Distributed Code for Semantic Relations Predicts Neural Similarity during Analogical Reasoning

Jeffrey N. Chiang, Yujia Peng, Hongjing Lu, Keith J. Holyoak, and Martin M. Monti

Abstract

■ The ability to generate and process semantic relations is central to many aspects of human cognition. Theorists have long debated whether such relations are coarsely coded as links in a semantic network or finely coded as distributed patterns over some core set of abstract relations. The form and content of the conceptual and neural representations of semantic relations are yet to be empirically established. Using sequential presentation of verbal analogies, we compared neural activities in making analogy judgments with predictions derived from alternative computational models of relational dissimilarity to adjudicate among rival accounts of how semantic relations are coded and compared in the brain. We found that a frontoparietal network encodes the three relation types included in the design. A computational model based on semantic relations coded as distributed representations over a pool of abstract relations predicted neural activities for individual relations within the left superior parietal cortex and for second-order comparisons of relations within a broader left-lateralized network. ■

INTRODUCTION

The poet Samuel Taylor Coleridge claimed that the creative mind needs to become "accustomed to contemplate not *things* only, … but likewise and chiefly the *relations* of things…" (Coleridge, 1810/1969, p. 451). Because relations provide basic building blocks for language and thought, they are central for a range of cognitive tasks. A prime example is the critical role of relation representations in analogical reasoning (Holyoak, 2012), a mental process that impacts human activities as diverse as metaphor comprehension (Holyoak, 2019), mathematics education (Richland, Zur, & Holyoak, 2007), scientific discovery (Dunbar & Klahr, 2012), and engineering design (Chan & Schunn, 2015). However, although the importance of relations is widely recognized, no consensus has emerged regarding the form of relation representations in the mind and brain.

For the past half century, cognitive scientists exploring human semantic memory have sought to identify the nature of the code for the first-order relations between two concepts (for reviews, see Jones, Willits, & Dennis, 2015; Holyoak, 2008). Two longstanding views, mainly based on data from speeded verification of category–membership relations (e.g., deciding as rapidly as possible whether a rose *is a* flower), continue to be influential. One approach, originating in computer science (Collins & Quillian, 1969), treats relations as being coarsely coded, with labeled unitary links between localist nodes representing concepts (e.g., an "*is a*" link connecting rose to flower). Relation verification is viewed as an all-or-none process of retrieving the relevant link. For example, the word pair *rich–poor* might trigger retrieval of the relation type "opposite" to form the symbolic representation *opposite* (*rich*, *poor*). Current computational models of analogy based on traditional symbolic knowledge representations (Forbus, Ferguson, Lovett, & Gentner, 2017) continue to assume relations are coded as localist links.

In contrast, an alternative view hypothesizes that the meanings of relations are more finely coded by means of operations performed on featural representations of entities (Smith, Shoben, & Rips, 1974; Meyer, 1970). In support of the latter view, analyses of verification time based on speed–accuracy decomposition have revealed that relation information accrues continuously over time, rather than being retrieved in an all-or-none fashion (Kounios, Montgomery, & Smith, 1994). Moreover, much like object categories (Rosch, 1975), examples of semantic relations exhibit a typicality gradient (e.g., *hot–cold* is considered a better example of "opposite" than is *warm–cool*; Jurgens, Turney, Mohammad, & Holyoak, 2012). There is continuing debate as to whether the relation between a pair of concepts is coarsely coded as a general relation type or whether the relation is more finely coded based on the features of the concepts it links (Popov, Hristova, & Anders, 2017).

Not only can word pairs instantiate a particular relation to different degrees, as suggested by previous research on relation typicality, but many word pairs seem to instantiate

This article is part of a Special Focus, Relational Reasoning, deriving from a symposium at the 2019 annual meeting of the Cognitive Neuroscience Society, organized by Silvia Bunge and Keith Holyoak.

University of California, Los Angeles

multiple relations to some degree. For example, the concepts *hill–mountain* primarily instantiate the relation of "similar" (both are types of high geological formations), but they also, to some degree, instantiate "contrast" (differing in height). These systematic and graded variations pose challenges for the second-order relation comparisons required to solve analogy problems, suggesting that analogical validity may itself be a matter of degree, varying with some measure of relation dissimilarity.

Previous work has identified regions within a left-lateralized frontoparietal network that support component processes involved in analogical reasoning. In particular, subareas of parietal cortex appear to support the encoding of individual relations (Wendelken, 2015), whereas the rostrolateral pFC (RLPFC) appears critical in second-order relational comparisons (Hobeika, Diard-Detoeuf, Garcin, Levy, & Volle, 2016; Green, Kraemer, Fugelsang, Gray, & Dunbar, 2010; Bunge, Helskog, & Wendelken, 2009; Wendelken, Bunge, & Carter, 2008; Bunge, Wendelken, Badre, & Wagner, 2005; for a review, see Holyoak & Monti, this issue). However, it remains unclear what content of relation representations is encoded in the brain and compared during analogical reasoning. To address questions about the specific nature and content of relation representations in the brain, it is important to obtain neural evidence based on item-level analyses. Such detailed evidence has the potential to identify properties of relation representation that yield graded variations in the representations of individual relations and in second-order relational comparisons. By performing item-level analyses, we can compare neural activities with predictions derived from alternative computational models to adjudicate among rival accounts of how semantic relations are coded and compared in the brain. To test the proposed computational code for semantic relations and their comparison, various models were used to predict degrees of relation dissimilarity between word pairs. Model predictions were correlated with patterns of neural dissimilarity across word pairs and were used to predict neural activities in making an analogy judgment.

Computational Models of Relation Representation and Comparison in Analogy

Here, we test alternative computational models of relation representation, combining recent advances in machine learning and cognitive science with neuroimaging. Following Coleridge, to represent relations between things, it is first necessary to have representations of those "things"—in the case of semantic relations, we first need semantic representations of individual words. To represent word meanings, we adopt word embeddings produced by a recent machine-learning model, Word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). This model applies a predictive learning algorithm to a large text corpus (e.g., Google News) to create high-dimensional semantic vectors for individual words. Vectors generated by Word2vec and similar models have been shown to accurately capture human judgments of semantic similarities among words (Zhila, Yih, Meek, Zweig, & Mikolov, 2013) and have also been used to create a neural decoder to predict patterns of brain activity produced in response to sentences (Pereira et al., 2018).

Although major computational models of analogical reasoning, such as SME (Forbus et al., 2017) and LISA (Hummel & Holyoak, 2005), critically depend on assumptions about relation representations, most such models do not specify a mechanism by which relations could be learned from nonrelational inputs. The DORA model (Doumas, Hummel, & Sandhofer, 2008) does address relation learning but has not been applied to semantic vectors as inputs. In the present article, we assessed three computational models based on semantic vectors.

Two of these models derive dissimilarity predictions directly from Word2vec vectors for the individual words in a pair. These two models differ in their assumptions about how (or whether) the relation between the two words is represented. Under Word2vec-concat, the meaning of the words within a pair is a simple aggregate of the semantic vectors of the two individual words. The dissimilarity between any two word pairs is computed by the cosine distance between the two concatenated vectors. This model is nonrelational, instead capturing semantic dissimilarity across pairs based solely on the meanings of the individual words. Word2vec-concat serves to identify patterns of dissimilarity based on lexical semantics, separate from any representation of the relation between the two words within each pair.[1]

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 dissimilarity of relations is assessed by the cosine distance between the two difference vectors that form the analogy. This model, which has been directly applied to analogy problems in work on machine learning (Zhila et al., 2013), codes relations only implicitly (i.e., as a difference vector computed from individual words). Word2vec-diff is able to solve some verbal analogy problems based on relatively specific relations (e.g., *king*:*queen*::*man*:*woman*), although its success is limited (Linzen, 2016). To the best of our knowledge, the model has not been tested with analogies based on abstract semantic relations of the sort used in this study.

