Decoding Brain Activity Associated with Literal and Metaphoric Sentence Comprehension Using Distributional Semantic Models

Vesna G. Djokic[†] **Jean Maillard**[‡] **Luana Bulat**[‡] **Ekaterina Shutova**[†]

[†]ILLC, University of Amsterdam, The Netherlands
[‡]Dept. of Computer Science & Technology, University of Cambridge, United Kingdom
vesna@imsquared.eu, jean@maillard.it,
ltf24@cam.ac.uk, e.shutova@uva.nl

Abstract

Recent years have seen a growing interest within the natural language processing (NLP) community in evaluating the ability of semantic models to capture human meaning representation in the brain. Existing research has mainly focused on applying semantic models to decode brain activity patterns associated with the meaning of individual words, and, more recently, this approach has been extended to sentences and larger text fragments. Our work is the first to investigate metaphor processing in the brain in this context. We evaluate a range of semantic models (word embeddings, compositional, and visual models) in their ability to decode brain activity associated with reading of both literal and metaphoric sentences. Our results suggest that compositional models and word embeddings are able to capture differences in the processing of literal and metaphoric sentences, providing support for the idea that the literal meaning is not fully accessible during familiar metaphor comprehension.

1 Introduction

Distributional semantics aims to represent the meaning of linguistic fragments as high-dimensional dense vectors. It has been successfully used to model the meaning of individual words in semantic similarity and analogy tasks (Mikolov et al., 2013; Pennington et al., 2014); as well as the meaning of larger linguistic units in a variety of tasks, such as translation (Bahdanau et al., 2014) and natural language inference (Bowman et al., 2015). Recent research has also demonstrated the ability of distributional models to predict patterns of brain activity associated with the meaning of words, obtained via functional magnetic resonance imaging (fMRI) (Mitchell et al., 2008; Devereux et al., 2010; Pereira et al., 2013). Following in their steps, Anderson et al. (2017b) have investigated visually grounded semantic models in this context. They found that while both visual and text-based models can equally decode concrete words, text-based models show an overall advantage over visual models when decoding more abstract words.

Other research has shown that data-driven semantic models can also successfully predict patterns of brain activity associated with the processing of sentences (Pereira et al., 2018) and larger narrative text passages (Wehbe et al., 2014; Huth et al., 2016). Recently, Jain and Huth (2018) investigated long short-term memory (LSTM) recurrent neural networks and showed that semantic models that incorporate larger-sized context windows outperform those with smaller-sized context windows, as well as the baseline bag-of-words model, in predicting brain activity associated with narrative listening. This suggests that compositional semantic models are sufficiently advanced to study the impact of linguistic context on semantic representation in the brain. In this paper, we investigate the extent to which lexical and compositional semantic models are able to capture differences in human meaning representations, resulting from meaning disambiguation of literal and metaphoric uses of words in context.

Metaphoric uses of words involve a transfer of meaning, arising through semantic composition (Mohammad et al., 2016). For instance, the meaning of the verb *push* is not intrinsically metaphorical; yet it receives a metaphorical interpretation

Transactions of the Association for Computational Linguistics, vol. 8, pp. 231–246, 2020. https://doi.org/10.1162/tacl_a_00307
Action Editor: Walter Daelemans. Submission batch: 11/2019; Published 4/2020.

when we talk about *pushing agendas*, *pushing drugs*, or *pushing ourselves*. Theories of metaphor comprehension differ in terms of the kinds of processes (and stages) involved in arriving at the metaphorical interpretation, mainly whether or not the abstract meaning is indirectly accessed via processing the literal meaning first or directly accessible largely bypassing the literal meaning (Bambini et al., 2016). To this extent, the role that access to and retrieval of the literal meaning plays during metaphor processing is often debated. On the one hand, metaphor comprehension involves juxtaposing two unlike things and this may invite a search for common relational structure through a process of direct comparison. Inferences flow from the vehicle to the topic giving rise to the metaphoric interpretation (Gentner and Bowdle, 2005). In a slightly different vein, Lakoff (1980) suggest that metaphor comprehension involves systematic mappings (between a concrete domain onto another typically more abstract domain) that become established through co-occurrences over the course of experience. This draws on mental imagery or the re-activation of neural representations involved during primary experience (i.e., sensorimotor simulation) allowing appropriate inferences to be made. Other theories, however, suggest that the literal meaning in metaphor comprehension may be largely bypassed if the abstract meaning is directly or immediately accessible involving more categorical processing (Glucksberg, 2003). For example, the word used metaphorically could be immediately recognized as belonging to an abstract superordinate category of which both the vehicle and topic belong. Alternatively, it has been suggested that more familiar metaphors involve categorical processing, while comparatively novel metaphor will involve initially greater processing of the literal meaning (Desai et al., 2011).

To contribute to our understanding of metaphor comprehension, including the accessibility of the literal meaning, we investigate whether semantic models are able to decode patterns of brain activity associated with literal and metaphoric sentence comprehension, using the fMRI dataset of Djokic et al. (forthcoming). This dataset contains neural activity associated with the processing of both literal and *familiar* metaphorical uses of hand-action verbs (such as *push, grasp, squeeze*, etc.) in the context of their nominal object. We experiment with several kinds of semantic models: (1) *word-based models*, namely, word embeddings of the verb and the nominal object; (2) *compositional models*, namely, vector addition and an LSTM recurrent neural network; and (3) *visual models*, learning visual representations of the verb and its nominal object. This choice of models allows us to investigate: (1) the role of the verb and its nominal object (captured by their respective word embeddings) in the interpretation of literal and metaphoric sentences; (2) the extent to which compositional models capture the patterns of human meaning representation in case of literal and metaphoric use; and (3) the role of visual information in literal and metaphor interpretation. We test these models in their ability to decode brain activity associated with literal and metaphoric sentence comprehension, using the similarity decoding method of Anderson et al. (2016). We perform decoding at the whole brain level, as well as within specific regions implicated in linguistic, motor and visual processing.

Our results demonstrate that several of our semantic models are able to predict patterns of brain activity associated with the meaning of literal and metaphorical sentences. We find that (1) compositional semantic models are superior in decoding both literal and metaphorical sentences as compared to the lexical (i.e., word-based) models; (2) semantic representations of the verb are superior compared to that of the nominal object in decoding literal phrases, whereas semantic representations of the object are superior to that of the verb in decoding metaphorical phrases; and (3) linguistic models capture both language-related and sensorimotor representations for literal sentences—in contrast, for metaphoric sentences, linguistic models capture language-related representations and the visual models captured sensorimotor representations in the brain. Although the results do not offer straightforward conclusions regarding the role of the literal meaning in metaphor comprehension, they provide some support to the idea that lexical-semantic relations associated with the literal meaning are not fully accessible during familiar metaphor comprehension, particularly within action-related brain regions.

