Connecting Concepts in the Brain: Mapping Cortical Representations of Semantic Relations

Yizhen Zhang[1,3], Kuan Han[1,3], Robert Worth[4], Zhongming Liu[*1,2,3]

[1]School of Electrical and Computer Engineering, Purdue University, West Lafayette

[2]Weldon School of Biomedical Engineering, Purdue University, West Lafayette

[3]Purdue Institute of Integrative Neuroscience, Purdue University, West Lafayette

[4]Department of Mathematical Sciences, Indiana University-Purdue University Indianapolis

[*]**Correspondence**

Zhongming Liu, PhD

Assistant Professor of Biomedical Engineering

Assistant Professor of Electrical and Computer Engineering

College of Engineering, Purdue University

206 S. Martin Jischke Dr.

West Lafayette, IN 47907, USA

Phone: +1 765 496 1872

Fax: +1 765 496 1459

Email: zmliu@purdue.edu

Abstract

In the brain, the semantic system is thought to store concepts. However, little is known about how it connects different concepts and infers semantic relations. To address this question, we collected hours of functional magnetic resonance imaging (fMRI) data from human subjects listening to natural stories. We developed a predictive model of the voxel-wise response, and further applied it to thousands of new words. We found that both semantic categories and relations were represented by spatially overlapping cortical networks, instead of anatomically segregated regions. Importantly, many such semantic relations that reflected conceptual progression from concreteness to abstractness were represented by a similar cortical pattern of anti-correlation between the default mode network and the frontoparietal attention network. Our results suggest that the human brain represents a continuous semantic space and uses distributed networks to encode not only concepts but also relationships between concepts. In particular, the default mode network plays a central role in semantic processing for abstraction of concepts across various domains.

Significance

Natural language comprehension requires that brains not only store concepts but also connect them to one another. But how does the brain relate one concept to another? To answer this question, we use a data-driven approach to model cortical responses to natural stories, and to study how the brain represents the semantic relations between thousands of words. Our results show that distributed and anti-correlated cortical networks represent semantic relations. In particular, the anti-correlation between the default mode network and the frontoparietal attention network represents the cortical signature common to semantic relations that reflect abstraction of concepts across various domains. This finding suggests an active role of the default mode network in semantic cognition, instead of being merely “task-negative”.

Introduction

Humans can describe the potentially infinite features of the world and communicate with others using a finite number of words. To make this possible, our brains need to encode semantics [1], infer concepts from experiences [2], relate one concept to another [3, 4], and learn new concepts [5]. Central to these cognitive functions is the brain's semantic system [6]. It is spread widely over many regions in the association cortex [7-9], and it also partially overlaps with the default-mode network [10]. Based on piecemeal evidence from brain imaging studies [11, 12] and patients with focal lesions [13], individual regions in the semantic system are thought to represent distinct categories or domains of concepts [11, 13]. Recent studies further suggested that concepts are represented in the brain by grounding their attributes in perception, action, and emotion systems [14, 15].

However, little is known about how the brain connects concepts and infers semantic relations [16, 17]. As concepts are related to one another in the real world, cortical regions that represent concepts are also connected, allowing them to communicate and work together as networks [18]. It is thus likely that the brain represents semantic relations as emerging patterns of network interaction [19]. Moreover, since different types of concepts may express similar relations, it is also possible that the cortical representation of a semantic relation may transcend any specific conceptual domain. Testing these hypotheses requires a comprehensive study of the semantic system as a set of distributed networks, as opposed to a set of isolated regions. Being comprehensive, the study should also survey cortical responses to a sufficiently large number of words from a wide variety of conceptual domains [1], ideally using naturalistic stimuli [20].

Similar to a prior work [1], we developed a predictive model of human fMRI responses given >11 hours of natural speech stimuli. In this model, individual words and their pairwise relationships were both represented as vectors in a continuous semantic space [21], which was learned from a large corpus[1] and was linearly mapped onto the brain's semantic system. Applying this model to many thousands of words and pairs of words, we have demonstrated how cortical networks represent semantic categories and semantic relations, respectively. Our results also shed new light on the role of the default mode network in semantic processing.

Results

Word embeddings explained cortical responses during natural speech comprehension

To extract semantic features from words, we used a word2vec model trained to predict the nearby words of every word in large corpora [21]. Through word2vec, we could represent any word as a vector in a 300-dimensional semantic space. Of this vector representation (or word embedding), every dimension encoded a distinct semantic feature learned entirely by data-driven methods [21], instead of by human intuition or linguistic rules [1, 22, 23]. To relate this semantic space to its cortical

[1] https://code.google.com/archive/p/word2vec/

representation, we defined a voxel-wise encoding model [24] – a multiple linear regression model that expressed each voxel's response as a weighted sum of semantic features [1] (Fig. S1).

To estimate the voxel-wise encoding model, we acquired whole-brain fMRI data from 19 native English speakers listening to different audio stories (from *The Moth Radio Hour*[2]), which in total lasted 11 hours and included 45,075 words (or 5,109 words if duplicates were excluded). The voxel-wise encoding model was estimated based on the fMRI data concatenated across all stories and subjects. By 20-fold cross-validation [25], the model-predicted response was significantly correlated with the measured fMRI response (false discovery rate or q<0.05) for voxels throughout the brain's semantic system [6]. The predictable voxels highlighted a map of semantic representation (Fig. S2). It was broadly distributed across the two hemispheres, as opposed to only the left hemisphere, which has conventionally been thought to dominate language processing and comprehension [26].