The third computational model, Bayesian Analogy with Relational Transformations (BART; Lu, Wu, & Holyoak, 2019; Lu, Chen, & Holyoak, 2012), assumes that specific semantic relations between words are coded as distributed representations over a set of abstract relations, specified in a taxonomy founded on linguistic and psychological evidence (Bejar, Chaffin, & Embretson, 1991). This taxonomy includes 10 general types of relations (e.g., similar, contrast, cause–purpose), each of which has several subtypes, resulting in 79 semantic relations. BART is trained with a small number of word pairs (~20 pairs) as positive

examples of each specific relation in the taxonomy (Jurgens et al., 2012). After learning the set of 79 abstract relations from example word pairs coded as semantic vectors derived from Word2vec, for any word pair, BART can estimate the probability that the word pair instantiates each learned relation, which constitutes a distributed representation of the specific relation between the two words. BART's relation vectors enable computations of second-order relational dissimilarity between word pairs, providing a direct basis for solving verbal analogies in the form *A*:*B*::*C*:*D* (e.g., *old*:*young*::*hot*:*cold*). Behavioral evidence indicates that BART can solve a set of simple verbal analogies with a degree of accuracy comparable to humans (Lu et al., 2019).

We employed a sequential event-related fMRI design (DeWolf, Chiang, Bassok, Holyoak, & Monti, 2016), in which participants judged the validity of *A:B*::*C:D* analogies involving three types of abstract relations (similar, contrast, and cause–purpose). This design aimed to separate the construction of first-order relations (i.e., relations between words in a pair) from the second-order assessment of dissimilarity between relations (i.e., the degree of analogical match between *A:B* and *C:D* relations).

To test the neural plausibility of the three computational models, we analyze the *A:B* and *C:D* phases of each analogy with a (dis)similarity analysis assessing the degree to which each computational model matches the observed neural representations of first-order relations (i.e., *A:B*) and the observed neural responses to second-order relational distance (i.e., *A:B* vs. *C:D*). The *A:B* phase provides a relatively pure measure of neural activity involved in coding the individual *A* and *B* words and the *A:B* relation. To arbitrate between the three alternative models (as well as a baseline model based on relation types alone), we probe the representation of this relation using first a multivariate decoding analysis, followed by a multivariate representational similarity analysis (RSA). The *C:D* phase includes the neural computation required to compare the two relations (as well as neural activity required to maintain the *A:B* relation and to represent the *C:D* relation). We examine this representation using a voxelwise correlation analysis, assessing the degree to which hypothesized second-order relational distance resembles neural activity. If semantic relations have distributed representations based on the taxonomy of abstract relations, we should find brain regions in which BART is the best predictor of neural similarity. In contrast, if relations are coded as atomic units, then similarity of two word pairs will only depend on whether they instantiate the same or different relation types.

METHODS

Participants

Sixteen participants (eight women) were recruited at the University of California, Los Angeles (UCLA) through a flyer distributed in the Psychology Department. Participants signed informed consent before the experimental session and were paid $50 for their participation in the 1-hr study, in compliance with the procedures accepted by the local institutional review board. The study was approved, including informed consent procedures, by the UCLA Office of the Human Research Protection Program.

Stimuli and Design

The stimuli were a set of analogy problems constructed from word pairs taken from a normed set of examples of abstract relations (Jurgens et al., 2012). The full norms include examples of word pairs instantiating 10 general types of relations, each including 5–10 more specific relations, for a total of 79 distinct relations. For this study, we focused on three general relation types (chosen as especially familiar) with three specific relations drawn from each, for a total of nine relations: similar (synonym, attribute similarity, change), contrast (contrary, directional, pseudoantonym), and cause–purpose (cause:effect, cause:compensatory action, activity:goal).

For each relation, we selected 16 word pairs high in typicality as assessed by human judgments (Jurgens et al., 2012), yielding 48 word pairs per relation type and 144 pairs in total. Examples of the word pairs used are shown in Table 1. In selecting word pairs to construct analogy problems, 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

Table 1. Examples of Word Pairs Used to Generate Analogy Problems, Organized by General Relation Type and Subtype

Similar		
Synonym	*Attribute Similarity*	*Change*
big:large	book:magazine	acceleration:speed
boat:ship	chair:sofa	darken:color
Contrast		
Contrary	*Directional*	*Pseudoantonym*
accept:reject	ahead:behind	bright:dull
big:small	below:above	day:evening
Cause–purpose		
Cause:effect	*Cause:compensatory action*	*Activity:goal*
accident:damage	anger:yell	advertise:promote
bath:cleanliness	coldness:shiver	cook:eat

passed our selection criteria, we also included some pairs that had been used as "seed" examples to elicit word pairs from humans (Jurgens et al., 2012). These were considered excellent examples.

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 of 72 analogy problems, for each of the nine specific relations, 8 of the 16 pairs were assigned to the *A*:*B* role and four were assigned 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 four-term analogy problems were created by random pairing of word pairs. For each specific relation, four problems were valid and four were invalid. Within a set of 72 analogy 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. Across four sets of problems, 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 × 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 2 sec and then the *C*:*D* pair for 2 sec (with an average 0.5-sec 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. Participants were instructed that all word pairs represented a meaningful relation but were not made aware of the structure of the relations. The *A*:*B* phase provided a measure of neural activity involved in coding the individual *A* and *B* words and the *A*:*B* relation. The *C*:*D* phase included the neural computation required to compare the two relations (as well as neural activity required to maintain the *A*:*B* relation and to represent the *C*:*D* relation).

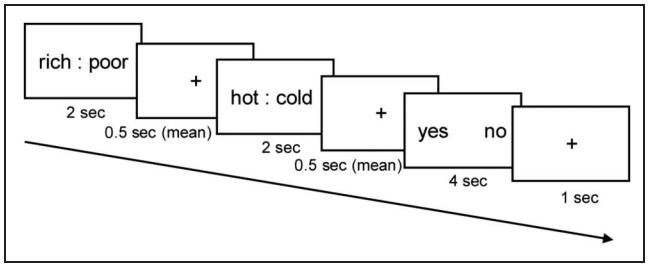

Figure 1. Timing of events on each trial. In a rapid event-related fMRI design, healthy young adults were asked to evaluate two pairs of semantic concepts. Participants were shown two word pairs, first an *A*:*B* pair for 2 sec and then a *C*:*D* pair for 2 sec after a jitter, and finally a cue to make 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. The *A*:*B* phase provides a relatively pure measure of neural activity involved in coding the *A*:*B* relation. The *C*:*D* phase includes the neural computation required to compare the two relations (as well as neural activity required to maintain the *A*:*B* relation and to represent the *C*:*D* relation).

fMRI Data Acquisition

Data were acquired on a 3-T Siemens PRISMA MRI scanner at the OneMind Staglin IMHRO Center for Cognitive Neuroscience at UCLA. Structural data were acquired using a T1-weighted sequence (magnetization prepared rapid gradient echo, repetition time = 1900 msec, echo time = 2.26 msec, voxel size = 1 mm^3 isovoxel). BOLD data were acquired with a T2*-weighted gradient recall echo sequence (repetition time = 1000 msec, echo time = 37 msec, 60 interleaved slices [2-mm gap], voxel size = 2 × 2 × 2 mm, 6× multiband acceleration).

fMRI Preprocessing

Data preprocessing was carried out using FMRIB Software Library (FSL; Jenkinson, Beckmann, Behrens, Woolrich, &

Smith, 2012). Preprocessing steps included motion correction, slice-timing correction (using Fourier-space time-series phase shifting), spatial smoothing using a Gaussian kernel of 5-mm FWHM, and high-pass temporal filtering (Gaussian-weighted least-squares straight line fitting, with $\sigma = 50.0$ sec). Spatial smoothing was omitted from the above preprocessing steps for all analyses (RSA and voxelwise correlation) to preserve spatial heterogeneities.