2 Related Work

2.1 Decoding Brain Activity

Mitchell et al. (2008) were the first to show that distributional representations of concrete nouns

built from co-occurrence counts with 25 experiential verbs could predict brain activity elicited by these nouns. Later studies used the fMRI data of Mitchell et al. (2008) as a benchmark for testing a range of semantic models including topic model-based semantic features learned from Wikipedia (Pereira et al., 2013), feature-norm based semantic features (Devereux et al., 2010), and skip-gram word embeddings (Bulat et al., 2017). Anderson et al. (2013) demonstrate that visually grounded semantic models can also decode brain activity associated with concrete words and show the best results using multimodal models. Additionally, Anderson et al. (2015) show that text-based models are superior in predicting brain activity of concrete words in brain areas related to linguistic processing, and the visual models in those related to visual processing. Lastly, Anderson et al. (2017b) use image and text-based semantic models to decode an fMRI dataset containing nouns with varying degree of concreteness. They show that text-based models have an advantage decoding the more abstract words over the visual models, supporting the view that concrete concepts involve linguistic and visual codes, while abstract concepts mainly linguistic codes (Paivio, 1971).

Subsequent studies have focused on evaluating the ability of distributional semantic models to encode brain activity elicited by larger text fragments. Pereira et al. (2018) showed that a regression model trained to map between word embeddings and the fMRI patterns of words could predict neural representations for unseen sentences. Adding to this, both Wehbe et al. (2014) and Huth et al. (2016) showed that distributional semantic models could predict neural activity associated with narrative comprehension. For instance, Wehbe et al. (2014) showed that a regression model that learned a mapping between several story features (distributional semantics, syntax, and discourse-related) and fMRI patterns associated with narrative reading could distinguish between two stories. These findings suggest that encoding models using word embeddings as features can predict brain activity associated with larger linguistic units. Other researchers have evaluated models that more directly consider the role played by the linguistic context and syntax (Anderson et al., 2019; Jain and Huth, 2018). Jain and Huth (2018) showed that a regression-based model mapping between fMRI patterns associated with narrative listening and contextual features obtained from an LSTM language model outperformed the bag-of-words model. Moreover, the performance increased when using LSTMs with larger context-windows.

In parallel to this work, several other works have been successful in decoding word-level and sentential meanings using semantic models based on human behavioral data. Chang et al. (2010) use taxonomic encodings of McRae et al. (2005), while Fernandino et al. (2015) use semantic models based on human-elicited salience scores for sensorimotor attributes to decode neural activity associated with concrete concepts. Interestingly, the latter report that their model is unable to decode brain activity associated with the meaning of more abstract concepts. Lastly, other research has achieved similar success in decoding sentential meanings using neuro-cognitively driven features that more closely reflect human experience (Anderson et al., 2017a; Wang et al., 2017; Just et al., 2017). For example, Anderson et al. (2017a) showed that a multiple-regression model trained to map between 65-dimensional experiential attribute model of word meaning (e.g., motor, spatial, social-emotional) and the fMRI activations associated with words could predict neural activation of unseen sentences. These findings highlight the importance of considering the neurocognitive constraints on semantic representation in the brain.

2.2 Semantic Representation in the Brain

Semantic processing is thought to depend on a number of brain regions functioning in concert as a unified semantic network linking language, memory, and modality-specific systems in the brain (Binder et al., 2009). Xu et al. (2016) provide evidence in support of at least three functionally segregated systems that together comprise such a semantic network. A left-lateralized language-based system spanning frontal-temporal (e.g., left inferior frontal gyrus [LIFG], left posterior middle temporal gyrus [LMTP]), but also parietal areas, is associated with lexical-semantics and syntactic processing. It preferentially responds to language tasks when compared to non-linguistic tasks of similar complexity (Fedorenko et al., 2011). Notably, both Devereux et al. (2014) and Anderson et al. (2015) found that linguistic models could decode concrete concepts within brain areas in this system, mainly the LMTP. Importantly,

this system works in tandem with a memory-based simulation system that interacts directly with medial-temporal areas critical in memory (and multimodal) processing. The memory-based simulation system retrieves memory images relevant to a concept and includes occipital areas such as the superior lateral occipital cortex, implicated in visual processing and which Anderson et al. (2015) showed could decode concrete concepts with visual models. This system also recruits modality-specific information. In line with this, Carota et al. (2017) showed that the semantic similarity of text-based models correlates with fMRI patterns of action words not only in language-related areas, but also in motor areas (left precentral gyrus [LPG], left premotor cortex [LPM]). Lastly, a fronto-parietal semantic control system manages interactions between these two systems, such as directing attention to different aspects of meaning depending on the linguistic context.

Prior neuroimaging experiments show that concrete concepts activate the relevant modality-specific systems in the brain (Barsalou, 2008, 2009), while the processing of abstract concepts has been found to engage mainly language-related brain regions in the left hemisphere and areas implicated in cognitive control (Binder et al., 2005; Sabsevitz et al., 2005). Relatedly, action-related words and literal phrases activate motor regions (e.g., to access motoric features of verbs) (Pulvermuller, 2005; Kemmerer et al., 2008). In contrast, the degree to which action-related metaphors engage motor brain regions appears to depend on novelty, with more familiar metaphors (Desai et al., 2011) showing little to no activity in motor areas. In sum, concrete language involves modality-specific and language-related brain regions, while abstract language mainly language areas (Hoffman et al., 2015).

To assess the role of linguistic versus visual information in literal and metaphor decoding, we investigated the extent to which our semantic models were able to decode literal and metaphoric sentences not only across the whole brain (and brain's lobes), but also within specific brain regions of interest (ROIs) implicated in visual, action, and language processing. The visual ROIs include high-level visual brain regions (left lateral occipital temporal cortex, left ventral temporal cortex), part of the ventral visual stream implicated in object recognition (Bugatus et al., 2017). The action ROIs include sensorimotor brain regions (LPG, LPM) implicated in action-semantics (Kemmerer et al., 2008). Lastly, the language-related ROIs include areas of the classic language network (LIFG, LMTP) implicated in lexico-semantic and syntactic processing (Foderenko et al., 2012).

We expect to find that lexical and compositional semantic models can capture differences in the processing of literal and metaphoric language in the brain. In line with the idea that literal language co-occurs more directly with our everyday perceptual experience, we expect that visual models will show an overall advantage in literal but perhaps not metaphor decoding across the whole brain (particularly within Occipital and Temporal lobes) and in visual (action) ROIs compared to language-related ROIs. In contrast, for metaphor decoding we expect that linguistic models will mainly show an advantage in language-related ROIs compared with visual (and action) ROIs due to their more abstract nature. Lastly, we expect compositional models to be superior to lexical models in metaphor decoding, which relies on semantic composition for meaning disambiguation in context. This allows investigating whether metaphor comprehension involves lingering access to the literal meaning including more grounded visual and sensorimotor representations.