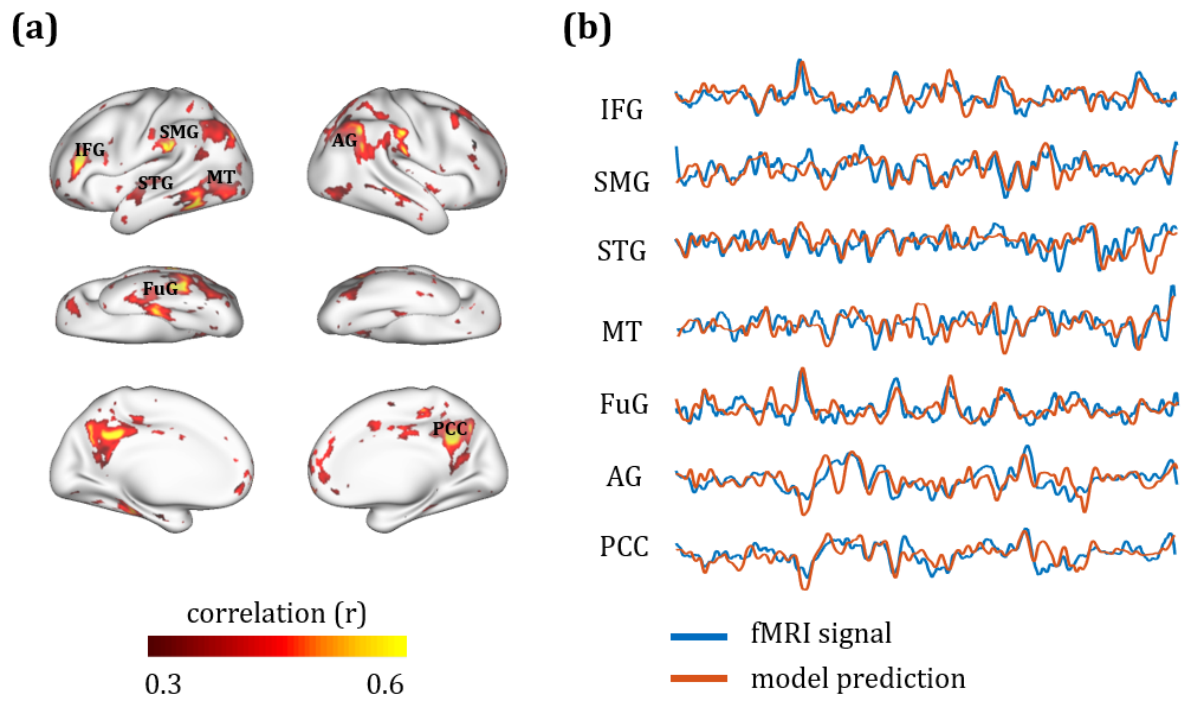

Fig. 1. Prediction accuracy of the voxel-wise encoding model for testing story. (a). The voxel-wise correlation between fMRI signals and model predictions for testing story. Statistical significance was assessed by a permutation test (FDR<0.05). (b). Time series of fMRI signals (blue) and model predictions (red) for selected regions. IFG, inferior frontal gyrus; SMG, supramarginal gyrus; STG, superior temporal gyrus; MT, middle temporal visual area; FuG, fusiform gyrus; AG, angular gyrus; PCC, posterior cingulate gyrus.

[2] https://themoth.org/radio-hour

The estimated encoding model was found to be generalizable to other words and sentences beyond those used for model estimation. Given an independent (untrained) testing story, the encoding model was able to reliably predict the evoked responses at localized cortical regions in the superior temporal gyrus (STG), middle temporal gyrus (MTG), angular gyrus (AG), supramarginal gyrus (SMG), posterior cingulate cortex (PCC), inferior prefrontal cortex (iPFC), superior prefrontal cortex (sPFC), and medial prefrontal cortex (mPFC) (Fig. 1a). However, these regions showed different response dynamics given the same story, suggesting their highly distinctive roles in semantic processing (Fig. 1b). Therefore, the word2vec-based encoding model was able to capture the quantitative and generalizable relationships between cortical responses and word attributes in a naturalistic context.

Semantic categories were represented by distributed cortical networks

Since it was generalizable to new words and sentences, we applied the estimated encoding model to >28,000 words from nine categories: "*tool*", "*human*", "*plant*", "*animal*", "*place*", "*communication*", "*emotion*", "*change*", "*quantity*" (Table S1), as defined in *WordNet* [27] and are representative of different conceptual domains. Within each category, we averaged the encoding-model-predicted responses given every word and mapped the statistically significant voxels (one sample t-test, FDR<0.01). We found that individual categories were represented by spatially overlapping and distributed cortical patterns (Fig. S3). For example, the category "*tool*" was represented by the SMG, posterior MTG (pMTG), fusiform gyrus (FuG) and inferior frontal gyrus (IFG); this representation was more pronounced in the left hemisphere than the right hemisphere. Words categorized as "*human*", "*plant*", and "*animal*" were also represented more by the left hemisphere than the right hemisphere. The category "place" was represented by bilateral parahippocampal gyrus (PhG), dorsolateral prefrontal cortex (dLPFC), and AG. In contrast, "*communication*", "*emotion*", "*change*", and "*quantity*", showed stronger representations in the right hemisphere than in the left hemisphere.

To each predictable voxel, we assigned a single category that gave rise to the strongest voxel response, thus dividing the brain's semantic system into category-labeled parcels (Fig. 2a). This analysis served to compare the category-wise representations in a "winners-take-all" manner. The resulting parcellation also revealed that none of the categories studied was represented by a single region. In addition, the distinction in left/right lateralization was likely attributable to the varying degree of concreteness for the words from individual categories. The concepts lateralized to the left hemisphere appeared relatively more concrete or exteroceptive, whereas those lateralized to the right hemisphere were more abstract or interoceptive (Fig. 2b). This intuitive interpretation was supported by quantitative rating of concreteness (from 1 to 5) for every word in each category [28]. The concreteness rating was high (between 4 and 5) for the categories lateralized to the left hemisphere, whereas it tended to be lower yet more variable for those categories dominated by the right hemisphere (Fig. 2c).

