Modeling affirmative and negated action processing in the brain with lexical and compositional semantic models

Vesna G. Djokic[♠] **Jean Maillard**[♣] **Luana Bulat**[♣] **Ekaterina Shutova**[♢]
[♠]Department of Neuroscience, University of Southern California, USA
[♣]Dept. of Computer Science & Technology, University of Cambridge, United Kingdom
[♢]ILLC, University of Amsterdam, The Netherlands
vesna@imsquared.eu, jean@maillard.it,
ltf24@cam.ac.uk, e.shutova@uva.nl

Abstract

Recent work shows that distributional semantic models can be used to decode patterns of brain activity associated with individual words and sentence meanings. However, it is yet unclear to what extent such models can be used to study and decode fMRI patterns associated with specific aspects of semantic composition such as the negation function. In this paper, we apply lexical and compositional semantic models to decode fMRI patterns associated with negated and affirmative sentences containing hand-action verbs. Our results show reduced decoding (correlation) of sentences where the verb is in the negated context, as compared to the affirmative one, within brain regions implicated in action-semantic processing. This supports behavioral and brain imaging studies, suggesting that negation involves reduced access to aspects of the affirmative mental representation. The results pave the way for testing alternate semantic models of negation against human semantic processing in the brain.

1 Introduction

Computational semantic models are increasingly being evaluated in their ability to capture aspects of human semantic processing, including similarity and association judgments (De Deyne et al., 2016) and semantic representation in the brain (Bulat et al., 2017). Prior work shows that distributional semantic models (DSMs) are able to decode functional magnetic resonance imaging (fMRI) patterns associated with the meaning of concrete words (Anderson et al., 2013). Relevant to our work, Carota et al. (2017) showed that the similarity structure of DSMs for action words correlates with that of fMRI patterns in brain regions implicated in action-semantic processing.

More recent studies have also investigated the ability of DSMs to predict fMRI patterns of sentential meanings (Pereira et al., 2018) and larger narrative text passages (Wehbe et al., 2014; Huth et al., 2016). They have shown that encoding models based on word embeddings are able to capture subtle aspects of sentence meaning in the brain, even when these models are oblivious of word order and syntactic structure. While promising, none of this research has so far systematically investigated specific semantic composition phenomena and processing at the syntax-semantic interface, such as that of the negation function.

Negation is a fundamental abstraction necessary for efficient reasoning and communication (Horn, 1989). Although it is typically marked syntactically, the semantics of negation in natural language usage has proven to be rather challenging to pinpoint (Speranza and Horn, 2010). In logical negation, the negation operator has been succinctly described as a truth-functional operation, reversing the truth value of a sentence. On the other hand, from a pragmatic point of view, the primary function of negation is to direct attention to an alternative meaning and can thus be, more generally, compared to our ability for counterfactual thinking (Hasson and Glucksberg, 2006). It is also often assumed that negation entails affirmation (as it is always positive by default), yet the extent to which the the affirmative situation need be processes is debated (Orenes et al., 2014). Despite the intuition that negated meanings are indeed quite distinct from their affirmative counterparts, there is still no comprehensive account of how the brain represents negated entities.

Neuroscientific studies on negation have predominantly focused on studying negation of action-related sentences and suggest that negation blocks access to aspects of the affirmative representation (Papeo et al., 2016). For exam-

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5155–5165
Florence, Italy, July 28 - August 2, 2019.

ple, negation of action-related sentences or imperatives involves decreased activity in motor systems of the brain implicated in action semantics when compared to the affirmative context (Tettamanti et al., 2008; Tomasino et al., 2010). However, overall reduced activation does not necessarily equate to a lack of information across patterns of activated or deactivated voxels in a brain region (Kriegeskorte et al., 2008). More importantly, the degree to which negation of action-related sentences impacts access to lexico-semantic representations and semantic similarity in the brain is not yet well understood. To contribute to our understanding of negation and its modeling, we investigate the extent to which lexical and compositional semantic models can decode fMRI patterns of negated and affirmative action sentences in the brain using similarity-based decoding (Anderson et al., 2016). We also test the extent to which the representational similarity structure (Kriegeskorte et al., 2008) of DSMs of action-verbs correlates with that of fMRI patterns associated with negated versus affirmative sentences containing hand-action verbs. We focus on motor areas and classical language-related brain regions implicated in action-semantic processing (e.g., understanding action words and sentences) (Pulvermuller, 2005; Kemmerer, 2015).

DSMs have proven successful in modeling aspects of semantic composition in the context of the natural language inference task (Bowman et al., 2015b). Although the modeling of logical negation using DSMs is wrought with challenges (Kruszewski et al., 2017), current state-of-the-art neural network based models appear to capture elements of markedness asymmetry in negation (Li et al., 2016) and, presumably, implicitly model negation at some level. In our experiments, we investigate the extent to which DSMs are able to decode (correlate with) fMRI patterns associated with the reading of sentences containing negated and affirmative action verbs. We experiment with (1) word-level representations of action verbs; and (2) compositional semantic models (based on addition of word-level representations and long short-term memory (LSTM) networks).

