Modelling brain activity associated with metaphor processing with distributional semantic models

Vesna G. Djokic (vesna@imsquared.eu)
ILLC, University of Amsterdam, The Netherlands

Ekaterina Shutova (e.shutova@uva.nl)
ILLC, University of Amsterdam, The Netherlands

Abstract

In this study we investigate how lexical-semantic relations associated with the literal meaning (and abstract meaning) are being accessed across the brain during familiar metaphor comprehension. We utilize a data-driven whole-brain searchlight similarity-decoding analysis. We contrast decoding metaphoric phrases ("she's *grasping* the idea") using distributional semantic models of the verb in the phrase (VERB model) versus that of the more abstract verb-sense (PARAPHRASE VERB model) obtained from literal paraphrases of the metaphoric phrases ("she's *understanding* the idea"). We showed successful decoding with the VERB model across frontal, temporal and parietal lobes mainly within areas of the language and default-mode networks. In contrast, decoding with the PARAPHRASE VERB model was restricted to frontal-temporal lobes within areas of the language-network which overlapped to some extent with significant decoding with the VERB model. Overall, the results suggest that lexical-semantic relations closely associated with the abstract meaning in metaphor processing are largely localized to language and amodal (multimodal) semantic memory systems of the brain, while those more associated with the literal meaning are processed across a distributed semantic network including areas implicated in mental imagery and social-cognition.

Keywords: metaphor; abstraction; distributional semantics

Introduction

Metaphor comprehension involves a mapping from one domain of experience onto another and draws on inferential processes in order to derive the speaker's intended meaning. This could involve linking language-based semantic interpretation with memory images or sensorimotor simulations (Lakoff & Johnson, 1980). Nevertheless, how semantic information is selectively accessed and processed during metaphor comprehension remains to be fully elucidated. Specifically, how lexical-semantic relations associated with the literal meaning (and abstract meaning) are accessed across the brain during metaphor comprehension is not well understood. In this paper we leverage distributional semantic information (or co-occurrence patterns) of verbs to map how lexical-semantic relations associated with both literal and metaphoric uses of verbs correlate with activity across the brain during the processing of metaphoric actions (e.g. "She's *grasping* the idea").

Distributional semantics defines the meaning of words and phrases as a function of the linguistic contexts in which they are used in large text corpora (Turney & Pantel, 2010). The resulting vector-based representations have been successfully applied in a number of complex language-based tasks (e.g., language translation and inference) and, recently, they have been leveraged to further understand semantic representation and processing in the brain. For instance, a number of studies show that distributional semantic models can decode (and predict) neural activity associated with the processing of words (Mitchell et al., 2008), sentences (Pereira et al., 2018), but also larger narrative texts (Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016). However, it is still far from clear how these models achieve semantic composition nor how distributional semantics can inform our understanding of neural processes associated with the building of sentence meaning from individual words and phrases, a central feature of language.

Metaphor comprehension may prove to be a powerful test case for bridging distributional semantic models of language understanding with neurocognitive work on semantic processing in the brain. Metaphor typically involves the construal of a more abstract concept in terms of a concrete one, (e.g., "She's *pushing* the agenda"), however, the accessibility and processing of the literal meaning during metaphor comprehension is debated. In the indirect view (Grice, 1975; Searle, 1979), the metaphoric meaning is accessed indirectly by first processing the literal meaning (or "context-free" meaning). In contrast, proponents of the direct view (Gibbs, 1994; Glucksberg & Keysar, 1990) argue that the metaphoric meaning may be immediately available and the literal meaning largely by-passed.

Still, others propose views somewhere in-between direct and indirect views in which factors such saliency (Giora, 1997) or familiarity (Gentner & Bowdle, 2005) may dictate the extent that the literal meaning is processed initially or in-parallel to the abstract meaning. For example, novel metaphors may involve the direct comparison of elements in the source and target domains such that structural alignment of higher-order relations between domains can be discovered (Gentner & Clement, 1988). However, as metaphors become familiar through repeated usage they can be processed directly as categorizations as their meanings become increasingly lexicalized.

