, Anderson et al. 2017) to the neural decoding of sentences using parallelized combinations of multiple models, brain regions and participants. Combination of multiple participants' neural data was achieved by "ensemble averaging" of decoding decisions. This sets the similarity-decoding method apart from "hyper alignment" methods (Haxby et al. 2011, Guntupalli et al. 2016) that represent neural responses using a common representational space. A disadvantage of integrating decoding decisions is that this does not generate predictions of individual voxel's activity (unlike regression-based *encoding*). Similarity-based approaches can be configured to estimate voxel activity by applying the correlation coefficients comprising similarity vectors (e.g. **Figure 1 stage 3)** as weights in a weighted average of corresponding brain activation patterns (as described by Anderson et al. 2016b). However, we leave a comparative investigation of this over to future work. We did not run the current analysis using similarity-based encoding in part to avoid the additional data normalization step that would have been required to combine data (see **Figure 1** caption). For the *decoding* case at hand the similarity-based approach performs competitively with ridge regression (better for the text-based model and worse for the experiential model, **Figure 10**) whilst cutting out overheads associated with repeating the analyses with different regularization penalties, and picking the appropriate one.

Cross participant neural decoding was introduced as a method to estimate an upper bound on decoding accuracy achievable with (group-level) models. This followed the reasoning that on average

the most accurate neural decoder will be based on neural data. Indeed, for LSTS, LSTG, LOTFFG, and LMOG decoding accuracy was significantly greater for the cross-participant approach, and there were no ROIs for which accuracy was significantly worse. Practically speaking this result was however not guaranteed. Had the fMRI data been too noisy and the group been too small, the model-based approach could have yielded stronger decoding. It is also important to recall that the upper bound estimate provided by cross participant decoding does not apply to decoding semantics per se but to decoding the entire linguistic processing stream from stimulus perception to semantic interpretation. Also, the cross-participant approach does not apply to decoding person specific aspects of semantic representation, so there may well be decodable neural signal left over that could only be revealed by personal information.

The upper bound decoding estimate was used to identify that weakly decoded sentences tended to contain abstract words, which suggests that the neural data contains undecoded aspects of abstract conceptual representations. This presents a challenge for future modeling to improve on models of abstract concepts. Given limitations in current understanding of abstract knowledge representation there may be an interesting opportunity to move forward here in a different way, by incorporating features of brain activation into artificial semantic models and in so doing provide a new way for neuroscience to feed back to AI (see also Fyshe et al. 2014 and Hassabis et al. 2017).

One limitation of the current study is the assumption that the additional neural activation decoded by the experiential model reflects semantic information that cannot be extracted from natural language data. This is not strictly guaranteed and it is possible that future text-based approaches will account for the signal decoded by the experiential model. A limitation of the current experiential approach is the assumption that experiential knowledge can be comprehensively estimated through introspective ratings of the relationship between concepts and putative neural systems. Indeed, we have revealed evidence that the text-based language model captures information the experiential model did not, however there may be other semantic features that cannot be verbally described and/or introspectively

accessed, in which case models that are truly grounded in modal information (e.g. Bruni and Baroni, 2014; Anderson et al. 2015) may come to the fore. Ultimately the answers to the above questions will be borne out through future work that incorporates different modalities of information into semantic models (e.g. Andrews et al. 2009, Bruni et al. 2014, Kiela and Clark 2017), and compares this to brain data (e.g. Anderson et al. 2013, Anderson et al. 2015, Anderson et al. 2017, Bulat et al. 2017). The current study has contributed methods that we hope will assist in this enterprise.

In conclusion, the current study has provided initial evidence that linguistic and non-linguistic experiential knowledge can be detected in sentence-level brain activation by extending a similarity-based framework to exploit respective models in fMRI decoding. It has also presented a cross participant decoding method which has demonstrated that a substantial amount of neural signal remains unexplained. This decoding gap is likely to be filled by modeling advances that take word order, syntax, morphology and polysemy into account in semantic composition and begin to accommodate pragmatic inferences and theory of mind. For the future, in all of these endeavors we contend that model-based approaches that integrate information across multiple modalities of experience will be necessary for the fullest interpretation of neural activation patterns associated with meaning.

References

Abnar S, Ahmed R, Mijnheer M, Zuidema W. (2018). Experiential, Distributional and Dependency-based Word Embeddings have Complementary Roles in Decoding Brain Activity. Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018), Salt Lake City, Utah, USA: Association for Computational Linguistics. pp 57–66.

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

Andrews M, Frank S, Vigliocco G. (2014). Reconciling embodied and distributional accounts of meaning in language. Top Cogn Sci. 6:359–370.

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

Anderson AJ, Bruni E, Lopopolo A, Poesio M, Baroni M. (2015). Reading visually embodied meaning from the brain: visually grounded computational models decode visual-object mental imagery induced by written text. Neuroimage. 120:309–322.

Anderson AJ, Zinzser BD, Raizada, RDS. (2016b). 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, RDS. (2016). Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. Cerebral Cortex. doi: 10.1093/cercor/bhw240.

Anderson AJ, Kiela D, Clark S, Poesio M. (2017). Visually Grounded and Textual Semantic Models Differentially Decode Brain Activity Associated with Concrete and Abstract Nouns. Transactions of the Association for Computational Linguistics. 5, 17-30.

Anderson AJ, Lalor EC, Lin F, Binder JR, Fernandino L, Raizada RDS, Grimm S, Wang X. (2018). Multiple regions of a cortical network commonly encode the meaning of words in multiple grammatical positions of read sentences. Cerebral Cortex. doi: 10.1093/bhy110

Anderson AJ, Lin F. (2019). How pattern information analyses of semantic brain activity elicited in language comprehension could contribute to the early identification of Alzheimer's Disease. NeuroImage: Clinical 22, 101788.