In agreement with previous work, our results show that distributional representations of action verbs (and to some extent verb-object phrases) show reduced decoding for negated versus affirmative action sentences. This is also reflected as a reduced correlation between the similarity structure of DSMs of action verbs and fMRI patterns of negated as compared to affirmative action sentences. Importantly, we show for the first time that negation impacts semantic similarity in motor areas, but also to some extent language-related brain regions. These findings lend further support to the hypothesis that negation may involve reduced access to aspects of the affirmative mental representation.

2 Related Work

Decoding brain activity Mitchell et al. (2008) were the first to show that DSMs based on co-occurrence counts with 25 sensorimotor verbs (e.g. see, hear, taste) can predict fMRI patterns associated with the meaning of concrete nouns. Later research has demonstrated that a range of DSMs can decode fMRI patterns of concrete nouns (Murphy et al., 2012; Anderson et al., 2013; Bulat et al., 2017) and, more recently, abstract nouns (Anderson et al., 2017). Most relevant to our study, Carota et al. (2017) showed that the similarity structure of a Latent Semantic Analysis (LSA) model for action words (nouns and verbs) correlates with that of fMRI patterns in motor areas (left precentral gyrus (LPG)) and classical language-related brain regions (left inferior frontal gyrus (LIFG), left posterior middle temporal gyurs (LMTP)) implicated in lexico-semantic processing (Binder et al., 2009).

Moving beyond words, other studies have shown that DSMs can predict brain activity patterns associated with larger linguistic units (Wehbe et al., 2014; Huth et al., 2016; Pereira et al., 2018). For example, Pereira et al. (2018) showed that a regression model mapping between fMRI patterns of words and their word embeddings could synthesize vector representations for novel sentences that correlate with the average of the word embeddings of the sentence. Working with larger text fragments, Wehbe et al. (2014) and Huth et al. (2016) have been able to predict neural activity associated with the processing of narratives in the brain using encoding models with word embeddings (also syntactic markers) as features. Although these findings suggest that DSMs are able to predict fMRI patterns associated with the processing of compositional meanings, they do not reveal to what extent the models capture specific compositional phenomena nor the specific impact

of linguistic context on semantic representation in the brain. Our work extends this line of research to study individual aspects of semantic composition, focusing on the negation function.

Modeling negation in NLP Kruszewski et al. (2017) contrast *logical negation*, which captures the idea of the complement of a set, with *conversational negation*, the phenomenon by which negation identifies a set of alternative plausible utterances: i.e., the assertion "this is not a dog" suggests that the speaker may have been talking about other mammals, but is unlikely to have been talking about a skyscraper. They argue that distributional semantics is a good fit to model conversational negation. Their focus is on compositional distributional methods, which model the negation of nouns via linear transformations. This approach, unlike those used in the present work, relies on the availability of parsed training data.

The effect of negation has also been studied in recurrent neural network models for sentiment classification: Li et al. (2016) observe that their LSTM model does not simply learn a fixed transformation for "not", but rather manages to capture differences in the composition of different words; while Wang et al. (2015) study the behaviour of the LSTM gates in response to negation, showing the network's ability to simulate complex linguistic phenomena. Both groups of authors, like us, focus on LSTM networks, but their models were trained on a sentiment analysis task. We chose a natural language inference task, as it has over an order of magnitude more training data, and requires models to learn a full range of logical and commonsense inferences (Bowman et al., 2015a).

Neurocognitive processing of negation Neuroimaging studies show that negated hand action sentences (e.g., *Now I don't push the button*) and negative imperatives (e.g., *Don't write*) involve decreased activity in motor systems of the brain compared to the same sentences in the affirmative context (Tettamanti et al., 2008; Tomasino et al., 2010). Importantly, Papeo et al. (2016) using Transcranial Magnetic Stimulation (TMS) provide evidence that negation of action-related imperatives involves an immediate reduction of motor (cortical-spinal) excitability for negated compared to affirmative sentences as early as at the initial semantic access stage. Interestingly, the authors show that this suppression does not necessarily reflect neural inhibition in motor areas in contrast to prior studies suggesting a link between action negation and the inhibition of actions (de Vega et al., 2016).