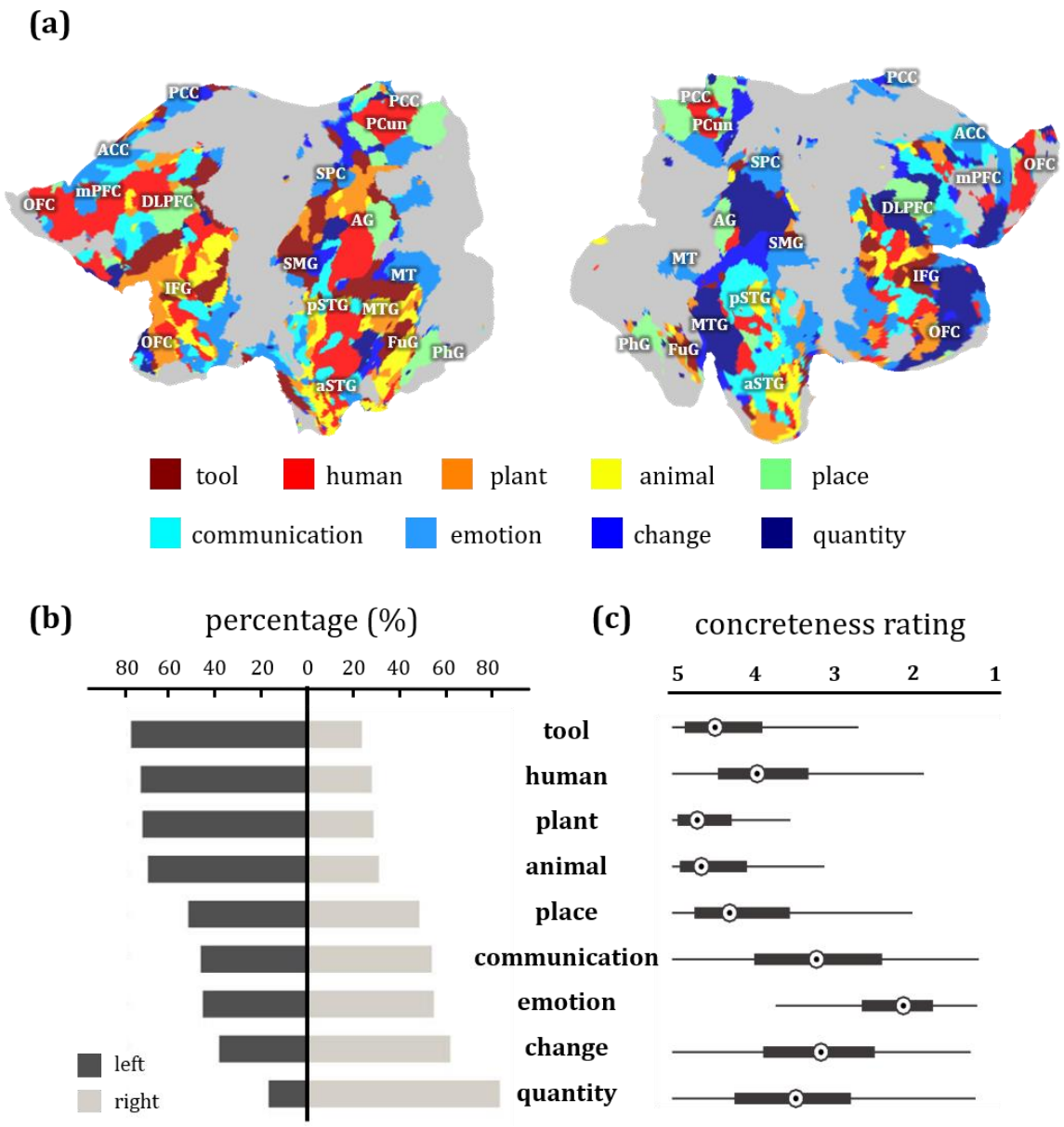

Fig.2. Cortical representation of semantic category. (a). Category-labeled parcellation based on voxel-wise selectivity using a "winners-take-all" strategy. (b). Cortical lateralization of categorical representations. For each category, the percentage value was calculated by counting the number of voxels on each hemisphere that were labeled by that category. (c). The concreteness rating of words in each category. The central mark indicates the median, and the box edges indicate the 25th and 75th percentiles respectively.

Semantic relationships were encoded by anticorrelated cortical networks

Through the word2vec model, we could also represent semantic relationships as vectors in the semantic space [29]. Specifically, we retrieved the semantic relationship between any pair of words from their

element-wise vector difference in word embedding, and further used the encoding model to convert this vector difference to a cortical map that represented the corresponding word-to-word relationship. Word-pair samples were selected from SemEval-2012 Task 2 dataset [30], in which thousands of word pairs were grouped into ten semantic relationships through crowdsourcing[3].

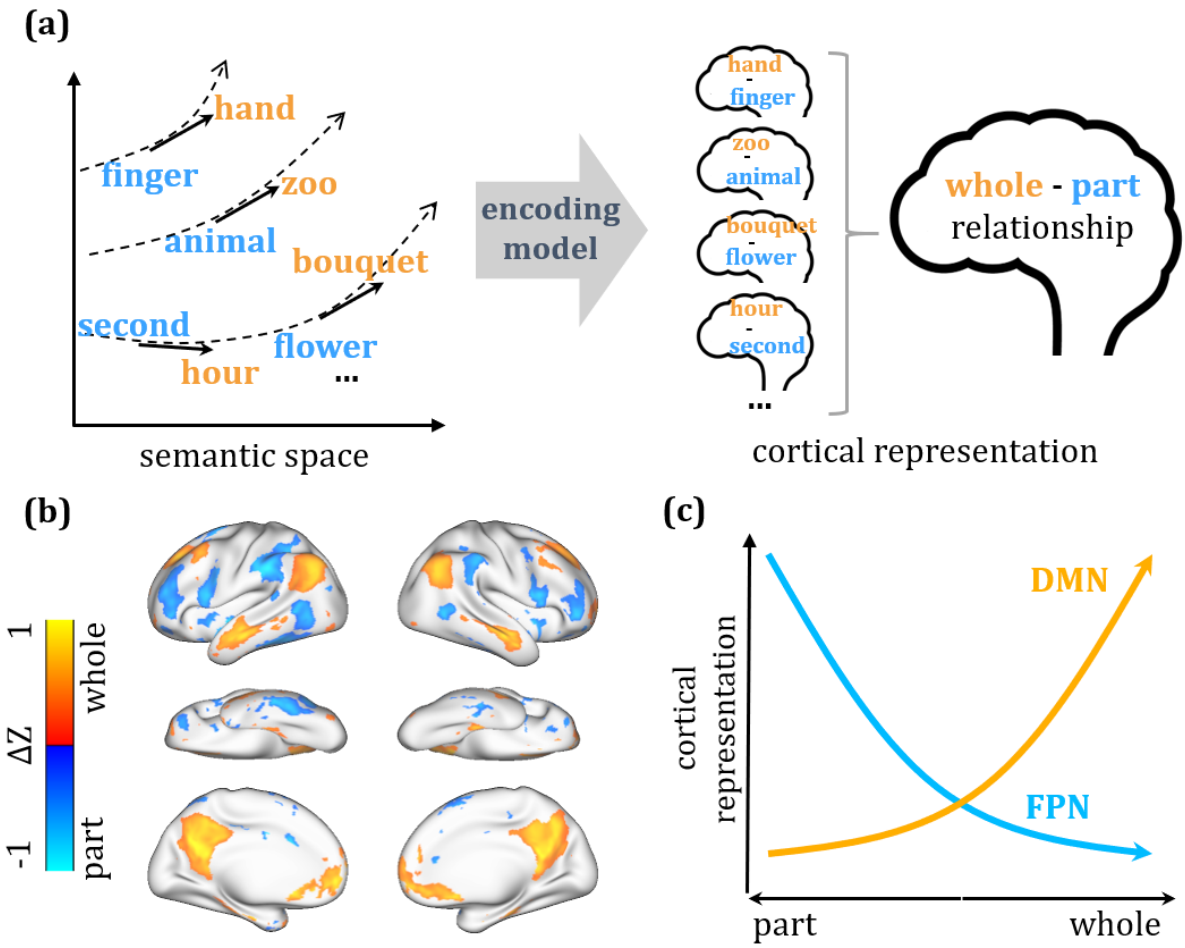

Fig.3. Mapping cortical representation of the whole-part relationship. (a). The illustration of mapping the whole-part relationship from semantic space to the human brain through the voxel-wise encoding model. We viewed the whole-part relationship as a vector field over the semantic space. This relationship field was sampled by the vector difference of each word pair holding such a relationship (left). The cortical representation of this vector difference was predicted by the voxel-wise encoding model. Cortical representation of the whole-part relationship was then obtained by averaging representations of all word pairs (right). (b). Cortical representation of the whole-part relationship. The statistical significance was assessed by a paired t-test (178 word pairs, FDR<0.01). (c). An interpretation of the cortical networks that encode the whole-part relationship. The cortical representation increased in DMN and decreased in FPN as concepts progressed from part to whole.