Barsalou LW. (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22(4), 637–660.

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

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, Desai RH. (2011). The neurobiology of semantic memory. Trends Cogn Sci. 15(11):527–536.

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

Bulat L, Clark S, Shutova E. (2017). Speaking, Seeing, Understanding: Correlating Semantic Models with Conceptual Representation in the Brain. Proc. Conf. Emp. Meth. Nat. Lang. Proc. 1532–1543 (Association for Computational Linguistics, Doha, Qatar, 2014).

Bruffaerts R, De Deyne S, Meersmans K, Liuzzi AG, Storms G Vandenberghe R. 2019. Redefining the resolution of semantic knowledge in the brain: advances made by the introduction of models of semantics in neuroimaging. Neuroscience and Biobehavioral Reviews. https://doi.org/10.1016/j.neubiorev.2019.05.015

Bruni E, Tran N, Baroni M. (2014). Multimodal distributional semantics. J. Artif. Intell. Res. 49, 1–47.

Brysbaert M, New B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41: 977–990.

Brysbaert M, Warriner AB, Kuperman V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. Behav. Res. Methods 46, 904–911.

Benjamini Y, Yekutieli D (2001). The control of the false discovery rate in multiple testing under dependency. Annals of Statistics. 29 (4): 1165–1188.

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

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, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

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

Cree GS, McRae K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). J Exp Psychol Gen. 132(2):163–201.

Chang KM, Mitchell TM, Just MA. (2010). Quantitative modeling of the neural representations of objects: How semantic feature norms can account for fMRI activation. Neuroimage: Special Issue on Multivariate Decoding and Brain Reading. 56:716–727.

Connell L. (2007). Representing object colour in language comprehension. Cognition 102: 476–485.

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

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, Volume 1 (Long and Short Papers) 2019 Jun (pp. 4171-4186).

Dove G. (2014). Thinking in Words: Language as an Embodied Medium of Thought. Topics in Cognitive Science 6, 371–389.

Dunlap WP, Cortina JM, Vaslow JB, Burke MJ. 1996. Meta-analysis of experiments with matched groups or repeated measures designs. Psychological methods. 1(2):170.

Fedorenko E, Thompson-Schill SL. (2014). Reworking the language network. Trends in cognitive sciences 18 (3), 120-126.

Fedorenko E, Scott TL, Brunner P, Coon WG, Pritchett B, Schalk G, Kanwisher N. (2016). Neural correlate of the construction of sentence meaning. Proceedings of the National Academy of Sciences 113 (41), E6256-E6262.

Fernandino L, Conant LL, Binder JR, Blindauer K, Hiner B, Spangler K, Desai RH. (2013) Where is the action? Action sentence processing in Parkinson's disease. Neuropsychologia. 51(8):1510-7.

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

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

Fu R, Guo J, Qin B, Che W, Wang H, Liu T. (2014). Learning semantic hierarchies via word embeddings. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland: Association for Computational Linguistics. pp 1199–1209.

Fyshe A, Talukdar PP, Murphy B, Mitchell TM. (2014). Interpretable Semantic Vectors from a Joint Model of Brain- and Text-Based Meaning. Proceedings of the Meeting of the Association for Computational Linguistics. Baltimore, Maryland: Association for Computational Linguistics. pp 489–499.

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

Glenberg AM, Kaschak MP. (2002). Grounding language in action. Psychonomic Bulletin & Review, 9, 558–565.

Glenberg AM, Sato M, Cattaneo L, Riggio L, Palumbo D, Buccino G. (2008). Processing abstract language modulates motor system activity. Quarterly Journal of Experimental Psychology, 61, 905–919.

Glenberg A. (2010). Embodiment as a unifying perspective for psychology. Wiley Interdisciplinary Reviews: Cognitive Science, 1(4), 586–596.

Guntupalli JS, Hanke M, Halchenko YO, Connolly AC, Ramadge PJ, Haxby JV. (2016). A model of representational spaces in human cortex. Cereb. Cortex, 26, 2919-2934.

Hamilton LS, Huth AG. (2018). The revolution will not be controlled: Natural stimuli in speech neuroscience. Language, Cognition and Neuroscience. 21:1-0.

Handjaras G, Leo A, Cecchetti L, Papale P, Lenci A, Marotta G, Pietrini P, Ricciardi E. (2017). Modality-independent encoding of individual concepts in the left parietal cortex. Neuropsychologia. 105. 39-49.

Handjaras G, Ricciardi E, Leo A, Lenci A, Cecchetti L, Cosottini M, Marotta G, Pietrini P. (2016). How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge. NeuroImage, 135, 232-242.

Harnad, S. (1990). The symbol grounding problem.Physica D: Nonlinear Phenomena,42(1-3), 335–346.

Hassabis D, Kumaran D, Summerfield C, Botvinick M. (2017). Neuroscience-inspired artificial intelligence. Neuron 95 (2), 245-258.

Hasson U, Egidi G, Marelli M, Willems RM. (2018). Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. Cognition. 180:135-57.

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

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.

de Heer WA, Huth AG, Griffiths TL, Gallant JL, Theunissen FE. (2017). The hierarchical cortical organization of human speech processing. 2017. J. Neurosci., 3267-3216.

Hoerl AE, Kennard RW. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

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.

Huth, AG, Lee T, Nishimoto S, Bilenko, NY, Vu AT, Gallant JL. (2016b). Decoding the semantic content of natural movies from human brain activity. Frontiers in systems neuroscience, 10, 81.

