Cerebral Cortex, December 2018;28: 4305–4318

doi: 10.1093/cercor/bhx283
Advance Access Publication Date: 23 November 2017
Original Article

ORIGINAL ARTICLE

Organizational Principles of Abstract Words in the Human Brain

Xiaosha Wang[1,†], Wei Wu[1,†], Zhenhua Ling[2], Yangwen Xu[1], Yuxing Fang[1], Xiaoying Wang[1], Jeffrey R. Binder[3], Weiwei Men[4,5], Jia-Hong Gao[4,5,6] and Yanchao Bi[1]

[1]National Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, China, [2]National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei 230027, China, [3]Departments of Neurology and Biophysics, Medical College of Wisconsin, Milwaukee, WI 53226, USA, [4]Center for MRI Research, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China, [5]Beijing City Key Lab for Medical Physics and Engineering, Institute of Heavy Ion Physics, School of Physics, Peking University, Beijing 100871, China and [6]McGovern Institute for Brain Research, Peking University, Beijing 100871, China

Address correspondence to Yanchao Bi, National Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, China. Email: ybi@bnu.edu.cn.

[†]These authors contributed equally to the work.

Abstract

Abstract words constitute nearly half of the human lexicon and are critically associated with human abstract thoughts, yet little is known about how they are represented in the brain. We tested the neural basis of 2 classical cognitive notions of abstract meaning representation: by linguistic contexts and by semantic features. We collected fMRI BOLD responses for 360 abstract words and built theoretical representational models from state-of-the-art corpus-based natural language processing models and behavioral ratings of semantic features. Representational similarity analyses revealed that both linguistic contextual and semantic feature similarity affected the representation of abstract concepts, but in distinct neural levels. The corpus-based similarity was coded in the high-level linguistic processing system, whereas semantic feature information was reflected in distributed brain regions and in the principal component space derived from whole-brain activation patterns. These findings highlight the multidimensional organization and the neural dissociation between linguistic contextual and featural aspects of abstract concepts.

Key words: abstract concepts, representational similarity analysis, semantic feature, semantic representation, word2vec

Introduction

Words denoting abstract concepts that do not have specific external referents, such as "TRUTH" and "BELIEF", constitute nearly half of the lexicon in most human languages and are the building blocks of human abstract thought and reasoning. Our knowledge of their representations in the brain is extremely limited (Barsalou 2008). The dominant research on conceptual representation is based on studies of objects and actions, with the current consensus being that concepts are organized along both modality-specific sensory/motor (Binder and Desai 2011; Meteyard et al. 2012) and domain-specific dimensions (e.g., animacy, Caramazza and Shelton 1998). The existing neural studies on abstract concepts have predominantly focused on their general differences from concrete concepts, revealing that abstract words tend to produce stronger activation than concrete words in the left anterior temporal cortex and left inferior frontal gyrus (Wang et al. 2010), without examining the conceptual or neural organizational principles within the semantic space of abstract concepts.

There are 2 major types of cognitive theories about abstract concept representations. A classical view is that abstract concepts are linguistically based. They are represented in a verbal format (the Dual Coding theory, Paivio 1986) and/or through contextual associations (the Context Availability hypothesis, Schwanenflugel and Shoben 1983). Consistent with this view, word properties in the linguistic contexts (word co-occurrence patterns in large language corpus) have been found to predict lexical decision latencies of abstract words in healthy individuals or interference effects in patients with brain damage (Recchia and Jones 2012; Hoffman 2015). The significant progress made in natural language processing to represent rich word meanings using various statistical learning algorithms of word co-occurrence, based on either the global linguistic context (e.g., the latent semantic analysis (LSA), Landauer and Dumais 1997) or the local context (e.g., word2vec, Mikolov et al. 2013), has also strengthened the feasibility of this linguistic corpus-based computational approach of abstract meaning representation.

An alternative hypothesis is that abstract word meanings are represented along multiple semantic features. This view was motivated by grounded or "embodied" approaches to cognition, which attempt to represent meaning through systems that are intrinsic parts of the brain, including sensory, motor, and affective systems (Barsalou 2008; Binder and Desai 2011; Meteyard et al. 2012; Pulvermuller 2013), and was put into practice in studies about concrete concepts using approaches such as feature generation (McRae et al. 2005). The plausibility of this view for abstract concepts has been supported by a recent series of behavioral studies demonstrating the crucial role of emotion in the representation of abstract words (Kousta et al. 2011; Vigliocco et al. 2014). Other studies have considered a wider range of semantic features (Crutch et al. 2013; Binder et al. 2016), including social interaction (Barsalou and Wiemer-Hastings 2005), showing that the high-dimensional semantic space of abstract words generated by the ratings of these semantic features is associated with the semantic interference effect both in patients with global aphasia (Crutch et al. 2013) and in healthy subjects (Primativo et al. 2016).