[3] https://sites.google.com/site/semeval2012task2/

For an initial exploration, we applied this analysis to 178 word pairs that all shared a "whole-part" relationship. For example, in four word pairs, ("*hand*", "*finger*"), ("*zoo*", "*animal*"), ("*hour*", "*second*"), and ("*bouquet*", "*flower*"), "*finger*" is part of "*hand*"; "*animal*" is part of "*zoo*"; "*second*" is part of "*hour*"; "*flower*" is part of "*bouquet*". Individually, the words from different pairs had different meanings and belonged to different semantic categories, e.g. "*finger*", "*animal*", "*second*", and "*flower*" shared little semantic similarity. Nevertheless, their pairwise relations all entailed the "whole-part" relationship, as illustrated in Fig. 3a. By using the encoding model, we mapped the pairwise word relationship onto the cortex, averaged the results across pairs, and highlighted the significant voxels (paired t-test, FDR<0.01). The resulting cortical representation ascribed the relationship between a pair of words to their representational contrast. We found that the "whole-part" relationship was mapped positively onto the default mode networks [31] (DMN, including AG, MTG, mPFC and PCC) and negatively onto the frontoparietal network [32, 33] (FPN, including LPFC, IPC and pMTG) (Fig. 3b). Taken together, DMN and FPN formed anticorrelated functional networks, encoding the "whole-part" relationship independent of the cortical representations of individual words that held this relationship. For this relationship, being "whole" manifested itself as decreasing representation in FPN and increasing representation in DMN, and vice versa for "part" (Fig. 3c).

Similarly, we also mapped the cortical representations of other semantic relationships, including "class-inclusion", "object-attribute", "agent-recipient", "space-associated", and "time-associated". Each of these relationships was encoded by a distinct set of anticorrelated networks (Fig. 4). The "class-inclusion" relationship, e.g. ("*color*", "*green*") where "*color*" includes "*green*", was mapped positively onto AG and MTG and negatively onto IFG and STG (Fig. 4b). The "object vs. attribute" relationship, e.g. ("*fire*", "*hot*") where "*fire*" is "*hot*", was mapped onto bilaterally asymmetric networks (Fig. 4c); the right hemisphere encoded "*attributes*", whereas the left hemisphere encoded the "*objects*" that possessed the "*attributes*". The "agent-recipient" relationship, e.g. ("*coach*", "*player*") where a "coach" teaches a "player", was represented by networks similar to the "whole-part" relationship (Fig. 4d), even though the two relationships were not intuitively consistent. The "space-associated" relationship, e.g. ("*library*", "*book*") where "*book*" is an associated item in a "*library*", was mapped positively onto AG and PCC and negatively onto STG (Fig. 4e). Lastly, the "time-associated" relationship, e.g. ("*morning*", "*sunrise*") where "*sunrise*" is a phenomenon associated with "*morning*", was represented by a bilaterally asymmetric network (Fig. 4f); the right FPN encodes the "time", whereas the left hemisphere encoded the associated terms.

For each semantic relationship, we aggregated its cortical representation by regions, and visualized it as a graph in which each node corresponded to a region (Fig. S4). The graph showed the representational geometry for each semantic relationship, allowing us to better appreciate its pattern (e.g. bilateral (a)symmetry) and to compare different semantic relationships. Noticeably, regions in DMN were positive for most of the relationships, suggesting that those relationships share common features pertaining to the functional role of DMN. Homologous regions from both hemispheres were found to contribute equally to some semantic relationships (e.g. "whole-part", "agent-recipient"), but not to the others (e.g. "object-attribute", "class-inclusion", "space-associated", "time-associated").

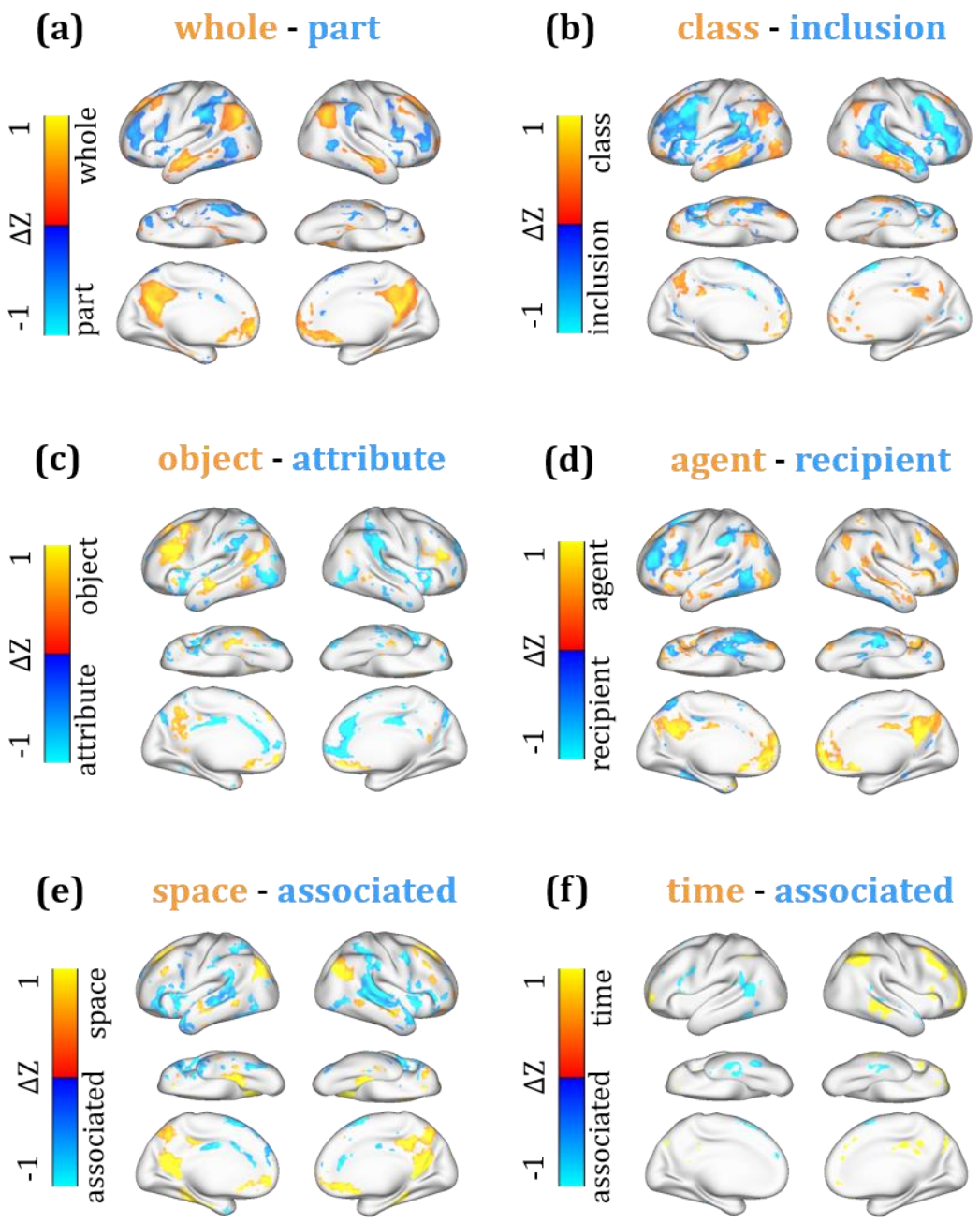

Fig.4. Cortical representations of each semantic relationship (FDR<0.01). (a). whole-part relationship (178 word pairs, e.g. "*hand*"-"*finger*"). (b). class-inclusion relationship (113 word pairs, e.g. "*color*"-"*green*"). (c). object-attribute relationship (63 word pairs, e.g. "*fire*"-"*hot*"). (d). agent-recipient (106 word pairs, e.g. "*coach*"-"*player*"). (e). space-associated relationship (58 word pairs, e.g. "*library*"-"*book*"). (f). time-associated relationship (44 word pairs, e.g. "*morning*"-"*sunrise*"). Detailed information (number of paired samples, number of significant voxels and word pair examples) is listed in SI (Table S2).