Taking an embodied perspective, conceptual metaphor theory finds that metaphor processing may depend on pre-stored conceptual mappings that are learned throughout the course of experience (Lakoff & Johnson, 1980). In this view concrete meanings play a direct role in structuring the more abstract concepts. However, the extent to which language users need to access more grounded representations is debated. Taken together, the extent to which aspects of the literal meaning (lexical-semantic or more embodied representations) are ac-

Object	Sentence
The scientific idea	She's *grasping* the idea
The biology concept	He's *grasping* the concept
The poem's meaning	She's *grasping* the meaning
The math topic	He's *grasping* the topic

Figure 1: Sample stimuli for the verb *grasp*

cessed and processed across the brain during the construction of the metaphoric meaning remains largely underspecified.

In this study, we investigate how distributional information associated with literal compared to the more abstract meaning can be mapped across the brain during metaphor comprehension. Specifically, we look at decoding metaphoric uses of verbs ("She's grasping the idea") in the brain using distributional semantic models of the verb ("grasp") and also that of the more abstract meaning ("understand") obtained from the literal paraphrases of the metaphoric phrases ("She's understanding the idea"). To do this, we use a data-driven whole-brain searchlight similarity-decoding analysis. We assume that lexical-semantic relations associated with the more abstract verb obtained from the literal paraphrase of the metaphor should closely overlap with the relations associated with the abstract meaning of the metaphorically used verbs. We find that the abstract meaning in metaphor processing is largely localized to the language-network and areas implicated in amodal (multimodal) semantic memory, while the literal meaning is processed across a more distributed semantic network including within the default-mode network implicated in mental imagery, episodic memory retrieval, and social cognition.

Materials

fMRI data

We re-analyzed the data collected from the fMRI experiment of (Djokic, Wehling, & Aziz-Zadeh, in press), who investigated force-dynamics of metaphoric uses of hand-action verbs in the brain. Below we provide an overview of the relevant information for our specific study.

Participants 10 right-handed, native English speakers (age range 18-25, 4 females and 6 males) with no history or neurological illness and normal vision participated in the study.

Stimuli Stimuli consisted of 120 metaphoric sentences. A total of 30 unique hand-action verbs were used to create all sentences each repeated four times. All sentences were in the 3rd person singular, present tense, progressive, see Figure 1. Stimuli were normed for length and familiarity.

Experimental Paradigm The fMRI task was divided into 4 runs each lasting 8.5 minutes. Participants saw a total of 120 metaphoric sentences and 8 catch trials. During each run participants saw 30 metaphoric sentences and 2 catch trials. Sentences containing the same verb were presented once in each run. The object of the sentence was presented on the screen for 1.8 seconds followed by an inter-trial interval of 0.2 seconds and then the sentence was presented for 6 seconds followed by a rest period of 8 seconds in an event-related design. During each rest period participants simply fixated a cross on a gray screen. All the stimuli were presented on a computer screen using Matlab [1]. Stimulus presentation was randomized across subjects.

fMRI data acquisition A Siemens MAGNETOM Trio 3T System was used to acquire fMRI images using a 32-channel head matrix coil. The T1-weighted structural anatomical scans (MPRAGE) were obtained using a T1-weighted magnetization prepared rapid gradient echo protocol with TR=1950 ms, TE=2.26 ms, flip angle of 10 degrees, 208 coronal slices, and resolution of 1mm with 256 x 256 mm matrix. The functional images (37 contiguous axial slices) were obtained using a T2* weighted single-shot gradient-recalled echo-planar sequence (EPI) using blood oxygenation-level-dependent contrast in interleaved mode with TR=2000 ms, TE=30 ms, flip angle 90 degrees and 3.5 mm resolution with 64 x 64 mm matrix.

Semantic models

For all of our linguistic models we used the GloVe word embeddings (Pennington, Socher, & Manning, 2014). We use the 100-dimensional word vectors, which were trained by the authors of the study on Wikipedia and the Gigaword corpus.

VERB In the VERB model the stimulus phrases are represented as the individual *D*-dimensional word embeddings for their verb.