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

Kaschak MP, Madden CP, Therriault DJ, Yaxley RH, Aveyard ME, Blanchard AA, Zwaan RA. (2005). Perception of motion affects language processing. Cognition, 94, B79–B89.

Kaschak MP, Zwaan RA, Aveyard M, Yaxley RH. (2006). Perception of auditory motion affects language processing. Cognitive Science, 30(4), 733–744.

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

Kiela D, Clark S. (2017). Learning Neural Audio Embeddings for Grounded Semantics in Auditory Perception. Journal of Artificial Intelligence Research. doi.org/10.1613/jair.5665.

Kriegeskorte N, Mur M, Bandettini P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience, 2, 4.

Kousta ST, Vigliocco G, Vinson DP, Andrews M, Del Campo E. (2011). The representation of abstract words: why emotion matters. Journal of Experimental Psychology: General 140 (1), 14.

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(2), 211–240.

Lambon Ralph MA, Jefferies E, Patterson K, Rogers TT. (2017) The neural and computational bases of semantic cognition. Nature Reviews Neuroscience. 2017 Jan;18(1):42.

Louwerse MM, Jeuniaux P. (2010). The linguistic and embodied nature of conceptual processing. Cognition. 114:96–104.

Louwerse MM. (2018). Knowing the Meaning of a Word by the Linguistic and Perceptual Company It Keeps. Topics in Cognitive Science. doi:10.1111/tops.12349

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

Lynott D, Connell L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. Behavior research methods 45 (2), 516-526.

Mahon BZ, Caramazza A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. Journal of physiology-Paris 102 (1-3), 59-70.

Mahon BZ. (2015). What is embodied about cognition? Language, cognition and neuroscience 30 (4), 420-429.

Mikolov T, Chen K, Corrado G, Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mitchell J, Lapata M. (2010). Composition in distributional models of semantics. Cogn. Sci. 34, 1388–1429.19.

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

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

Paivio A. (1971). Imagery and verbal processes. Holt, Rinehart, and Winston, New York.

Patterson K, Nestor PJ, Rogers TT. (2007) Where do you know what you know? The representation of semantic knowledge in the human brain. Nature Reviews Neuroscience. 8(12):976.

Pennington J, Socher R, Manning CD. (2014). GloVe: Global Vectors for Word Representation. In Proc. Conf. Emp. Meth. Nat. Lang. Proc. Doha, Qatar: Association for Computational Linguistics. pp 1532–1543.

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. Nature Communications 9 (963)

Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. (2018). Deep contextualized word representations. In Proceedings of NAACL-HLT 2018 (pp. 2227-2237).

Popov V, Ostarek M, Tenison C. (2018). Practices and Pitfalls in Inferring Neural Representations. NeuroImage, 174, 340-351. doi:10.1016/j.neuroimage.2018.03.041.

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

Riordan B, Jones MN. (2010). Redundancy in perceptual and linguistic experience: Comparing feature based and distributional models of semantic information. Topics in Cognitive Science, 3, 303–345.

Roller S, Erk K, Boleda G. Inclusive yet selective: Supervised distributional hypernymy detection. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers 2014 Aug (pp. 1025-1036).

Speed LJ, Vigliocco G. (2014). Eye movements reveal the dynamic simulation of speed in language. Cognitive science 38 (2), 367-382.

Stanfield RA, Zwaan RA. (2001). The effect of implied orientation derived from verbal context on picture recognition. Psychological Science, 12, 153–156.

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

Subramanian S. Trischler A, Bengio Y, Pal, C. J. (2018). Learning general purpose distributed sentence representations via large scale multi-task learning. In Proceedings of the 2018 Inter- national Conference on Learning Representations (ICLR).

Sun J, Wang S, Zhang J, Zong C. (2019). Towards sentence-level brain decoding with distributed representations. In Proceedings of the AAAI Conference on Artificial Intelligence 2019 Jul 17 (Vol. 33, pp. 7047-7054).

Talairach J, Tournoux P. (1988). Co-planar stereotaxic atlas of the human brain. 3-Dimensional proportional system: an approach to cerebral imaging. New York: Thieme.

Turney P, Pantel P. (2010). From frequency to meaning: Vector space models of semantics. J. Artif. Intell. Res. 37, 141–188.

Utsumi A. (2018). A Neurobiologically Motivated Analysis of Distributional Semantic Models. arXiv:1802.01830.

Vigliocco G, Meteyard L, Andrews M, Kousta S. (2009). Toward a theory of semantic representation. Language and Cognition 1 (2), 219-247.

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

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

Vinson DP, Vigliocco G, Cappa S, Siri S. (2003). The breakdown of semantic knowledge: insights from a statistical model of meaning representation. Brain Lang. 86(3):347–365.

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. Human brain mapping 38 (10), 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. Cerebral Cortex. 28(12):4305-18.

Wehbe L, Murphy B, Talukdar P, Fyshe A, Ramdas A, Mitchell T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. PloS one. 9(11), e11257.

Willems RM, Toni I, Hagoort P, Casasanto D. (2010). Neural dissociations between action verb understanding and motor imagery. Journal of Cognitive Neuroscience, 22(10), 2387–2400.

Winter B, Bergen B. (2012). Language comprehenders represent object distance both visually and auditorily. Language and Cognition, 4(1), 1–16.

