Neural and computational mechanisms of analogical reasoning

*Jeffrey N. Chiang[1] Yujia Peng[1] Hongjing Lu[1,2]
Keith J. Holyoak[1,3] Martin M. Monti[1,3,4]

[1]Department of Psychology
[2]Department of Statistics
[3]Brain Research Institute
[4]Department of Neurosurgery
405 Hilgard Ave.
University of California, Los Angeles
Los Angeles, CA 90095-1563

Summary

High-level cognition inevitably involves multiple component processes, which are difficult to distinguish at the neural level. We apply *model-guided componential analysis* to disaggregate components of verbal analogical reasoning, a hallmark of human intelligence. This approach integrates a sequential task design with representational and encoding analyses of fMRI data. The analyses were guided by three computational models of lexical and relation semantics that vary in the specificity of their relation representations. *Word2vec-concat* is nonrelational (based solely on individual word meanings); *Word2vec-diff* computes the generic relation between any word pair; and *BART* derives relational similarity from a set of learned abstract semantic relations (e.g., *synonym*, *antonym*, *cause-effect*). The predictions derived from BART, based on its learned relations, showed the strongest correlation with neural activity in regions including the left posterior parietal cortex (during both relation representation and relation comparison) and rostrolateral prefrontal cortex (during relation comparison). Model-guided componential analysis shows promise as an approach to discovering the neural basis of propositional thought.

Keywords: relations, analogy, reasoning, rostrolateral PFC, posterior parietal cortex, Word2vec, BART, Representational Similarity Analysis (max=10)

Introduction

A key component of human reasoning, creativity, and problem solving is the ability to draw inferences based not only on individual concepts, but also on *relations* between concepts. A paradigmatic example of relational processing is analogical reasoning (Holyoak, 2012). Reasoning by analogy is pervasive in daily life, and verbal analogies are closely linked to psychometric measures of intelligence; however, it remains unclear how the human brain achieves the essential component processes. To determine whether or not a verbal analogy in the canonical form *A* is to *B* as *C* is to *D* (denoted *A*:*B*::*C*:*D*; e.g., *old:young* :: *hot:cold*) is valid, multiple component processes must operate. First, it is necessary to encode the meanings of the individual words. Second, it is necessary to form an active representation of the semantic relation(s) between *A* and *B* (e.g., *contrast*), and that between *C* and *D*-—a process made challenging by the fact that the relations (unlike the individual words) are not present in the input. Finally, it is necessary to compare the two relations to assess the second-order similarity of the relation between *A* and *B* to that between *C* and *D*, to determine whether they are the same (a valid analogy) or different (invalid).

Progress has been made in understanding how semantic representations of individual words are instantiated in a widespread brain network (Binder et al., 2009), with an intricate (but systematic) topographic gradation of semantic sensitivity across different regions of cortex (Huth et al., 2016; Carota et al., 2017; Pereira et al., 2018). Progress has also been made in pinpointing the neural mechanisms of relational comparisons, typically localized to a left-lateralized fronto-parietal network (Vendetti & Bunge, 2014; Wendelken, Ferrer, Whitaker, & Bunge, 2016). This network includes left rostrolateral prefrontal cortex (RLPFC) in Brodmann areas (BA) 10 and 47, a region known to respond in a graded manner to increasing relational complexity (Bunge, Helskog, & Wendelken, 2009; Christoff et al., 2001) and semantic distance between pairs of concepts (i.e., between *A:B* and *C:D*; Green et al., 2010, 2012; for meta-analyses see Hobeika et al., 2016; Vartanian, 2012). Parietal regions, particularly around the intraparietal sulcus (IPS) and supramarginal gyrus (SMG), have also been associated with relational reasoning across multiple domains and tasks, including relation comparison during analogical reasoning (for a meta-analysis see Wendelken, 2015). In addition, left lateral frontal areas (BA 44, 45, 46, and 6) appear to be involved in retrieving relational and abstract information in the context of verbal analogies (Bunge, Wendelken, Badre, & Wagner, 2004; Krawczyk, 2012; Rosa, Catricalà, Canini, Vigliocco, & Cappa, 2018).

However, the cognitive processes that translate the meanings of individual words into semantic relations remain a mystery. The mechanisms that set the stage for relation comparison—forming representations of the semantic relations between individual words—has received almost no attention. Hence, it is unclear how non-relational single-word representations (broadly represented across cortex) give rise to semantic relations that are not directly provided in the input (e.g., how the words *old* and *young* generate an abstract relation representation such as *contrast*). This question has not been clearly addressed in neural studies, in part due to the fact that it is difficult to distinguish the representation of the relation itself from that of the individual concepts being related.

In the present paper, we adopt an approach we term *model-based componential analysis*, which leverages recent advances in computational modeling to provide theoretical tools for localizing and characterizing representations of abstract semantic relations, as well as isolating component cognitive processes involved in analogy. Specifically, employing a sequential event-related fMRI design (see DeWolf, Chiang, Bassok, Holyoak, & Monti, 2016), we separate the construction of first-order relations (i.e., relations between words in a pair) from the second-order assessment of similarity between relations (i.e., the analogical match between *A:B* and *C:D* relations) (see Figure 1).

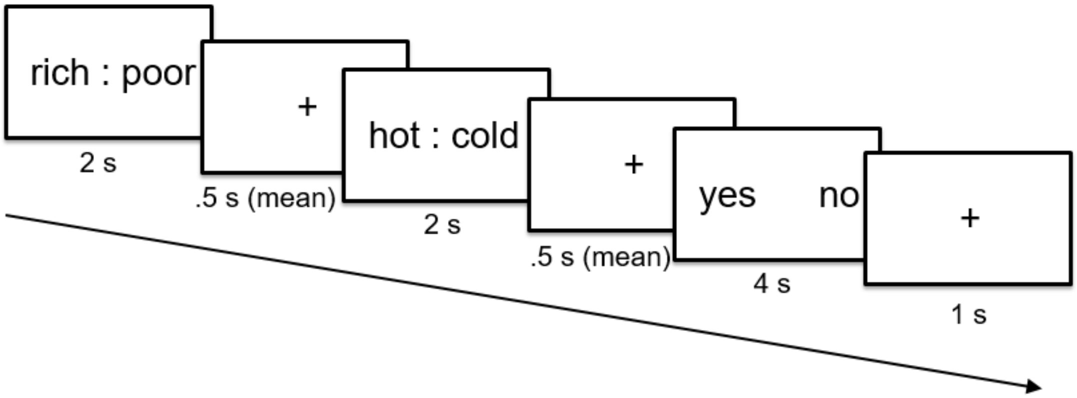

Figure 1. (a) Timing of events on each trial. Participants were shown two word pairs, first an *A:B* pair for 2 seconds, then a *C:D* pair for 2 seconds after a jitter, and finally a cue to make a yes/no decision about the validity of the analogy. Participants responded by pressing a button box, where the location of "yes" and "no" buttons varied from trial to trial, making it impossible to plan a specific motor response until the first two phases had been completed. In a rapid event-related fMRI design, 16 healthy volunteers were asked to evaluate two pairs of semantic concepts. Each analogy was presented as two pairs of words, an *A:B* pair (e.g., *rich:poor*) followed by a *C:D* pair (e.g., *hot:cold*) exemplifying the same relation (here *contrast*) as the *A:B* pair (valid analogy), or else a *C':D'* pair (e.g., *loss:grief*) exemplifying a different relation (here *cause-purpose*) (invalid analogy). All analogies were based on word pairs exemplifying three abstract relation types (*similar*, *contrast*, and *cause-purpose*) from the Jurgens et al. (2012) norms (see Supplemental Material, Table S4). The *A:B* phase provided a relatively pure measure of neural activity involved in coding the individual *A* and *B* words and the *A:B* relation, whereas the *C:D* phase included neural activity required to maintain the *A:B* relation, represent the *C:D* relation, and compare the two.

To guide our investigation, we employed three computational models that each instantiate a distinct approach to the possible form of relation representations in the brain (Figure 2). These models (see Online Methods) all take as inputs semantic representations of individual words, termed *word embeddings*, which are feature vectors created using a deep-learning algorithm, *Word2vec* (Mikolov et al., 2013; Le & Mikolov, 2014; Zhila et al., 2013). After learning to predict words in local contexts taken from a large text corpus, Word2vec yields a 300-dimensional vector as the semantic representation for an individual word. Recent work has shown that these representations can predict the similarity of neural responses to individual words (Anderson et al., 2017; Pereira et al., 2018). However, additional computations are necessary to model relation representations.