Beta-series (Rissman, Gazzaley, & D'Esposito, 2004) parameter estimates were derived using the least-squares separate approach (Mumford, Turner, Ashby, & Poldrack, 2012). The least-squares separate algorithm iteratively estimates parameters for an event using a general linear model (GLM) including a regressor for that event as well as another regressor for all other events. This procedure was used to estimate beta parameters for all *A*:*B* and *C*:*D* word pairs, which were used as features in the dissimilarity and voxelwise correlation analyses.

Dissimilarity Analyses

RSA (Nili et al., 2014; Kriegeskorte & Kievit, 2013; Kriegeskorte, Mur, & Bandettini, 2008; Kriegeskorte, Goebel, & Bandettini, 2006) was used to characterize the similarities of neural responses across pairs during the *A*:*B* phase. 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 (see Figure 2 for intercorrelations among RDMs). 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. Resulting correlation values were registered to the Montreal Neurological Institute template for group analysis, and statistical significance of positive values was assessed using FSL *randomise* (Winkler, Ridgway, Webster, Smith, & Nichols, 2014; Smith & Nichols, 2009). 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).

For the *C*:*D* phase (second-order relation comparison), a univariate dissimilarity analysis was performed. All the models of analogical comparison considered in this article make the general prediction that the conceptual difficulty of deciding the validity of an analogy will be related to the relation-based dissimilarity of the *A*:*B* and *C*:*D* word pairs, with greater dissimilarity making the decision more difficult. In this analysis, only trials consisting of valid analogies

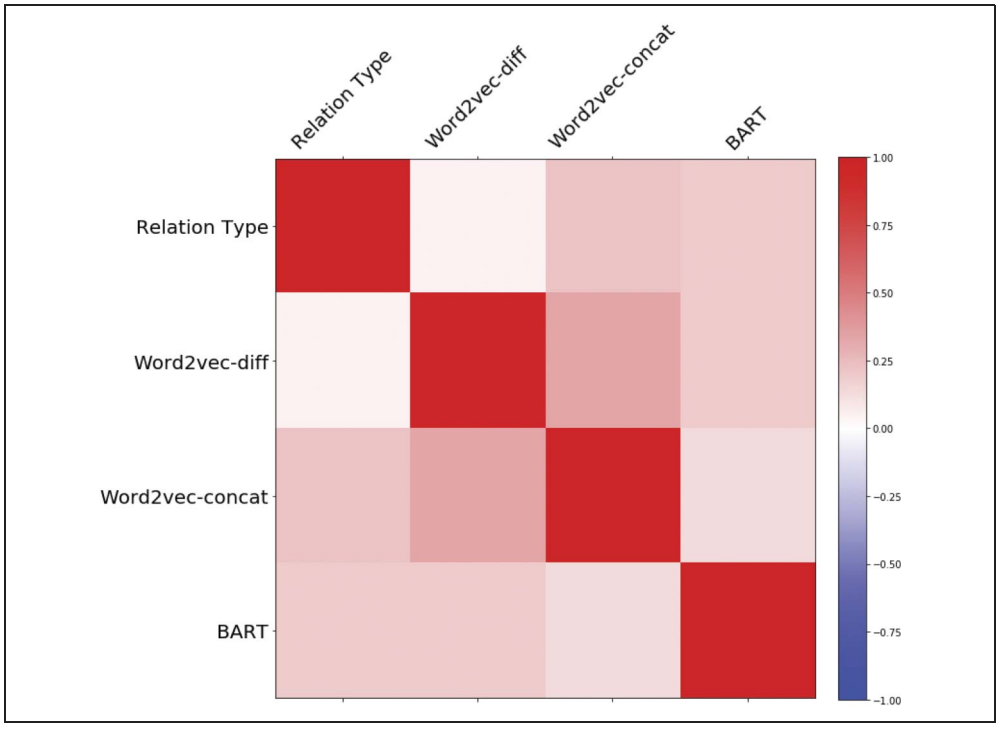

Figure 2. Correlations among different theoretical RDMs.

(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 article, 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 voxel 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 at $q = 0.05$.

Data and Code Availability

Raw and preprocessed NIFTI files, as well as experiment timing files, are available at openneuro.org. Code for the BART model can be downloaded from cvl.psych.ucla.edu/BART2code.zip. Code for the experiment and all custom analyses can be found at github.com/njchiang/analogy-fmri.

RESULTS

Behavioral Results

Mean proportion correct in solving analogy problems was 0.82 ($SD = 0.07$) across all conditions, an accuracy level well above chance ($p < .001$). A repeated-measures ANOVA was conducted on performance accuracy across the three abstract relation types for *A*:*B*. Problems using the "contrast" relation yielded the highest accuracy ($M = 0.87$, $SD = 0.08$), followed by "cause–purpose" ($M = 0.82$, $SD = 0.07$) and "similar" ($M = 0.78$, $SD = 0.09$). Bonferroni-corrected t tests indicated that "contrast" problems were more accurate than either of the other relation types ($p < .005$).

RT was calculated as the time from the appearance of the response cue (i.e., "yes" and "no" indicators after the *C*:*D* phase) to the button press. Only RTs for accurate trials were analyzed. Mean RT was 913 msec ($SD = 255$ msec) across all relation types. No reliable RT differences were found among the three types.

To examine how well the different models can account for human behavioral performance, we derived predictions of accuracy on the analogy task for each of the three models. For BART (which is based on vectors of probability values), we applied a nonlinear cubic power transformation to down-weight contributions to the decision stage from the large number of dimensions with low probabilities. As the human task involved yes/no judgments, a decision module is required to derive such judgments from the vectors produced by each model. We used the same decision module for all three models. For each model, relational dissimilarity between the *A*:*B* and *C*:*D* word pairs in an experimental trial was calculated using cosine distance between the vectors. Each model's yes/no response was determined by whether the cosine distance was less/greater than a decision threshold. This threshold was selected by a search to maximize each model's accuracy.

The BART model yielded mean accuracy of 0.823, very similar to human-level performance (0.822). Both the Word2vec-diff and Word2vec-concat models yielded accuracy levels near chance (0.576 and 0.583, respectively), substantially lower than human performance. At the level of individual relation types, the BART model yielded a proportion correct of 0.802 for "similar," 0.896 for "contrast," and 0.708 for "cause–purpose." BART's accuracy was close to the human level for the former two relations but less accurate than human performance for the "cause–purpose" relation, suggesting that humans may benefit from a deeper understanding of causal relations (e.g., knowledge of how causality is related to interventions; Waldmann, 2017).