PARAPHRASE VERB In the PARAPHRASE VERB model the stimulus phrases are represented as the individual *D*-dimensional word embeddings for the verb obtained from their literal paraphrase.

Literal paraphrases of the metaphor were created by the authors of this study and normed for familiarity by an independent set of participants in a separate behavioral experiment.

Methods

fMRI data preprocessing and response estimation

All preprocessing and statistical analysis were carried out using FSL and PyMVPA [2]. All runs were concatenated to form the language dataset for each subject. Functional data were co-registered with the MPRAGE structural image of each subject. Preprocessing included slice timing correction, high-pass filtering (90 secs), and motion correction to the mean image using FSL's MCFLIRT. Following this, each dataset was linearly detrended. The response-amplitude (Beta values) were estimated using the General Linear Model (GLM) for each individual stimulus presentation in an event-related design. This gave voxel-wise Beta maps for each stimulus presentation that were then normalized to z-scores. We calculated neural verb estimates by averaging the voxel-wise z-score maps of all sentences containing the same verb. This gave thirty unique neural verb estimates. We used these neural verb estimates to

[1] Psychophysics toolbox 3, www.psychtoolbox.org

[2] PyMVPA 0.6, http://www.pymvpa.org/

perform similarity-decoding across the whole-brain using the VERB and PARAPHRASE VERB models, separately.

Similarity-decoding

We used similarity-decoding (Anderson, Zinszer, & Raizada, 2016), an extension of representational similarity analysis (RSA) (Kriegeskorte, Marieke, & Peter, 2008). We perform similarity-decoding separately for the VERB model and the PARAPHRASE VERB model. As in RSA, similarity-decoding involves comparing neural and model similarity matrices (Pearson's correlation is used as a distance metric). Neural similarity matrices are created by calculating all pairwise similarities using neural estimates of each verb, while the model similarity matrices are created by calculating all pairwise similarities using the word embedding vectors of each verb (either from the VERB model or PARAPHRASE VERB model). Neural and model similarity-matrices are then compared by performing leave-two-out decoding allowing the classification of individual terms. Specifically, a single pair of verbs is selected at a time out of the total number of possible verb pair combinations (k = 30 verbs is 435). Each verb is represented as a vector of its pairwise similarities with all other verbs (similarities with the pairs themselves are removed to not bias decoding). Neural and model similarity verb vectors are thus extracted from the respective columns of the neural and model similarity matrices. The neural and model similarity vectors of the pair of verbs are then correlated using the correct labeling assignment, but also the incorrect labeling assignment. If the sum of correlation coefficients with the correct labeling is higher than with the incorrect label assignment, this is counted as a correct classification and incorrect otherwise. The final decoding accuracy is calculated as the number of correct classifications over the total number of possible pairs.

Whole-brain searchlight analysis We performed similarity-decoding using a searchlight analysis across the brain. Specifically, this involves placing a sphere with a 4 voxel radius (a cluster of 257 voxels) centered on each voxel of the brain in each subject's native space and performing similarity-decoding within this region of interest. This gives a classification accuracy score for each voxel across the brain.

Statistical Significance Statistical evaluation was performed using non-parametric cluster-thresholding of searchlight-based group analysis (Stelzer, Chen, & Turner, 2013) using PyMVPA. This involves performing within-subject permutations by shuffling class labels in order to obtain 100 random accuracy maps per subject. A bootstrap procedure is then used to calculate 10000 random group-average accuracy maps. The 10000 thresholded group-average accuracy maps (uncorrected voxel threshold of $p < 0.001$) give a null distribution of chance cluster sizes. Statistically significant searchlight-based group results are reported at an uncorrected cluster forming threshold of $p < 0.001$ and a family-wise error rate for multiple comparison correction of cluster size probabilities $p < 0.05$ using FDR correction.

Results

VERB model:

We performed a whole-brain similarity-decoding searchlight analysis to localize areas of the brain that can significantly decode metaphoric uses of verbs using the VERB model, see Table 1 and Figure 2.