3 Brain Imaging Data

Stimuli consisted of sentences divided into five main conditions: 40 affirmative literal, 40 negated literal, 40 affirmative metaphor, 40 negated metaphor, and 40 affirmative literal paraphrases of the metaphor.

Stimuli Stimuli consisted of sentences divided into five main conditions: 40 affirmative literal, 40 negated literal, 40 affirmative metaphor, 40 negated metaphor, and 40 affirmative literal paraphrases of the metaphor (used as control). A total of 31 unique hand- action verbs were used (9 verbs were re-used twice per condition). For each verb, the authors created four conditions: affirmative literal, affirmative metaphoric, negated literal, and negated metaphoric. All sentences were in the third person singular, present tense, progressive, see Figure 1. Stimuli were created by the authors and normed for psycholinguistic variables (i.e., length, familiarity, concreteness) by an

Condition	Sentence
Affirm. Literal	She's *pushing* the wheelbarrow
Negated Literal	He's not *pushing* the cart
Affirm. Metaphor	She's *pushing* the agenda
Negated Metaphor	He's not *pushing* the idea
Affirm. Paraphrase	She's promoting the agenda

Figure 1: Sample stimuli for the verb *push*.

independent set of participants in a behavioral experiment.

Participants Fifteen adults (8 women, ages 18 to 35) were involved in the fMRI study. All participants were right-handed, native English speakers.

Experimental Paradigm Participants were instructed to passively read the object of the sentence (e.g., ''the yellow lemon''), briefly shown on screen first, followed by the sentence (e.g., ''She's squeezing the lemon''). The object was shown on screen for 2 s, followed by a 0.5 s interval, then the sentence was presented for 4 s followed by a rest of 8 s. A total of 5 runs were completed, each lasting 10.15 minutes (3 participants only completed 4 runs). Stimulus presentation was randomized across participants.

fMRI data acquisition fMRI images were acquired with a Siemens MAGNETOM Trio 3T System with a 32-channel head matrix coil. High-resolution anatomical scans were acquired with a structural T1-weighted magnetization prepared rapid gradient echo (MPRAGE) with TR $=$ 1950 ms, TE $=$ 2.26 ms, flip angle 10%, 256 $\times$ 256 mm matrix, 1 mm resolution, and 208 coronal slices. Whole brain functional images were obtained with a T2* weighted single-shot gradient-recalled echoplanar imaging, echo-planar sequence (EPI) using blood oxygenation-level-dependent contrast with TR = 2000 ms, TE = 30 ms, flip angle 90 degrees, 64$\times$64 mm matrix, 3.5 mm resolution. Each functional image consisted of 37 contiguous axial slices, acquired in interleaved mode.

4 Semantic Models

4.1 Linguistic Models

All our linguistic models are based on GloVe (Pennington et al., 2014) 100-dimensional (*dim*) word vectors provided by the authors, trained on Wikipedia and the Gigaword corpus.[1] We investigate the following semantic models:

Individual Word Vectors In this model, stimulus phrases are represented as the individual D-*dim* word embeddings for their verb and direct object. We will refer to these models as VERB and OBJECT, respectively.

Concatenation We then experiment with modelling phrase meanings as the $2D$-*dim* concatenation of their verb and direct object embeddings (VERBOBJECT).

Addition This model takes the embeddings $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ for the words of the stimulus phrase, and computes the stimulus phrase representation as their average: $\boldsymbol{h} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_i$.

LSTM As a more sophisticated compositional model, we take the LSTM recurrent neural network architecture of Hochreiter and Schmidhuber (1997). We trained the LSTM on a natural language inference task (Bowman et al., 2015), as it is a complex semantic task where we expect rich meaning representations to play an important role. Given two sentences, the goal of natural language inference is to decide whether the first *entails* or *contradicts* the second, or whether they are *unrelated*. We used the LSTM to compute compositional representations for each sentence, and then used a single-layer perceptron classifier (Bowman et al., 2016) to predict the correct relationship. The inputs to the LSTM were the same 100-*dim* GloVe embeddings used for the other models, and were updated during training. The model was optimized using Adam (Kingma and Ba, 2014). We extracted 100-*dim* vector representations from the hidden state of the LSTM for the verb-object phrases in our stimulus set.

4.2 Visually Grounded Models

We use the MMfeat toolkit (Kiela, 2016) to obtain *visual representations* in line with Anderson et al. (2017b). We retrieved 10 images for each word or phrase in our dataset using Google Image Search. We then extracted an embedding for each of the images from a deep convolutional neural network that was trained on the ImageNet classification task (Russakovsky et al., 2015). We used an architecture consisting of five convolutional

[1] https://nlp.stanford.edu/projects/glove/.

layers, followed by two fully connected rectified linear unit layers and a softmax layer for classification. To obtain an embedding for a given image we performed a forward pass through the network and extracted the 4096-*dim* fully connected layer that precedes the softmax layer. The visual representation of a word or a phrase is computed as the mean of its 10 individual image representations.

We experiment with word-based models (VISUAL VERB and VISUAL OBJECT) and the following three visual compositional models:

Concatenation This model represents the stimulus phrase as the concatenation of the two *D-dim* visual representations for the verb and the object (VISUAL VERBOBJECT).

Addition We take the average of the visual representations for the verb and object to give the representation for the phrase (VISUAL ADDITION).

Phrase We obtain visual representations for the phrase, by querying Google images for the verb-object phrase directly (VISUAL PHRASE).

5 Decoding Brain Activity

5.1 fMRI Data Processing

For our experiments we limited analysis to the 12 individuals who completed all runs. The runs were combined across time to form each participant's dataset and preprocessed (high-pass filtered, motion-corrected, linearly detrended) with FSL.[2]

General Linear Modeling After fMRI preprocessing, we selected sentences within the affirmative literal and affirmative metaphoric conditions representative of the 31 unique verbs as conditions of interest for all our experiments. We fit a model of the hemodynamic response function to each stimulus presentation using a univariate general linear model with PyMVPA.[3] The entire stimulus presentation was modeled as an event lasting 6 s after taking into account the hemodynamic lag of 4 s. The model parameters (Beta weights) were normalized to Z-scores. Each stimulus presentation was then represented as a single volume containing voxel-wise Z-score maps for each of the 31 affirmative literal and 31 metaphoric sentences. The affirmative literal or metaphoric neural estimates were then used to perform similarity-based decoding, separately.