Discussion

Using fMRI data from subjects listening to natural speech stimuli, we established a predictive model to map the cortical representations of semantic categories and relations. We found that semantic categories were not represented by separate cortical regions but by spatially overlapping cortical networks, mostly involving multimodal association areas. Although both cerebral hemispheres supported semantic representations, the left hemisphere was more selective for concrete concepts whereas the right hemisphere was more selective for abstract concepts. Importantly, semantic relations were represented by anti-correlated cortical networks. Many such semantic relations that reflected conceptual abstraction were represented positively by the default-mode network and negatively by the attention network. Together, these findings suggest that the human brain represents a continuous semantic space. To support conceptual inference and reasoning, the brain uses distributed cortical networks to encode not only concepts but also relationships between concepts. In particular, the default-mode network plays an active role in semantic processing for abstraction of concepts.

Word2vec defines a continuous semantic space

In this study, we embed concepts in a continuous semantic space [21]. Although we use words to study concepts, words and concepts are different. The number of concepts expressible by humans is much larger than the number of words available to express them. We view words vs. concepts as discrete vs. continuous samples from the semantic space. The vocabulary of human speakers is finite. However, concepts are infinite. Varying a concept may create a new concept or arrive at a different concept. This allows concepts to be usable to describe the world, as is understood by the brain and, to a lesser extent, as is expressed in language.

Moreover, concepts are not isolated but related to one another. Since we view concepts as points in the semantic space, we consider conceptual relationships to be continuous vector fields that exist in the same space. A given concept as a position in the semantic space may experience multiple fields, and different positions may experience the same field. Thus, a concept may relate to other concepts in various ways, and different pairs of concepts may hold the same relationship [4]. Because semantics and concepts reflect how humans understand and describe the world, we speculate that the brain not only encodes such a continuous semantic space [1] but also encodes semantic relations as its vector fields, in order to support conceptual inference and reasoning.

Machine learning leverages the notion of continuous semantic space for natural language processing [21, 34, 35], and provides a new way to model and reconstruct neural responses [1, 24, 36-38]. For example, word2vec can represent millions of words as vectors in a lower-dimensional semantic space [21]. Two aspects about word2vec have motivated us to use it for this study. First, words with similar meanings share similar linguistic contexts and have similar vector representations in the semantic space [39]. Second, the relationship between two words is captured by their vector difference and is transferable to a different word. For an illustrative example, "(*man* - *women*) + *queen*" results in a vector close to "*king*" [29]. As such, word2vec defines a continuous semantic space and preserves both word meanings and word-to-word relationships.

In addition, word2vec learns the semantic space from large corpora in a data-driven manner [21]. This is different from defining the semantic space based on keywords that are hand selected [22], frequently used [1], minimally grounded [40], or neurobiologically relevant [23, 41]. Although these models are more intuitive and easier to interpret, they are arguably subjective and may not be able to describe the complete semantic space. We prefer word2vec as a model of word embedding, because it leverages big data to learn natural language statistics without any human bias.

Distributed cortical representations of semantics

Does the brain encode a continuous semantic space as is obtained by word2vec? Since word2vec is not constrained by any neurobiological knowledge, we do not expect it to encode the same semantic space as does the brain. Instead, we hypothesize that the two semantic spaces, encoded by the model vs. the brain, are similar up to linear projection (i.e. transformation through linear encoding).

Our results support this hypothesis and reveal a distributed semantic network (Fig. S2). This network resembles the semantic system previously mapped by Binder et al. [6], albeit with noticeable distinctions. As in that paper, our results also highlight a similar set of semantics-encoded regions, such as AG, SMG, MTG, PCC, IFC, mPFC, dlPFC, FuG and PhG (Fig. S2). Most of these regions are associated with high-level integrative processes that transcend any single modality [8, 9]. However, our semantic network is bilateral rather than being dominated by the left hemisphere as suggested by Binder et al. [6]. To partially reconcile this distinction, it is worth noting that the activation foci analyzed by Binder et al. are indeed distributed on both hemispheres (see Figure 2 in [6]). In addition, the two hemispheres seem to be selective for different aspects of semantics. Unlike prior findings using contrast analysis [42, 43], we found the left hemisphere encodes exteroceptive and concrete concepts, whereas the right hemisphere encodes interoceptive and abstract concepts (Fig. 2). Nevertheless, we do not claim a rigid dichotomy.

Our semantic system also appears very similar to that reported by Huth et al. [1]. This similarity is perhaps not surprising, because both studies use similar natural speech stimuli and encoding models. However, unlike Huth et al. [1], we do not emphasize semantic selectivity of each region or tile the cortex into regions associated with distinct conceptual domains. On the contrary, we have found that none of the conceptual categories addressed in this study is represented by a single region (Fig. 2). Instead, individual categories are represented by spatially distributed and partly overlapping cortical networks (Fig. S3). Each network presumably integrates various domain-defining attributes that are represented by various distributed cortical regions [11, 14, 15]. Therefore, we advocate efforts to address semantic selectivity by means of cortical networks, as opposed to regions [18, 19].

Cortical representations of semantic relationships

In the semantic space, the relationship between words or concepts is represented by their vector difference, of which the direction and magnitude indicate different aspects of the underlying semantic relationship. Let us take, for example, the pair of words ("*minute*", "*day*"). Of their vector difference, the direction indicates a "part to whole" relationship, and the magnitude indicates the offset along this

direction. Starting from "*minute*" and relative to "*day*", a larger offset leads toward "*month*" or "*year*", a smaller offset leads toward "*hour*", and a negative offset leads toward "*second*". Furthermore, vector representations of semantic relationship are transferable to different word pairs.

The ability of using word2vec to capture, transfer and generalize semantic relationships lends support to the notion of continuous vector fields in the semantic space. Each field governs one type of semantic relationship and applies broadly to various concepts or even domains of concepts. If we visualize a vector field as many field lines, then the points (i.e. concepts) that each field line passes through are related to one another by the same semantic relationship. Speculatively, the fields are continuous and smooth in the semantic space but are not necessarily linear.

In a nominal relationship (e.g. "whole-part"), each word pair takes a sample from the underlying vector field (as illustrated by Fig. 3a). If the samples taken by different word pairs always align with the same vector field, we expect a common cortical representation to be shared by all word pairs; hence, we interpret it as the neural substrate of this semantic relationship. However, not all human-defined semantic relations have an underlying vector field that can be represented as a consistent cortical region or network (Table S2).