Do either of these 2 models characterize the representation of abstract concepts in the brain, and how? The few existing neuroimaging studies on abstract meaning have not answered this question. The finding that greater activity in language-related regions, including left anterior temporal cortex and left inferior frontal gyrus, is elicited by abstract words than by concrete words has been interpreted as supportive of the linguistic representation account (Binder et al. 2009; Wang et al. 2010). It has been further shown that abstract words produce stronger activation in an emotion-processing brain region compared to concrete words, a finding that has been interpreted to support the feature-based representation account (Vigliocco et al. 2014). These interpretations rely strongly on assumptions about the functions of these implicated brain regions and are therefore prone to the limitations of reverse inference. Even if these assumptions hold, they do not reveal the type of computations used to represent linguistic or feature properties of abstract concepts.

In this study, we examined the specific effects of these 2 cognitive variables—language-corpus-based linguistic contexts and semantic feature decompositions—on the neural representation of abstract concepts. By "specific" we meant the effects of either measure beyond those that could be also explained by the other, given that these 2 measures are likely to be correlated to some extent, as words sharing semantic features may be more likely to appear in similar linguistic contexts. We sampled 360 abstract words that covered a wide range of frequency and semantic content. Language-corpus-based (Landauer and Dumais 1997; Mikolov et al. 2013) and semantic-feature-based (Crutch et al. 2013; Primativo et al. 2016) word distance measures were obtained. Although indeed correlated, both measures made unique contributions to the subjectively rated semantic distance between abstract words. Blood-oxygen-level-dependent (BOLD) fMRI responses to each of these 360 abstract words in a familiarity judgment task were collected using the condition-rich event-related design (Kriegeskorte, Mur and Bandettini 2008). Representational similarity analysis (RSA) (Kriegeskorte and Kievit 2013) was conducted to evaluate the specific correspondence between each cognitive measure and neural response patterns with the other cognitive measure controlled. Given the paucity of our knowledge of how the brain represents abstract words, and previous investigations of semantic representations at both small (Peelen and Caramazza 2012; Fairhall and Caramazza 2013; Clarke and Tyler 2014) and large (Huth et al. 2012, 2016) neural scales, we performed RSA at 3 neural levels: the regional level, using the voxel-wise whole-brain searchlight approach; the system level, using established language and semantic masks mainly distributed in the left temporal, frontal, and parietal regions; and the whole-brain level, using principal component analysis (PCA) over the whole-brain response patterns.

To illustrate the different predictions, consider the abstract words "to accumulate", "experience", and "to add". In a given region, if the neural representation is sensitive to word co-occurrence in a language corpus, "to accumulate" and "experience" would be coded by more similar neural patterns, resulting in significant correlation between the neural space and the language-corpus-based space. If the neural representation is sensitive to semantic features, "to accumulate" and "to add" would be coded by more similar neural patterns, resulting in significant correlation between neural space and semantic feature space.

Materials and Methods

Participants

Sixty-eight healthy college students (mean age = 22.5 years; range: 18–29 years) were recruited for behavioral ratings of semantic features and subjective semantic distance. Another 6 participants (3 females; mean age = 25 years; range: 23–29 years) completed the fMRI experiments, with each participating in 4 scanning sessions. All participants were right-handed, healthy, native speakers of Chinese with no history of neurological or psychiatric disorders. They were compensated for their participation and gave informed consent to the experimental protocol approved by the Human Subject Review Committee at Peking University.

Stimuli

The stimuli were 360 abstract, 2-character, bisyllabic Chinese words. Stimuli were first taken from previous studies on

English abstract words (Crutch et al. 2013; Troche et al. 2014), which contained 200 abstract nouns that were sampled from the low-imageability tail (<450) of the MRC Psycholinguistic Database. We then added all the direct hypernyms of the first meaning for each English word, and translated all of the stimuli into Chinese. After removing phrases and untranslatable words, we kept 360 Chinese words as the final stimulus set. These words covered a wide range of word frequency (Sun et al. 1997) (range: 0–2385 per 1.8 million, mean log word frequency = 1.45) and visual complexity (6–27 strokes and 2–9 radicals per word). About 200 words were primarily used as nouns, 120 as verbs and the remaining 40 as adjectives and adverbs. A complete list of the stimuli is shown in the Appendix.

Building Representational Dissimilarity Matrices of Abstract Words

We obtained 3 types of representational dissimilarity matrices (RDMs) of abstract words based on the language-corpus-based word co-occurrence, ratings of multiple semantic features and subjective ratings of semantic distance between word pairs. An RDM is a symmetric $n \times n$ matrix, with each column/row referring to 1 experimental condition and hence n the total number of experimental conditions (in this study, 360 abstract words). Each cell in the off-diagonal elements contains a number indicating the distance for each pair of words in a given measure.