Voxel Selection We performed feature selection by selecting the top 35% of voxels that showed the highest sensitivity (F-statistics) using a univariate ANOVA as a feature-wise measure with two groups: the 31 affirmative literal sentences versus 31 affirmative metaphoric sentences. F-statistics were computed for each feature as the standard fraction of between and within group variances using PyMVPA. This selected voxels sensitive to univariate activation differences between literal and metaphoric categories.

5.2 Defining Regions of Interest

Following Anderson et al. (2013), we performed decoding at the whole-brain level and across four gross anatomical divisions: the frontal, temporal, occipital, and parietal lobes. The masks were created using the Montreal Neurological Institute (MNI) Structural Atlas in FSL. We also defined the following a priori ROIs to compare the performance of literal and metaphoric decoding in visual and sensorimotor brain regions vs. language-related brain areas implicated in lexical-semantic processing: (1) visual ROIs (left lateral occipital temporal cortex [LLOCT], left ventral temporal cortex (LVT)); (2) action ROIs (LPG, LPM); (3) language-related ROIs (LMTP, LIFG). The LLOTC and LVT were created manually in FSL using the anatomical landmarks of Bugatus et al. (2017). The LPG and LPM were created using the Juelich Histological Atlas thresholded at 25% in FSL. The LMTP and LIFG were created using the Harvard-Oxford Cortical Probabilistic Atlas thresholded at 25% in FSL. Masks were transformed from MNI standard space into the participant's functional space.

5.3 Similarity-Based Decoding

We use similarity-based decoding (Anderson et al., 2016) to evaluate to what extent the representations produced by our semantic models are able to decode brain activity patterns associated with our stimuli. We first compute two similarity matrices (k stimuli $\times$ k stimuli), containing similarities between all stimulus phrases in the dataset: the model similarity matrix (where similarities are computed using the semantic model vectors) and the brain similarity matrix (where similarities are computed using the brain activity vectors). The

[2] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki.
[3] http://www.pymvpa.org/.

MODELS	Frontal	Parietal	Temporal	Occipital	Whole-Brain
OBJECT	0.55	0.53	0.50	0.40	0.50
VERB	0.67	**0.69***	0.62	0.65	0.63
VERBOBJECT	0.57	0.57	0.51	0.55	0.52
ADDITION	**0.72***	**0.72***	**0.69***	**0.70***	**0.69***
LSTM	0.58	0.50	0.55	0.56	0.53
VISUAL OBJECT	0.60	0.51	**0.70***	**0.69***	0.66
VISUAL VERB	0.41	0.44	0.59	0.66	0.50
VISUAL VERBOBJECT	0.49	0.49	0.61	0.63	0.55
VISUAL ADDITION	0.58	0.48	**0.68***	0.65	0.57
VISUAL PHRASE	0.56	0.52	0.44	0.42	0.45

Table 1: *Literal Decoding*. Leave-2-out decoding accuracies, significant values ($p < 0.05$) surviving FDR correction for multiple comparisons indicated in **bold** by an asterisk (*).

similarities were computed using Pearson correlation coefficient as a measure. We then perform the decoding using a leave-two-out decoding scheme in this similarity space. Specifically, from the set of all possible pairs of stimuli (the number of possible pairs for k = 31 stimuli is 465), a single pair is selected at a time. Model similarity-codes are obtained for each stimulus in the pair by extracting the relevant column vectors for those stimuli from the model similarity-matrix. In the same way, neural similarity-codes are extracted from the neural similarity-matrix. Correlations with the stimuli pairs themselves are removed to not bias decoding. The model similarity-codes of the two held-out stimuli are correlated with their respective neural similarity-codes. If the correct labeling scheme produces a higher sum of correlation coefficients than the incorrect labeling scheme, this is counted as a correct classification, and otherwise as incorrect. When this procedure is completed for all possible held-out pairs, the number of correct classifications over the total number of possible pairs yields a decoding accuracy. We perform group-level similarity-decoding by first averaging the neural similarity-codes across participants to yield group-level neural similarity-codes across participants to yield group-level neural similarity-codes equivalent to a fixed-effects analysis as in Anderson et al. (2016). The group-level neural similarity-codes and model similarity-codes are then used to perform leave-two-out decoding as described above.

5.4 Statistical Significance

Statistical significance was carried out as in Anderson et al. (2016) using a non-parametric permutation test. The null-hypothesis is that there is no correspondence between the model-based similarity-codes and the group-level neural similarity codes. The null-distribution was estimated using a permutation scheme. We randomly shuffled the rows and columns of the model-based similarity matrix, leaving the neural similarity-matrix fixed. Following each permutation (n = 10,000), we perform group-level similarity-decoding obtaining 10,000 decoding accuracies we would expect by chance using random labeling. The probability (p-value) of obtaining a decoding accuracy under the null-distribution is then at least as large as the observed accuracy score. We correct for the number of statistical tests performed using False-Discovery-Rate (FDR) with a corrected error probability threshold of $p = 0.05$ (Benjamini and Hochberg, 1995).

6 Experiments and Results

We use group-level similarity-decoding to decode brain activity associated with literal and metaphoric sentences using each of our semantic models.

We perform decoding at the sentence level for literal and metaphor conditions (affirmative only), separately. Decoding was performed at the whole-brain level and across the brain's lobes, as well as within a priori defined ROIs implicated in visual, action and language-related processing.

6.1 Linguistic Models

Literal sentences When decoding literal sentences with linguistic models across the brain's lobes we found significant decoding accuracies surviving FDR correction for multiple testing for the ADDITION and VERB models, see Table 1. A

MODELS	Frontal	Parietal	Temporal	Occipital	Whole-Brain
OBJECT	0.61	0.62	**0.68***	0.52	0.58
VERB	0.53	0.45	0.62	0.41	0.49
VERBOBJECT	**0.68***	0.53	**0.72***	0.67	0.60
ADDITION	**0.71***	0.63	**0.77***	**0.75***	**0.71***
LSTM	0.58	0.51	0.62	0.53	0.55
VISUAL OBJECT	0.59	0.59	0.60	0.42	0.56
VISUAL VERB	**0.70***	0.66	0.67	0.58	0.63
VISUAL VERBOBJECT	0.67	0.64	0.64	0.49	0.62
VISUAL ADDITION	0.59	0.62	0.60	0.57	0.58
VISUAL PHRASE	**0.70***	0.63	0.63	0.67	0.66

Table 2: *Metaphor Decoding*. Leave-2-out decoding accuracies, significant values ($p < 0.05$) surviving FDR correction for multiple comparisons are indicated in **bold** by an asterisk (*).

two-way ANOVA without replication showed a main effect for model $F(4,16) = 38.22$, $p < 0.001$ but not brain lobe. Post-hoc two-tailed t-tests surviving correction for multiple testing confirmed a significant advantage for the ADDITION model over other models. Lastly, the VERB model showed a significant decoding advantage over all other models except the ADDITION model. A post-hoc unpaired t-test confirmed a significant advantage for the VERB model in literal versus metaphor decoding ($t = 3.97$, $p < 0.01$, $df = 8$). The results suggest that the ADDITION and VERB models are superior compared to other models in decoding literal sentences. Furthermore, they suggest that the VERB model more closely captures the variance associated with the literal compared to metaphoric category.