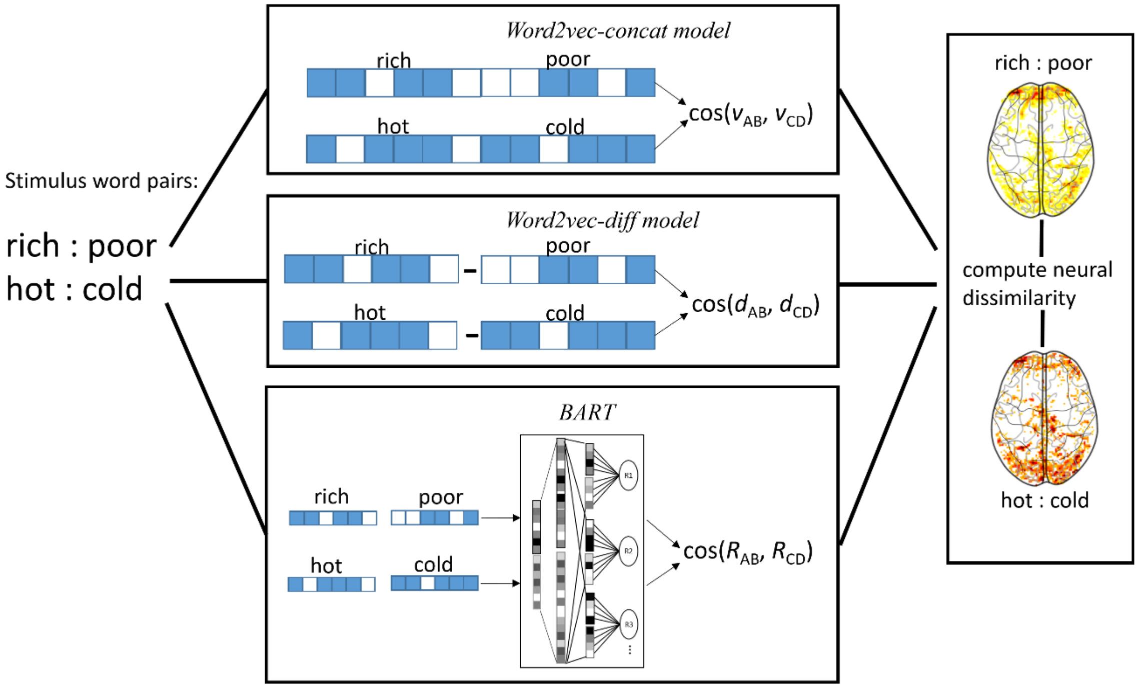

Figure 2. Model-guided approach to discovering neural signatures of specific relations. For any two word pairs (e.g., *rich:poor*, *hot:cold*), three alternative models are used to predict dissimilarity based on the cosine distance between the representations of each individual word pair, using 300-dimensional Word2vec vectors as inputs (left). *Word2vec-concat* (nonrelational) concatenates the vectors for individual words in a pair; *Word-2vec-diff* (generic relation) defines the relation as the difference vector; *BART* (specific relations) creates a new relational vector for each pair based on previously-learned relations. The neural response to each word pair (right) is obtained, allowing a calculation of dissimilarity between patterns of voxels. Neural dissimilarities are compared with computational predictions in order to arbitrate between alternative models.

Two of the models used in the present study (*Word2vec-concat* and *Word2vec-diff*) are based directly on the Word2vec output (Mikolov et al., 2013). The third model is a recent version of *Bayesian Analogy with Relational Transformations* (*BART*; Lu, Wu, & Holyoak, 2019; see also Lu, Chen, & Holyoak, 2012). Each model makes different predictions about how pairs of words, and the relation between them, might be represented. Under *Word2vec-concat*, the meaning of the two words within a pair is a simple aggregate of the semantic vectors of the two individual words. This model is nonrelational, instead capturing semantic similarity across pairs based solely on the meanings of the individual words. This model allows us to identify patterns of similarity based on lexical semantics, separate from any representation of the relation between the two words within each pair. The similarity between any two word pairs is computed by the cosine distance between the two concatenated vectors.

Under *Word2vec-diff*, the first-order relation between two words is defined in a generic fashion as the *difference* between the semantic vectors of each word within a pair; second-order similarity of relations is assessed by the cosine distance between the two difference vectors that form the analogy. This difference-vector method has been used by deep learning models to achieve some degree of success in solving verbal analogies (e.g., Mikolov et al., 2013).

The third model, BART, aims to represent the specific semantic relation(s) between each pair of words (Lu et al., 2019). BART is trained with a small number of word pairs (~20 pairs) as positive examples of each specific relation, acquiring representations of a set of specific relations (e.g., *synonym*, *antonym*, *cause-function*) that could link any two words. Then for any word pair, the model can estimate the probability that the word pair instantiates each learned relation. The resulting vector of relation probabilities provides a distributed representation of the specific relation between the two words. To solve a verbal analogy between two word pairs, BART assesses their second-order relation similarity based on the cosine distance between the two distributed patterns of relations. Using behavioral measures, BART has been shown to be more successful than Word2vec-diff in predicting human judgments of relation typicality and in solving verbal analogies (Lu et al., 2019).

To test the three models as predictors of neural processing during distinct component processes of analogical reasoning, we applied fMRI methods in conjunction with several analytic techniques that comprise model-based componential analysis. First, on each trial we employed a sequential presentation of components of a verbal analogy problem (first the *A:B* pair, followed separately by the *C:D* pair) in order to distinguish the process of relation representation from that of second-order relation comparison. Second, we employed multivariate pattern analysis to assess the degree to which brain activity during the encoding of semantic relations (i.e., *A:B* phase) and the second-order relational comparison (i.e., *C:D* phase) reflect specific semantic relations. Finally, to disentangle the component processes of analogical reasoning, we performed a multivariate Representational Similarity Analysis (RSA; Kriegeskorte & Kievit, 2013; Kriegeskorte, Mur, & Bandettini, 2008; Nili et al., 2014) on the *A:B* phase, coupled with a univariate correlational analysis of model-based relational dissimilarity in the *C:D* phase.

Our overarching aim was not only to separate broad component processes of analogical reasoning, but also to determine the form of the information used to code individual semantic relations at the neural level. Each of the three computational models generates distinct predictions for similarity of neural processing during both relation formation (*A:B* phase) and relation comparison (*C:D* phase). Word2vec-concat bases its predictions solely on the meanings of the individual words in a pair. This model would be expected to predict neural patterns in brain areas previously linked to the semantics of individual words (notably left-lateralized frontal, temporal, and parietal cortices). Word2vec-diff and BART generate competing predictions about both the representation of relations and about the degree of relation similarity between two word pairs that enter into an analogical comparison. Word2vec-diff derives its predictions from a generic measure of the difference in meaning between the two words in a pair, whereas BART derives its predictions from learned representations of a set of specific semantic relations. Given the superior performance of BART on behavioral tests of human relation judgments, including verbal analogies (Lu et al., 2019), we hypothesized that BART (rather than Word2vec-diff) would be the better predictor of neural activity associated both with representation of first-order relations and with second-order comparisons of relations. Based on previous work reviewed above, the areas most likely to exhibit sensitivity to similarity of relations during the comparison process were expected to be the left RLPFC and left IPS.

Results

Univariate analyses: Localization of relation representation and comparison

In order to identify regions engaged in relation representation and relation comparison, we computed the univariate main effects (see Online Methods) for the *A:B* and *C:D* phases (i.e., *A:B* – rest, *C:D* – rest) using FSL. As shown in Figure 3 (left), in the *A:B* stage, related word pairs elicited mostly left-lateralized frontal and temporal activity, bilateral parietal activity, and activity in the occipital lobe (see Supplementary Material, Table S1, for detailed list). The *C:D* stage, compared to simple fixation, recruited many of the same regions as did the *A:B* stage (likely involved in processing each word of the *C:D* pair and encoding their semantic relation), as well as unique activations likely involved in second-order relation assessment for relation comparison. Specifically, the *A:B* and *C:D* stimuli shared activations in the inferior lateral occipital cortex (BA19), fusiform gyrus (BA37), and left frontal regions spanning the rostrolateral prefrontal cortex (BA10 and BA47). In addition, processing *C:D* word pairs uniquely led to a greater BOLD response in left inferior frontal gyrus (*pars triangularis, pars* *opercularis;* BA44, 45) as well as bilateral superior parietal cortex (in BA7; for a full list see Supplementary Material, Table S2).