In the left hemisphere significant clusters were found across frontal, parietal, and temporal lobes. We found three frontal clusters. The first frontal cluster was found in the left dorsal medial prefrontal cortex with peak decoding accuracy in the left frontal pole extending to the superior frontal gyrus (SFG) and paracingulate cortex. A second frontal cluster was found in left ventral medial prefrontal cortex, with peak decoding accuracy localized to the left lateral orbital frontal cortex (OFC) extending to the frontal pole and temporal pole. The third frontal cluster showed a peak decoding accuracy in the left precentral gyrus extending to middle frontal gyrus (MFG) and inferior frontal gyrus (IFG), pars opercularis.

Two clusters were found in the left parietal lobe. This included a cluster with a peak decoding accuracy localized to the left medial superior parietal lobule (SPL), mainly the left precuneus extending to the posterior cingulate cortex (PCC). A second parietal cluster was localized to areas of the left inferior parietal lobule (IPL), mainly the left angular gyrus (AG) extending to the posterior supramarginal gyrus (SMG) and superior lateral occipital cortex (LOC). Lastly, a small cluster was also found in the left posterior superior temporal gyrus (STG) within the left temporal lobe.

In the right hemisphere of the brain we found significant decoding in frontal and temporal lobes. This included a large cluster within right dorsolateral prefrontal cortex (DLPFC) with a peak decoding accuracy localized to the right frontal pole extending to the MFG and IFG, pars triangularis. A smaller cluster was also found in more midline regions with a peak decoding accuracy localized to the right supplementary motor cortex (SMA) extending to the anterior cingulate cortex (ACC). Lastly, a small cluster was also found in the right anterior temporal lobe with peak decoding accuracy localized to the right temporal fusiform cortex and extending to the anterior parahippocampal gyrus and temporal pole.

PARAPHRASE VERB model:

We performed a whole-brain similarity-decoding searchlight analysis to localize areas of the brain that can significantly decode metaphoric uses of verbs using the PARAPHRASE VERB model, see Table 2 and Figure 2.

In the left hemisphere significant clusters were found across the frontal and temporal lobes. We found two frontal clusters. The first frontal cluster had peak decoding accuracy localized to the left IFG, pars opercularis extending to frontal/central operculum cortex, anterior insula, but also slightly to the IFG, pars triangularis. The second left frontal cluster was localized to the left MFG extending to the IFG, pars opercularis and slightly to the precentral gyrus. In the left temporal lobe a single cluster was found with peak decoding accuracy localized

to the left posterior middle temporal gyrus (MTG) extending to the posterior STG.

Lastly, in the right hemisphere significant clusters were found across the frontal and temporal lobes. In the right temporal lobe the largest cluster was found in the right anterior temporal lobe with a peak decoding accuracy within the the right anterior temporal fusiform cortex extending to the right posterior temporal fusiform cortex, anterior inferior temporal gyrus (ITG), anterior MTG, anterior parahippocampal gyrus, temporal pole, and also to areas of the ventromedial prefrontal cortex (VMPFC), mainly the medial OFC. In the right frontal lobe a smaller cluster was also found in ventrolateral prefrontal cortex (VLPFC) with peak decoding accuracy in the right frontal pole extending to the lateral OFC.

Overlap between VERB model and PARAPHRASE VERB model

We found brain regions showing significant clusters when decoding with both the VERB and PARAPHRASE VERB models across the whole brain, see Figure 2. Overlapping significant clusters were found in the left frontal lobe and right temporal lobe. Specifically, in the left frontal lobe we found overlapping clusters mainly in the left IFG, pars opercularis, but also to some extent in areas of left MFG, left precentral gyrus, frontal/central operculum, and anterior insula. In the right temporal lobe we found overlapping clusters in the right anterior temporal lobe mainly in the right temporal fusiform cortex and parahippocampal gyrus, but also extending slightly to the temporal pole.

Discussion

We used a whole-brain similarity-decoding searchlight analysis to investigate how lexical-semantic relations associated with the literal meaning (and abstract meaning) correlate with brain activity across the brain during familiar metaphor comprehension. In order to identify areas of the brain sensitive to lexical-semantic relations associated with action-verbs used in a literal versus metaphoric context we contrasted decoding metaphoric phrases using word embeddings of the VERB versus that of the abstract verb-sense or PARAPHRASE VERB obtained from literal paraphrases of the metaphoric phrases.