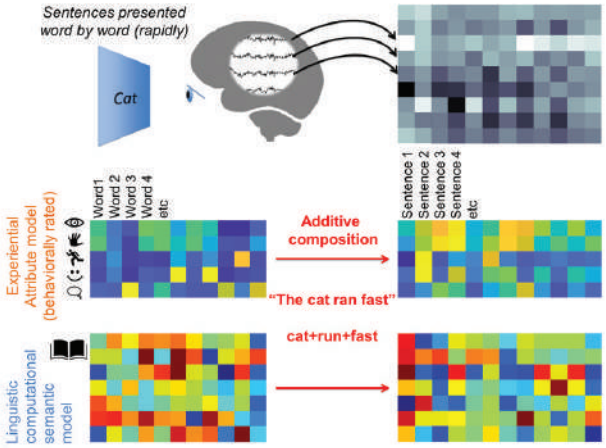
Yang Y, Wang J, Bailer C, Cherkassky V, Just MA. (2017). Commonality of neural representations of sentences across languages: Predicting brain activation during Portuguese sentence comprehension using an English-based model of brain function. NeuroImage (146):658-666.

Zwaan RA, Stanfield RA, Yaxley RH. (2002). Language comprehenders mentally represent the shapes of objects. Psychol. Sci. 13 (2), 168–171.

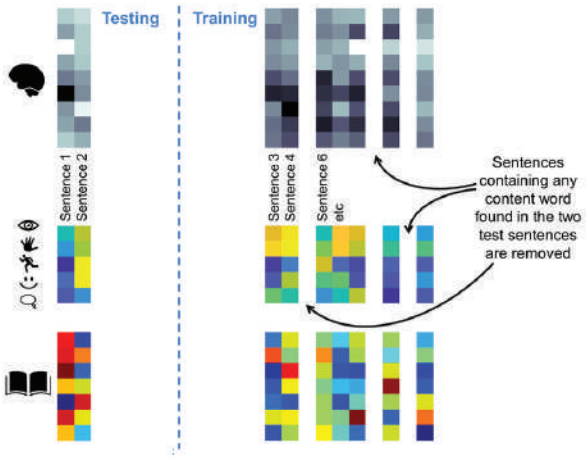
Zwaan RA, Pecher D. (2012). Revisiting mental simulation in language comprehension: Six replication attempts. PLoS ONE, 7, 12.

Zwaan RA. (2014). Embodiment and language comprehension: Reframing the discussion. Trends in Cognitive Sciences, 18(5), 229–234.

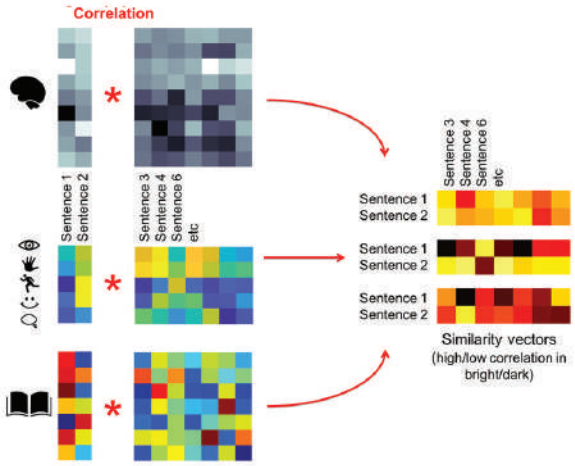
1. Neural, experiential and text-based sentences



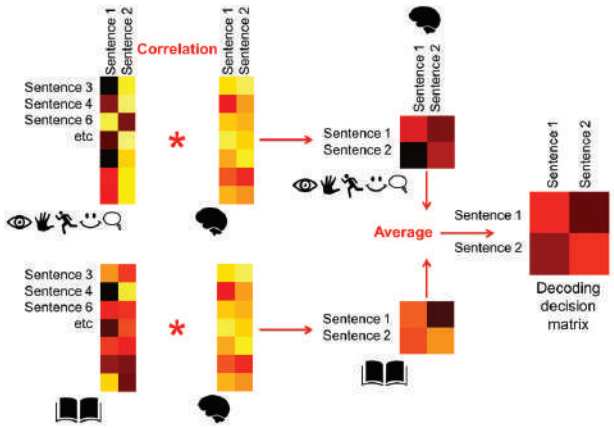
2. Cross validation training/test split



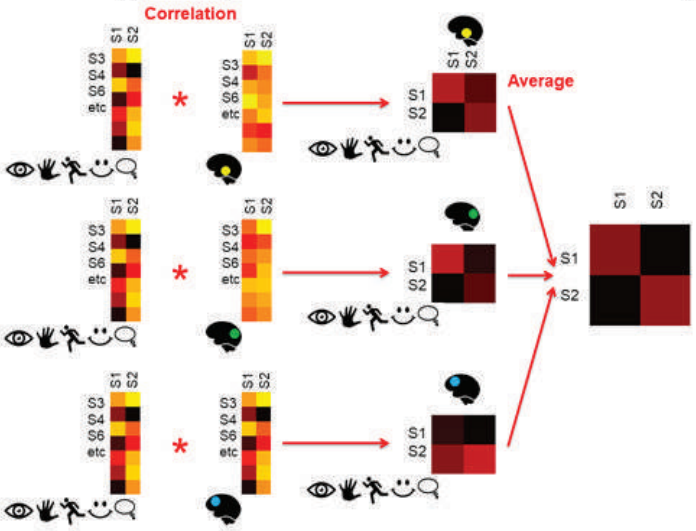
3. Re-representation in common similarity space