As shown in Figure 3 (right), the (univariate) comparison of *C:D* and *A:B* phases revealed a fronto-parietal network, mainly left lateralized, presumably involved in the process of second-order relation comparison, which recruits additional regions beyond those related to processing individual words and their semantic relation(s). Specifically, the contrast uncovered significant clusters in the left RLPFC (BAs 10, 47), replicating prior results implicating this region in complex relational comparisons (Christoff et al., 2001; Bunge et al., 2009), as well as in the left inferior frontal gyrus (BAs 44 and 45), bilateral posterior parietal (BA7) and occipital cortices (BA19).

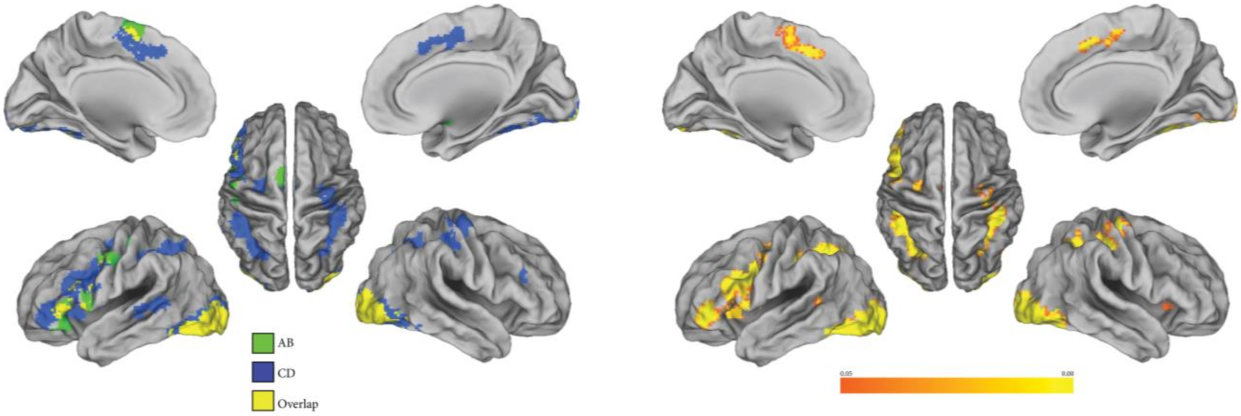

Figure 3. Univariate analysis results. Left: Main effects of *A:B* and *C:D* phases of trials. Cluster were obtained by contrasting each phase (i.e., *A:B*, *C:D*) to simple fixation. Right: *C:D* – *A:B* univariate contrast. Regions in which activity while reading the *C:D* word pair was greater than when reading the *A:B* word pair. Depicted group-level activations were obtained with a non-parametric permutation approach (FSL randomize), significance was set at *p*=0.05 FWER, cluster corrected with threshold-free cluster enhancement (TFCE; Smith & Nichols, 2009).

Multivariate classification analyses: Decoding abstract relations in the brain

To characterize the representations of abstract semantic relations in the brain, we first performed a searchlight classification analysis (Kriegeskorte, Goebel, & Bandettini, 2006), to find areas capable of systematically distinguishing different abstract relations on the basis of their spatial pattern of activations during the *A:B* and *C:D* phases.

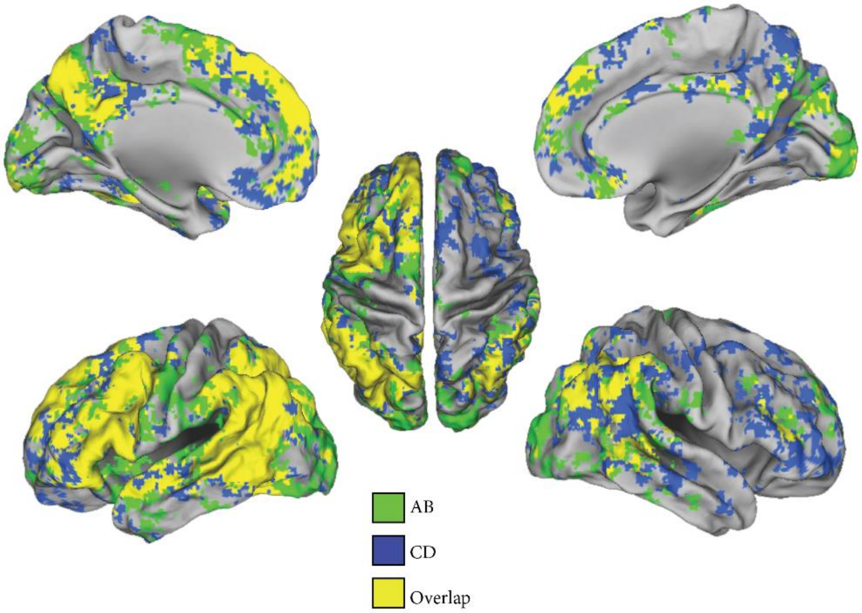

Figure 4. Searchlight results. Regions in which the three general semantic relations could be discriminated from each other during different phases of the analogy task.

As shown in Figure 4, the general semantic relations instantiated by the word pairs (*similar*, *contrast*, and *cause-purpose*) could be reliably distinguished above chance level during the *A:B* phase in large areas of the brain, including frontal and temporal cortices (most pronounced in the left hemisphere), and bilateral parietal cortices, as assessed by a Wilcoxon signed-rank test with TFCE cluster correction (Smith & Nichols, 2009). The three semantic relations could also be reliably distinguished during the *C:D* trials, in many of the same regions, as well as in additional regions in the right hemisphere (particularly across frontal and temporal cortices). Overall, the overlap in regions capable of distinguishing the three semantic relations across both *A:B* and *C:D* phases (areas in yellow in Figure 3) matches very closely the areas previously suggested to be part of the semantic representation system for single words (Binder et al., 2009; Carota et al., 2017; de Heer et al., 2017; Huth et al., 2016), and also includes the parietal regions previously associated more specifically with relational reasoning (Wendelken, 2015).

Representational Similarity Analysis (A:B phase): Representing semantic relations

In order to assess the computations reflected within the regions highlighted above, we employed Representational Similarity Analysis (RSA; Kriegeskorte & Kievit, 2013; Kriegeskorte et al., 2008; Nili et al., 2014). For each individual participant, theoretical patterns of dissimilarity across word pairs predicted by each of three computational models (*Word2vec-concat*, *Word2vec-diff*, *BART*) were compared to the empirical pattern of dissimilarity across word pairs observed in neural activity patterns in pre-selected regions of interest (ROI; see Online Methods) during the *A:B* phase. These analyses were conducted at the level of individual word pairs; hence the size of each Representational Dissimilarity Matrix (RDM) was 144 x 144 (see Figure 5).

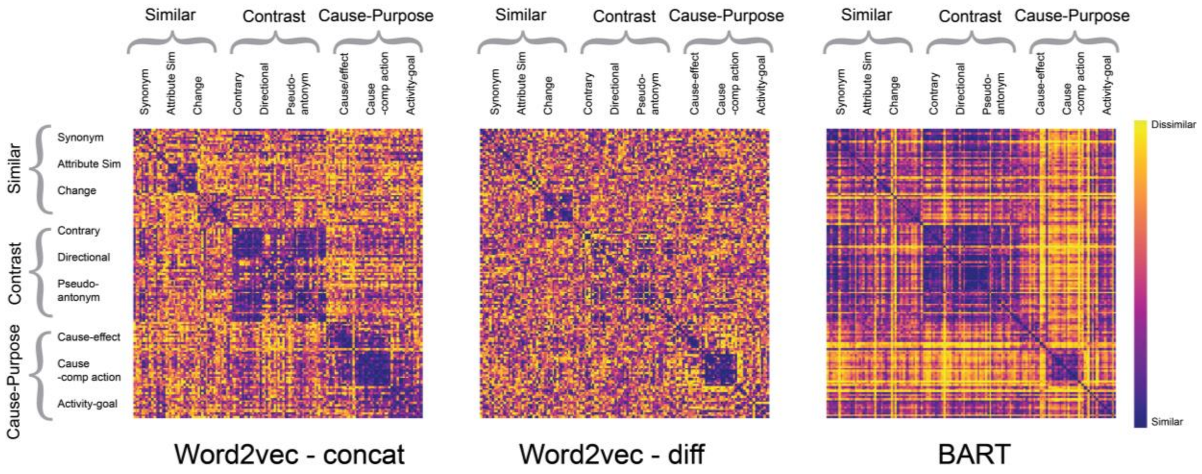

Figure 5. Theoretical Representational Dissimilarity Matrices (RDMs) of size 144 x 144 (i.e., based on individual word pairs) derived from three computational models. Theoretical RDMs capturing the cosine distance between the vector representation for each word pair were correlated with empirical RDMs derived from brain activity patterns within each regions of interest (ROI), using Spearman's rho and corrected for the false discovery rate (FDR) at the 0.05 level. ROIs included BAs 9, 10m, 10l, 44, 45, 46 and 47, in frontal cortex; and the intraparietal sulcus (IPS), superior parietal lobule (SPL), angular gyrus (AG) and posterior supramarginal gyrus (pSMG), in parietal cortex. ROIs were selected on the basis of previous literature and defined on the basis of anatomical atlases (see Online Methods for details).