The results showed successful decoding with the VERB model predominantly across brain regions in the language and default-mode networks. In contrast, we found successful decoding with the PARAPHRASE VERB model largely across brain regions in the language-network. Xu, Lin, Han, He, and Bi (2016) provide evidence that the semantic system is comprised of at least three modules that work in concert during semantic processing, mainly (1) a perisylvian language network (PSN) associated with lexical-semantic processing, (2) default-mode network (DMN) implicated in memory-based simulation, (3) a frontal-parietal network (FPN) involved in semantic control. In this context, the results suggest that the abstract meaning in metaphor processing is largely localized to language and amodal (multimodal) semantic memory systems of the brain, while the literal meaning is processed across a more distributed semantic network including areas implicated in memory-based simulation or the re-enactment of conceptual knowledge drawing on mental imagery, episodic memory retrieval, and aspects of social cognition (Xu et al., 2016). Importantly, our results align with neuroscientific work showing that abstract concept processing depends to a larger extent on language-related brain regions (Hoffman, Binney, & Lambon Ralph, 2015), while concrete compared to abstract concepts show greater activation in the DMN (Binder, Westbury, Possing, McKiernan, & Medler, 2005).

When decoding with the VERB model we found significant clusters spanning left-lateralized classical language areas, including in areas of the left IFG (Broca's area), left MFG, left SFG, left posterior STG (Wernicke's area), and to some extent the left temporal pole. Recent work using more sensitive individual subject analysis (Fedorenko, Behra, & Kanwisher, 2011), additionally implicate the full extent of the left temporal lobe, the left lateral OFC and the left AG in language-based compared to non-language based tasks of similar cognitive effort (Fedorenko et al., 2011). To this extent, it is likely that the significant clusters we found in the left lateral OFC and left AG when decoding with the VERB model also reflect higher-level language processing.

Critically, Xu et al. (2016) suggest that the left AG actually may act as an integration hub linking information across the PSN, DMN, and the FPN during semantic processing, but with the left posterior AG more closely associated with the DMN. This aligns with recent work suggesting that the language and default-mode networks occur side by side as part of parallel distributed association networks that work together to accomplish higher-order cognition (Braga, DiNicola, & Buckner, 2019).

We found significant clusters within a number of brain regions typically associated with the DMN when decoding with VERB model. The DMN shows decreased activity during goal-oriented tasks and increased activity during the resting-state, and typically includes the bilateral AG, PCC, VMPFC, and DMPFC (Binder & Desai, 2011). However, Xu et al. (2016) also include areas of the superior LOC and areas of the fusiform and parahippocampal gyrus. We found a significant cluster with peak decoding accuracy in the left angular gyrus, extending posteriorly to the left superior LOC. We also found clusters in other areas of the DMN including the precuneus and PCC, VMPFC, and DMPFC. While areas of the left anterior AG, VMPFC and DMPFC have also been implicated in higher-level language processing (Fedorenko et al., 2011; Xu et al., 2016), the significant clusters in the left posterior AG, precuneus and PCC likely are more specific to memory-based simulation. The DMN has been associated with memory-based simulation relevant to a range of tasks including visual-spatial navigation, recall of autobiographical memories, as well as mentalizing (Xu et al., 2016).

When decoding with the VERB model we also found a cluster with peak decoding accuracy localized to areas of the left precentral gyrus. However, this was very close to the poste-

Anatomical Brain Region	Voxels	Coordinate (x,y,z)			Accuracy	p-value
L Frontal Pole	2144	-24	50	32	0.601	2.0×10^{-7}
R Frontal Pole	1513	30	42	36	0.605	3.9×10^{-7}
L Precentral Gyrus	1368	-44	2	34	0.600	5.9×10^{-7}
L Precuneus	396	-8	-38	46	0.587	7.8×10^{-7}
L Frontal Orbital Cortex/Frontal Pole	289	-36	38	-14	0.587	9.8×10^{-7}
L Angular Gyrus	270	-56	-54	26	0.585	1.2×10^{-6}
R Supplementary Motor Cortex	231	6	-12	48	0.576	1.4×10^{-6}
R Temporal Fusiform Cortex, anterior	203	34	0	-34	0.573	1.6×10^{-6}
L Superior Temporal Gyrus, posterior	68	-64	-34	16	0.575	1.8×10^{-6}

Table 1: Results of similarity-decoding searchlight with the VERB model. MNI coordinates and significance levels shown for the peak voxel in each cluster.