Language-corpus-based RDM

This RDM reflects the distance between word vectors of co-occurrence patterns over a large corpus of text. We adopted 2 distinct, widely used algorithms: the Google word2vec tool and LSA. The word2vec algorithm in this study computes continuous vector representations of words based on the skip-gram architecture (Mikolov et al. 2013). With the Baidu Baike corpus containing approximately 1 billion word tokens, a vocabulary of the most frequent 249 222 words was first constructed via the Stanford parser. The word2vec tool was then used to train vector representations of words (https://code.google.com/p/word2vec/) with the following parameters: window size = 5, sub-sampling rate = 10^{-4}, negative sample number = 5, learning rate = 0.025, dimension number = 300. The LSA distance (http://www.lsa.url.tw/modules/lsa/) is measured through second-order co-occurrence. The Academia Sinica Balanced Corpus of Modern Chinese of 11 245 330 word tokens across 19 247 documents were used, and co-occurrence vectors were factored using singular value decomposition (employing 300 factors) to reduce the high dimensionality of the corpus. Note that 5 abstract words we used were not included in the LSA corpus and therefore were excluded from the LSA distance calculation. For both algorithms, the semantic distance was measured as 1 minus the cosine angle between feature vectors of each word pair. In the main analyses, we combined the distance metrics of the 2 algorithms to yield a single language-corpus-based RDM by computing the z scores across all of the word pairs for each distance metric and then averaging the 2 z scores for each word pair.

Semantic Feature RDM

Following Crutch and colleagues (Crutch et al. 2013; Troche et al. 2014), in separate norming sessions, we collected ratings on the relatedness of particular semantic features to the meaning of each abstract word on a 7-point Likert scale. Specifically, Crutch and colleagues gleaned from the literature a total of 12 semantic features that have been empirically/theoretically thought to affect the representations of some, if not all, abstract concepts: social interaction, morality, thought, emotion, valence (referred to as polarity in the Crutch et al. studies), time, space, quantity, sensation, action, ease of teaching, and ease of modifying. We made 2 modifications to this feature list. First, we added 1 semantic feature, arousal, which serves as another basic dimension of emotion that is at least partially dissociable from valence (Russell 1980; Kuperman et al. 2014). Second, we replaced ease of modifying with context availability, because our finding that the 2 indices showed only a weak correlation (r = 0.276) indicated that the former was not a valid representative of the latter, as Crutch and colleagues have proposed. Rating instructions for the 12 semantic features included in the Crutch et al. studies were taken from Troche et al. (2014). Rating instructions for arousal were taken from Bradley and Lang (1999), and for contextual availability from Clark and Paivio (2004). For each feature, approximately 17 ± 1.6 participants (range: 16–21, female ratio approximately 50%) were recruited. Data were collected via an online survey (http://www.sojump.com/), in which participants logged in and completed ratings. The inter-rater reliability for each rating was high (range of intra-class coefficients (ICC): 0.696–0.976). Based on the ratings of the 13 features specified above, each abstract word could be described as a 13-dimensional vector with each dimension indicating the extent to which the meaning of the given word is associated with each semantic feature. The feature-based RDM was then constructed by calculating the Euclidean distance between feature vectors of each word pair (Crutch et al. 2013).

Subjectively Rated RDM

We asked a group of healthy college students to rate the subjective distance of 100 abstract concept pairs randomly selected from the original 200 abstract words used in the studies of Crutch and colleagues. Pairwise combination of the 100 words resulted in 4950 word pairs, which were randomly divided into 8 rating sessions, each of which contained 615 or 620 word pairs. Sixteen participants (8 females) were recruited for each session and were asked to rate how close the 2 words were in meaning using a 7-point Likert scale (7 for the closest). The inter-rater reliability was high for each subset (ICC range: 0.773–0.877). The rating-based RDM was computed as 7 minus the averaged rating scores of 16 participants for each word pair, which resulted in a symmetric 100 × 100 matrix. We also included a common set of 10 additional word pairs in each subset and found that the ICC for the averaged ratings of these word pairs was 0.998, indicating excellent inter-subset reliability.