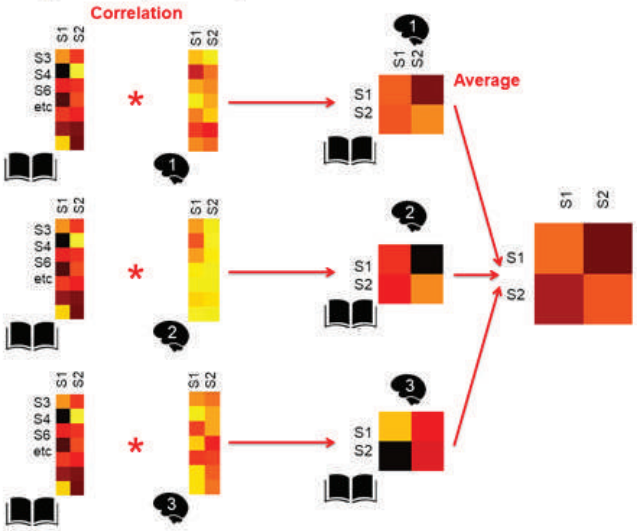
4. Multimodal model-based decoding



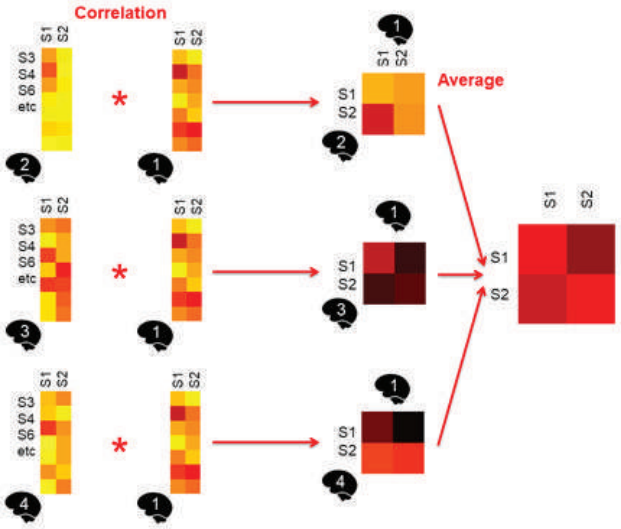
Combining multiple ROIs in model-based decoding