Anatomical Brain Region	Voxels	Coordinate (x,y,z)			Accuracy	p-value
R Temporal Fusiform Cortex, anterior	2271	34	2	-38	0.600	2.0×10^{-7}
L Inferior Frontal Gyrus, pars opercularis	527	-58	14	10	0.583	3.9×10^{-7}
L Middle Frontal Gyrus	445	-48	18	36	0.595	5.9×10^{-7}
R Frontal Pole	267	38	44	-16	0.594	7.8×10^{-7}
L Middle Temporal Gyrus, posterior	115	-66	-34	-8	0.577	9.8×10^{-7}

Table 2: Results of similarity-decoding searchlight with the PARAPHRASE VERB model. MNI coordinates and significance levels shown for the peak voxel in each cluster.

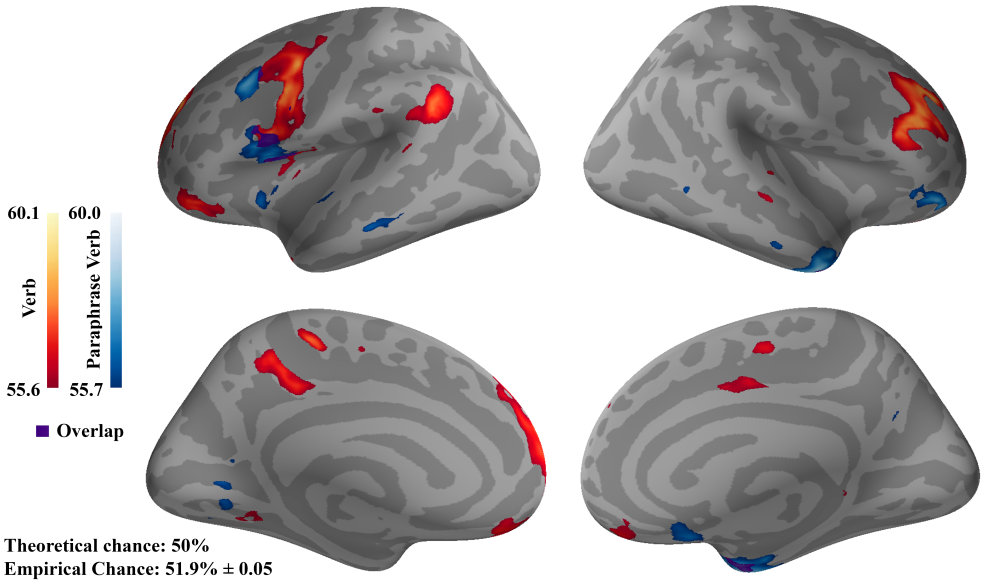

Figure 2: Significant decoding of metaphoric uses of lexical-verbs across the brain with the VERB model [red] and PARAPHRASE VERB model [blue]. Regions of overlap are show in purple. Results presented at ($p < 0.001$) uncorrected, ($p < 0.05$) FWE corrected cluster threshold.

rior areas of the left MFG suggesting that this cluster likely reflects language processing. Still, this cluster did extend to areas of the ventral premotor cortex that have also been implicated in motor imagery and action-related verb and sentence processing (Pulvermuller, 2005). Prior neuroscientific work suggests that processing of action-verbs (and sentences containing action-verbs) recruits both primary motor and premotor areas (Pulvermuller, 2005). Future work will need to more fully understand how lexical-semantic relations associated with the VERB model correlate with motor areas during the processing of literal versus metaphoric uses of action-verbs.