Univariate Analyses of *A*:*B* and *C*:*D* Phases

We first performed univariate analyses to identify the brain regions active during the *A*:*B* and *C*:*D* phases of each trial (including both valid and invalid trials). The general relation type was coded separately for each phase. 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 noninterest, as well as motion parameters. The GLM analysis was carried out using FSL FEAT (Jenkinson et al., 2012; Smith et al., 2004). Before univariate analysis, 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 FWHM, and high-pass temporal filtering (Gaussian-weighted least-squares straight line fitting, with $\sigma = 50.0$ sec). Data from individual runs were aggregated employing a mixed effects model (i.e., employing both within- and between-participant variance) and using automatic outlier detection. Statistical significance for univariate analyses was assessed using FSL *randomise* with threshold-free cluster enhancement (TFCE) cluster correction (Winkler et al., 2014; Smith & Nichols, 2009).

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 Figure 3A). 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* phases shared activations in the inferior lateral occipital cortex (BA 19),

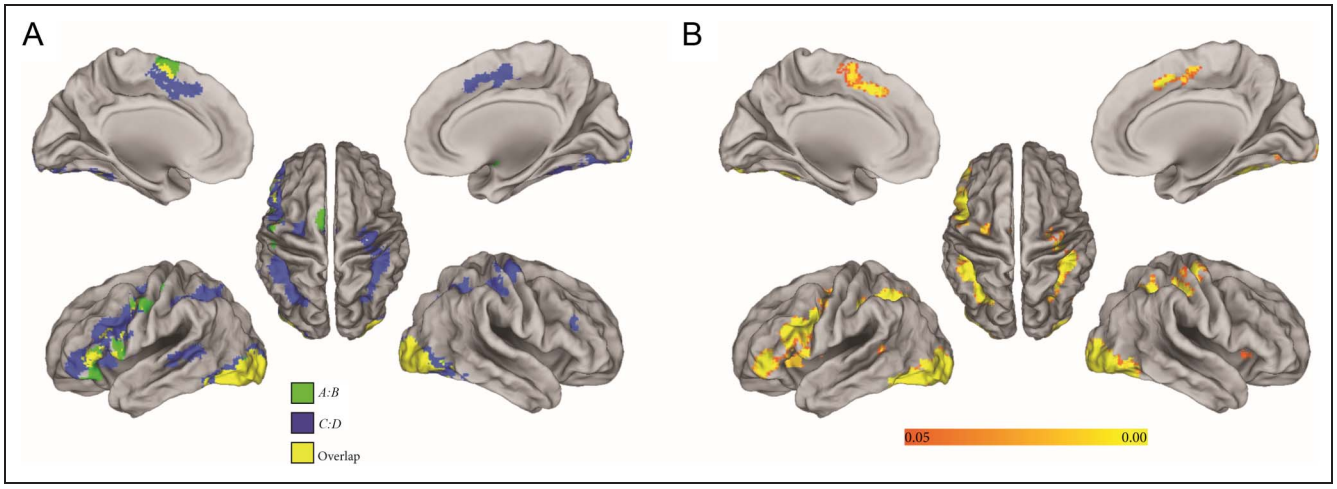

Figure 3. Univariate analysis results. (A) Main effects of the *A*:*B* and *C*:*D* phases of trials. Clusters were obtained by contrasting each phase (i.e., *A*:*B* and *C*:*D*) to simple fixation. (B) *C*:*D* − *A*:*B* univariate contrast. Regions in which activity while reading the *C*:*D* word pair was greater than when reading the *A*:*B* word pair. Depicted group-level activations were obtained with a nonparametric permutation approach (FSL *randomise*); significance was set at $p = .05$ family-wise error rate, cluster-corrected with TFCE (Smith & Nichols, 2009).

fusiform gyrus (BA 37), and left frontal regions spanning the RLPFC (BAs 10 and 47). In addition, processing *C*:*D* word pairs uniquely led to a greater BOLD response in the left inferior frontal gyrus (pars triangularis, pars opercularis; BAs 44 and 45) as well as bilateral superior parietal cortex (inBA 7).

As shown in Figure 3B, the univariate comparison of *C*:*D* minus *A*:*B* revealed a frontoparietal network, mainly left lateralized. Specifically, the contrast uncovered significant clusters in the left RLPFC (BAs 10 and 47), replicating prior results implicating this region in complex relational comparisons (Bunge et al., 2009; Christoff et al., 2001), as well as in the left inferior frontal gyrus (BAs 44 and 45) and bilateral posterior parietal (BA 7) and occipital (BA 19) cortices.

Decoding Neural Activity Patterns to Classify Relation Types

To characterize the representations of abstract semantic relations in the brain, we conducted a multivariate pattern analysis (MVPA; e.g., Haxby et al., 2001) using a searchlight method (Kriegeskorte et al., 2006). Classifiers were trained to distinguish between the three general relation types (similar, contrast, and 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 5-mm-radius sphere and a linear SVM (Abraham et al., 2014; Pedregosa et al., 2011).

This MVPA revealed left-lateralized areas of the brain capable of distinguishing different types of abstract relations on the basis of activation patterns across both the *A*:*B* and *C*:*D* phases. In addition, during the second-order comparison (i.e., the *C*:*D* phase), the three abstract relations could also be distinguished in the left rostrolateral and right frontotemporal cortices.[2] As shown in Figure 4, distributed areas of the brain are involved in decoding semantic relation types (similar, contrast, and cause–purpose). Areas in color achieved above-chance classification performance ($p < .01$), as assessed by a Wilcoxon signed-rank test with TFCE cluster correction (Smith & Nichols, 2009).

During the *A*:*B* phase, the active regions for relation classification include frontal and temporal cortices (most pronounced in the left hemisphere) as well as bilateral parietal cortices. During the *C*:*D* phase, the three relation types can also be distinguished in many of the same regions and also 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 the *A*:*B* and *C*:*D* phases (areas in yellow in Figure 4) includes areas previously

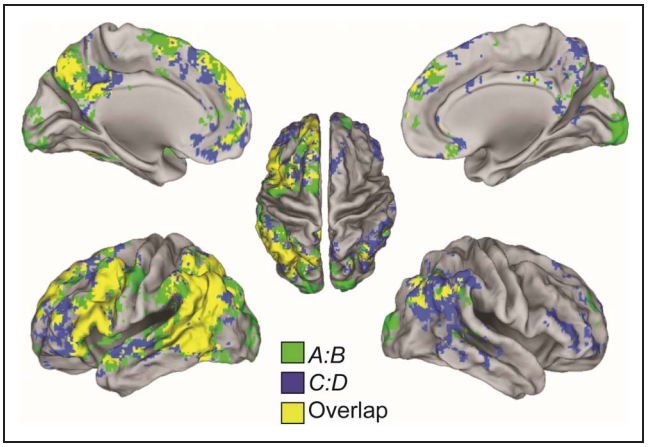

Figure 4. MVPA searchlight results. Regions in which the three general semantic relations could be discriminated above chance during different phases of the analogy task (corrected $p < .01$, assessed using FSL *randomise* with TFCE cluster correction for multiple comparisons).