Neural RDMs derived from the left BA44 and BA45, pSMG, SPL, and the toMTG were significantly correlated with all three models (see Figure 6). This finding suggests that these regions represent relational information as well as lexical semantics, consistent with prior reports that these regions are involved in the formation and/or retrieval of relational information (Krawczyk, 2012; Vendetti & Bunge, 2014; Wendelken, Ferrer, Whitaker, & Bunge, 2016). Interestingly, neural similarity in the inferior parietal sulcus (IPS) significantly correlated with similarity patterns derived from the two models that capture relation similarity (Word2vec-diff: Mean Spearman's rho r = 0.005, p = 0.002 with FDR correction, and BART, r = 0.009, p = 0.015), but not with the RDM derived from the Word2vec-concat model, which only captures semantics of individual words. This finding is consistent with prior studies and meta-analyses indicating that the IPS region (and others within the posterior parietal cortex) plays an important role in relational representation (Wendelken, 2015). In addition, within the left parietal cortex, the BART-derived RDM yielded a significantly stronger correlation with the neural RDM than did that derived from Word2vec-diff in two areas (IPS: $t(15) = 2.39$, $p = 0.033$ with FDR correction; SPL: $t(15) = 2.38$, $p = 0.047$). Notably, none of the theoretical RDMs correlated with the neural patterns of activity in rostrolateral PFC (BA10m, BA10l), consistent with the hypothesis that the RLPFC is primarily engaged when making second-order *comparisons* between relations, rather than representing single relations.

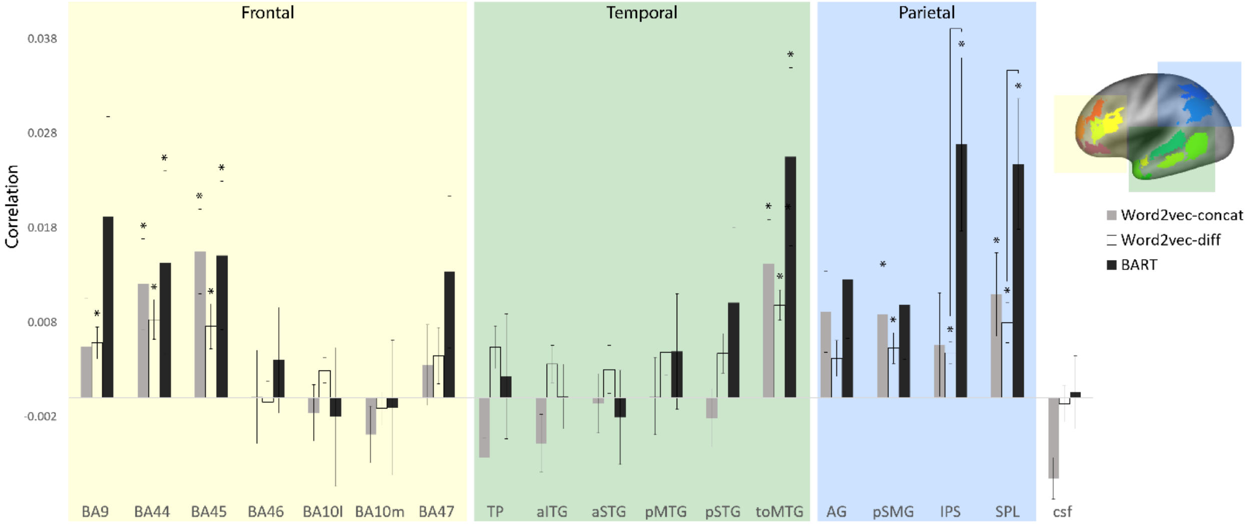

Figure 6. Results from RSA analysis for representation of abstract semantic relations during *A:B* phase. Spearman correlations between ROI-derived neural RDMs and theoretical RDMs (all of size 144 x 144) derived from the computational models. Difference bars indicate that the correlation values were significantly different.

Univariate relational dissimilarity analysis (C:D phase): Second-order relation comparison

With respect to the phase in which the *A:B* and *C:D* relations are compared to assess the validity of the analogy (i.e., *C:D* phase), all models considered here make the general prediction that the difficulty of identifying a valid analogy is proportional to the (word or relation-based) similarity of the *A:B* and *C:D* word pairs, with greater dissimilarity making the analogy harder to verify. In order to derive a measure of *relational dissimilarity* from each of the three models, for every valid *A:B::C:D* analogy (144 problems in total) we calculated the cosine distance between the representations of *A:B* and *C:D* specified by each model, with higher cosine distance implying greater dissimilarity between the two pairs of words. For each individual participant, the theoretical relational dissimilarity scores derived from each model were then correlated (using Spearman's rho) with observed mean ROI activity during the *C:D* phase of each valid analogy.

As shown in Figure 7, comparison of theoretical relational dissimilarity and neural activity shows that prefrontal ROIs were significantly correlated with both the predictions of Word2vec-concat (BA44: $r = 0.045$, p (FDR corrected) = 0.03; BA45: $r = 0.068$, $p = 0.003$; BA47: $r = 0.052$, $p = 0.026$; BA10l: $r = 0.057$, $p = 0.009$), and BART (BA9: $r = 0.06$, $p = 0.022$; BA44: $r = 0.105$, $p = 0.003$; BA45: $r = 0.135$, $p < 0.001$; BA10m: $r = 0.052$, $p = 0.022$; BA10l: $r = 0.069$, $p = 0.006$), but not with the predictions of Word2vec-diff.

A follow-up regression analysis (see Online Methods) revealed that in each of these prefrontal ROIs, the BART model explained unique variance in brain activities after accounting for the influence of word meanings as measured by the predictions of Word2vec-concat. The analysis revealed significant semi-partial correlations (BA44: $r = 0.109$, p (FDR-corrected) = 0.015, BA45: $r = 0.136$, $p = 0.005$, BA 10l: $r = 0.07$, $p = 0.01$, BA 47: $r = 0.139$, $p = 0.006$). The stepwise regression with a reversed order revealed that Word2vec-concat did not predict variance beyond that attributable to BART. In addition, the BART model, but neither of the alternative ones, was significantly correlated with activity in subregions of parietal cortex (AG: $r = 0.085$, $p < 0.001$; pSMG: $r = 0.055$, $p = 0.014$; and trending in the IPS, $r = 0.0521$, $p = 0.081$). BART's predictions of relational dissimilarity were also correlated with activation levels in temporal cortex (pSTG: $r = 0.088$, $p = 0.004$; pMTG: $r = 0.076$, $p = 0.007$; toMTG: $r = 0.078$, p=0.003; aSTG: r = 0. 082, $p = 0.003$; TP: $r = 0.079$, $p = 0.001$), perhaps reflecting the need to represent and maintain multiple semantic relations during the comparison process. In general, the process of relation comparison appears to be dominated by relational dissimilarity as measured by the BART model, which explicitly represents relations that support analogical inference.