Lastly, we also found significant clusters when decoding with the VERB model in areas of the right anterior temporal lobe. Xu et al. (2016) suggest that the left anterior temporal lobe serves as the seat of amodal (multimodal) semantic memory and acts as a connector hub integrating information between the FPN and DMN. Given recent work showing that the bilateral anterior temporal lobes are implicated in the storage of amodal (and multimodal) semantic memory (Rice, Caswell, Moore, Hoffman, & Lambon Ralph, 2018), it is possible that the right anterior temporal lobe plays a similar function during metaphor comprehension. Critically, we also found significant clusters in the right DLPFC (also SMA and ACC) implicated in working memory and conflict monitoring, which will need to be further investigated in future work.

In contrast to the more distributed semantic network we observed for the VERB model, lexical-semantic relations associated with the PARAPHRASE VERB model correlated predominantly with left-lateralized frontal temporal regions associated with the classical language network. We found significant clusters across areas of the left IFG, left MFG, left SFG, left posterior MTG, and left STG. Interestingly, Xu et al. (2016) suggest that the left posterior MTG may link the PSN and FPN, possibly suggesting that the abstract meaning relies to a greater extent on the FPN, but this will also need to be further investigated. Importantly, we also found a significant cluster localized to areas of the right anterior temporal lobe that has been implicated in amodal (multimodal) conceptual knowledge storage. The results suggest that the abstract meaning in metaphor processing is largely relegated to areas of the brain involved in higher-level language processing.

Overlapping areas of significant decoding between the VERB model and PARAPHRASE VERB model occurred mainly within classical language-related areas and brain regions implicated in amodal (multimodal) conceptual knowledge storage. Specifically, overlap was found for clusters with peak decoding accuracies localized to the left IFG and the right anterior temporal lobe within the anterior temporal fusiform cortex and parahippocampal gyrus. Taken together, the results suggest that lexical-semantic relations more closely associated with the abstract meaning in metaphor processing are largely localized to language and amodal (multimodal) semantic memory systems of the brain, while those more associated with the literal meaning are processed across a more distributed semantic network including areas implicated in mental imagery and social-cognition. This suggests that aspects of the literal meaning are being processed, possibly those most relevant to social-emotional experience.

Critically, there are important limitations to the current study. We assume that the semantic neighborhood of the abstract verb-sense captured by the VERB PARAPHRASE model should be more closely associated with that of the abstract meaning (or metaphoric meaning) than the literal meaning. However, the VERB PARAPHRASE model likely only represents a subset of the lexical-semantic relations relevant to the abstract or metaphoric meaning. In the direct view, the nominal metaphor "my lawyer is a shark" is understood directly via the creation of an abstract superordinate category "predatory creatures" of which the vehicle "shark" is a prototypical member (Glucksberg & Keysar, 1990). Importantly, this ad-hoc category is not truly lexicalized (i.e., there is no linguistic phrase that captures its full semantic content) and, therefore, it may be argued that the abstract verb-sense from the literal paraphrase provides only a rough approximation to this ad-hoc category. The metaphoric meaning may best be captured as a conceptual blend between the source and target. Kintsch (2000) provide a computational model in line with the categorization view in which they model the metaphoric meaning as a conceptual blend of the source and target by focusing on the words at the intersection of their semantic neighborhoods. To this extent, future work will need to experiment with different compositional semantic models of metaphor that more explicitly model the interaction between the source and target and test these against word-level vectors.

In this study we sought to understand how distributional information associated with literal compared to the more abstract meaning can be mapped across the brain during metaphor comprehension. We found evidence to suggest that lexical-semantic relations more closely associated with the abstract meaning in metaphor processing are largely localized to language and amodal (multimodal) semantic memory systems of the brain, while those more associated with the literal meaning are processed across a more distributed semantic network including areas implicated in mental imagery and social cognition. Future work will need to more carefully model what aspects of the literal meaning are necessary to metaphor compared to literal sentence processing and how that information is being selected for and processed in the brain. Ultimately, understanding what features of the literal meaning are being filtered and how (either via inhibition or some other mechanism) during metaphor comprehension will allow us to discern the relative contribution of the literal meaning in the construction of the metaphoric meaning. Testing distinct processing models of metaphor in the brain that specify the interaction of source and target across time may further help adjudicate between different putative temporal stages involved in metaphor processing, such as processing of the literal meaning at an early stage (Weiland, Bambini, & Schumacher, 2014) or via a parallel processing route (Cartson, 2010). Finally, it has been suggested that a comprehensive theory of metaphor may