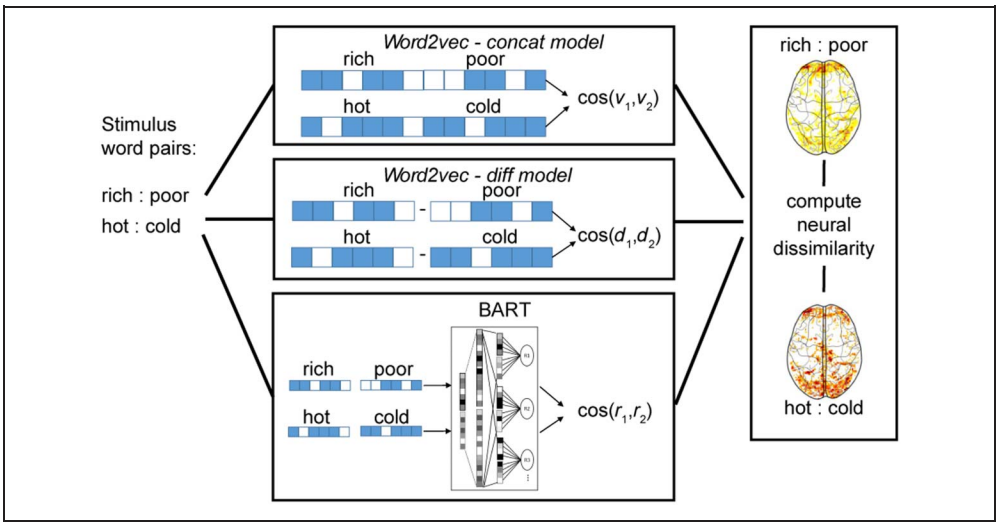

Figure 5. RSA approach to discovering neural signatures of specific relations. For any two word pairs shown during the *A*:*B* phase (e.g., *rich*:*poor*, *hot*:*cold*), three alternative models are used to predict dissimilarity based on the cosine distance between the representations of each individual 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, and 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 to arbitrate between alternative models.

proposed to underlie the semantic representation system for individual words (Carota et al., 2017; de Heer, Huth, Griffiths, Gallant, & Theunissen, 2017; Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016; Binder, Desai, Graves, & Conant, 2009). The MVPA also highlights the important role of parietal regions associated more specifically with relational reasoning (Wendelken, 2015).

First-order Relations (*A*:*B* Phase)

The MVPA reported above involved training classifiers to use neural activity to distinguish among relation types, making use of any and all properties of individual words and/or relations that may reliably influence brain signals, without any guidance from computational models. We then moved on to perform analyses that use computational models to predict neural activity, with a particular focus on alternative representations of relations per se. To assess and contrast the neural plausibility of the first-order relational representations specified by each of the three models (Word2vec-diff, Word2vec-concat, and BART), we performed an RSA (Kriegeskorte et al., 2008). Specifically, we compared the matrix of trial-by-trial dissimilarity across word pairs derived from the BOLD signal during the *A*:*B* phase (i.e., the empirical RDM) to that predicted by each of the three computational models (see Figure 5). We also included in the analysis a fourth relation-type model (i.e., the design matrix) to serve as a simple baseline model distinguishing the three general relation types (i.e., similar, contrast, and cause–purpose; see Figure 6).

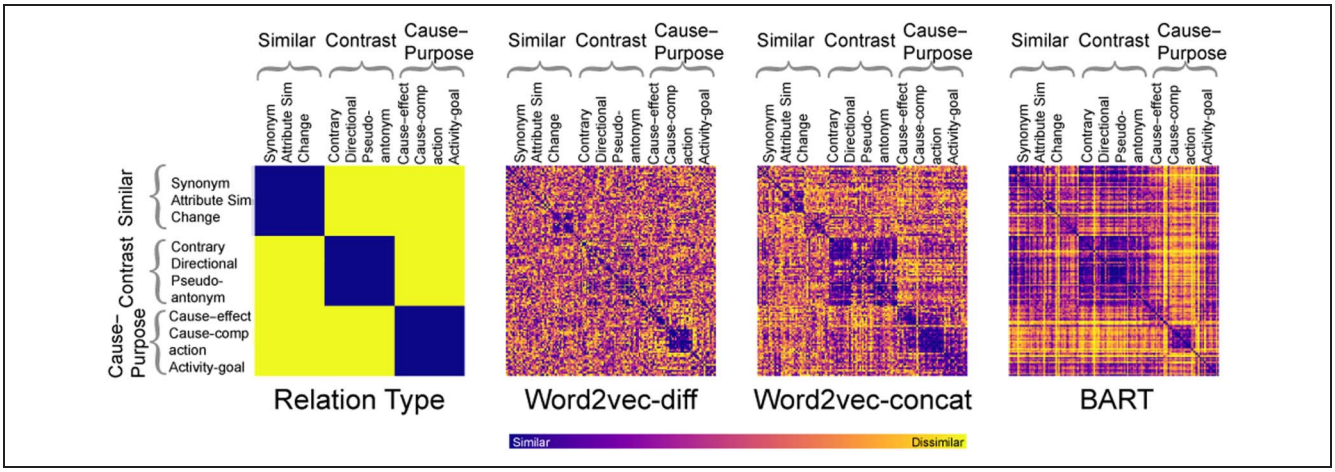

Figure 6. Theoretical RDMs. The RDMs derived from the three computational models are of size 144 × 144 (i.e., based on individual word pairs). Theoretical RDMs capturing the cosine distance between the vector representation for each word pair were correlated with empirical RDMs derived from brain activity patterns.

We performed RSAs using a whole-brain searchlight approach. Among the four models that were tested, only the RDM derived from BART yielded significant correlations with neural RDMs (Figure 7A). These correlations primarily involved the left superior parietal lobe and left intraparietal sulcus. In approximately the same regions, the correlation for BART was significantly greater than those for either of the other computational models (Figure 7B) or for the relation-type baseline model (Figure 7C).

Second-order Relational Processing (*C*:*D* Phase)

To investigate second-order relational comparisons, we performed a form of dissimilarity analysis in which we contrasted the three models by calculating, for each, a measure of relational dissimilarity between first-order relations. BART and the two Word2vec models make the general prediction that the conceptual difficulty of identifying a valid analogy is proportional to the (word or relation-based) dissimilarity of the *A*:*B* and *C*:*D* word pairs, with greater dissimilarity making the analogy harder to verify. Specifically, 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 relational dissimilarity scores derived from each model were then correlated (using Spearman's rho) with observed activity during the *C*:*D* phase of each valid analogy. (The relation-type model was inapplicable because valid analogies by definition have the same relation for *A*:*B* and *C*:*D*.) This *C*:*D*-phase analysis was conducted using a whole-brain approach. For each voxel, we computed the rank correlation between voxel activity and relational dissimilarity.

Only BART yielded significant correlations as assessed by *randomise* with TFCE correction. Relational dissimilarity measures as calculated by BART correlated with voxel activity in left-lateralized frontal, temporal, and parietal sites (see Figure 8). Specifically, BART correlated with voxel activity within ventrolateral pFC including the pars opercularis, triangularis, and orbitalis of the inferior frontal gyrus; dorsolateral pFC spanning BAs 8, 9, and 46 in the middle frontal gyrus; and RLPFC in the BA 10 portions of the inferior and middle frontal gyri, as well as in lateral premotor cortex in BA 6, and in the medial aspect of the superior frontal gyrus in BAs 6, 8, and 9. In addition, significant correlations were also detected in parietal areas in the intraparietal sulcus spanning BAs 40 and 7 and in temporal areas spanning BAs 21 and 22.