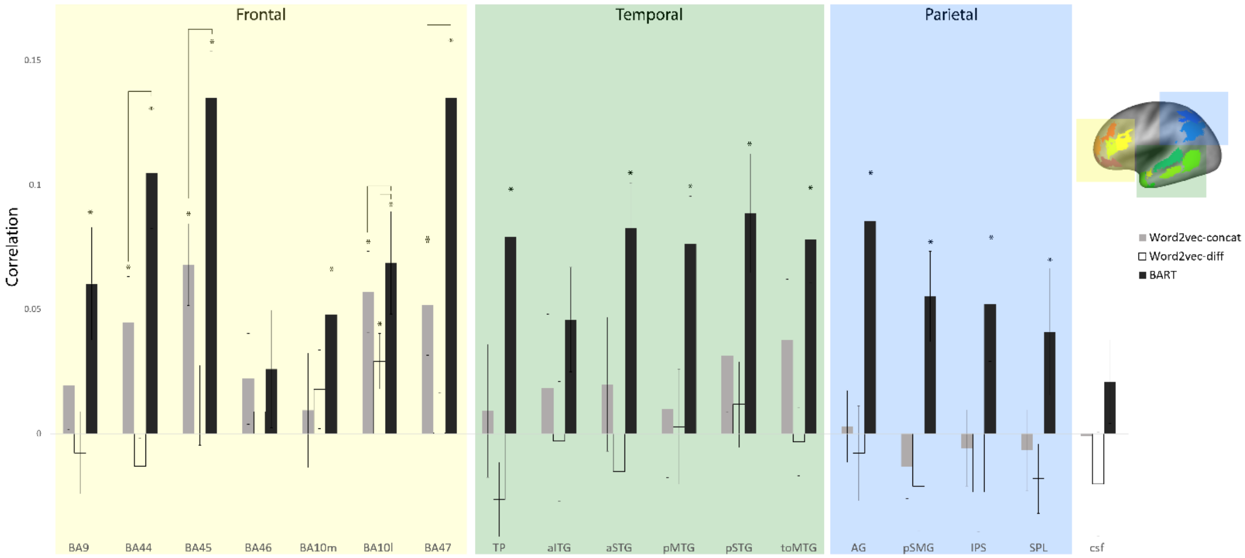

Figure 7. Relation comparison in the brain. Correlation between model-derived relational dissimilarity and mean activity in each ROI during the *C:D* phase. Difference bars are only shown for regions that were correlated with more than one model, and indicate that one model explained significantly more variance (see Online Methods) than the other.

Discussion

The present study combined computational approaches and neuroimaging in a model-guided componential analysis of the representation and comparison of abstract semantic relationships in the brain. By using a sequential presentation with clear temporal phases (DeWolf et al., 2016), we were able to decouple the neural activity associated with encoding the representation of the individual words in a pair and the relation between them from that associated with the comparison of two relations (while also separating these high-level reasoning processes from planning for a motor response). Furthermore, we were able to distinguish between alternative accounts of how semantic relations are coded and compared in the brain. Our analyses were guided by three computational models developed in the fields of machine learning (Mikolov et al., 2013) and cognitive science (Lu et al., 2019). The three models were provided with identical inputs (300-dimensional vectors representing the meaning of individual words) and generated the same final measure of similarity (cosine distance); however, each instantiates a distinct measure of similarity between word pairs. One model (Word2vec-concat) is nonrelational, simply predicting similarity based on the meanings of individual words in pairs. The other two (Word2vec-diff and BART) each predict similarity based on relations between the words, but differ with respect to whether it is assumed that relational representations can be captured by a generic operation (Word2vec-diff), or are acquired through learning specific relations (BART).

Our findings establish a close link between neural representations for meanings of individual words and neural representations for relations instantiated by a pair of words. We found that the process of solving verbal analogies based on abstract relations activated brain regions that spanned the bilateral prefrontal, temporal, and parietal cortices. It appears that abstract semantic relations are encoded at least in part by recruiting the same broad semantic network that has been identified in studies of the semantic representations of individual words (Binder et al., 2009; Carota et al., 2017; de Heer et al., 2017; Huth et al., 2016).

In addition to identifying the semantic network recruited in solving verbal analogy problems, tests of the predictions of the computational models provided a more complete picture of the neural computations underlying basic components of analogical reasoning. By applying Representational Similarity Analysis (Kriegeskorte et al., 2008) to neural activity in different ROIs during the *A:B* phase (i.e., the period when an active representation of the semantic relation(s) between the words in a single pair is being generated), we tested the effectiveness of the three models in predicting empirical patterns of similarity among the BOLD signals for individual word pairs. Similarities of relations between word pairs derived from the BART model (on the basis of learned relation representations) provided the strongest predictors of similarities among neural responses in multiple ROIs within left dorsolateral PFC (but not RLPFC), the temporo-occipital junction, and parietal cortex. At the same time, similarities derived from Word2vec-concat (based on the meanings of individual words) were reliably related to neural patterns within dorsolateral PFC and parietal cortex. These findings are consistent with the hypothesis that generation of an active representation of the semantic relation between a pair of words involves processing of the meanings of the individual words, as well as of the relation between them.

An ROI-based univariate analysis of relational dissimilarity was performed on the *C:D* phase of each trial for valid analogies. This analysis was designed to relate the computational models to patterns of neural activity elicited by comparison of the *A:B* and *C:D* word pairs in the process of assessing whether the analogy was valid. For each model, a measure of relational dissimilarity between the *A:B* and *C:D* relations was correlated with overall neural activity in each ROI. The BART model generated the strongest and most reliable correlations for several ROIs within the frontal, parietal, and temporal cortices. The significant association between the predictions of the BART model and the activity pattern detected in posterior temporal ROIs is consistent with previous work showing that these regions play an important role in comparing relationships that link triplets of elements (actor, agent, patient) across different sentence structures (Monti, Parsons, & Osherson, 2009, 2012). Our findings for frontal and parietal cortices support the hypothesis that second-order relation comparisons are based on a network that transmits abstract representations formed in the parietal cortex to the RLPFC, where abstract relational reasoning is performed (Wendelken, 2015)..

The fact that similarity measures derived from the BART model yielded stronger and more reliable predictions of relational processing—both of individual relations, and of comparisons between relations—than did the Word2vec-diff model is consistent with computational evidence favoring the former model as an account of human relational judgments. Based on human behavioral results, BART predicts judged typicality of relation examples more accurately than does Word2vec-diff, and comes closer to achieving human levels of accuracy in solving verbal analogies (Lu et al., 2019). The relative success of the BART model in predicting patterns of neural activity is directly relevant to a debate as to whether or not individual semantic relations have explicit representations (for discussion see Popov, Hristova, & Anders, 2017). Whereas Word2vec-diff provides only a generic representation of relational similarity (i.e., the difference vector between semantic vectors for two words), BART learns specific representations of individual semantic relations, which then collectively provide a distributed representation of the relations(s) linking any word pair. The neural evidence favoring the BART model of relation similarity thus supports the hypothesis that the brain encodes explicit representations of individual semantic relations, such as *synonym*, *antonym*, and *cause-effect*. Our findings thus have clear implications for the nature of the neural code for abstract semantic relations.

The present study focused on abstract semantic relations. These are particularly important because a pool of abstract relations provides basic elements that can be used to represent more specific relations (Lu et al., 2019). However, further research will be required to determine the extent to which the neural basis for relational reasoning may differ for more concrete semantic and visuospatial relations (e.g., inferring that grasping a hammer enables it to be lifted). More generally, future studies may benefit from applying the overall strategy of model-guided componential analysis. This approach has the potential to be used to analyze patterns of neural activity underlying semantic representations of information units more complex than individual words. Careful task design (e.g., presenting a problem in sequential phases) can be used to separate key component processes. Alternative computational models can then be used to generate item-level predictions of neural similarity, which can be tested by methods such as Representational Similarity Analysis. This approach shows promise in making it possible to decouple component processes and to identify specific representations involved in high-level reasoning. Future work should aim to develop and test well-specified models of how propositions and larger knowledge units are represented in the brain and used to reason.

Online Methods

Participants

Sixteen participants (8 female) were recruited at the University of California, Los Angeles (UCLA) through a flyer distributed in the Psychology department. Participants signed informed consent prior to the experimental session, and were paid $50 for their participation in the 1-hour study, in compliance with the procedures accepted by the local institutional review board (IRB).

Stimuli

The stimuli were a set of analogy problems constructed from word pairs taken from a normed set of examples of abstract relations (Jurgens, Turney, Mohammad, & Holyoak, 2012). These norms were in turn based on a linguistic taxonomy of semantic relations (Bejar, Chaffin, & Embretson, 2012). The full norms include examples of word pairs instantiating ten general types of relations, each including five to ten more specific relations, for a total of 79 distinct relations. For the present study, we focused on three relation types with three specific relations drawn from each, for a total of nine relations: *similar* (*synonym*, *attribute similarity*, *change*); *contrast* (*contrary*, *directional*, *pseudoantonym*); *cause-purpose* (*cause:effect*, *cause:compensatory action*, *activity:goal*). For each relation, we selected 16 word pairs from among the most highly rated (i.e., most prototypical) examples. In making this selection we avoided duplicate pairs that were simple reversals (e.g., *happy-sad* and *sad-happy*), choosing in such cases the pair with the higher typicality rating. Pairs that included conspicuously long or low-frequency words were also excluded. Because for some subcategories it proved difficult to identify 16 pairs that passed our selection criteria, we also included some pairs that Jurgens et al. (2012) had used as “seed” examples to elicit word pairs from humans. These were considered excellent examples (most taken from Bejar et al., 2012). The full list of word pairs is provided in Supplementary Materials, Table 4.