include combining processes associated with categorization (conceptual blending) with those involved in analogical reasoning such as structure mapping (Holyoak & Stamenkovic, 2018). Indeed, metaphor comprehension depends on the ability of language users to go beyond the literal meaning of the words given and this may draw on pragmatic inferencing or the ability to integrate lexical-semantics with context and world knowledge (Weiland et al., 2014). The present results point to the need to further understand the way the language-network may flexibly interact with the memory-based simulation system to accomplish this during metaphor comprehension.

References

Anderson, A. J., Zinszer, B. D., & Raizada, R. D. (2016). Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *NeuroImage*, *128*, 44–53.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–36.

Binder, J. R., Westbury, C. F., Possing, E. T., McKiernan, K. A., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience*, *17*(6), 905–17.

Braga, R. M., DiNicola, L. M., & Buckner, R. L. (2019). Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. *bioRxiv*.

Cartson, R. (2010). Metaphor: Ad hoc concepts, literal meaning, and mental images. *Proceedings of the Aristotelian Society*, *110*, 295-321.

Djokic, V. G., Wehling, E. E., & Aziz-Zadeh, L. (in press). Distinguishing metaphors that differ in their encoded force dynamics.

Fedorenko, E., Behra, M. K., & Kanwisher, N. (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.

Gentner, D., & Bowdle, B. F. (2005). The career of metaphor. *Psychological Review*, *112*(1), 193–216.

Gentner, D., & Clement, C. (1988). *Evidence for relational selectivity in the interpretation of analogy and metaphor* (Vol. 22). In G. H. Bower (Ed.), Advances in the psychology of learning and motivation. New York: Academic Press.

Gibbs, R. W. (1994). *The poetics of mind: Figurative thought, language and understanding*. Cambridge: Cambridge University Press.

Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis. *Cognitive Linguistics*, *8*, 183-206.

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, *97*, 3-18.

Grice, H. P. (1975). Logic and conversation. *Syntax Semant.*, *3*, 41-58.

Hoffman, P., Binney, R. J., & Lambon Ralph, M. A. (2015). Differing contributions of inferior prefrontal and anterior temporal cortex to concrete and abstract conceptual knowledge. *Cortex*, *63*, 250–66.

Holyoak, K., & Stamenkovic, D. (2018). Metaphor comprehension: A critical review of theories and evidence. *Psychological Bulletin*, *144*(6), 641-671.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin and Review*, *7*, 257-266.

Kriegeskorte, N., Marieke, M., & Peter, B. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Front. in Systems Neurosci.*, *2*(4), 4.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191–1195.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP* (pp. 1532–1543). ACL.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., ... Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*, 963.

Pulvermuller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, *6*, 576-582.

Rice, G. E., Caswell, H., Moore, P., Hoffman, P., & Lambon Ralph, M. A. (2018). The Roles of Left Versus Right Anterior Temporal Lobes in Semantic Memory: A Neuropsychological Comparison of Postsurgical Temporal Lobe Epilepsy Patients. *Cerebral Cortex*, *28*(4), 1487-1501.

Searle, J. R. (1979). *"Metaphor"*. In Metaphor and Thought, Andrew Ortony (Ed.). Cambridge: Cambridge University.

Stelzer, J., Chen, Y., & Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. *Neuroimage*, *65*, 69-82.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37*(1), 141–188.

Weiland, H., Bambini, V., & Schumacher, P. (2014). The role of literal meaning in figurative language comprehension: Evidence from masked priming ERP. *Front. Hum. Neurosci.*, *9*, 583.

Xu, Y., Lin, Q., Han, Z., He, Y., & Bi, Y. (2016). Intrinsic functional network architecture of human semantic processing: Modules and hubs. *NeuroImage*, *132*, 542–55.