As a follow-up, a semipartial correlation analysis was performed to test whether BART captured additional information relative to the Word2vec-derived models. Relational dissimilarity scores derived from Word2vec-concat and Word2vec-diff were first regressed out of fMRI-based dissimilarity scores, and the resulting residuals were then correlated with relational dissimilarity predictions derived from BART. The same procedure was performed with the group-averaged trial-by-trial accuracy, to examine the effect of task difficulty. Essentially, the same areas shown in Figure 8 (left frontoparietal network as well as temporal regions) exhibited statistically significant semipartial correlations with BART as assessed by *randomise* with

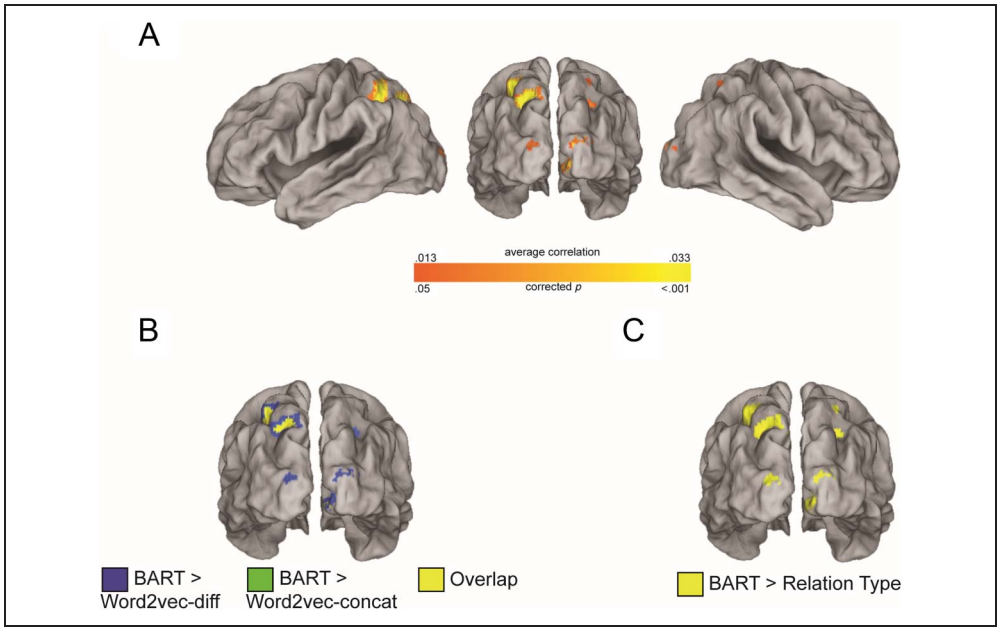

Figure 7. Searchlight results for RSAs testing alternative models as predictors of neural dissimilarity during the *A*:*B* phase for 144 word pairs instantiating abstract semantic relations. (A) Lateral and posterior views of areas in which the BART model based on distributed relation representations was significantly correlated with neural RDM. None of the other three models yielded areas with significant correlations. (B) Posterior view of areas in which correlation of BART with neural RDM was significantly greater than correlation for each of the alternative computational models. (C) Posterior view of areas in which correlation of BART with neural RDM was significantly greater than that for the baseline model, which assumes discrete codes for relations. Colored regions represent searchlight sphere centers that were significant as assessed by FSL *randomise* with TFCE cluster correction for multiple comparisons (corrected $p < .05$).

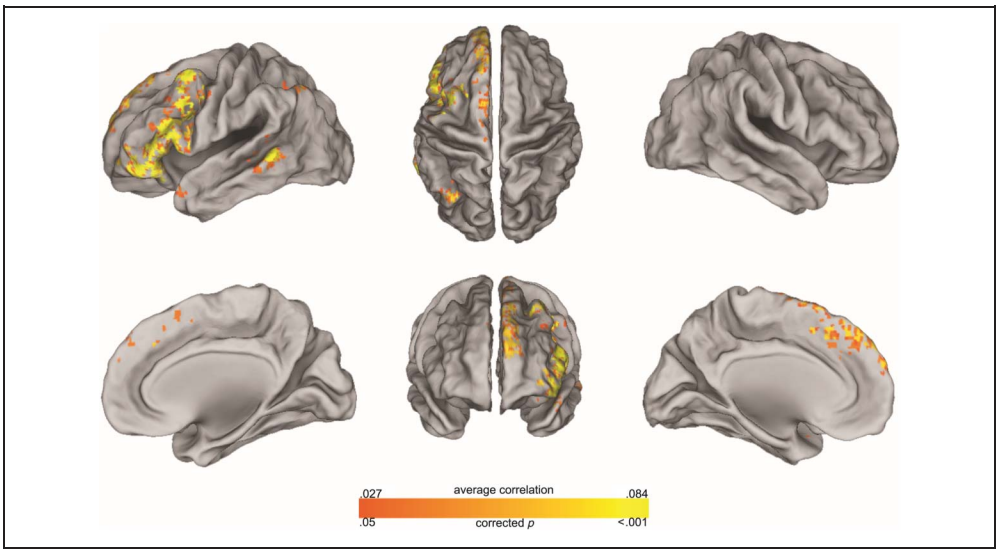

Figure 8. Correlation between model-derived relational dissimilarity between *A*:*B* and *C*:*D* relations and trial parameter estimates during the *C*:*D* phase. Average Spearman correlation between BART-derived relational dissimilarity between *A*:*B* and *C*:*D* relations and trial parameter estimates during the *C*:*D* phase. Only the relational dissimilarity measure predicted by BART was significantly correlated with trial parameter estimates. Regions significantly correlated with BART also show significant correlations after accounting for variance accounted for by both the Word2vec models and also after accounting for trial difficulty (estimated by mean accuracy). Significance was assessed using FSL *randomise* with TFCE cluster correction for multiple comparisons (corrected $p < .05$). Top row: lateral and dorsal views. Bottom row: medial and anterior views.

TFCE correction. The reverse analysis was also performed. No regions showed a significant impact of the Word2vec models after controlling for variance predicted by BART.

Finally, a similar semipartial correlation analysis was performed controlling for trial-by-trial accuracy (mean accuracy for each item across all participants), again for valid analogies only. Rather than solely reflecting the conceptual difficulty of identifying a valid analogy, mean accuracy is a coarser measure of overall task difficulty, because errors may arise at multiple processing stages (e.g., word identification or motor responses). After partialing out the variance predicted by accuracy, the same areas shown in Figure 8 exhibited statistically significant semipartial correlations with BART as assessed by *randomise* with TFCE correction. The reverse analysis yielded no areas that were reliably predicted by accuracy after controlling for the variance predicted by BART.