Several semantic relationships are mapped onto anticorrelated networks (Fig. 4). This is reasonable since a semantic relationship is directional and its cortical representation should indicate the direction. For instance, the direction from "part" to "whole" is shown as positivity in DMN and negativity in FPN (Fig. 3), whereas the reverse direction (i.e. from "whole" to "part") is shown as the opposite pattern. This specific pattern of anticorrelated networks that encode semantic relationships are likely intrinsic and observable even in task-free conditions [44].

The role of default-mode network in semantic processing

Our results suggest that DMN is involved in cortical processing of not only concepts but also semantic relations. This finding underscores the fact that the DMN plays an active role in language and cognition [10, 45-48], rather than only a task-negative and default mode of brain function [31]. In particular, several semantic relationships, such as "whole-part" and "class-inclusion", are all mapped onto DMN (Fig. 4), suggesting that DMN is likely associated with the semantic regularity common to those relationships. Indeed, words being "whole" or "class" are more abstract and general, whereas those being "part" or "inclusion" are relatively more detailed and specific. As such, these word relationships all indicate (to a varying degree) conceptual abstraction. This progression involves DMN, increasing or decreasing its activity as a concept (of various types) becomes more abstract or specific, respectively. Moreover, concrete concepts, e.g. "*tool*", "*plant*", "*animal*", are represented by regions outside the DMN, whereas more abstract concepts, e.g. "*communication*", "*emotion*", "*quantity*", are represented by cortical regions that reside in, or at least overlap with, DMN (Fig. S3).

These observations lead us to speculate that DMN underlies a cognitive process for abstraction of concepts. This interpretation is consistent with findings from several prior studies [48, 49]. For example, Spunt et al. has shown that conceptualizing the same action at an increasingly higher level of abstraction gives rise to an increasingly greater responses at regions within DMN [49]. Sormaz et al. has shown evidence that activity patterns in DMN during cognitive tasks are associated with whether thoughts are

detailed, rather than whether they are task related or unrelated [48]. In contrast to DMN, another network, FPN, seems to play an opposite role in semantic processing. FPN is often activated by attention-demanding tasks and is intrinsically anti-correlated with DMN [44]. Our results suggest that FPN is increasingly activated when the brain is engaged in conceptual specification.

Materials and Methods

Subjects and experiments

19 human subjects (11 females, age 24.4±4.8, all right-handed) participated in this study. All subjects provided informed written consent according to a research protocol approved by the Institutional Review Board at Purdue University. While being scanned for fMRI, each subject was listening to several audio stories collected from *The Moth Radio Hour* (https://themoth.org/radio-hour)and presented through binaural MR-compatible headphones (Silent Scan Audio Systems, Avotec, Stuart, FL). A single story was presented in each fMRI session (6 mins 48 secs ± 1 min 58 secs). For each story, two repeated sessions were performed for the same subject.

Different audio stories were used for training vs. testing the encoding model (details about the model are described in subsequent subsections). For training, different subjects listened to different sets of stories. When combined across subjects, the stories used for training amounted to a total of 5 hours 33 mins. For testing, every subject listened to the same single story for 6 mins 53 secs; this story was different from those used for training.

Data acquisition and processing

T_1 and T_2-weighted MRI and fMRI data were acquired in a 3T MRI system (Siemens, Magnetom Prisma, Germany) with a 64-channel receive-only phased-array head/neck coil. The fMRI data were acquired with 2 mm isotropic spatial resolution and 0.72 s temporal resolution by using a gradient-recalled echo-planar imaging sequence (multiband = 8, 72 interleaved axial slices, TR = 720 ms, TE = 31 ms, flip angle = 52°, field of view = 21 × 21 cm^2).

Since our imaging protocol was similar to what was used in the human connectome project (HCP), our MRI and fMRI data were preprocessed by using the minimal preprocessing pipeline established for the HCP. After preprocessing, the images from individual subjects were co-registered onto a common cortical surface template (see details in [50]). Then the fMRI data were spatially smoothed by using a gaussian surface smoothing kernel with a 2mm standard deviation.

For each subject, the voxel-wise fMRI signal was standardized (i.e. zero mean and unitary standard deviation) within each session and was averaged across repeated sessions. Then the fMRI data were concatenated across different sessions and subjects.

Modeling and sampling the semantic space

To represent words as vectors, we used a pre-trained word2vec model [21]. Briefly, this model was a shallow neural network trained to predict the neighboring words of every word in the Google News dataset, including about 100 billion words. After training, the model was able to convert any English word to a vector embedded in a 300-dimensional semantic space. Note that the basis functions learned with word2vec should not be interpreted individually, but collectively as a space. Arbitrary rotation of the semantic space would end up with an equivalent space, even though it may be spanned by entirely different semantic features. The model was also able to extract the semantic relationship between words by simple vector operations [29]. In an attempt to sample the semantic space, we intentionally chose audio stories of diverse topics. Individual words were extracted from audio stories using *Speechmatics* (https://www.speechmatics.com/), and then were converted to vectors through word2vec.

Voxel-wise encoding model

We mapped the semantic space, as modeled by word2vec, to the cortex through voxel-wise linear encoding models, as explored in previous studies [1, 24, 37, 38]. For each voxel, we modeled its response as a linear combination of all features in the semantic space.

$$x_i = a_i + \boldsymbol{b_i}\boldsymbol{y} + \varepsilon_i, \qquad (1)$$

where x_i is the response at the i-th voxel, $\boldsymbol{y}$ is the word embedding represented as a 300-dimensional column vector with each element corresponding to one axis (or feature) in the semantic space, $\boldsymbol{b_i}$ is a row vector of regression coefficients, a_i is the bias term, and ε_i is the error or noise.

Training the encoding model with cross-validation

We used the (word, data) samples from the training stories to estimate the encoding model. As words occurred sequentially in the audio story, each word was given a duration based on when it started and ended in the audio story. A story was represented by a time series of word embedding sampled every 0.1 second. For each feature in the word embedding, its time-series signal was further convolved with a canonical hemodynamic response function (HRF) to account for the temporal delay and smoothing due to neurovascular coupling [51]. The HRF-convolved feature-wise representation was standardized and down-sampled to match the sampling rate of fMRI.

It follows that the response of the i-th voxel at time t was expressed as Eq. (2)

$$x_i(t) = a_i + \boldsymbol{b_i}\boldsymbol{y}(t) + \varepsilon_i(t), \qquad (2)$$

We estimated the coefficients $(a_i, \boldsymbol{b_i})$ given time samples of $(x_i, \boldsymbol{y})$ by using least-squares estimation with L_2-norm regularization. That is, to minimize the following loss function defined separately for each voxel.

$$L_i = \frac{1}{T}\sum_{t=1}^{T}\left(x_i(t) - a_i - \boldsymbol{b_i}\boldsymbol{y}(t)\right)^2 + \lambda_i \parallel \boldsymbol{b_i} \parallel_2^2 , \qquad (3)$$

where T is the number of temporal samples, and λ_i is the regularization parameter for the i-th voxel.