Metaphoric Sentences When decoding metaphor with linguistic models we found significant decoding accuracies for the ADDITION, VERBOBJECT, and OBJECT models, mainly in the Temporal lobe, see Table 2. A two-way ANOVA showed a main effect of model $F(4,16) = 18.77$, $p < 0.001$, and brain lobe $F(4,16) = 7.58$, $p < 0.01$. Post-hoc t-tests showed a significant advantage for the ADDITION model over other models. We also found that the VERBOBJECT model significantly outperformed the LSTM ($t = 3.89$, $p < 0.05$, $df = 4$) and VERB ($t = 4.36$, $p < 0.05$, $df = 4$) models, while the OBJECT model also outperformed the VERB model ($t = 5.42$, $p < 0.01$, $df = 4$). Thus, models that incorporate the object directly (i.e., OBJECT, VERBOBJECT), outperform the VERB model. A post-hoc unpaired t-test confirmed that the performance of the OBJECT model was higher in metaphor versus literal decoding ($t = 2.88$, $p < 0.05$, $df = 8$). The results suggest that the ADDITION, VERBOBJECT, and OBJECT models are superior compared with other models in decoding metaphoric sentences and, furthermore, that the OBJECT model more closely captures the variance associated with the metaphor versus literal category.

Lastly, additional post-hoc t-tests showed that the Temporal lobe significantly outperformed other lobes (except the Occipital lobe) across the models. This suggests an advantage for linguistic models in temporal areas, possibly pointing to an increased dependence on memory and language processing associated with medial and lateral temporal areas, respectively.

6.2 Visual Models

Literal Sentences When decoding literal sentences with visual models, we found significant decoding accuracies for the VISUAL OBJECT and VISUAL ADDITION models, mainly in Occipital and Temporal lobes, see Table 1. A two-way ANOVA showed a main effect of model, but this did not survive multiple-testing correction. The results suggest that visual models can decode brain activity associated with concrete concepts only in occipital-temporal areas, part of the ventral visual stream, possibly pointing to increased reliance on these areas for object recognition, but see ROI analysis in section 6.4.

Metaphoric sentences When decoding metaphoric sentences with visual models in the brain, we found significant decoding accuracies for both the VISUAL VERB and VISUAL PHRASE model in the

	Visual		Action-related		Language-related	
MODELS	**LLOCT**	**LVT**	**LPG**	**LPM**	**LMTP**	**LIFG**
OBJECT	0.59	0.59	0.52	0.62	0.45	0.47
VERB	**0.73***	**0.72***	**0.68***	**0.77***	**0.69***	0.60
VERBOBJECT	0.63	0.61	0.59	**0.71***	0.47	0.51
ADDITION	**0.74***	**0.76***	**0.74***	**0.78***	**0.69***	**0.73***
LSTM	0.53	0.62	0.58	0.56	**0.68***	0.61
VISUAL OBJECT	**0.68***	**0.71***	0.50	0.42	**0.67***	0.57
VISUAL VERB	0.55	0.62	0.43	0.48	0.49	0.40
VISUAL VERBOBJECT	0.56	0.62	0.47	0.45	0.51	0.41
VISUAL ADDITION	0.48	**0.69***	0.56	0.42	0.65	0.49
VISUAL PHRASE	0.40	0.40	0.41	0.54	0.44	0.44

Table 3: *Region of Interest: Literal Decoding*. Leave-2-out decoding accuracies, significant values ($p > 0.05$) surviving FDR correction for multiple comparisons for the ROIs indicated in **bold** by an asterisk (*).

Frontal lobe, see Table 2. A two-way ANOVA showed a main effect of model $F(4,16) = 6.12$, $p < 0.01$ and brain lobe $F(4,16) = 5.21$, $p < 0.01$. Post-hoc t-tests showed that both the VISUAL VERB ($t = 5.40$, $p < 0.01$, $df = 4$) and VISUAL VERBOBJECT ($t = 8.49$, $p < 0.01$, $df = 4$) models outperformed the VISUAL OBJECT model across the lobes. This suggests that visual information about the verb is more relevant to metaphor decoding than that of the object. Relatedly, when comparing the performance of visual and linguistic models across the lobes, we found that the OBJECT model significantly outperformed the VISUAL OBJECT model across the lobes, surviving correction for multiple comparisons. In sum, these results suggest that visual information corresponds more strongly to the concrete verb, whereas linguistic information corresponds more strongly with the abstract object in metaphor decoding, but see ROI analysis section 6.3. We found a main effect of brain lobe that did not survive multiple-testing correction.

6.3 Region of Interest (ROI) analysis

Literal Sentences When comparing the performance of linguistic models across the ROIs, we found that the performance of linguistic models within language-related ROIs was on par with that within vision and action ROIs, see Table 3. This suggests that the linguistic models may be capturing sensorimotor and visual representations in the brain during literal sentence processing. Adding to this, we observed that linguistic models significantly outperformed visual models in action ROIs ($t = 6.83$, $p < 0.001$, $df = 9$), suggesting that the linguistic models are more closely able to capture the motoric features and action semantics relevant to literal sentence processing when compared even to the more visually grounded models. The results suggest that the visual models may correlate with information in action-related brain regions (e.g., sensorimotor representations). In sum, the results suggest that literal sentence processing involves both language-related and perceptual/sensorimotor representations (relevant to action semantics) in the brain.

Metaphoric Sentences When comparing the performance of linguistic models across the ROIs (see Table 4), we observed that linguistic models were superior in decoding metaphoric sentences in language-related ROIs compared to visual ($t = 3.11$, $p < 0.05$, $df = 9$) and action ROIs ($t = 2.97$, $p < 0.05$, $df = 9$). This suggests that linguistic models mainly capture language-related representations in the brain during metaphor processing. Interestingly, we did observe that the visual models significantly outperformed the linguistic models in action related ROIs ($t = 3.91$, $p < 0.01$, $df = 9$) for metaphor decoding. Relatedly, we also observed that visual models were superior in decoding metaphoric sentences in action compared with language-related ROIs ($t = 3.06$, $p < 0.05$, $df = 9$), in contrast to literal sentences as described above.