Counterbalancing to Form Analogy Problems

Using the 144 (16 examples × 9 specific relations) distinct word pairs selected as described above, we formed pairs of pairs to create verbal analogy problems in the form *A:B* ::

C:D (valid) or else *A:B* :: *C':D'* (invalid), where all pairs were drawn from the pool of 144. For the invalid pairs, the *C':D'* pair was drawn from a different relation type than was *A:B*. We avoided creating invalid items using different specific relations within the same general relation type (e.g., specific relations *contrary* and *pseudoantonym*, both subtypes of *contrast*) because pilot work suggested that such "near-miss" problems would lead to excessive errors. At the same time, *C':D'* pairs always instantiated a natural semantic relation (rather than being semantically anomalous), forcing participants to consider the paired relations carefully in judging validity of the analogies.

Counterbalancing was used to create four complete sets of analogy problems. To form an individual set, for each of the nine specific relations, eight of the 16 pairs were assigned to the *A:B* role and four to the *C:D* role. The remaining four pairs were assigned to the *C':D'* role associated with *A:B* pairs for four of the six specific relations representing the two remaining general relation types. Assignments to the *C:D* role were random subject to the above restriction. Subject to all of the above restrictions, specific 4-term analogy problems were created by random pairing of word pairs. Each set thus consisted of 72 analogy problems (9 specific relations x 8 problems each). For each specific relation, four problems were valid and four were invalid. Within a set of 72 problems, each of the 144 word pairs occurred twice in the *A:B* role and once in each of the *C:D* and *C':D'* roles. The same procedure was used to create a total of four sets, each with 72 problems distributed as described above. Across all four sets, each of the 144 word pairs appeared in each role with the same proportions (i.e., twice as often as *A:B* than as *C:D* or C'D'). The four sets, with a total of 288 problems (4 sets x 72 problems each), were treated as four blocks administered to each participant. The procedure for problem generation ensured that any individual analogy problem occurred only once in the set of 288 problems. The order of problems was randomized within each block, and the order of the four blocks was counterbalanced across participants. The overall aim of this procedure for problem creation was to ensure that data analyses could be based on neural patterns associated with each of the 16 word pairs representing each of the nine specific relations (144 pairs in total), in each of the three possible roles (*A:B*, *C:D*, *C':D'*), while avoiding any confounding between specific pairs and roles. Finally, each of these four sets was further split into two sets of 36 for presentation convenience.

Procedure

The experiment was administered using PsychoPy2 (Peirce, 2009). On each trial (see Figure 1), participants were first shown the *A:B* word pair for 2s, then the *C:D* pair for 2s (with an average .5s jitter in between). The words "yes" or "no" then appeared on the left and right of the screen, indicating the assignment of two response buttons used to indicate whether or not the two pairs represented the same relation. Critically, the assignment of "yes" and "no" buttons was randomly varied, ensuring that participants could not begin planning a motor response during the earlier phases of the trial.

fMRI Data Acquisition

Data were acquired on a 3 Tesla Siemens Prisma Magnetic Resonance Imaging (MRI) scanner at the Staglin IMHRO Center for Cognitive Neuroscience at UCLA. Structural data were acquired using a T1-weighted sequence (MPRAGE, TR = 1,900 ms, TE = 2.26 ms, voxel size 1 mm^3 isovoxel). Blood oxygenation level dependent (BOLD) data were acquired with a T2*-weighted Gradient Recall Echo sequence (TR = 1,000 ms, TE = 37 ms, 60 interleaved slices (2mm gap), voxel size 2x2x2 mm, 6x multiband acceleration).

fMRI Preprocessing

Data preprocessing was carried out using FSL (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2011; Smith et al., 2004). Prior to univariate analyses, data underwent preprocessing steps including motion correction, slice-timing correction (using Fourier-space time-series phase-shifting), spatial smoothing using a Gaussian kernel of 5 mm full-width half-max, and highpass temporal filtering (Gaussian-weighted least-squares straight line fitting, with s=50.0s). Data from each individual run were analyzed employing a univariate general linear model approach (Monti, 2011) inclusive of a pre-whitening correction for autocorrelation.

Spatial smoothing was omitted from the above preprocessing steps for classification and Representational Similarity Analysis in order to preserve spatial heterogeneities. Beta-series (Rissman, Gazzaley, & D’Esposito, 2004) parameter estimates were derived using the Least Squares-Separate, LS-S approach (Mumford, Turner, Ashby, & Poldrack, 2012). The LS-S algorithm iteratively estimates parameters for each trial using a general linear model including a regressor for that trial as well as another regressor for all other trials.

Univariate Analyses

The general relation type was coded separately for the *A:B* and *C:D* phases of each trial (including both valid and invalid trials) . A univariate analysis using the GLM approach was performed to identify regions engaged in representing semantic relations. The response phase of each trial was included as a condition of non-interest, as well as motion parameters. The GLM analysis was carried out using FSL FEAT (Smith et al., 2004, Jenkinson et al., 2011). Data from individual runs were aggregated employing a mixed effects model (i.e., employing both the within- and between-subject variance), and using automatic outlier detection.

Statistical significance for univariate analyses were assessed using FSL randomize with TFCE cluster correction (Smith & Nichols, 2009; Winkler, Ridgway, Webster, Smith, &

Nichols, 2014). The following contrasts are displayed in Figure 3 of the main text: (left) *A:B* – rest, *C:D* – rest; (right) *C:D* – *A:B*.

Classification Analyses

Classifiers were trained to distinguish between the three general relation types (*similar*, *contrast*, *cause-purpose*), and were evaluated using a leave-one-run-out cross-validation approach (see (Etzel & Braver, 2013). For each participant, two such classifications were run: one on the *A:B* phase and one on the *C:D* phase (including both valid and invalid trials). We used a 5mm radius sphere and a linear SVM (Abraham et al., 2014; Pedregosa et al., 2011). Statistical significance was assessed using FSL randomise with TFCE cluster correction (Smith & Nichols, 2009; Winkler, Ridgway, Webster, Smith, & Nichols, 2014).

ROI Selection

We selected a number of ROIs related to relational and semantic representation using the Juelich, Sallet, Neubert, and Harvard Oxford atlases in FSL (Desikan et al., 2006). The variety of atlases was used so ROIs would be roughly the same size, and so that ROIs would cover previously-reported coordinates based on the following meta analyses of studies of relational reasoning: Krawczyk (2012), Hobeika et al. (2016), Wendelken (2015), and Wendelken et al. (2016). The ROIs associated with relational reasoning fell within the left lateral fronto-parietal network defined by the aforementioned reviews. This included Brodmann area (BA) 10, which was separated into a medial BA10 (BA10m) defined by the Sallet dorsal frontal connectivity parcellation, and a lateral BA10 (BA10l) defined by the Neubert ventral frontal connectivity parcellation. These two ROIs, together with BA47 (see below), were selected to fully cover the area of previously-reported activations in rlPFC. We also selected areas from the ventrolateral and dorsolateral PFC, including BA9 (defined using the Sallet frontal connectivity parcellation) and BAs 44 and 45 (defined by the Harvard Oxford atlas). In the parietal cortex, we used the

Juelich histological atlas. We created the IPS ROI by taking the union of all IPS subdivisions (Wendelken, 2015). We also selected the superior frontal lobe (SPL 7A), and two subdivisions of the inferior parietal lobe (PFm and PGa) corresponding to the angular gyrus and posterior supramarginal gyrus.

In addition to the above ROIs, we selected additional regions associated with semantic representation. These were BA 47 (selected using the Neubert frontal connectivity parcellation), and broad regions of the temporal cortex (temporal pole, aSTG, aMTG, aITG, pMTG, pSTG, toMTG) corresponding to regions selected in a recent study of semantic representation (Carota et al., 2017) and mentioned by another review of the neural distribution of semantics (Binder et al., 2009).

As a control, cerebral spinal fluid (CSF), a region that would not plausibly be involved in processing of abstract semantic relations, was included as a "region of disinterest". A spherical CSF ROI was manually drawn for each participant using the anatomical T1 image and registered to functional space.