DISCUSSION

This study combined computational modeling with neuroimaging to investigate the representation and comparison of abstract semantic relations in the brain. We used a sequential presentation of verbal analogies with clear temporal phases to examine the neural activity associated with (1) representing the individual words in a pair and the relation between them and (2) comparing two first-order relations (while also separating these high-level reasoning processes from planning for a motor response). By testing alternative computational models of relational dissimilarity, we were able to distinguish between rival accounts of how semantic relations are coded and compared in the brain. The BART model, which postulates that semantic relations between words are coded as a distributed representation based on a taxonomy of abstract relations, was able to predict patterns of neural activity during analogical reasoning that could not be explained by alternative models. During the phase in which a single relation is being encoded (*A*:*B* phase), the BART model was the most effective predictor of patterns of neural activity in the left superior parietal cortex, a region previously associated with relation representation (Wendelken, 2015; Wendelken et al., 2008). During the phase in which relations are compared to verify whether the analogy is valid (*C*:*D* phase), BART was the most effective predictor of neural activity in multiple prefrontal ROIs, including areas in the left RLPFC previously linked to higher-order relational comparisons (Green et al., 2010; Bunge et al., 2005, 2009). Although the left RLPFC has received the most attention in the literature, the present findings are consistent with previous evidence that analogical reasoning depends on a broader left frontoparietal network (for a review, see Holyoak & Monti, this issue).

The present findings support three major conclusions. First, analogical reasoning depends on fine coding of semantic relations with distributed representations, primarily supported by the left superior parietal cortex. Second, the content of these distributed representations can be learned from a small number of examples instantiating abstract relations, as operationalized in a computational model, BART. This model not only accounts for behavioral accuracy in solving verbal analogy problems but also yields measures of item-level relation dissimilarity that correlate with dissimilarity of neural responses in frontoparietal regions. This evidence for distributed coding of relations is inconsistent with models of analogical reasoning based on localist relation representations (e.g., Forbus et al., 2017). Third, during verification of an analogy,

neural activities in frontal areas (including the left RLPFC) as well as parietal and temporal regions exhibit a graded response to the degree of relational dissimilarity between the two word pairs forming the analogy. The graded neural responses in analogy-selective frontal regions can be predicted by the BART model based on its distributed relation representations.

The fact that dissimilarity measures derived from the BART model yielded stronger and more reliable predictions of relational processing—for both first- and second-order relations—than did the Word2vec-diff model is consistent with computational evidence favoring the former model as an account of human relational judgments (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 (Popov et al., 2017). Whereas Word2vec-diff provides only a generic and implicit representation of relational dissimilarity (i.e., the difference vector between semantic vectors for two words), BART learns representations of individual semantic relations, which in the context of analogical reasoning collectively provide a distributed representation of the relations(s) linking any word pair. The neural evidence favoring BART as a model of relation dissimilarity thus supports the hypothesis that the brain encodes semantic relations between words as distributed representations across abstract semantic relations, such as the specific relations "synonym," "antonym," and "cause–effect." By coupling computational modeling with analyses of dissimilarity in neural activity, it proved possible to resolve a major theoretical issue concerning the representation of semantic relations.

This study focused on abstract semantic relations. These are particularly important because a pool of abstract relations may provide basic elements that can be used to represent more specific relations. 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 item-level analyses of neural patterns. 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 activity using both RSA and other analytic techniques, such as neural encoding analyses (Huth et al., 2016; Mitchell et al., 2008). This research strategy shows promise in decoupling component processes and in identifying specific representations underlying high-level reasoning. Future work should aim to develop and test well-specified computational models of how propositions and larger knowledge units are represented in the brain and used to reason.

Acknowledgments

Preparation of this article was supported by National Science Foundation grant BCS-1827374 to H. L. and K. J. H., a UCLA Academic Senate grant to K. J. H., and a Department of Defense 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), and an earlier version of the article was posted on bioRxiv in April 2019.

Reprint requests should be sent to Jeffrey N. Chiang, Department of Psychology, University of California, Los Angeles, 405 Hilgard Ave., Los Angeles, CA 90095-1563, or via e-mail: njchiang@ucla.edu.

Notes

1. We also tested a similar model, Word2vec-sum (Pereira et al., 2018; Mikolov et al., 2013), which aggregates a word pair via vector addition rather than concatenation. All results obtained using Word2vec-sum were virtually identical to those based on Word2vec-concat; hence, we only report results for Word2vec-concat.
2. Significant decoding ability was also observed in early visual cortex. Although visual properties of the stimuli were not precisely quantified in this experiment, we ran an RSA using number of characters in word pairs as a proxy for perceptual differences. We observed a significant correlation with this word length model in the early visual cortex region.

REFERENCES

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., et al. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*, 14. **DOI:** https://doi.org/10.3389/fninf.2014.00014, **PMID:** 24600388, **PMCID:** PMC3930868

Bejar, I. I., Chaffin, R., & Embretson, S. E. (1991). *Cognitive and psychometric analysis of analogical problem solving*. New York: Springer-Verlag. **DOI:** https://doi.org/10.1007/978-1-4613-9690-1

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*, 2767–2796. **DOI:** https://doi.org/10.1093/cercor/bhp055, **PMID:** 19329570, **PMCID:** PMC2774390

Bunge, S. A., Helskog, E. H., & Wendelken, C. (2009). Left, but not right, rostrolateral prefrontal cortex meets a stringent test of the relational integration hypothesis. *Neuroimage*, *46*, 338–342. **DOI:** https://doi.org/10.1016/j.neuroimage.2009.01.064, **PMID:** 19457362, **PMCID:** PMC2864011

Bunge, S. A., Wendelken, C., Badre, D., & Wagner, A. D. (2005). Analogical reasoning and prefrontal cortex: Evidence for separable retrieval and integration mechanisms. *Cerebral Cortex*, *15*, 239–249. **DOI:** https://doi.org/10.1093/cercor/bhh126, **PMID:** 15238433

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*, 294–309. **DOI:** https://doi.org/10.1093/cercor/bhw379, **PMID:** 28077514, **PMCID:** PMC6044349

Chan, J., & Schunn, C. (2015). The impact of analogies on creative concept generation: Lessons from an *in vivo* study in engineering design. *Cognitive Science*, *39*, 126–155. **DOI:** https://doi.org/10.1111/cogs.12127, **PMID:** 24835377

Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., et al. (2001). Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *Neuroimage*, *14*, 1136–1149. **DOI:** https://doi.org/10.1006/nimg.2001.0922, **PMID:** 11697945

Coleridge, S. T. (1810/1969). *The collected works of Samuel Taylor Coleridge, Volume 4 (Part II)*. Princeton, NJ: Princeton University Press. **DOI:** https://doi.org/10.1515/9781400874965

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240–247. **DOI:** https://doi.org/10.1016/S0022-5371(69)80069-1

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*, 6539–6557. **DOI:** https://doi.org/10.1523/JNEUROSCI.3267-16.2017, **PMID:** 28588065, **PMCID:** PMC5511884

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. **DOI:** https://doi.org/10.1016/j.neuroimage.2016.07.052, **PMID:** 27474523

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*, 1–43. **DOI:** https://doi.org/10.1037/0033-295X.115.1.1, **PMID:** 18211183

Dunbar, K. N., & Klahr, D. (2012). Scientific thinking and reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 701–718). New York: Oxford University Press. **DOI:** https://doi.org/10.1093/oxfordhb/9780199734689.013.0035

Etzel, J. A., & Braver, T. S. (2013). MVPA permutation schemes: Permutation testing in the land of cross-validation. In *2013 International Workshop on Pattern Recognition in Neuroimaging* (pp. 140–143). Philadelphia, PA: IEEE. **DOI:** https://doi.org/10.1109/PRNI.2013.44