Partial Correlation Analysis

Spearman correlations were computed among the 3 types of RDMs. We also included a familiarity RDM as a nuisance variable. Familiarity ratings for all the word stimuli were obtained from 17 college students (9 females) using a 7-point Likert scale (7 for the most familiar), and the absolute values of the difference in the group-averaged familiarity scores between word pairs were used to construct the familiarity RDM. To explore the unique effects of the 2 aspects of abstract concepts in explaining the real semantic space, we then performed partial Spearman correlations, for example, computing the correlation between the subjectively rated semantic distances and 1 aspect of semantic distance (e.g., language-corpus-based measure), with the other aspect controlled for (in this case, the semantic feature-based distance). Ninety-eight words (producing 4753 pairwise distance scores) with complete data for behavioral ratings and language-corpus-based word co-occurrence were used in this analysis.

fMRI Data Collection

During the fMRI experiment, participants performed a familiarity judgment task in which they were asked to decide whether the word was familiar or unfamiliar to them by pressing 1 of the 2 response buttons using their right index or middle fingers. This task was adopted to avoid biases towards any specific aspects of abstract word meanings. We adopted a condition-rich rapid event-related design (Kriegeskorte, Mur and Bandettini 2008) with each word stimulus as an experimental condition, with words displayed for 500 ms and followed by a 1500-ms fixation period. Words were presented in the SONG font and subtended approximately 8.19° × 3.38° of visual angle. A red dot always appeared in the center of the screen for subjects to fixate on.

Each subject participated in 6 9-min runs per scanning session. In addition to 10-s fixation at the beginning and the end, each run consisted of 180 word trials presented exactly once and 60 interspersed null trials, whose orders were pseudorandomized so that no 2 characters occurred on consecutive word trials more than once and that both the first and the last presentation were word trials. Different trial orders were used per subject. The 180 words in each run were randomly selected from the stimulus set and were repeated in 2 sessions, that is, 12 runs. To obtain the fMRI data of 360 words, each subject participated in 4 sessions (24 runs) on 4 separate days.

MRI Parameters

Whole-brain imaging was performed on a Siemens PRISMA 3 Tesla MR scanner at the Center for MRI Research, Peking University. The functional data were acquired with a simultaneous multislice (SMS) sequence supplied by Siemens (slice planes scanned along the rectal gyrus, 64 slices, phase encoding direction from posterior to anterior, repetition time (TR) = 2000 ms, echo time (TE) = 30 ms, multi-band factor = 2, flip angle (FA) = 90°, field of view (FOV) = 224 mm × 224 mm, slice thickness = 2 mm, gap = 0.2 mm, voxel size = 2 × 2 × 2 mm). Each run consisted of 250 volumes. In addition, a high-resolution anatomical scan was acquired using the magnetization-prepared rapid acquisition gradient echo (MPRAGE) sequence for anatomical reference (192 sagittal slices, TR = 2530 ms, TE = 2.98 ms, FA = 7°, FOV = 224 × 256, voxel size = 0.5 × 0.5 × 1 mm, interpolated).

fMRI Data Preprocessing

Functional brain volumes were preprocessed and analyzed using statistical parametric mapping software (SPM12; Wellcome Department of Cognitive Neurology, London, UK). One run of one subject was excluded from analysis due to technical errors during data acquisition. After the first 4 volumes of each run were discarded, the functional data were corrected for slice timing and head motion. No participant showed excessive head movement (<1.1 mm or 1.2°) in any of the scanning sessions. Alignment of volumes from 4 scan sessions was then performed by coregistering volumes of the last 3 scan sessions, via their own mean functional images, to the mean functional image of the first session. The preprocessing time courses in the subjects' native space, without any spatial normalization or smoothing, were used for the following activation pattern estimation. For the purposes of image transformation between the MNI and native space, we coregistered the mean functional image of the first session to the structural image and obtained the original and inverse normalization parameters from the SPM12 segmentation tool.

Obtaining Whole-brain Activation Patterns of Each Abstract Word

For each word stimulus, a general linear model (GLM) was estimated for each voxel, which included for each run 1 regressor containing the onsets of a given word and another regressor containing the onsets of all other word trials, both convolved with the canonical hemodynamic response function. This approach has been shown to be more representative of the true activation magnitudes unique to each trial type than other model estimation methods for rapid event-related designs (Mumford et al. 2012). Also included in the GLM were 6 head motion parameters and a global mean predictor of each run. The high-pass filter was set at 128 s. After model estimation, the whole-brain beta-weight image for each word was produced by contrasting the given word versus the baseline.