These findings seem in some regards contrary to the predictions of linguistic theories of negation. For example, it has been suggested that, at some level, negation must involve processing of the affirmative situation followed by either its modification or rejection (Russell, 1948). Specifically, Kaup et al. (2007) suggest that the abstract syntactic negation marker may act to reverse the truth value of a sentence through a two-step simulation process involving first, a simulation of the affirmative situation, and, subsequently, a simulation of the actual state of affairs, leading eventually to the suppression of the affirmative situation. While a few behavioral studies have found evidence in favor of the idea that negation involves a simulation of the affirmative situation (Kaup et al., 2007), it has been argued that these effects may be the result of task-induced cognitive strategies (Papeo et al., 2016). On the whole, behavioral and neuroscientific findings do not paint a complete picture of negation, but they suggest that access to some aspects of the affirmative semantic representation in the brain are being immediately reduced (or blocked). Given the above, we might expect to see significant differences in the way in which the semantic similarity of DSM models for action-words and sentences is reflected across the brain areas implicated in action-semantics when comparing affirmative and negated actions.

3 Brain Imaging Data

We use the fMRI data by Djokic et al. (forthcoming), who investigated negation of literal and metaphoric actions in the brain.

Participants Fifteen healthy adults (8 female, ages 18 to 35) took part in the study. All subjects were right-handed, native English speakers.

Stimuli Thirty-one unique hand-action verbs were used to create 40 affirmative literal (AL), 40 negated literal (NL), 40 affirmative metaphor (AM), and 40 negated metaphor (NM). Each verb was repeated once for each condition, except 9 verbs which were repeated twice for each condition. Additionally, 40 affirmative literal paraphrases of the metaphor were created. All sentences are in the 3rd person singular, present tense, progressive (Figure 1). Stimuli were created by

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

Figure 1: Sample stimuli for the verb *push*

the authors of the study and normed for psycholinguistic variables in a separate experiment.

Experimental Paradigm Subjects 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'). Catch trials were included that contained a semantically incongruent object (e.g., 'the wooden table', 'She's eating the table'). Participant's recall of catch trials (and non-catch) trials was tested to ensure participants were paying attention. 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 subjects only completed 4 runs). Stimulus presentation was pseudo-randomized (i.e., such that sentences with the same verb were not shown in succession).

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 × 256 mm matrix, 1 mm resolution, and 208 coronal slices. Whole brain functional images were obtained with a T2* weighted single-shot gradient-recalled echo-planar sequence (EPI) using blood oxygenation-level-dependent contrast with TR=2000 ms, TE=30 ms, flip angle 90°, 64 × 64 mm matrix, 3.5 mm resolution. Each functional image consisted of 37 contiguous axial slices, acquired in interleaved mode.

4 Semantic models

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

Verb In this model, stimulus phrases are represented as the individual D-dimensional word embeddings of their verb.

Addition This model takes the embeddings of the verb and object of the phrase, and computes the phrase representation as their average.

LSTM As a more sophisticated compositional model, we take the long short-term memory (LSTM) recurrent neural network architecture (Hochreiter and Schmidhuber, 1997). Due to the lack of a large training set, directly training the LSTM model for our specific task (i.e. brain decoding) was not possible. Instead, we trained the LSTM on a natural language inference task (Bowman et al., 2015a), 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 hidden representations for each sentence, and then used a single-layer perceptron classifier as in Bowman (2016) to predict the correct relationship. The inputs were the same 100-dimensional word embeddings used for the other models, and were updated during training. The model was optimised using Adam (Kingma and Ba, 2014). We extracted the 100-dimensional hidden representations learnt by the LSTM for the verb-object phrases in our stimulus set.

5 Brain activity decoding

5.1 fMRI data preprocessing

We restricted analysis to the 12 subjects that completed all runs (3 out of 15 subjects scanned only completed 4 out of 5 runs). The runs were combined across time to form each subject's dataset. The functional data was co-registered with the MPRAGE structural image, high-pass filtered (90 secs) and motion corrected to the middle slice using the fMRI software FSL[2]. Lastly each dataset was linearly detrended and (baseline) normalized per run using PyMVPA[3].

5.2 Estimation of fMRI Patterns

GLM Modeling The Blood oxygenation level dependent (BOLD) signal response was estimated

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

[2] Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB's) Software Library, https://fsl.fmrib.ox.ac.uk/fsl

[3] http://www.pymvpa.org/

using the general linear model (GLM) with the hemodynamic response function (HRF) regressor with PyMVPA. The entire stimulus duration for each object and action-related sentence was modeled as an event lasting six seconds (3 TRs) after taking into account the hemodynamic lag. This gave a response amplitude (Beta) estimate for each sentence resulting in voxel-wise Beta maps that were normalized to Z-scores.

Verbs Estimated fMRI patterns were calculated for each of the thirty-one action-verbs by combining action-related sentences with the same action-verb across all stimuli, irrespective of sentence context (**All Verbs**). Estimated fMRI patterns for action-verbs presented in an affirmative context (**Aff Verbs**) were obtained by combining only affirmative sentences containing the same action-verbs. Similarly, fMRI estimates for action-verbs in a negative context (**Neg Verbs**) were obtained by combining negative sentences containing the same action-verbs. In all three cases, estimated brain responses for sentences containing the same action-verbs were averaged together across runs to yield voxel-wise Z-score maps for each of the thirty-one verb presentations and used to perform similarity-based analysis within each subject's native functional space. We performed voxel selection by selecting the top fifteen percent of voxels that had the highest correlation stability across runs using **All Verbs**.