We applied 10-fold generalized cross-validation [25] to determine the regularization parameter. Specifically, the training data were divided evenly into 10 subsets, of which nine were used for model estimation and one was used for model validation. The validation was repeated 10 times such that each subset was once for validation. In each time, the correlation between the predicted and measured fMRI responses was calculated and used to evaluate the validation accuracy. The average validation accuracy across all 10 times was considered as the cross-validation accuracy. We chose the regularization parameter that yielded the highest cross-validation accuracy.

With the optimized regularization parameter, we further evaluated the encoding model using 20-fold cross-validation, which yielded 20 correlation coefficients for each voxel. The statistical significance was evaluated by applying one-sample t-test to the z-transformed correlation coefficients, with the false discovery rate (FDR or q) < 0.05. Finally, we used the optimized regularization parameter and all training data for model estimation, ending up with the finalized model parameters denoted as $(\widehat{a_i}, \widehat{\boldsymbol{b_i}})$.

Testing the encoding model

After training the encoding model, we tested it against a new (testing) story that was not used for model training. For this purpose, the estimated encoding model was applied, voxel by voxel, to the testing story, generating a model prediction of the fMRI response to the testing story.

$$\widehat{x_i}(t) = \widehat{a_i} + \widehat{\boldsymbol{b_i}}\boldsymbol{y}(t) , \ (4)$$

where $\boldsymbol{y}(t)$ is the HRF-convolved time series of word embedding extracted from the testing story.

To evaluate the encoding performance, we calculated the correlation between the predicted fMRI response $\widehat{x_i}$ and the actually measured fMRI response x_i. To evaluate the statistical significance, we used a block-wise permutation test [52] (20-sec window size; 100,000 permutations) with FDR<0.05.

Mapping cortical representation of semantic categories

Using it as a predictive mode, we further applied the estimated encoding model to a large vocabulary set [28], in which every word had been rated for concreteness, ranging from 1 (most abstract) to 5 (most concrete). Specifically, we used 28,003 words from nine categories: "*tool*", "*human*", "*plant*", "*animal*", "*place*", "*communication*", "*emotion*", "*change*", "*quantity*" (Table S1). For each word, we used word2vec to compute its vector representation, and then used the voxel-wise encoding model to map its cortical representation.

As words were grouped by categories, we sought the common cortical representation shared by those in the same category. For this purpose, we averaged the cortical representation of every word in each category, and thresholded the average representation based on its statistical significance (one-sample t-

test, FDR<0.01). We also evaluated the semantic selectivity of each voxel, i.e. how the voxel was more selective to one category than the others. For a coarse measure of categorical selectivity, we identified, separately for each voxel, a single category that resulted in the strongest voxel response among all nine categories, and assigned that voxel to the identified category (winners take all). After labeling voxels by category, we evaluated whether a given category was differentially represented by the left vs. right hemisphere, by counting the number of voxels on each hemisphere that were labeled by that category.

Mapping cortical representation of semantic relationships

Note that word2vec preserves semantic relationships which can then be readily extracted by simple vector arithmetic [29]. For example, an arithmetic expression of ("*hand*" - "*finger*" + "*second*"), based on their vector representations, leads to a vector representation close to that of "*hour*" (cosine similarity). In this example, the subtraction serves to extract the relationship between "*hand*" and "*finger*", which is intuitively interpretable as a "whole-part" relationship as a "*finger*" is part of a "*hand*". The addition serves to transfer this extracted relationship to another word "*second*", ending up with the word "*hour*", which makes sense since a "*second*" is indeed part of an "*hour*".

As a linear transformation, the encoding model could be applied to not only the vector representation of a single word but also the arithmetic difference in vector representation between a pair of words, which represented their semantic relationship. Therefore, we used the encoding model to predict the cortical representations of several semantic relationships. Specifically, word pairs that held specific relationships are selected from SemEval-2012 Task 2 dataset [30] (https://sites.google.com/site/semeval2012task2/), which contains ten general semantic relationships. These word-pair samples were collected and evaluated by large numbers of human responses (Amazon Mechanical Turk: https://www.mturk.com/). We excluded "*reference*" relation and separated "*space-time*" relation into "*space-associated*" and "*time-associated*". We then manually chose the most illustrative word pairs (positively scored in this dataset) for ten semantic relations, including "*whole-part*" (178 pairs), "*class-inclusion*" (113), "*object-attribute*" (63), "*agent-recipient*" (106), "*space-associated*" (58), "*time-associated*" (44), "*similar*"(160), "*contrast*"(162), "*object-nonattribute*"(69), "*cause-effect*" (107). Detailed information of these semantic relations was listed in Table S2.

For each semantic relationship, we converted every word pair into two vectors and calculated their difference by element-wise subtraction. Then we converted this vector difference to a cortical pattern through the voxel-wise encoding model. Lastly, we averaged the resulting cortical pattern across all word pairs that held the same relationship, and thresholded the average cortical representation by its statistical significance (paired t-test, FDR<0.01).

Acknowledgment

The authors thank David Kemmerer for helpful discussions and comments to this paper.

Reference

[1] A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature,* vol. 532, no. 7600, p. 453, 2016.

[2] E. Yee and S. L. Thompson-Schill, "Putting concepts into context," *Psychonomic bulletin & review,* vol. 23, no. 4, pp. 1015-1027, 2016.

[3] K. J. Holyoak, "Analogy and relational reasoning," *The Oxford handbook of thinking and reasoning,* pp. 234-259, 2012.

[4] D. Mirman, J.-F. Landrigan, and A. E. Britt, "Taxonomic and thematic semantic systems," *Psychological bulletin,* vol. 143, no. 5, p. 499, 2017.

[5] A. J. Bauer and M. A. Just, "Monitoring the growth of the neural representations of new animal concepts," *Human brain mapping,* vol. 36, no. 8, pp. 3213-3226, 2015.

[6] J. R. Binder, R. H. Desai, W. W. Graves, and L. L. Conant, "Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies," *Cerebral Cortex,* vol. 19, no. 12, pp. 2767-2796, 2009.

[7] M. A. L. Ralph, E. Jefferies, K. Patterson, and T. T. Rogers, "The neural and computational bases of semantic cognition," *Nature Reviews Neuroscience,* vol. 18, no. 1, p. 42, 2017.

[8] J. R. Binder, "In defense of abstract conceptual representations," *Psychonomic bulletin & review,* vol. 23, no. 4, pp. 1096-1108, 2016.

[9] K. Patterson and M. A. L. Ralph, "The hub-and-spoke hypothesis of semantic memory," in *Neurobiology of Language*: Elsevier, 2015, pp. 765-775.

[10] G. F. Humphreys, P. Hoffman, M. Visser, R. J. Binney, and M. A. L. Ralph, "Establishing task-and modality-dependent dissociations between the semantic and default mode networks," *Proceedings of the National Academy of Sciences,* p. 201422760, 2015.

[11] A. Martin, "The representation of object concepts in the brain," *Annu. Rev. Psychol.,* vol. 58, pp. 25-45, 2007.

[12] M. Kiefer and F. Pulvermüller, "Conceptual representations in mind and brain: theoretical developments, current evidence and future directions," *cortex,* vol. 48, no. 7, pp. 805-825, 2012.