Regional-level RSA: Voxel-wise Whole-brain Searchlight

To determine the semantic content in the activation patterns across voxels, we performed a spheric searchlight analysis (Kriegeskorte et al. 2006) with the following procedures. First, the activation maps for each word in each subject's native space were normalized to the MNI space using the normalization parameters from the SPM12 segmentation tool. The resampling voxel size was 2 × 2 × 2 mm. Second, gray matter voxels were defined as those with a probability higher than 0.4 in the SPM gray matter mask (a total of 157 904 voxels). For each gray matter voxel in each subject, we extracted the activation values of all of the 360 abstract words from a spheric region of interest (ROI; radius 6 mm, corresponding to 123 voxels) and computed the neural RDM as 1-Pearson correlations of all word pairs over all voxels within the spheric ROI. Note that this computation was restricted to ROIs containing at least 30 voxels. Third, the neural RDMs at a given voxel were z-transformed and averaged across all 6 subjects and the resulting group-level neural RDM was compared with semantic RDMs using Spearman's rank correlations, producing a correlation coefficient for this voxel. By moving the searchlight center throughout the cortex, we finally obtained whole-brain r-maps that contained 151 836 correlation coefficients. The significance thresholds of these maps were determined using FDR $q = 0.01$ on the corresponding p-maps (1-tailed, indicating that correlation is significantly greater than zero), combined with a minimum cluster size of 800 mm^3 (100 voxels). Negative correlations were not considered due to lack of a priori expectation for their interpretation (Chikazoe et al. 2014). To remove the potential influence of familiarity and visual complexity, we conducted partial correlation analyses by taking familiarity and pixelwise RDMs as nuisance variables. The pixelwise RDM was generated by computing the 1-pixelwise correlation between pairs of the black-and-white silhouettes of word forms (Peelen and Caramazza 2012). With the pixelwise RDM, we also conducted a searchlight analysis to validate our data preprocessing and analysis methods.

System-level RSA

Two system-level masks were defined to investigate the potential contribution of high-level linguistic and semantic areas to the linguistic contextual and feature aspects of abstract concepts. The high-level linguistic mask was obtained from a group-level language localizer via the contrast between sentences and nonword lists in 220 participants (Fedorenko et al. 2010) (https://evlab.mit.edu/funcloc/download-parcels), including regions in the left hemisphere covering lateral temporal regions, temporoparietal junction, inferior frontal gyrus, and precentral gyrus. The second mask, implicated in semantic processing, was identified in a meta-analysis across 120 functional neuroimaging studies (Binder et al. 2009). Specifically, the mask was obtained by transforming the thresholded activation likelihood estimate map using all semantic contrasts (Fig. 3 in Binder et al. 2009) from the Talairach space into the MNI space using the "tal2icbm" transformation (Lancaster et al. 2007). This semantic processing mask has been found to intrinsically comprise 3 stable modules by graph-theoretical analyses of the resting-state functional connectivity patterns (Xu et al. 2016). The masks of these modules consisted of sets of voxels assigned with the module membership at sparsity = 0.4 (Fig. 4, Validation Analysis 3 in Xu et al., 2016).

With these masks of interest, system-level RSA was performed using the following procedures (Kriegeskorte, Mur, Ruff, et al. 2008): 1) reverse-normalizing each mask defined in the MNI space to each subject's native space; 2) extracting activation patterns in a given mask to each of 360 abstract words and subtracting the word-general activation pattern (calculated by averaging all but the target 1 word patterns) from each word pattern; 3) computing neural RDMs as the 1-Pearson correlation between all word pairs in each subject and averaging the z-transformed neural RDMs across all 6 subjects to obtain a group-level neural RDM; and 4) conducting Spearman correlation analyses between neural RDMs and language-corpus-based and feature-based RDMs, respectively.

Whole-brain-level RSA—PCA of the Whole-brain Patterns

To examine the types of semantic information about abstract concepts embedded at the whole-brain level, we used PCA to uncover the underlying neural dimensions from the whole-brain activation patterns (Huth et al. 2012, 2016) and then adopted the RSA approach to assess the correlation between neural space built from the first few neural principal components (PCs) and the abstract conceptual space estimated using various distance measures.

Voxel Selection

Because not every voxel in the brain is expected to be involved in our fMRI task, we included voxels yielding stable response to a given word for PCA. Within the gray matter mask in each subject's native space, we computed for each voxel the "stability score" (Mitchell et al. 2008), which measures the consistency of each voxel's responses to all the word stimuli across a total of 12 repetitions. We first extracted the beta image for each repetition of each word from the GLM we built previously and then assigned each voxel a 12 × 360 matrix, where the entry at row *i*, column *j*, was the value of this voxel during the *i*th repetition of the *j*th word. The stability score for this voxel was then computed as the average pairwise correlation over all pairs of rows in this matrix. We selected voxels whose stability score was higher than 0.01 in the main results. This yielded 7592 voxels in Subject 1, 8777 voxels in Subject 2, 8769 voxels in Subject 3, 10 953 voxels in Subject 4, 5993 voxels in Subject 5, and 7881 voxels in Subject 6. We found that varying the stability threshold changed the number of voxels analyzed, but did not strongly affect the neural PC interpretation.