A post-hoc unpaired t-test confirmed that the performance of visual models in action ROIs

	Visual		Action-related		Language-Related	
MODELS	**LLOCT**	**LVT**	**LPG**	**LPM**	**LMTP**	**LIFG**
OBJECT	0.62	0.63	0.46	0.53	**0.67***	**0.67***
VERB	0.54	0.58	0.56	0.60	**0.75***	0.57
VERBOBJECT	0.63	0.62	0.49	0.54	**0.77***	0.64
ADDITION	**0.69***	0.65	0.51	0.59	**0.73***	**0.70***
LSTM	0.56	0.47	0.54	0.64	0.61	0.51
VISUAL OBJECT	0.63	0.56	**0.68***	0.64	**0.73***	0.52
VISUAL VERB	0.65	0.66	**0.69***	**0.75***	0.55	0.63
VISUAL VERBOBJECT	**0.70***	0.65	**0.68***	**0.74***	0.62	**0.67***
VISUAL ADDITION	**0.74***	0.57	**0.69***	0.63	0.62	0.59
VISUAL PHRASE	0.58	0.62	0.61	0.54	0.60	0.54

Table 4: *Region of Interest: Metaphor Decoding*. Leave-2-out decoding accuracies, significant values ($p < 0.05$) surviving FDR correction for multiple comparisons for the ROIs indicated in **bold** by an asterisk (*).

was significantly higher in metaphor versus literal decoding ($t = 8.92$, $p < 0.001$, $df = 18$). The results suggest that the visual models may correlate with information in action-related brain (e.g., sensorimotor representations). The significant values reported are those that survived correction for multiple comparisons in the ROI analysis.

7 Discussion

Addition vs. LSTM We found that the ADDITION model outperformed both lexical models and the VERBOBJECT model. This suggests that compositional semantic models that average semantic representations of the individual words in a phrase can decode brain activity associated with sentential meanings, irrespective of whether action-verbs are used in a literal or metaphoric context. The findings complement prior work showing that regression-based models that use word embeddings as features can predict brain activity associated with larger linguistic units (Wehbe et al., 2014; Huth et al., 2016; Pereira et al., 2018).

The LSTM, however, did not outperform the other models. This is surprising given prior work showing that contextual representations from an unsupervised LSTM language model outperform the bag-of-words model (Jain and Huth, 2018). The authors show increasing performance gains using representations from the second layer with longer context lengths (i.e., > 3 words). However, using the representations from the last layer together with a shorter context window sometimes showed inferior performance compared to the word-embedding encoding model. The latter finding is more closely aligned with our own parameters and findings. It is possible that the LSTM model shows the largest performance gain over the bag-of-words model when predicting brain activity associated with narrative listening (i.e., where the subject must keep track of entities and events over longer periods). In contrast, our sentence comprehension task depends on the next word for meaning disambiguation. It is also possible that semantic models trained in the NLI task may not be ideally suited for capturing differences in literal and metaphor processing.

The Role of the Verb and the Object We found that the VERB model outperformed the other models (except the ADDITION model) in literal decoding. In contrast, in metaphor decoding we observed that models that incorporate the object directly (i.e., VERBOBJECT and OBJECT models) outperformed the VERB model. Moreover, the performance of the VERB model was higher in literal versus metaphor decoding, while we found the opposite pattern in metaphor decoding where the OBJECT model had an advantage. It is possible that the VERB model more closely captures the variance associated with the overall concrete meaning in the brain. In support of this, the performance of the linguistic models including that of the VERB model was higher in action-related brain regions in literal compared to metaphoric decoding. On the other hand, the OBJECT model may best capture the variance associated with the overall abstract meaning in the brain. The objects (topic) in

metaphoric sentences tend to be more abstract and capture the overall aboutness of the metaphoric meaning to a greater extent than the verb (vehicle). In support of this, in metaphor decoding the linguistic models exhibited a higher performance in language-related areas than within visual and action-related areas. Critically, we restricted analysis to voxels showing maximum variance between the univariate brain response of literal and metaphoric categories. Thus, the results mainly highlight models that can decode literal and metaphoric sentences to the extent that they are able to identify the largest differences between literal and metaphor processing in the brain, more generally. Therefore, the results do not necessarily suggest that the VERB model, for example, is not an adequate representation for metaphoric sentences, just that when distinguishing literal and metaphoric processing in the brain it more closely aligns with representations for literal sentences.

Alternatively, the VERB model may be superior in capturing the variance associated with the literal case, in particular compared to the OBJECT model, as the verbs were found to be significantly more frequent than their arguments for literal sentences in the training corpus. However, we also found that the metaphoric uses of the verbs are significantly more frequent than the literal uses in the training corpus likely due to the fact that written language often reflects more abstract topics. However, we found that the VERB model showed higher performance in literal compared to metaphoric decoding suggesting that frequency of usage in the corpus does not always impact decoding as might be expected. Importantly, the literal and metaphorical sentences did not differ in familiarity (i.e., subjective frequency) nor did we find significant differences in the cloze probabilities between the literal and metaphoric phrases in the training corpus suggesting this broader factor is not at play.

Taken together, the results with the linguistic models suggest that one of the main ways lexical-semantic similarity differs in literal versus metaphor processing in the brain is along the concrete versus abstract dimension, as we might expect. The results are in line with prior neuroscientific studies showing that concrete concepts recruit more sensorimotor areas, whereas abstract concepts rely more heavily on language-related brain regions (Hoffman et al., 2015). More specifically, the findings are in agreement with the idea that action-related words and sentences are embedded in action-perception circuits in the brain due to co-occurrences between the words and the action-percepts they denote (Pulvermuller, 2005). However, the extent to which action-perception circuits are recruited may be modulated by the linguistic context (Desai et al., 2013).

These results also shed light on possible factors underlying the performance advantage we observed for the ADDITION model over the lexical models (and VERBOBJECT model). The ADDITION model enhances common features present in the individual word embeddings of the verb and the object. Therefore, given the preference we observed for the VERB over the OBJECT in literal decoding (and vice versa for metaphor decoding), this suggests that adding the complimentary embedding largely enhances lexical-semantic relations already present in either the VERB or OBJECT alone rather than provide other significant dimensions of variance, per se. For literal decoding, the OBJECT may enhance variance already associated with the VERB by narrowing the range of relevant object-directed actions (e.g., actions on inanimate versus animate objects) highlighting more concrete information. In contrast, for metaphor decoding it is more likely that the VERB enhances variance associated with the OBJECT by narrowing in on abstract uses as opposed to literal uses of each object (e.g., ''writing the poem'' versus ''grasping the poem''), highlighting more abstract information in the process. It should be noted that this effect may be due to the fact that we used familiar metaphors well represented in the corpus, which will need to be investigated in future work.