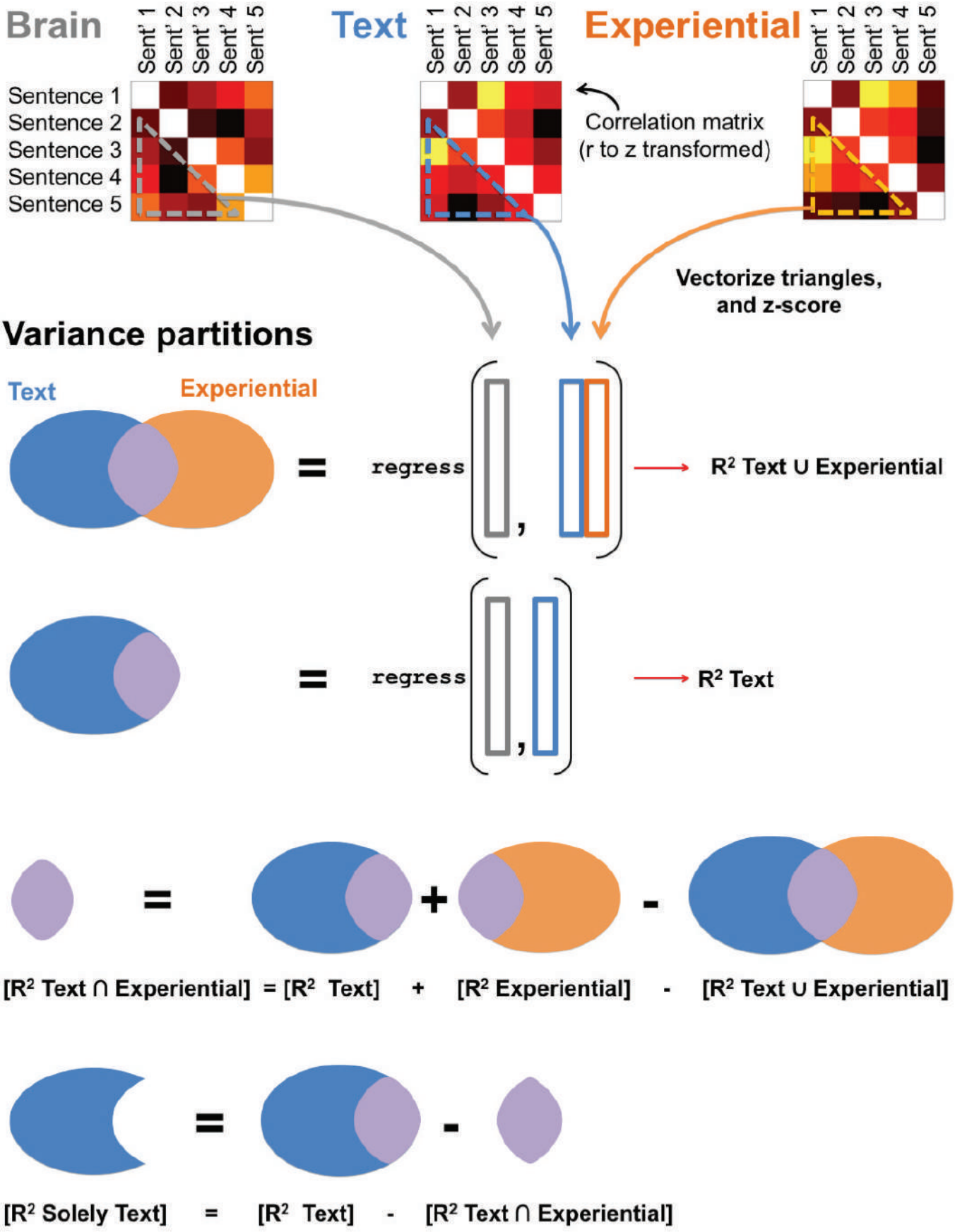


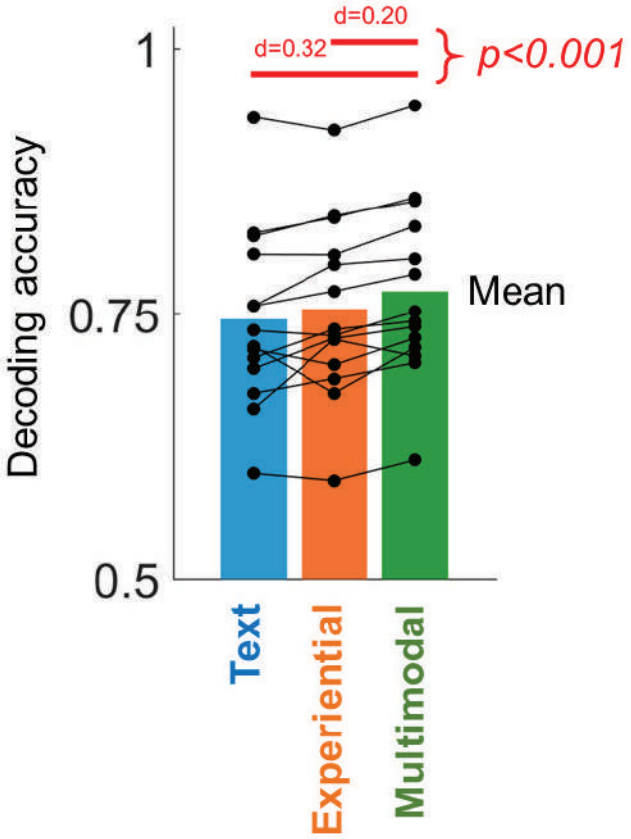
Combining multiple subjects in model-based decoding