Principal Component Analysis

To investigate the underlying neural dimensions of abstract concepts that are shared across individuals, following Huth et al. (2012), we pooled voxels from all 6 subjects (49 965 voxels) and applied PCA to the combined data, resulting in 360 PCs. The number of neural PCs was determined by the elbow of a scree plot illustrating the variance explained by each PC. As shown in Figure 4*d*, the elbow point was found at the third component; therefore, the first 2 PCs were taken as the potential organizing dimensions of abstract concepts. For the purpose of illustration, the projection of PC scores in the native space was normalized to the MNI space and spatially smoothed with a Gaussian kernel at 4 mm full width at half maximum (FWHM).

RSA and Correlation Analysis

We used RSA to evaluate whether and how linguistic contexts and semantic features of abstract concepts are represented in the first few neural PCs. As the neural PC1 is expected to contain visual information of word stimuli (see results), we constructed a neural-PC-derived RDM by first regressing out visual complexity (numbers of strokes and radicals) from the PC1 loadings, standardizing both the residue PC1 loadings and the original PC2 loadings, and finally computing the Euclidean distance between word pairs in these 2 PCs. The Spearman's rank partial correlation was then computed between the neural RDM and semantic RDMs with familiarity as a nuisance variable. Additional validation analyses were carried out using the neural RDM derived from raw PC loadings, in which familiarity and pixelwise RDMs were simultaneously controlled for as covariates in RSA. Finally, to understand the information conveyed in each neural PC, we conducted partial Pearson correlations between neural PC loadings and the ratings of the 13 semantic features collected above with familiarity ratings as a nuisance variable.

Results

The analysis scheme is presented in Figure 1. We sampled the language-corpus-based, behavioral rating, and BOLD fMRI data of 360 abstract words. We first established that both language-corpus-based and semantic-feature-based word distance measures independently explain significant portions of variance in the subjectively rated semantic distance. RSA was then conducted between the 2 aspects of abstract concepts and neural response patterns on 3 different scales of brain organization.

Language-corpus-based and Behavioral Rating Results

Representational dissimilarity matrices (RDMs) of 360 abstract words were constructed based on the pairwise (64 620 pairs in total) distance derived from word co-occurrence in large language corpora and semantic feature ratings, respectively. Therefore, we obtained, for each measure, a symmetric matrix with each column/row referring to an abstract word and each cell in the off-diagonal elements containing a distance value for a word pair in that measure. Multidimensional scaling