Stimulus Phrases Estimated fMRI patterns for individual action sentences in each condition (affirmative literal (AL), affirmative metaphor (AM), negated literal (NL), and negated metaphor (NM)), were calculated, separately, by modeling unique action sentences within a condition as separate events. Analysis was restricted to only sentences within each condition representative of the 31 unique verbs. We performed voxel selection by selecting the top fifteen percent of voxels with the greatest correlation stability across runs between sentences in the specific condition being modeled.

5.3 Definition of Regions of Interest

We selected a priori regions of interest (ROIs) implicated in action semantics to perform our analysis. This includes 1) left precentral gyrus (**LPG**), implicated in sensorimotor processing (i.e., motoric features) (Pulvermuller, 2005); 2) left middle temporal gyrus, posterior (**LMTP**); 3) left inferior frontal gyrus (**LIFG**), the latter two implicated in language processing (i.e., lexical-semantics/syntax) (Fedorenko et al., 2011). ROIs were created using the Harvard-Oxford Cortical Structural Probabilistic Atlases thresholded at 25% in FSL. Masks were transformed from the Montreal Neurological Institute (MNI) standard space into the subject's native functional space.

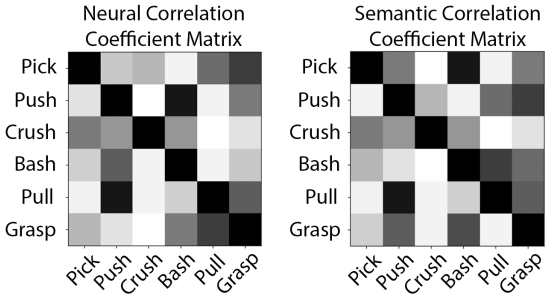

Figure 2: Neural and semantic correlation coefficient matrices. In the study the number of verbs is 31.

5.4 Representational Similarity Analysis

Representational similarity analysis (RSA) is a multivariate approach to fMRI data analysis and avoids model over-fitting and dependence on learning parameters when dealing with high-dimensional data (Kriegeskorte et al., 2008). It calculates a global measure comparing the similarity structures of neural and model-based stimuli representations. The neural and semantic model vectors are first transformed into an abstracted similarity space by computing a similarity matrix from the brain activity vectors (N stimuli $\times$ N stimuli) and a similarity matrix from the semantic model-based vectors (N stimuli $\times$ N stimuli), as shown in Figure 2. The similarities are computed using Pearson correlation coefficient as a measure following Kriegeskorte et al. (2008). The elements in the neural and semantic correlation matrices are then converted into correlation distances $(1 - r)$, leaving zeros in the diagonal. The resulting matrices are referred to as representational dissimilarity matrices (RDMs) and indicate the degree to which conditions can be distinguished from each other (i.e., distance in high-dimensional similarity space). An overall (dis)similarity measure is given by the strength of Spearman's rank correlation between the vectorized lower below diagonal triangle of the model RDM and the vectorized lower below diagonal of the neural RDM giving an overall indication of the correspondence between the representational information carried in the brain and model. We used a one-sided Wilcoxon signed-

rank test to test whether correlations across subjects were significantly greater than zero. False-Discovery-Rate (FDR) (Benjamini and Hochberg, 1995) was used to correct for multiple testing.

5.5 Group-level Similarity-based Decoding

We used similarity-based decoding (Anderson et al., 2016), based on RSA, to investigate if our semantic models can decode fMRI patterns of action-related sentences. In similarity-based decoding, neural and semantic models are first each projected to a similarity space, in the same manner as in RSA, allowing decoding to be performed in a common unit space. Following Anderson et al. (2016), we perform leave-two-out decoding (for $n = 31$, possible pairs $= 465$). Given a pair of stimuli, the neural and semantic similarity codes for each stimulus are obtained by extracting the relevant labeled column vector from the neural similarity matrix and the semantic similarity matrix, respectively. These similarity codes are further reduced by removing the entries referring to that pair, to avoid auto-correlations. These reduced neural and semantic similarity codes are then correlated with each other. If the sum of the correlation coefficients of the correct labeling scheme (i.e. when the neural and semantic codes have the same label) has a higher sum of correlation coefficients than the incorrect labeling (i.e., when they don't match) this is counted as a correct classification, otherwise as incorrect. The decoding accuracy is calculated as the number of correct classifications over the number of possible pairs.

We performed group-level similarity-based decoding in which prior to the decoding step the neural similarity codes of each subject are averaged together to yield one single group-level neural similarity code, as in Anderson et al. (2016). Leave-two-out decoding was then performed using group-level neural similarity and model-based similarity codes as described above.