Details of Computational Models

All quantitative models used to create theoretical RDMs for the RSA analysis were based directly (Word2vec-concat, Word2vec-diff) or indirectly (BART) on the outputs of a machine learning model, Word2vec (Mikolov et al., 2013). This model takes a large text corpus (Google News) as input, examines distributional statistics relating each word to neighboring words in sentences (local context), and outputs a modular vector representation for each individual word, termed a *word embedding*. Word2vec vectors of length 300 were obtained for all words used in the present study. Word2vec-concat (the concatenation of the vectors for the two words in a pair) and Word2vec-diff (the difference vector derived from the two individual vectors) were calculated and used to create theoretical RDMs.

The BART model (*Bayesian Analogy with Relational Transformations*; Lu, Wu & Holyoak, 2019) was applied to learn specific relations between word pairs. BART takes as inputs pairs of positive and negative examples of a given relation, where each pair is represented by the concatenation of the Word2vec vector for each word. For example, a vector formed by concatenating the individual vectors for *love* and *hate* would constitute a positive example of the *antonymy* relation, but a negative example of the *category membership* relation. The model used supervised learning with 20 positive examples and a fixed set of 64-74 negative instances (the top example for each relation from each general category other than that of the target relation) to form weight distributions representing each of the 79 relations in the Jurgens et al. (2012) norms. For each word pair used in the study, these learned weights were used to calculate the posterior probability that the pair instantiated each of the 79 learned relations. The vector of length 79 formed by these posterior probabilities represented the specific relation between the two words in the pair. These vectors were used to create BART's theoretical RDMs.

Representational Similarity Analysis

Representational Similarity Analysis (RSA; Kriegeskorte & Kievit, 2013; Kriegeskorte, Mur, & Bandettini, 2008; Nili et al., 2014) was used to characterize the similarities of neural responses across pairs. RSA characterizes the representation in a brain region by a representational dissimilarity matrix (RDM), and compares this empirical matrix with a theoretical model. An RDM is a square symmetric matrix, with each entry referring to the dissimilarity between the activity patterns associated with two trials (e.g., entry (1,2) would represent the dissimilarity between activity patterns on trial 1 and trial 2). Procedurally, each element of the RDM is calculated as 1 minus the Pearson correlation between the beta-series for each pair of trials (Carota, Kriegeskorte, Nili, & Pulvermüller, 2017; Nili et al., 2014).

Hypothesis models were manually generated to reflect idealized RDMs expected given a theoretical representational space. We generated theoretical RDMs from each of the three computational models. Each model uses a different calculation to yield a feature vector characterizing a word pair; however, the RDM was calculated in the same way for all models, as the cosine distance between word-pair representations.

RDMs and hypothesis models were compared by calculating a “second-order similarity” (Nili et al., 2014), defined as the Spearman correlation coefficient between the two matrices. All analyses were carried out using Python, making extensive use of the machine learning packages Scikit-learn (Pedregosa et al., 2011) and NiLearn (Abraham et al., 2014). Data and analysis code are available upon request.

Univariate Relational Dissimilarity Analysis

All the models of analogical comparison considered in the present paper make the general prediction that the difficulty of deciding the validity of an analogy will be related to the relation-based similarity of the *A:B* and *C:D* word pairs, with greater similarity making the decision easier. In this analysis, only trials consisting of valid analogies (i.e., *A:B* :: *C:D*) were included so that the relation representations during the *C:D* phase would not be confounded by additional cognitive operations associated with processing a relation inconsistent with *A:B*. To derive a specific prediction from each of the three candidate models in the paper, for every valid analogy of word pairs, *A:B::C:D*, a *relational dissimilarity* measure was calculated by taking the cosine distance between the representations of *A:B* and of *C:D* specified by the model (i.e., higher cosine distance implies greater dissimilarity between the two pairs). These model-derived relational dissimilarity scores for each trial were then correlated (using Spearman’s rho) with mean ROI activity to identify brain regions that track relational dissimilarity according to the predictions from each of the alternative models. The resulting *p* values were adjusted for multiple comparisons by controlling the false discovery rate (FDR) at q = 0.05 and are reported in the main text.

Follow-up RegressionAanalysis

Within regions that were significantly correlated with both BART and Word2vec-concat models (Figure 7), the general trend was that BART-derived predictions showed greater correlation with ROI-derived relational similarity than with Word2vec-concat, the model based on word similarity. In these ROIs (BA44, BA45, BA47, BA10l), a semipartial correlation analysis was performed to determine whether these two models captured the same or different information. Word2vec-concat relational dissimilarity scores were first regressed out of the ROI-based similarity scores, and subsequently the resulting residuals were correlated with the relational dissimilarity predictions from BART. (See main text for results.) The reverse analysis was also performed, in which BART relational dissimilarity predictions were regressed out of ROI-based dissimilarity scores and then correlated with the Word2vec-concat predictions. No regions showed a significant impact of Word2vec-concat after controlling for the variance predicted by BART. That is, Word2vec-concat did not appear to capture any information about relational dissimilarity beyond that accounted for by the BART model. The associated p-values were corrected for the false discovery rate at q = 0.05 and the adjusted values are reported in the main text of the paper.

Acknowledgements

Preparation of this paper was supported by NSF Grant BCS-1827374 and a Google Faculty Research Award to H.L. and K.J.H., a UCLA Academic Senate grant to K.J.H., and a Department of Defense (DoD) grant through the National Defense Science and Engineering Graduate Fellowship Program to J.N.C. A preliminary report of this work was presented at the 2018 meeting of the Society for Neuroscience (San Diego, November).

Author Contributions

Conceptualization, K.J.H and M.M.M.; Methodology, J.N.C., Y.P., H.L., and M.M.M.; Formal Analysis, J.N.C. and H.L.; Investigation, J.N.C. and Y.P.; Writing – Original Draft, J.N.C. and K.J.H.; Writing – Review & Editing, J.N.C., Y.P., H.L., K.J.H., and M.M.M.; Supervision, M.M.M; Funding Acquisition, H.L. and K.J.H.

Declaration of Interests

The authors declare no competing interests.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., … Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics, 8*, 14. https://doi.org/10.3389/fninf.2014.00014

Anderson, A. J., Kiela, D., Clark, S., & Poesio, M. (2017). Visually grounded and textual semantic models differentially decode brain activity associated with concreter and abstract nouns. *Transactions of the Association for Computational Linguistics, 5*, 17-30.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences, 15*(11), 527–536. https://doi.org/10.1016/j.tics.2011.10.001

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex, 19*(12), 2767 96. https://doi.org/10.1093/cercor/bhp055

Bookheimer, S. (2002). Functional MRI of language: new approaches to understanding the cortical organization of semantic processing. *Annual Review of Neuroscience, 25*, 151–188. https://doi.org/10.1146/annurev.neuro.25.112701.142946

Bunge, S., Helskog, E., & Wendelken, C. (2009). Left, but not right, rostrolateral prefrontal cortex meets a stringent test of the relational integration hypothesis. *NeuroImage, 46*(1), 338-342.

Bunge, S., Wendelken, C., Badre, D. D., & Wagner, A. D. (2004). Analogical reasoning and prefrontal cortex: evidence for separable retrieval and integration mechanisms. *Cerebral Cortex,15*(3), 239-249. https://doi.org/10.1093/cercor/bhh126

Burks, J. D., Boettcher, L. B., Conner, A. K., Glenn, C. A., Bonney, P. A., Baker, C. M., … Sughrue, M. E. (2017). White matter connections of the inferior parietal lobule: A study of surgical anatomy. *Brain and Behavior, 7*(4), e00640. https://doi.org/10.1002/brb3.640

Carota, F., Kriegeskorte, N., Nili, H., & Pulvermüller, F. (2017). Representational similarity mapping of distributional semantics in left inferior frontal, middle temporal, and motor cortex. *Cerebral Cortex, 27*(1), 294–309. https://doi.org/10.1093/cercor/bhw379

Chang, C., & Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27,. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chen, D., Peterson, J. C., & Griffiths, T. L. (2017). Evaluating vector-space models of analogy. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 1746-1751). Austin, TX: *Cognitive Science Society*.

Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., & Gabrieli, J. (2001). Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *NeuroImage, 14*(5), 1136–1149. https://doi.org/10.1006/nimg.2001.0922

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *Journal of Neuroscience, 37*(27), 6539–6557. https://doi.org/10.1523/JNEUROSCI.3267-16.2017

DeWolf, M., Chiang, J. N., Bassok, M., Holyoak, K. J., & Monti, M. M. (2016). Neural representations of magnitude for natural and rational numbers. *NeuroImage, 141*, 304–312. https://doi.org/10.1016/j.neuroimage.2016.07.052

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172-179.