[13] B. Z. Mahon and A. Caramazza, "Concepts and categories: A cognitive neuropsychological perspective," *Annual review of psychology,* vol. 60, pp. 27-51, 2009.

[14] A. Martin, "GRAPES—Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain," *Psychonomic bulletin & review,* vol. 23, no. 4, pp. 979-990, 2016.

[15] L. W. Barsalou, "On staying grounded and avoiding quixotic dead ends," *Psychonomic bulletin & review,* vol. 23, no. 4, pp. 1122-1142, 2016.

[16] O. Sachs *et al.*, "Automatic processing of semantic relations in fMRI: neural activation during semantic priming of taxonomic and thematic categories," *Brain research,* vol. 1218, pp. 194-205, 2008.

[17] M. F. Schwartz *et al.*, "Neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain," *Proceedings of the National Academy of Sciences,* vol. 108, no. 20, pp. 8520-8524, 2011.

[18] F. Pulvermüller, "How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics," *Trends in cognitive sciences,* vol. 17, no. 9, pp. 458-470, 2013.

[19] P. Hagoort, "Nodes and networks in the neural architecture for language: Broca's region and beyond," *Current opinion in Neurobiology,* vol. 28, pp. 136-141, 2014.

[20] U. Hasson, R. Malach, and D. J. Heeger, "Reliability of cortical activity during natural stimulation," *Trends in cognitive sciences,* vol. 14, no. 1, pp. 40-48, 2010.
[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111-3119.
[22] T. M. Mitchell *et al.*, "Predicting human brain activity associated with the meanings of nouns," *science,* vol. 320, no. 5880, pp. 1191-1195, 2008.
[23] A. J. Anderson *et al.*, "Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation," *Cerebral Cortex,* vol. 27, no. 9, pp. 4379-4395, 2016.
[24] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant, "Encoding and decoding in fMRI," *Neuroimage,* vol. 56, no. 2, pp. 400-410, 2011.
[25] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics,* vol. 21, no. 2, pp. 215-223, 1979.
[26] S. Knecht *et al.*, "Language lateralization in healthy right-handers," *Brain,* vol. 123, no. 1, pp. 74-81, 2000.
[27] G. Miller, *WordNet: An electronic lexical database*. MIT press, 1998.
[28] M. Brysbaert, A. B. Warriner, and V. Kuperman, "Concreteness ratings for 40 thousand generally known English word lemmas," *Behavior research methods,* vol. 46, no. 3, pp. 904-911, 2014.
[29] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 746-751.
[30] D. A. Jurgens, P. D. Turney, S. M. Mohammad, and K. J. Holyoak, "Semeval-2012 task 2: Measuring degrees of relational similarity," in *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 2012: Association for Computational Linguistics, pp. 356-364.
[31] M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman, "A default mode of brain function," *Proceedings of the National Academy of Sciences,* vol. 98, no. 2, pp. 676-682, 2001.
[32] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature reviews neuroscience,* vol. 3, no. 3, p. 201, 2002.
[33] M. Scolari, K. N. Seidl-Rathkopf, and S. Kastner, "Functions of the human frontoparietal attention network: Evidence from neuroimaging," *Current opinion in behavioral sciences,* vol. 1, pp. 32-39, 2015.
[34] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532-1543.
[35] F. Pereira, S. Gershman, S. Ritter, and M. Botvinick, "A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data," *Cognitive neuropsychology,* vol. 33, no. 3-4, pp. 175-190, 2016.
[36] A. G. Huth, S. Nishimoto, A. T. Vu, and J. L. Gallant, "A continuous semantic space describes the representation of thousands of object and action categories across the human brain," *Neuron,* vol. 76, no. 6, pp. 1210-1224, 2012.
[37] H. Wen, J. Shi, Y. Zhang, K.-H. Lu, J. Cao, and Z. Liu, "Neural encoding and decoding with deep learning for dynamic natural vision," *Cerebral Cortex,* pp. 1-25, 2017.

[38] F. Pereira *et al.*, "Toward a universal decoder of linguistic meaning from brain activation," *Nature communications,* vol. 9, no. 1, p. 963, 2018.
[39] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Communications of the ACM,* vol. 8, no. 10, pp. 627-633, 1965.
[40] P. Vincent-Lamarre, A. B. Massé, M. Lopes, M. Lord, O. Marcotte, and S. Harnad, "The latent structure of dictionaries," *Topics in cognitive science,* vol. 8, no. 3, pp. 625-659, 2016.
[41] J. R. Binder *et al.*, "Toward a brain-based componential semantic representation," *Cognitive neuropsychology,* vol. 33, no. 3-4, pp. 130-174, 2016.
[42] J. R. Binder, C. F. Westbury, K. A. McKiernan, E. T. Possing, and D. A. Medler, "Distinct brain systems for processing concrete and abstract concepts," *Journal of cognitive neuroscience,* vol. 17, no. 6, pp. 905-917, 2005.
[43] J. Wang, J. A. Conder, D. N. Blitzer, and S. V. Shinkareva, "Neural representation of abstract and concrete concepts: A meta-analysis of neuroimaging studies," *Human brain mapping,* vol. 31, no. 10, pp. 1459-1468, 2010.
[44] M. D. Fox, A. Z. Snyder, J. L. Vincent, M. Corbetta, D. C. Van Essen, and M. E. Raichle, "The human brain is intrinsically organized into dynamic, anticorrelated functional networks," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, no. 27, pp. 9673-9678, 2005.
[45] J. R. Andrews-Hanna, J. S. Reidler, J. Sepulcre, R. Poulin, and R. L. Buckner, "Functional-anatomic fractionation of the brain's default network," *Neuron,* vol. 65, no. 4, pp. 550-562, 2010.
[46] R. N. Spreng, "The fallacy of a “task-negative” network," *Frontiers in psychology,* vol. 3, p. 145, 2012.
[47] E. Simony *et al.*, "Dynamic reconfiguration of the default mode network during narrative comprehension," *Nature communications,* vol. 7, p. 12141, 2016.
[48] M. Sormaz *et al.*, "Default mode network can support the level of detail in experience during active task states," *Proceedings of the National Academy of Sciences,* vol. 115, no. 37, pp. 9318-9323, 2018.
[49] R. P. Spunt, D. Kemmerer, and R. Adolphs, "The neural basis of conceptualizing the same action at different levels of abstraction," *Social cognitive and affective neuroscience,* vol. 11, no. 7, pp. 1141-1151, 2015.
[50] M. F. Glasser *et al.*, "The minimal preprocessing pipelines for the Human Connectome Project," *Neuroimage,* vol. 80, pp. 105-124, 2013.
[51] M. A. Lindquist, J. M. Loh, L. Y. Atlas, and T. D. Wager, "Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling," *Neuroimage,* vol. 45, no. 1, pp. S187-S198, 2009.
[52] D. Adolf, S. Weston, S. Baecke, M. Luchtmann, J. Bernarding, and S. Kropf, "Increasing the reliability of data analysis of functional magnetic resonance imaging by applying a new blockwise permutation method," *Frontiers in neuroinformatics,* vol. 8, p. 72, 2014.