Statistical significance of group-level decoding accuracies was assessed using permutation testing as in Anderson et al. (2016). The rows and columns of the model-based correlation matrix were shuffled to remove relationships between the stimulus label and its model-based similarity code, while the neural correlation matrix was held fixed. Classification accuracies were obtained using the randomly shuffled data. This procedure was repeated 10,000 times to obtain a null distribution of decoding accuracies, reflecting expected chance-level accuracies with random labeling. The null hypothesis is that there is no relationship between the model-based and the group-level neural similarity codes of our stimuli. The p-value for each accuracy was calculated as the proportion of scores equal to or larger than that accuracy score.

6 Experiments and Results

6.1 Verb Model

Representational Similarity Analysis We used RSA to obtain a measure of relatedness between our fMRI patterns for 31 verbs and the semantic similarity of the VERB model. We performed a condition-based analysis, comparing three types of neural estimates of the verbs: 1) **All Verbs**, 2) **Aff Verbs**, and 3) **Neg Verbs**. We correlated the RDMs for each condition of the neural estimates of the verbs (**All Verbs**, **Aff Verbs**, and **Neg Verbs**) separately with the RDM of the VERB model. Each analysis was performed within the a priori-defined ROIs (LPG, LIFG, and LMTP).

Significant correlations (greater than zero) across subjects were found between the dissimilarity structures of the neural estimates for **All Verbs** and the VERB model in the LPG ($r = 0.04, p < 0.01$), LIFG ($r = 0.04, p < 0.01$), but not the LMTP (Table 1). Similarly, the **Aff Verbs** neural estimates showed significant correlations with the VERB model in the LPG ($r = 0.04, p < 0.01$), LIFG ($r = 0.05, p < 0.01$) and not the LMTP. In contrast, we did not find that **Neg Verbs** triggered any significant correlations with the VERB model in the ROIs tested. Moreover, **Aff Verbs** showed greater overall correlations with the VERB model when compared to **Neg Verbs** (as assessed by two-tailed paired Wilcoxon Sign Rank test) within the LPG and the LIFG ($p < 0.05$), but not the LMTP. These results suggest that (1) the semantic similarity of the VERB model correlates with fMRI patterns of sentences containing the same action verb (irrespective of polarity) in motor (LPG) and the language-related brain region (LIFG) (2) neural estimates for **Neg Verbs** show a reduced sensitivity to the similarity structure of the VERB model compared to **Aff Verbs** in the same ROIs, mainly motor (LPG) and the language-related brain region (LIFG). This suggests that negation involves reduced access to sensorimotor and lexico-semantic representations associated with the affirmative representation.

Region	All	Aff	Neg
LPG	**0.04(0.00)**	**0.04(0.00)**	-0.01(0.83)
LIFG	**0.04(0.00)**	**0.05(0.00)**	0.00(0.21)
LMTP	0.01(0.24)	0.01(0.18)	0.01(0.07)

Table 1: RSA with VERB Model: Significant Spearman's rank correlation coefficients and p-value in bold.

Region	All	Aff	Neg
LPG	**0.09(0.02)**	**0.09(0.00)**	-0.03(0.77)
LIFG	**0.05(0.00)**	**0.08(0.00)**	0.04(0.11)
LMTP	**0.07(0.02)**	**0.10(0.00)**	0.01(0.21)

Table 2: RSA with VERB model for restricted set of verbs: Significant Spearman's rank correlation coefficients and p-value in bold.

We performed an additional analysis restricted to nine verbs, for which we had maximal number of sentences with these same verbs (giving improved signal to noise ratio). We observed a stronger but similar pattern with significant correlations for **All Verbs** in the LPG ($r = 0.09, p < 0.05$), LIFG ($r = 0.05, p < 0.01$), and also within the LMTP ($r = 0.07, p < 0.05$) (Table 2). Similarly, for **Aff Verbs** we found significant correlations across the LPG ($r = 0.09, p < 0.01$), LIFG ($r = 0.08, p < 0.01$), and also within the LMTP ($r = 0.10, p < 0.01$). These results are in line with work showing semantic category effects for action-words in brain regions implicated in action-semantics (Carota et al., 2017), extending this to action sentences. Similar to the previous analysis, we did not find any significant correlations with the **Neg Verbs** in any of the ROIs tested (Table 2). In the restricted analysis only the LPG ($p < 0.05$) (as opposed to both the LPG and LIFG) showed greater correlations for **Aff Verbs** than **Neg Verbs** in line with work showing that action negation impacts modal (e.g., motor) areas (Ghio et al., 2018).