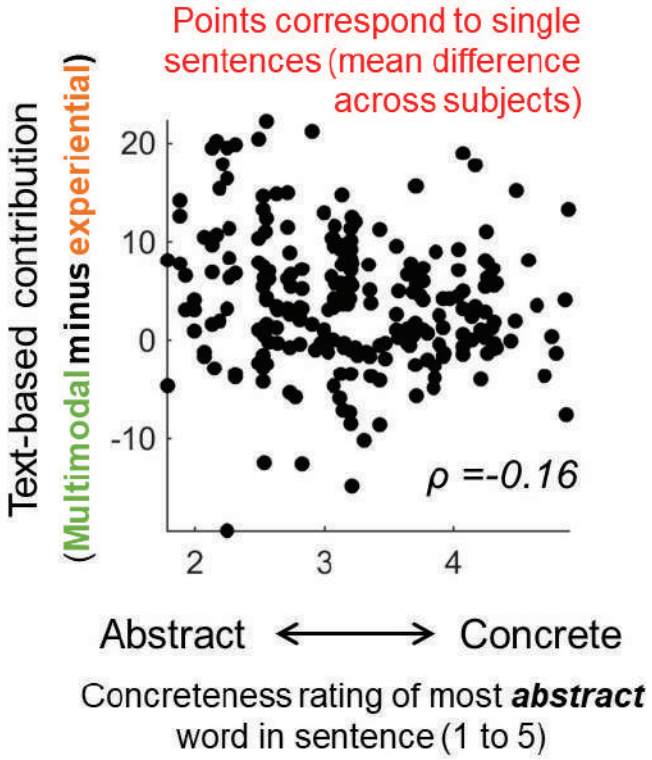


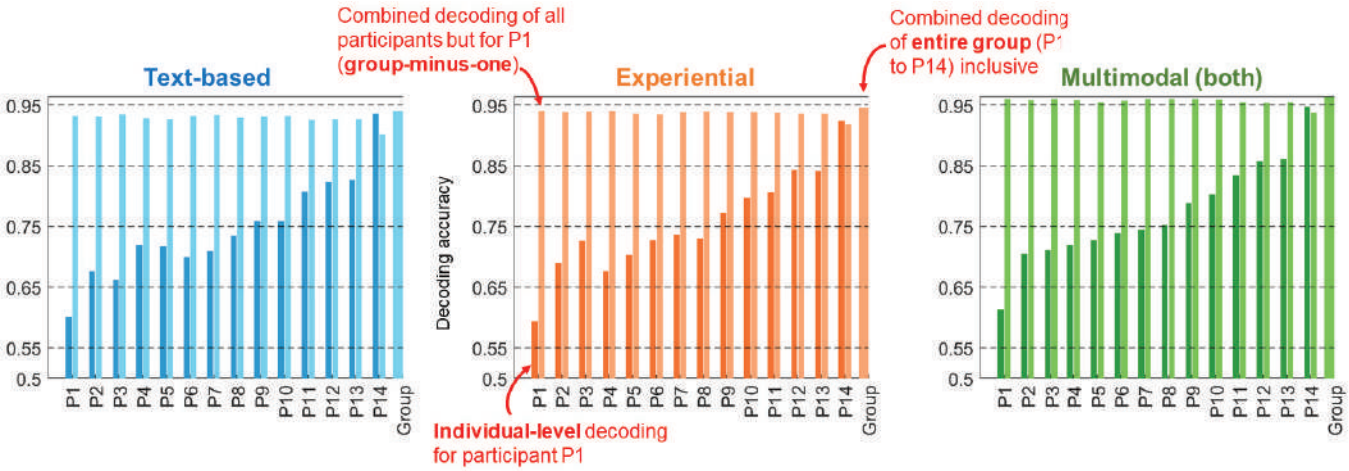
Cross subject brain-based decoding



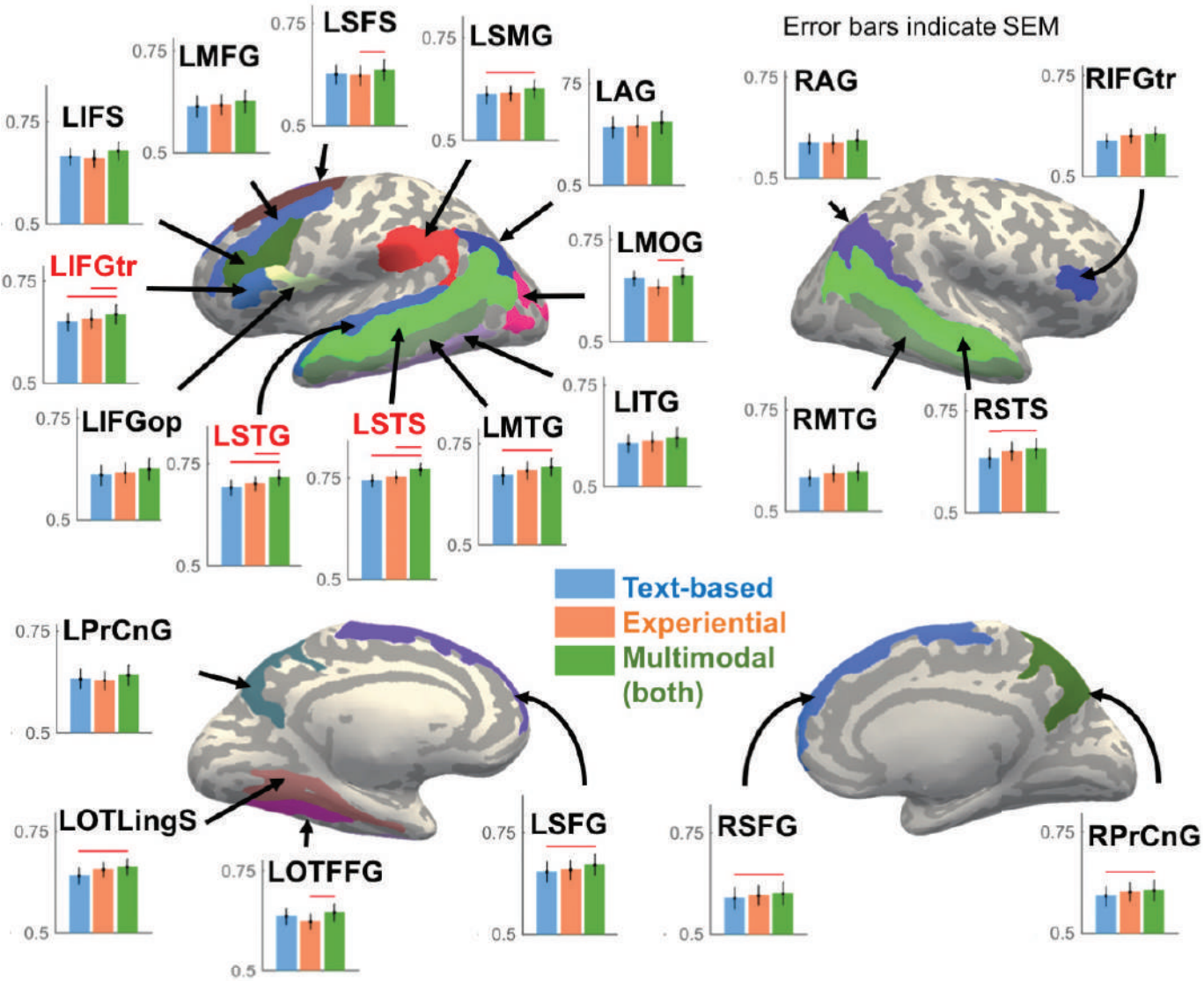




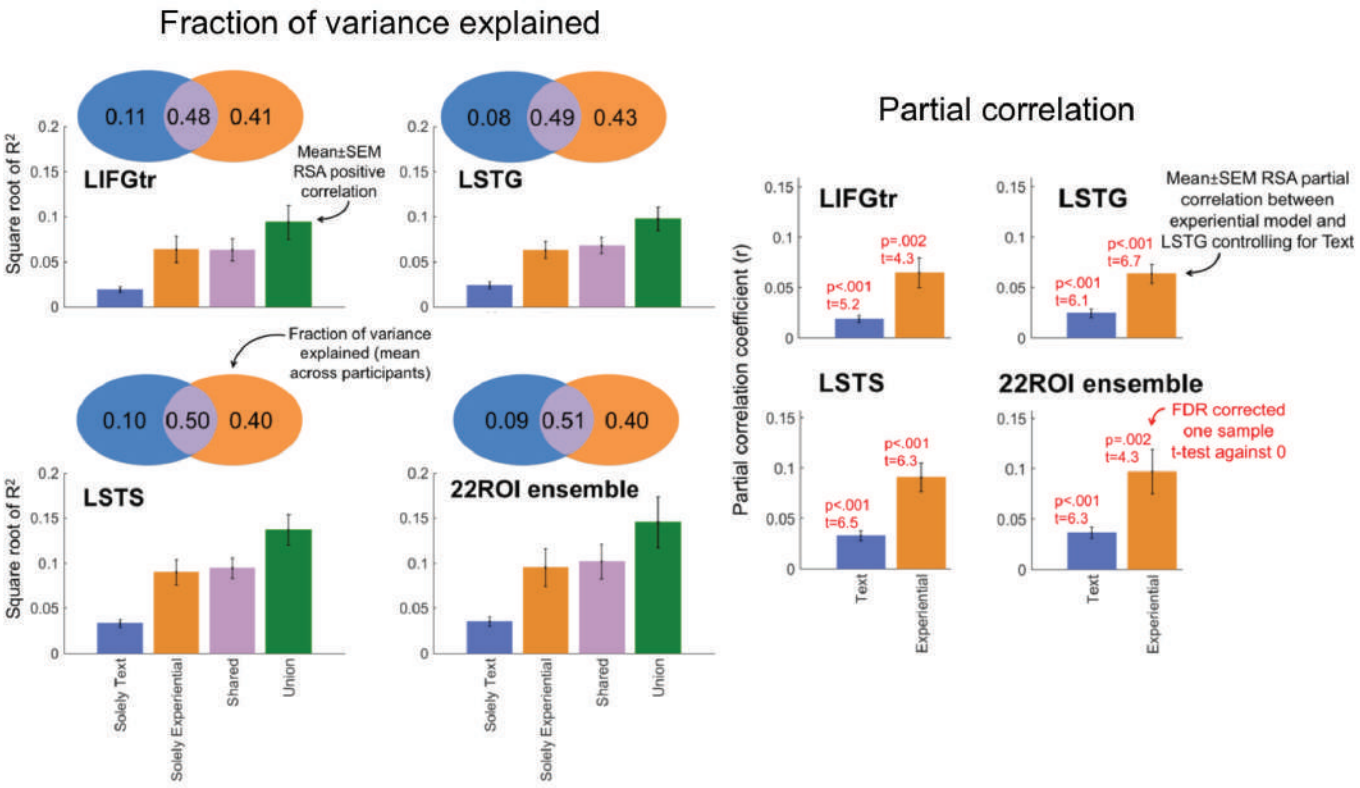


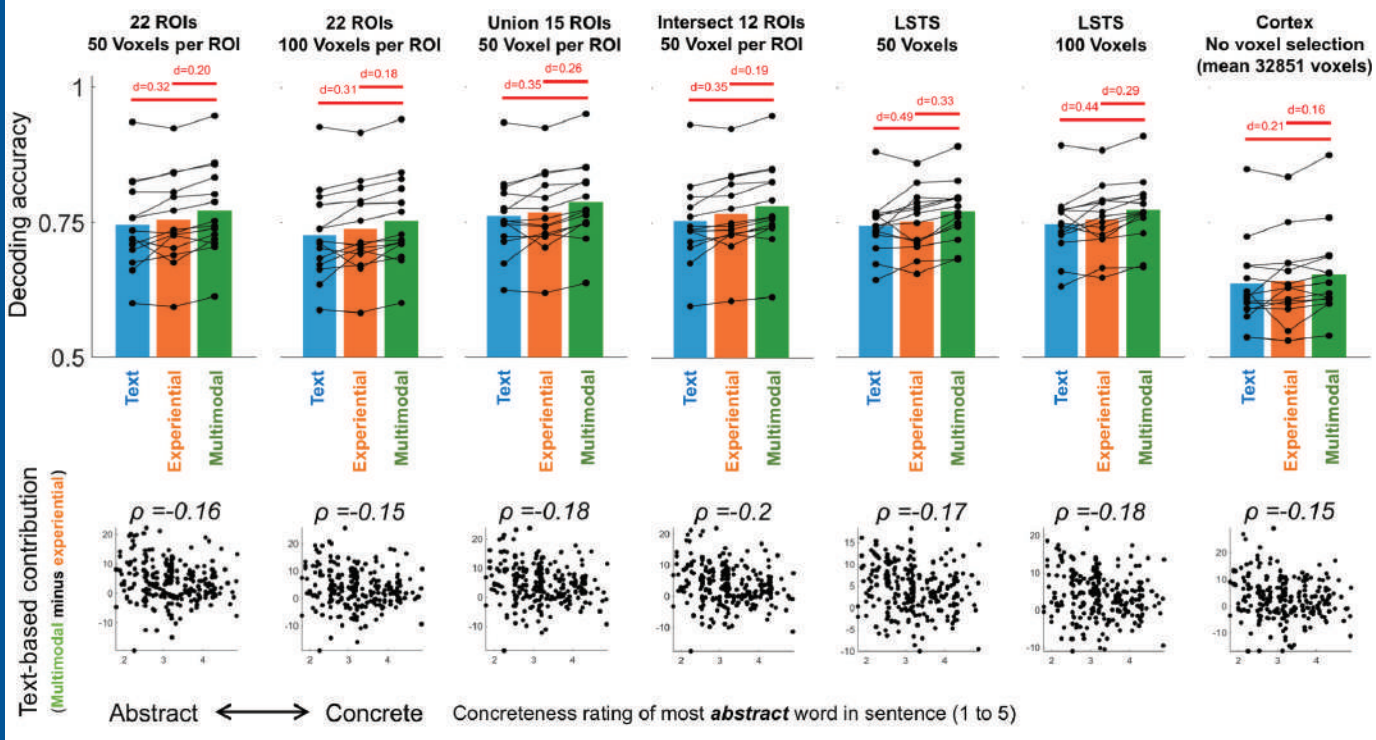


Note: each participant contributed complementary information (entire group accuracy surpassed all individuals, and all group-minus-one combinations)



Significant ($p<0.05$) difference between connected model pair (paired t-test, FDR corrected)





Note: Cross-subject (brain activation-based) decoding is liable to decode semantic and non-semantic (e.g. orthographic/syntactic) elements of neural activation