Ettinger-Veenstra, H., McAllister, A., Lundberg, P., Karlsson, T., & Engström, M. (2016). Higher language ability is related to angular gyrus activation increase during semantic processing, independent of sentence incongruency. *Frontiers in Human Neuroscience*, *10*, 110. https://doi.org/10.3389/fnhum.2016.00110

Etzel, J. A., & Braver, T. S. (2013). MVPA Permutation schemes: Permutation testing in the land of cross-validation. *IEEE*, 140–143. https://doi.org/10.1109/PRNI.2013.44

Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, *110*(41), 16616–16621. https://doi.org/10.1073/pnas.1315235110

Fedorenko, E., & Varley, R. (2016). Language and thought are not the same thing: Evidence from neuroimaging and neurological patients. *Annals of the New York Academy of Sciences*, *1369*(1), 132–153. https://doi.org/10.1111/nyas.13046

Ferreira, R. A., Göbel, S. M., Hymers, M., & Ellis, A. W. (2015). The neural correlates of semantic richness: evidence from an fMRI study of word learning. *Brain and Language*, *143*, 69–80. https://doi.org/10.1016/j.bandl.2015.02.005

Green, A. E., Fugelsang, J. A., Kraemer, D. J., Shamosh, N. A., & Dunbar, K. N. (2006). Frontopolar cortex mediates abstract integration in analogy. *Brain Research*, *1096*(1), 125–137. https://doi.org/10.1016/j.brainres.2006.04.024

Hobeika, L., Diard-Detoeuf, C., Garcin, B., Levy, R., & Volle, E. (2016). General and specialized brain correlates for analogical reasoning: A meta analysis of functional imaging studies. *Human Brain Mapping, 37*(5), 1953–1969. https://doi.org/10.1002/hbm.23149

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 234-259). New York: Oxford University Press.

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. https://doi.org/10.1038/nature17637

Jurgens, D., Turney, P., Mohammad, S. M., & Holyoak, K. J. (2012). Semeval-2012 task 2: Measuring degrees of relational similarity. *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, 356-364.

Knowlton, B. J., Morrison, R. G., Hummel, J. E., & Holyoak, K. J. (2012). A neurocomputational system for relational reasoning. *Trends in Cognitive Sciences, 16*(7), 373–381. https://doi.org/10.1016/j.tics.2012.06.002

Krawczyk, D. C. (2012). The cognition and neuroscience of relational reasoning. *Brain Research, 1428*, 13–23. https://doi.org/10.1016/j.brainres.2010.11.080

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences, USA, 103*(10), 3863 3868. https://doi.org/10.1073/pnas.0600244103

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N, Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2,* 1 – 28.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning*, *32*(2), 1188-1196.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review, 119*, 617-648.

Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences, USA*.

McCandliss, B., Cohen, L., & Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences, 7*(7), 293-299.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Monti, M. M., Parsons, L. M., & Osherson, D. N. (2009). The boundaries of language and thought in deductive inference. *Proceedings of the National Academy of Science, USA, 106*(30), 12554-12559.

Monti, M. M., Parsons, L. M., & Osherson, D. N. (2012). Thought beyond language: Neural disassociation of algebra and natural language. *Psychological Science*, *23*(8), 914-922.

Morcom, A.M., & Fletch, P.C. (2007). Does the brain have a baseline? Why we should be resisting a rest. *NeuroImage, 37*(4), 1073-1082. https://doi.org/10.1016/j.neuroimage.2007.06.019

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology, 10*(4), e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

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*(1), 963. https://doi.org/10.1038/s41467-018-03068-4

Popov, V., Hristova, P., & Anders, R. (2017). The relational luring effect: Retrieval of relational information during associative recognition. *Journal of Experimental Psychology: General, 146*(5), 722-745.

Price, Amy R, Peelle, J. E., Bonner, M. F., Grossman, M., & Hamilton, R. H. (2016). Causal evidence for a mechanism of semantic integration in the angular gyrus as revealed by high-definition transcranial direct current stimulation. *Journal of Neuroscience, 36*(13), 3829–3838. https://doi.org/10.1523/JNEUROSCI.3120-15.2016

Price, A R, Bonner, M., Peelle, J., & Grossman, M. (2015). Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *Journal of Neuroscience, 35*(7), 3276 3284. https://doi.org/10.1523/JNEUROSCI.3446-14.2015

Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience, 18*(1), 42–55. https://doi.org/10.1038/nrn.2016.150

Rosa, P. A., Catricalà, E., Canini, M., Vigliocco, G., & Cappa, S. F. (2018). The left inferior frontal gyrus: A neural crossroads between abstract and concrete knowledge. *NeuroImage, 175*, 449-459. https://doi.org/10.1016/j.neuroimage.2018.04.021

Smith, S., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage, 44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Vartanian, O. (2012). Dissociable neural systems for analogy and metaphor: Implications for the neuroscience of creativity. *British Journal of Psychology, 103*(3), 302–316. https://doi.org/10.1111/j.2044-8295.2011.02073.x

Vendetti, M. S., & Bunge, S. A. (2014). Evolutionary and developmental changes in the lateral frontoparietal network: A little goes a long way for higher-level cognition. *Neuron, 84*(5), 906–917. https://doi.org/10.1016/j.neuron.2014.09.035

Vigliocco, G., Kousta, S.-T. T., Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2014). The neural representation of abstract words: The role of emotion. *Cerebral Cortex*, *24*(7), 1767–1777. https://doi.org/10.1093/cercor/bht025

Waechter, R. L., Goel, V., Raymont, V., Kruger, F., & Grafman, J. (2013). Transitive inference reasoning is impaired by focal lesions in parietal cortex rather than rostrolateral prefrontal cortex. *Neuropsychologia, 51*(3), 464–471. https://doi.org/10.1016/j.neuropsychologia.2012.11.026

Wendelken, C. (2015). Meta-analysis: How does posterior parietal cortex contribute to reasoning? *Frontiers in Human Neuroscience, 8,* Article ID 1042. https://doi.org/10.3389/fnhum.2014.01042

Wendelken, C, Ferrer, E., Whitaker, K. J., & Bunge, S. A. (2016). Fronto-parietal network reconfiguration supports the development of reasoning ability. *Cerebral Cortex*, *26*(5), 2178–2190. https://doi.org/10.1093/cercor/bhv050

Wendelken, C, Bunge, S. A., & Carter, C. S. (2008). Maintaining structured information: an investigation into functions of parietal and lateral prefrontal cortices. *Neuropsychologia*, *46*(2): 665-678.

Wertheim, J., & Ragni, M. (2018). The neural correlates of relational reasoning: A meta-analysis of 47 functional magnetic resonance imaging studies. *Journal of Cognitive Neuroscience*, 1–15. https://doi.org/10.1162/jocn_a_01311

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80–83.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage, 92*, 381–397. https://doi.org/10.1016/j.neuroimage.2014.01.060

Zhila, A., Yih, W., Meek, C., Zweig, G., & Mikolov, T. (2013). Combining heterogeneous methods for measuring relational similarity. *Proceedings of NAACL-HLT*, 1000-1009.

References (Method-Specific)

Bejar, I. I., Chaffin, R., & Embretson, S. (2012). *Cognitive and psychometric analysis of analogical problem solving*. New York: Springer-Verlag.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2011). FSL. *NeuroImage, 62*(2), 782 90. https://doi.org/10.1016/j.neuroimage.2011.09.015

Monti, M. M. (2011). Statistical analysis of fMRI time-series: A critical review of the GLM approach. *Frontiers in Human Neuroscience, 5*, 28. https://doi.org/10.3389/fnhum.2011.00028

Mumford, J. A., Turner, B. O., Ashby, G. F., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage, 59*(3), 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics, 2,* 10. https://doi.org/10.3389/neuro.11.010.2008

Rissman, J., Gazzaley, A., & D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *NeuroImage, 23*(2), 752 63. https://doi.org/10.1016/j.neuroimage.2004.06.035

Smith, S., Jenkinson, M., Woolrich, M., Beckmann, C., Behrens, E. J. T., Johansen-Berg, H., … & Matthews, P. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage, 23*, S208–S219.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., … Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289. https://doi.org/10.1006/nimg.2001.0978