Group-level Similarity-based Decoding We also performed the same condition-based analysis with group-level similarity-based decoding allowing us to observe systematic patterns across subjects, more generally. Table 3 shows the decoding accuracy obtained for each ROI at the group-level in the condition-based analysis. Overall, findings are in line with the RSA results with significant decoding accuracies found for **All Verbs** in the LPG ($Acc = 0.72, p < 0.01$) and LIFG ($Acc = 0.64, p < 0.05$), as well as, similar significant decoding accuracies for **Aff Verbs** in the LPG ($Acc = 0.66, p < 0.05$) and LIFG ($Acc = 0.65, p < 0.05$). Although the **Neg Verbs** did not show significant decoding in the LPG and LIFG, we observed significant decoding within the LMTP for **Neg Verbs** ($Acc = 0.64, p < 0.05$). The above finding coupled with the fact that in the RSA analysis we never observed significant correlation differences between **Neg Verbs** and **Aff Verbs** in the LMTP, may suggest that this area is less impacted by polarity.

Region	All	Aff	Neg
LPG	**72(0.00)**	**66(0.01)**	53(0.33)
LIFG	**64(0.02)**	**65(0.01)**	42(0.77)
LMTP	51(0.37)	52(0.35)	**64(0.02)**

Table 3: Group-Level Similarity-based decoding with VERB. Significant accuracies (%) and p-value in bold.

6.2 Addition and LSTM Models

Group-level Similarity-Based Decoding As an exploratory component to our study we also performed group-level similarity-based decoding for the 31 sentences that each contained a unique verb for each condition type (i.e., AL, NL, AM, NM), separately, allowing us to assess the ability of compositional semantic models (ADDITION and LSTM models) to decode different kinds of negated and affirmative sentences. We observed that the ADDITION model showed significant decoding in the LPG ($Acc = 0.64, p < 0.05$) and LIFG ($Acc = 0.65, p < 0.05$) for the affirmative literal condition (AL) but not in the the negated condition (NL) (Table 4). Interestingly, while we found significant decoding accuracies for the affirmative metaphor condition (AM) in the LPG and LMTP, we also observed significant decoding accuracies for the negated metaphor condition (NM) within the LPG ($Acc = 0.70, p < 0.01$) and LIFG ($Acc = 0.64, p < 0.05$). For the LSTM model we showed significant decoding in the LPG for the affirmative literal condition (AL) ($Acc = 0.67, p < 0.05$) and affirmative metaphoric condition (AM) ($Acc = 0.73, p < 0.01$) but not for the negated conditions (NL, NM) (Table 5). Significant decoding was also found in the LMTP but only for the AM condition ($Acc = 0.70, p < 0.01$). The results suggest reduced decoding for the negated as compared to affirmative literal conditions primarily in sensorimotor brain areas in line with our previous RSA findings at the verb-level with more mixed results for the LIFG and LMTP. Given that we observed that the ADDITION model appears to be sensitive to negated metaphoric actions within the LPG and LIFG, suggests this may not be the

Region	AL	NL	AM	NM
LPG	**64(0.01)**	59(0.13)	**73(0.00)**	**70(0.00)**
LIFG	**65(0.01)**	49(0.55)	53(0.33)	**64(0.02)**
LMTP	58(0.15)	55(0.24)	**70(0.00)**	55(0.24)

Table 4: Group-Level Similarity-based decoding with ADDITION. Significant accuracies and p-value in bold.

Region	AL	NL	AM	NM
LPG	**67(0.01)**	60(0.10)	**71(0.00)**	56(0.20)
LIFG	50(0.48)	51(0.41)	61(0.08)	62(0.06)
LMTP	56(0.22)	48(0.58)	**75(0.00)**	54(0.34)

Table 5: Group-Level Similarity-based decoding with LSTM. Significant accuracies (%) and p-value in bold.

case for the negated metaphoric condition.

7 Discussion

Representational similarity analysis showed that the semantic similarity structure provided by the VERB model corresponded well with neural similarity of sentences containing the same action-verbs (**All Verbs**) within motor (LPG) and language-related brain regions (LIFG, LMTP), both implicated in action-semantic processing (Pulvermuller, 2005). Crucially, when looking at the specific impact of sentential context we found that the fMRI response patterns for negated action-verbs (**Neg Verbs**) showed significantly reduced correlations with the VERB model than the affirmative action-verbs (**Aff Verbs**) mainly in the LPG and LIFG. Similarly, when performing a group-level similarity-based decoding analysis, we also found evidence suggesting reduced decoding accuracies for **Neg Verbs** compared to **Aff Verbs** within the LPG and LIFG. Taken together, these findings provide support to previous neuroscientific studies that suggest that negation manifests foremost as reduced access to motor areas implicated in coding sensorimotor features of action verbs (Tettamanti et al., 2008; Tomasino et al., 2010; Papeo et al., 2016). However, they also provide compelling evidence in support of the idea that the modulatory impact of negation may extend to areas of the language-network. Lastly, our experiments with compositional models show that some of these effects may carry over to more complex models.