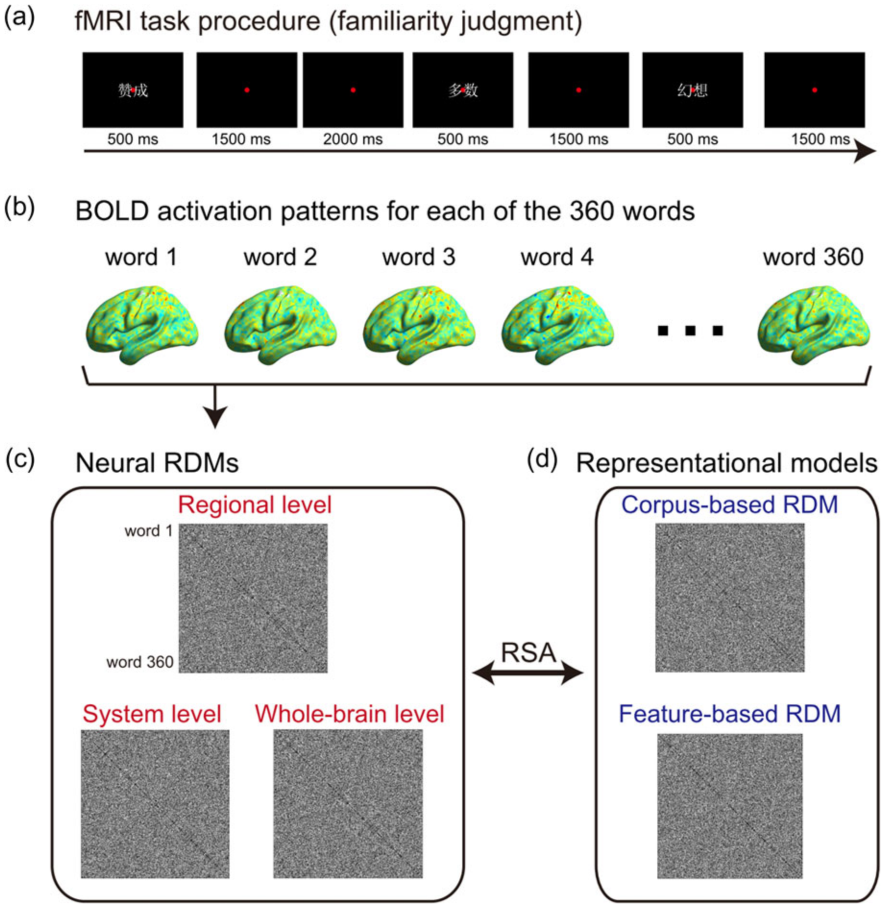


Figure 1. The flowchart of fMRI task and analysis procedures. (*a*) A condition-rich rapid event-related design was adopted for fMRI data collection, with each abstract word as an experimental condition and a total of 360 conditions. (*b*) The whole-brain activation patterns of each abstract word were obtained. The neuroimaging data are mapped on cortical surfaces using the BrainNet Viewer (Xia et al. 2013). (*c*) Neural representational dissimilarity matrices (RDM) were constructed by computing the correlation distance of the activation patterns between all word pairs at the 3 neural scales: at the regional level, RDMs are computed from a spherical region centering each voxel of the gray matter regions; at the system level, RDMs are computed from all the voxels within the established language or semantic regions; at the whole-brain level, an RDM is computed from the PCs of the whole-brain activation patterns. (*d*) Two representational models of abstract words were obtained: the corpus-based distance was derived from 2 popular algorithms of natural language processing (word2vec and LSA) to reflect both local and global linguistic contextual information; the feature-based distance was computed based on the ratings of 13 semantic features drawn from literature. Finally, RSAs were conducted to evaluate the correspondence between cognitive and neural RDMs of abstract concepts.

(MDS) was used to project 2 of these RDMs into a 2-dimensional space for visualization of distance among abstract words (Fig. 2). Subjective ratings of semantic distance between word pairs were collected for a subset of 100 words to reflect the ground truth semantic space. The Spearman correlations between these RDMs were then computed to examine the role of linguistic contexts and semantic features in the ground truth semantic space of abstract concepts (Fig. 3).

Language-corpus-based RDM

The linguistic-based semantic distance is obtained in the co-occurrence statistics embedded in large natural language corpora. We adopted 2 popular algorithms—word2vec and LSA, which compute word co-occurrence patterns in local or global linguistic context, respectively. Both algorithms generated a real-valued vector for each word. Word dissimilarity was thus defined as the cosine distance between word vectors. The RDMs based on each algorithm were significantly correlated with each other (r = 0.430, $P < 10^{-10}$) and were averaged to yield a composite language-corpus-based RDM for the main analyses. RSA results with each of these 2 RDMs are shown in Supplementary Figure S1 and Supplementary Table S1.

Semantic-feature-based RDM

Semantic distance based on multiple features was obtained following previous studies (Crutch et al. 2013), with minor modifications. For each of the 13 semantic features (i.e., social interaction, morality, thought, emotion, valence, arousal, time, space, quantity, sensation, action, ease of teaching and context availability), we collected ratings of its association to the meaning of each word on a 7-point Likert scale, obtaining a 13-dimensional vector for each word. The Euclidean distances between the feature vectors of all word pairs were calculated to construct the semantic feature RDM. It may be argued that 2 semantic features—ease of teaching and context availability—reflect more strongly verbal associations. We also constructed a semantic feature RDM taking them out and based on the remaining 11 features and found that these 2 semantic feature RDMs were almost perfectly correlated (r = 0.987, $P < 10^{-10}$). Thus the semantic feature RDM based on 13-dimensional vectors were used in the main analyses. RSA results using the 11