Visual Models We observed that the VISUAL OBJECT and VISUAL ADDITION models performed well in temporal-occipital areas. These results are in line with prior work showing that visual models can decode brain activity associated with concrete concepts in lateral occipital-temporal areas part of the ventral visual stream implicated in object recognition (Anderson et al., 2015). However, this was not specific to literal decoding. In fact, we observed that the VISUAL VERB and VISUAL VERBOBJECT models outperformed the VISUAL OBJECT model in metaphor decoding. Overall, we found that the visual models outperformed linguistic models in action-related ROIs in metaphor decoding. The performance of visual models in

action ROIs was also significantly higher in metaphor versus literal decoding. The latter suggests that the visual models correlate with sensorimotor features and may play a role in metaphor processing in the brain. This could possibly suggest that different aspects of the literal meaning of the verb (distinct from its prototypical or salient literal use) may play a role in metaphor processing in the brain. These less salient motoric aspects of the literal meaning captured by the visual verb models could reflect (a) more abstract sensorimotor representations such as information about higher-level action goals or (b) social-emotional factors associated with each action, such as information about people, bodies, or faces tied to interoceptive experience.

It could also be the case that these aspects of the literal meaning are not necessarily less salient or prototypical, but are simply distinct from the specific literal uses of verbs in our stimuli (which contained primarily verb predicates with inanimate objects as arguments). It is possible that verb predicates with *animate* objects as arguments involving social interactions may also be relevant to the metaphoric meaning. Indeed, an important embodied dimension of variance for abstract concepts is social-emotional information (Barsalou, 2009).

Additionally, it is possible that differences in overall visual statistics between our images for objects versus verbs across literal and metaphorical sentences may have biased decoding. Kiela et al. (2014) show that images for concrete objects are more internally homogenous (less dispersed) than that for abstract concepts, which may have impacted the performance of the Visual Object model in metaphor decoding. Importantly, however, differences in literal and metaphor decoding with the Visual Verb model should not necessarily be impacted by this as the verbs used were the same. Therefore, the fact that the visual models in action-related areas overall had higher decoding accuracies in metaphor compared to literal decoding suggests that this effect is not influenced by image dispersion. Rather this effect suggests that the Visual Verb may capture sensorimotor features relevant to metaphor decoding. Future studies will need to more carefully consider these possible confounding factors and possibly experiment with video data in place of images.

Accessibilty of the Literal Meaning When only looking at the linguistic models, the results appear largely in line with the direct view or a categorical processing of familiar metaphor in which the literal meaning is not fully accessible. The Verb model showed a clear advantage in literal compared to metaphor decoding. Moreover, the Verb model showed significant decoding accuracies in motor areas only in the literal but not metaphoric case, suggesting that the literal meaning is not being fully simulated in the metaphoric case. This aligns with neuroimaging work showing that literal versus familiar metaphoric actions more reliably activate motor areas (Desai et al., 2011).

Importantly, however, we found evidence that the Verb model showed some significant decoding accuracies for metaphor decoding in language-related brain regions (e.g., LMTP). Future work will need to determine whether this reflects distinct aspects of the literal meaning relevant to metaphor processing or reflects lexico-semantic information associated primarily with the more abstract sense of the verb. Adding to this, the poor temporal resolution of fMRI does not permit looking at different temporal processing stages and, therefore, cannot rule out the idea that the literal meaning is initially fully accessed and, subsequently, (partially) discarded or suppressed.

We also found further evidence to suggest that the linguistic context may modulate which representations associated with the verb are most accessible. Mainly, we found that visual models including the Visual Verb model were superior in decoding metaphoric versus literal sentences in action-related brain areas. This suggests that different aspects of the literal meaning (possibly less salient or prototypical literal meanings) may play a role in processing the metaphoric meaning. Thus, while the results do not definitively adjudicate between different putative stages of metaphor processing, they, nevertheless, inform our understanding of the debate in that they suggest that future studies will need to consider (control for) contextual effects of literality and their role in the study of metaphor comprehension. For instance, it may be useful to present subjects in the scanner with single words (grasp, push, etc.) to assess a prototypical brain response and then look at how different contexts (literal or metaphorical) modulate that response over time. This may reveal different kinds of processing

stages and the influence of bottom up (immediate and automatic) versus top down (context and inference-driven) influences at play during literal versus metaphor processing. This would permit more carefully assessing the role of the literal meaning in metaphor comprehension.

8 Conclusion and Future Directions

We presented the first study evaluating a range of semantic models in their ability to decode brain activity when reading literal and metaphoric sentences. We found evidence to suggest that compositional models can decode sentences irrespective of figurativeness in the brain and that at least for the linguistic models the VERB model may be more closely associated with the literal (concrete) meaning and the OBJECT model more closely associated with the metaphoric (abstract) meaning. This includes a closer relationship between the VERB model and action-related brain regions in the brain during literal sentence processing, in line with neuroimaging work showing that literal versus familiar metaphoric actions more reliably activate sensorimotor areas. This adds support to the idea that the literal meaning may not be as accessible for familiar metaphors. Taken together, the linguistic model results are in line with prior neuroscientific studies suggesting that differences between literal and metaphoric sentence processing align with concrete versus abstract concept processing in the brain, mainly with a greater reliance of concrete concepts on sensorimotor areas, while abstract concepts rely more heavily on language-related brain regions. Interestingly, however, the results with the visual models point to the need to also consider how metaphor (abstract language) may be grounded in more abstract knowledge about actions or social-interaction.

Future studies will need to further investigate the accessibility of the literal meaning (and abstract meaning) in metaphor comprehension using a larger dataset. For example, by considering a wider range of metaphors (e.g., metaphoric uses of objects) representing different semantic domains and different degrees of ambiguity. Also, it may be useful to consider event embeddings optimized towards learning representations of events and their thematic roles that may be better able to deal with different verb senses by learning non-linear compositions of predicates and their arguments (Tilk et al., 2016).

References

Andrew J. Anderson, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Mario Aguilar, Xixi Wang, Donias Doko, and Rajeev D.S. Raizada. 2017a. Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cerebral Cortex*, 27(9):4379–4395.

Andrew J. Anderson, Elia Bruni, Ulisse Bordignon, Massimo Poesio, and Marco Baroni. 2013. Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1960–1970. Seattle, Washington, USA. Association for Computational Linguistics.

Andrew J. Anderson, Elia Bruni, Alessandro Lopopolo, Massimo Poesio, and Marco Baroni. 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.

Andrew J. Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. 2017b. 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.

Andrew J. Anderson, Edmund C. Lalor, Feng Lin, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Rajeev D. S. Raizada, Scott Grimm, and Xixi Wang. 2019. Multiple regions of a cortical network commonly encode the meaning of words in multiple grammatical positions of read sentences. *Cerebral Cortex*, 29(6):2396–2411.