Our RSA findings for **All Verbs** (and also **Aff Verbs**) are consistent with the work of Carota et al. (2017) who showed that an LSA model reflecting semantic category information about both verbs and objects associated with actions (e.g., tools and foods) significantly correlated with the similarity of fMRI patterns for verbs and objects in the LPG and LIFG (and to a lesser extent the LMTP). When this analysis was restricted to only action verbs, the LIFG was predominantly sensitive to the semantic similarity of action verbs. It is likely that our results for **All Verbs** (irrespective of polarity) are more closely aligned with their results for verbs and objects associated with actions, given that our action verbs were presented in a sentence context that included information about the object.

Notably, we found a modulatory impact of negation in both sensorimotor (LPG) and to some extent the language-related brain region (LIFG). The LIFG has been implicated in lexical-semantic similarity in the brain but also in the selection of competing semantic alternatives (Thompson-Schill et al., 1997; Carota et al., 2017). For example, the LIFG may be important for event prediction, such as knowing which words (objects or tools) are implied by a given action verb (Carota et al., 2017). This provides further support to the hypothesis that negation involves reduced access to the affirmative mental representation. Importantly, this involves not only reduced access to motoric features, but also access to lexico-semantic relations in language-related brain regions.

The LMTP may have been less impacted by action negation as it is more closely associated with higher-level object processing (Devereux et al., 2013) and, therefore, possibly captures less of the overall semantic variance associated with any given action verb. Moreover, in our study we focused on neural estimates of action verbs irrespective of their specific objects. Thus, the LPG and LIFG may more closely reflect action-semantic variance and show a greater modulatory effect of negation. However, given that similarity-based analysis is sensitive to the semantic distance of the stimuli in question, future work should investigate polarity decoding with verb-object phrases with maximal semantic-variance (e.g., action verbs associated with distinct effectors and object-directed goals).

Lastly, when testing compositional models we also observed that significant decoding accuracies were predominantly found in motor areas (LPG) for affirmative conditions. Interestingly, we did observe an exception to this for negated action-verbs that were also used in a metaphorical con-

text, possibly suggesting that compositional models are better able to capture motor features associated with metaphorical meanings on the whole, but this would need further investigation.

Our main finding of a modulatory impact of negation on motor but also to some extent language-related brain regions is in line with the earlier work of Tettamanti et al. (2008) who found a reduction in activations within left-hemispheric frontal-temporal-parietal areas implicated in the representation of actions for negative compared to positive action sentences, but see (Ghio et al., 2018). Importantly, however, our results do not rule out the possibility that other brain regions may correlate with the VERB model. Recent neuroscientific work suggests that negation not only modulates modality-specific brain regions but also brain areas implicated in syntactic processing and cognitive control (Ghio et al., 2018). It is possible that prefrontal areas implicated in control and working memory may act as an intermediate stage in charge of assigning polarity and temporarily hold a representation of the affirmative situation. We are currently investigating this possibility through a whole-brain searchlight analysis, but note that the temporal resolution of fMRI may possibly hinder detection of any intermediate processing steps.

In this study we provide support for the idea that negation may be mediated in part by reducing (or blocking) access to aspects of the affirmative representation. This may provide a 'default' negation meaning (Papeo et al., 2016), as well as allow competing or cooperating semantic alternatives to emerge. On the other hand, it is also possible that the results reflect a more 'categorical' representation of negation and that the current semantic models are merely not a suitable represenation for the negated meaning. Future work will need to understand the mechanisms by which negation modulates semantic similarity and lexico-semantic relations in brain regions implicated in action-semantics and how this gives rise to a negated meaning. It would be interesting to test alternate models for negation that can simultaneously explain, for example, why the verb 'grasping' has a more crystallized meaning than its negation 'not grasping', whose meaning may also depend to a greater extent on the specific linguistic (or extralinguistic) context.

A fruitful avenue of research may be to investigate the extent to which contextual representations of LSTM models in the context of a sentiment classification task can be used to predict fMRI activations for positive versus negative affective phrases. Predicting sentiment is intimately tied to polarity (e.g., 'good' versus 'not good') and the relationship between affective words and their negated counterparts near orthogonal. Prior work shows the role of LSTM gates in modeling negation in sentiment prediction in part by locally minimizing the input of the negated affective word (Wang et al., 2015), providing insight into the role of learned contextual information in building the negated meaning. The sentiment test case may offer a means to measure how changes in contextual representations relevant to the semantic modeling of negation can contribute directly to predicting brain activity associated with negation processing.

Alternatively, Kruszewski et al. (2017) show that conversational negation can be modeled with a distributional approach, acting like a 'graded similarity function' that prompts a search for 'similar' alternative meanings. Although prior psycholinguistics work on negation consistently shows evidence to suggest that negation reduces access to the affirmative representation, at least one study showed that this is not the case for entities semantically related to the negated representation (MacDonald and Just, 1989). This more closely aligns with the idea that some dimensions of the affirmative representation are being processed while others reduced, possibly due to competing semantic alternatives. Thus, future work should also investigate whether modeling negation as a set of alternative meanings can further show the impact of negation on semantic representation in the brain.