Forbus, K. D., Ferguson, R. W., Lovett, A., & Gentner, D. (2017). Extending SME to handle large-scale cognitive modeling. *Cognitive Science*, *41*, 1152–1201. **DOI:** https://doi.org/10.1111/cogs.12377, **PMID:** 27322750

Green, A. E., Kraemer, D. J. M., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2010). Connecting long distance: Semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral Cortex*, *20*, 70–76. **DOI:** https://doi.org/10.1093/cercor/bhp081, **PMID:** 19383937

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*, 2425–2430. **DOI:** https://doi.org/10.1126/science.1063736, **PMID:** 11577229

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*, 1953–1969. **DOI:** https://doi.org/10.1002/hbm.23149, **PMID:** 27012301, **PMCID:** PMC6867453

Holyoak, K. J. (2008). Relations in semantic memory: Still puzzling after all these years. In M. A. Gluck, J. R. Anderson, & S. M. Kosslyn (Eds.), *Memory and mind: A festschrift for Gordon H. Bower* (pp. 141–158). Mahwah, NJ: Erlbaum.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 234–259). New York: Oxford University Press. **DOI:** https://doi.org/10.1093/oxfordhb/9780199734689.001.0001

Holyoak, K. J. (2019). *The spider's thread: Metaphor in mind, brain, and poetry*. Cambridge, MA: MIT Press. **DOI:** https://doi.org/10.7551/mitpress/11119.001.0001

Holyoak, K. J., & Monti, M. M. (this issue). Relational integration in the human brain: A review and synthesis.

Hummel, J. E., & Holyoak, K. J. (2005). Relational reasoning in a neurally-plausible cognitive architecture: An overview of the LISA project. *Current Directions in Psychological Science*, *14*, 153–157. **DOI:** https://doi.org/10.1111/j.0963-7214.2005.00350.x

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*, 453–458. **DOI:** https://doi.org/10.1038/nature17637, **PMID:** 27121839, **PMCID:** PMC4852309

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *Neuroimage*, *62*, 782–790. **DOI:** https://doi.org/10.1016/j.neuroimage.2011.09.015, **PMID:** 21979382

Jones, M. N., Willits, J., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford handbook of computational and mathematical psychology* (pp. 232–254). New York: Oxford University Press. **DOI:** https://doi.org/10.1093/oxfordhb/9780199957996.013.11

Jurgens, D. A., Turney, P. D., Mohammad, S. M., & Holyoak, K. J. (2012). SemEval-2012 Task 2: Measuring degrees of relational similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics* (pp. 356–364). Montreal, Canada: Association for Computational Linguistics.

Kounios, J., Montgomery, E. C., & Smith, R. W. (1994). Semantic memory and the granularity of semantic relations: Evidence from speed–accuracy decomposition. *Memory & Cognition*, *22*, 729–741. **DOI:** https://doi.org/10.3758/BF03209258, **PMID:** 7808282

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences, U.S.A.*, *103*, 3863–3868. **DOI:** https://doi.org/10.1073/pnas.0600244103, **PMID:** 16537458, **PMCID:** PMC1383651

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*, 401–412. **DOI:** https://doi.org/10.1016/j.tics.2013.06.007, **PMID:** 23876494, **PMCID:** PMC3730178

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. **DOI:** https://doi.org/10.3389/neuro.06.004.2008, **PMID:** 19104670, **PMCID:** PMC2605405

Linzen, T. (2016). Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP* (pp. 13–18). Berlin: Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W16-2503

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*, 617–648. **DOI:** https://doi.org/10.1037/a0028719, **PMID:** 22775500

Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences, U.S.A.*, *116*, 4176–4181. **DOI:** https://doi.org/10.1073/pnas.1814779116, **PMID:** 30770443, **PMCID:** PMC6410800

Meyer, D. E. (1970). On the representation and retrieval of stored semantic information. *Cognitive Psychology*, *1*, 242–299. **DOI:** https://doi.org/10.1016/0010-0285(70)90017-4

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.),

Proceedings of the 26th International Conference on Neural Information Processing Systems (pp. 3111–3119). Red Hook, NY: Curran Associates, Inc.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., et al. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*, 1191–1195. **DOI:** https://doi.org/10.1126/science.1152876, **PMID:** 18511683

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*, *59*, 2636–2643. **DOI:** https://doi.org/10.1016/j.neuroimage.2011.08.076, **PMID:** 21924359, **PMCID:** PMC3251697

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, *10*, e1003553. **DOI:** https://doi.org/10.1371/journal.pcbi.1003553, **PMID:** 24743308, **PMCID:** PMC3990488

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

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 10. **DOI:** https://doi.org/10.3389/neuro.11.010.2008, **PMID:** 19198666, **PMCID:** PMC2636899

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., et al. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*, 963. **DOI:** https://doi.org/10.1038/s41467-018-03068-4, **PMID:** 29511192, **PMCID:** PMC5840373

Popov, V., Hristova, P., & Anders, R. (2017). The relational luring effect: Retrieval of relational information during associative recognition. *Journal of Experimental Psychology: General*, *146*, 722–745. **DOI:** https://doi.org/10.1037/xge0000305, **PMID:** 28368198

Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive supports for analogies in the mathematics classroom. *Science*, *316*, 1128–1129. **DOI:** https://doi.org/10.1126/science.1142103, **PMID:** 17525320

Rissman, J., Gazzaley, A., & D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage*, *23*, 752–763. **DOI:** https://doi.org/10.1016/j.neuroimage.2004.06.035, **PMID:** 15488425

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*, 192–233. **DOI:** https://doi.org/10.1037/0096-3445.104.3.192

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*, 214–241. **DOI:** https://doi.org/10.1037/h0036351

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., et al. (2004). Advances in functional and structural MR image analysis and mplementation as FSL. *Neuroimage*, *23(Suppl. 1)*, S208–S219. **DOI:** https://doi.org/10.1016/j.neuroimage.2004.07.051, **PMID:** 15501092

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, *44*, 83–98. **DOI:** https://doi.org/10.1016/j.neuroimage.2008.03.061, **PMID:** 18501637

Waldmann, M. R. (2017). Causal reasoning: An introduction. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 1–9). New York: Oxford University Press. **DOI:** https://doi.org/10.1093/oxfordhb/9780199399550.013.1

Wendelken, C. (2015). Meta-analysis: How does posterior parietal cortex contribute to reasoning? *Frontiers in Human Neuroscience*, *8*, 1042. **DOI:** https://doi.org/10.3389/fnhum.2014.01042, **PMID:** 25653604, **PMCID:** PMC4301007

Wendelken, C., Bunge, S. A., & Carter, C. S. (2008). Maintaining structured information: An investigation into functions of parietal and lateral prefrontal cortices. *Neuropsychologia*, *46*, 665–678. **DOI:** https://doi.org/10.1016/j.neuropsychologia.2007.09.015, **PMID:** 18022652

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. **DOI:** https://doi.org/10.1016/j.neuroimage.2014.01.060, **PMID:** 24530839, **PMCID:** PMC4010955

Zhila, A., Yih, W.-T., Meek, C., Zweig, G., & Mikolov, T. (2013). Combining heterogenous models for measuring relational similarity. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1000–1009). Atlanta, GA: Association for Computational Linguistics.