Andrew J. Anderson, Benjamin D. Zinszer, and Rajeev D.S. Raizada. 2016. Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity

using stimulus-model-similarities. *NeuroImage*, 128:44–53.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Valentina Bambini, Chiara Bertini, Walter Schaeken, Alessandra Stella, and Francesco Di Russo. 2016. Disentangling metaphor from context: An ERP study. *Frontiers in Psychology*, 7:559.

Lawrence W. Barsalou. 2008. Grounded cognition. *Annual Review of Psychology*, 59(1):617–645. PMID: 17705682,

Lawrence W. Barsalou. 2009. Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 364(1521):1281–89.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Jeffrey R. Binder, Rutvik H. Desai, William W. Graves, and Lisa L. Conant. 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12):2767–96.

Jeffrey R. Binder, Chris F. Westbury, Edward T. Possing, Kristen A. McKiernan, and David A. Medler. 2005. Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience*, 17(6):905–17.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *CoRR*, abs/1508.05326.

Samuel R. Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning, and Christopher Potts. 2016, aug. A fast unified model for parsing and sentence understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1466–1477. Berlin, Germany. Association for Computational Linguistics.

Lior Bugatus, Kevin S. Weiner, and Kalanit Grill-Spector. 2017. Task alters category representations in prefrontal but not high-level visual cortex. *Neuroimage*, 155437–449.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017, September. Speaking, seeing, understanding: Correlating semantic models with conceptual representation in the brain. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102. Copenhagen, Denmark. Association for Computational Linguistics.

Francesca Carota, Nikolaus Kriegeskorte, Hamed Nili, and Friedemann Pulvermuller. 2017. Representational similarity mapping of distributional semantics in left inferior frontal, middle temporal, and motor cortex. *Cerebral Cortex*, 27(1):294–309.

Kai-min Kevin Chang, Tom Mitchell, and Marcel Adam Just. 2010. Quantitative modeling of the neural representation of objects: How semantic feature norms can account for fMRI activation. *Neuroimage: Special Issue on Multi-variate Deciding and Brain Reading*, 56(2):716–727.

Rutvik H. Desai, Jeffrey R. Binder, Lisa L. Conant, Quintino R. Mano, and Mark S. Seidenberg. 2011. The neural career of sensory-motor metaphors. *Journal of Cognitive Neuroscience*, 23(9):2376–86.

Rutvik H. Desai, Lisa L. Conant, Jeffrey R. Binder, Haeil Park, and Mark S. Seidenberg. 2013. A piece of the action: Modulation of sensory-motor regions by action idioms and metaphors. *NeuroImage*, 83:862–69.

Barry Devereux, Colin Kelly, and Anna Korhonen. 2010. Using fMRI activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, pages 70–78. Los Angeles, USA. Association for Computational Linguistics.

Barry J. Devereux, Lorraine K. Tyler, Jeroen Geertzen, and Billi Randall. 2014. The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, 46(4):1–9.

Vesna G. Djokic, Ekaterina Shutova, Elisabeth Wehling, Benjamin Bergen, and Lisa Aziz-Zadeh. forthcoming. Affirmation and negation of metaphorical actions in the brain.

Evalina Fedorenko, Alfonso Nieto-Castanon, and Nancy Kanwisher. 2012. Lexical and syntactic representations in the brain: An fMRI investigation with multivoxel pattern analyses. *Neuropsychologia*, 4(50):499–513.

Evelina Fedorenko, Michael K. Behra, and Nancy Kanwisher. 2011. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, 108(39):16428–33.

Leonardo Fernandino, Colin J. Humphries, Mark S. Seidenberg, William L. Gross, Lisa L. Conant, and Jeffrey R. Binder. 2015. Predicting brain activation patterns associated with individual lexical concepts based on five sensory-motor attributes. *Neuropsychologia*, 76:17–26.

Dedre Gentner and Brian F. Bowdle. 2005. The career of metaphor. *Psychological Review*, 112(1):193–216.

Sam Glucksberg. 2003. The psycholinguistics of metaphor. *Trends in Cognitive Sciences*, 2(7):92–96.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Paul Hoffman, Richard J. Binney, and Lambon Ralph Matthew A. 2015. Differing contributions of inferior prefrontal and anterior temporal cortex to concrete and abstract conceptual knowledge. *Cortex*, 63:250–66.

Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frederic E. Theunissen, and Jack L. Gallant. 2016. Natural speech reveals the semantic maps that title human cerebral cortex. *Nature*, 532(7600):453–458.

Shailee Jain and Alexander Huth. 2018, Incorporating context into language encoding models for fmri, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6628–6637, Curran Associates, Inc.

Marcel A. Just, Jing Wang, and Vladimir L. Cherkassy. 2017. Neural representations of the concepts in simple sentences: Concept activation prediction and context effects. *Neuroimage*, 157:511–520.

David Kemmerer, Javier G. Castillo, Thomas Talavage, Stephanie Patterson, and Cynthia Wiley. 2008. Neuroanatomical distribution of five semantic components of verbs evidence from fmri. *Brain Language*, 107(1):16–43.

Douwe Kiela. 2016. MMFeat: A toolkit for extracting multi-modal features. In *Proceedings of ACL-2016 System Demonstrations*, pages 55–60, Berlin, Germany. Association for Computational Linguistics.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 835–841. Baltimore, Maryland. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

George Lakoff. 1980. *Metaphors We Live By*, University of Chicago Press, Chicago.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv, abs/1301.3781v3

Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33. Berlin, Germany. Association for Computational Linguistics.

Allan Paivio. 1971. *Imagery and Verbal Processes*, Holt, Rinehart, & Winston, New York.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Doha, Qatar. Association for Computational Linguistics.

Francisco Pereira, Matthew Botvinick, and Greg Detre. 2013. Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial Intelligence*, 194:240–252.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9:963.

Friedemann Pulvermuller. 2005. Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, 6:576–582.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252.

David S. Sabsevitz, David A. Medler, Michael Seidenberg, and Jeffrey R. Binder. 2005. Modulation of the semantic system by word imageability. *NeuroImage*, 27(1):188–200.

Ottokar Tilk, Vera Demberg, Asad Sayeed, Dietrich Klakow, and Stefan Thater. 2016. Event participant modelling with neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 171–182. Austin, Texas. Association for Computational Linguistics.

Jing Wang, Vladimir L. Cherkassky, and Marcel A. Just. 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.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS ONE*, 9(1):e112575.

Yangwen Xu, Qixiang Lin, Zaizhu Han, Yong He, and Yanchao Bi. 2016. Intrinsic functional network architecture of human semantic processing: Modules and hubs. *NeuroImage*, 132:542–55.