8 Conclusion

In our work, we show for the first time that sensorimotor and to some extent language-related brain regions that correlate with distributional semantic models of action verbs may be impacted by negation. We also show that this effect may extend to more complex compositional models (in motor brain regions). Our work paves the way towards understanding the extent to which human meaning representation is impacted by negation. This finding can in turn inform the design of distributional models dealing with verb negation, for instance when modelling negation as a space of alternative meanings.

References

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 *EMNLP*, pages 1960–1970.

Andrew J Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5:17–30.

Andrew J Anderson, Benjamin D Zinszer, and Rajeev DS Raizada. 2016. Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *NeuroImage*, 128:44–53.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B.*, pages 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. *Cereb. Cortex*, 19(12):2767–96.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015a. 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. 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.

Samuel R Bowman, Christopher Potts, and Christopher D Manning. 2015b. Learning distributed word representations for natural logic reasoning. *Knowledge representation and reasoning: Integrating symbolic and neural approaches: Papers from the 2015 AAAI Spring Symposium.*, pages 289–300.

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

Francesa 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.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1861–1870.

Barry Devereux, Lorraine Tyler, Jeroen Geertzen, and Billi Randall. 2013. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, pages 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.

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.

Marta Ghio, Karolin Haegert, Matilde M Vaghi, and Marco Tettamanti. 2018. Sentential negation of abstract and concrete conceptual categories: a brain decoding multivariate pattern analysis study. *Philosophical Transactions B*, (373).

Uri Hasson and Sam Glucksberg. 2006. Does understanding negation entail affirmation? an examination of negated metaphors. *Journal of Pragmatics*, 38:1015–1032.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Laurence R Horn. 1989. *A Natural History of Negation*. University of Chicago Press, Chicago.

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 tile human cerebral cortex. *Nature*, 532:453.

Barbara Kaup, Richard H Yaxley, Carol J Madden, Rolf A Zwaan, and Jana Lüdtke. 2007. Experiential simulations of negated text information. *Quarterly Journal of Experimental Psychology*, 60(7):976–990.

David Kemmerer. 2015. Are the motor features of verb meanings represented in the precentral motor cortices? yes, but within the context of a flexible, multilevel architecture for conceptual knowledge. *Psychonomic Bulletin Review*, 22:1068–1075.

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

Nikolaus Kriegeskorte, Mur Marieke, and Peter Bandettini. 2008. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(4):4.

German Kruszewski, Denis Paperno, Raffaella Bernardi, and Marco Baroni. 2017. There is no logical negation here, but there are alternatives: Modeling conversational negation with distributional semantics. *Association for Computational Linguistics*, 42(4):637–660.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in nlp. In *HLT-NAACL*, pages 681–691.

Maryellen C MacDonald and Marcel A Just. 1989. Changes in activation levels with negation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 15(4):633–642.

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

Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 114–123. Association for Computational Linguistics.

Isabel Orenes, David Beltran, and Carlos Santamaria. 2014. How negation is understood: Evidence from the visual world paradigm. *Journal of Memory and Language*, 74:36–45.

Liuba Papeo, Jean-Remy Hochmann, and Lorella Battelli. 2016. The default computation of negated meanings. *Journal of Cognitive Neuroscience*, 28(12):1980–1986.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel 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.

Bertrand Russell. 1948. *Human knowledge: Its scope and limits*. Simon & Schuster, New York.

John L Speranza and Laurence R Horn. 2010. A brief history of negation. *Journal of Applied Logic*, 8(3):277–301.

Marco Tettamanti, Rosa Manenti, Pasquale A Della Rosa, Andrea Falini, Daniela Perani, Stefano F Cappa, and Andrea Moro. 2008. Negation in the brain: Modulating action representations. *NeuroImage*, 43(2):358–367.

Sharon L Thompson-Schill, Mark D'Esposito, Geoffrey K Aguirre, and Martha J Farah. 1997. Role of left prefrontal cortex in retrieval of semantic knowledge: a re-evaluation. *Proc Natl Acad Sci.*, 94:14792–14797.

Barbara Tomasino, Peter H Weiss, and Gereon R Fink. 2010. To move or not to move: Imperatives modulate action-related verb processing in the motor system. *Neuroscience*, 169(1):246–258.

Manuel de Vega, Yvrena Morera, Immaculada León, David Beltrán, Pialr Casado, and Manuel Martín-Loeches. 2016. Sentential negation might share neurophysiological mechanisms with action inhibition. Evidence from frontal theta rhythm. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 36(22):6002–6010.

Xin Wang, Yuanchao Liu, Chengjie SUN, Baoxun Wang, and Xiaolong Wang. 2015. Predicting polarities of tweets by composing word embeddings with long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1343–1353, Beijing, China.

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*, 9:11.