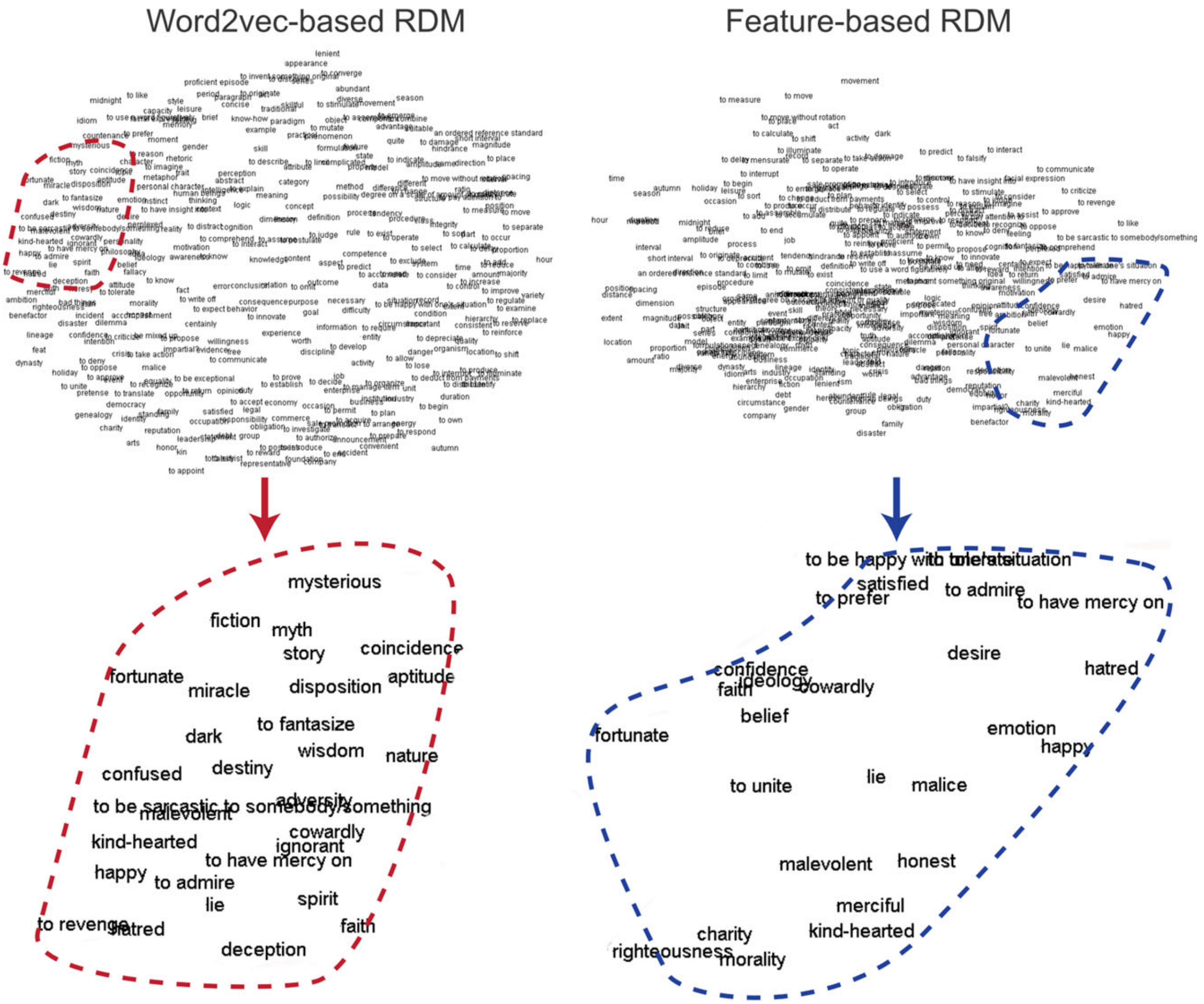


Figure 2. Visualization of semantic distances between abstract words based on word2vec and semantic features, respectively. Multidimensional scaling on the representational dissimilarity matrices was used for data reduction of the dissimilarity matrix onto 2 dimensions. Words placed close together indicated shorter semantic distance in a given metric. The words in circles are enlarged for illustration purposes.

semantic features, shown in Supplementary Figure S1 and Supplementary Table S1, were highly similar to those with the 13-dimensional semantic space.

There was a modest, yet highly significant, correlation between these 2 types of semantic distance measures of abstract concepts (Fig. 4*a*; $r = 0.297$, $P < 10^{-10}$). Both were significantly correlated with the subjective-rating-based semantic distance (Fig. 3; language-corpus-based distance: $r = 0.477$; semantic-feature-based distance: $r = 0.339$; $Ps < 10^{-10}$). To assess the unique effects of each representational model in explaining variance in the subjective semantic distance, we conducted a partial correlation controlling for the other cognitive dimension and found that the correlations remained highly significant for both measures (language-corpus-based distance with semantic-feature-based distance controlled: partial $r = 0.414$; semantic-feature-based distance with language-corpus-based distance controlled: partial $r = 0.226$; $Ps < 10^{-10}$). These results indicate that linguistic contexts and semantic features make unique contributions to the ground truth semantic space of abstract concepts.

Neural RDMs and RSA Results on 3 Neural Scales

The BOLD fMRI responses for each of the 360 visually presented abstract words were collected using the condition-rich event-related design (Kriegeskorte, Mur and Bandettini 2008) in 6 college subjects. Neural RDMs were constructed for each subject, using the word pairwise correlations of BOLD response patterns, at 3 different scales: the regional level, using the voxel-wise whole-brain searchlight approach; the system level, using established language and semantic masks; and the whole-brain level, using the PCA space derived from the whole-brain response patterns. The group-level neural RDMs were generated by averaging neural RDMs of all subjects and then correlated with the cognitive RDMs to assess their correspondence. Differences in familiarity and low-level visual features were treated as nuisance variables to control for in these analyses and that the main result patterns were similar when these variables were not included as covariates (Supplementary Fig. S2*a* and Supplementary Table S2). Note that we adopted the sample size following studies using similar condition-rich design to

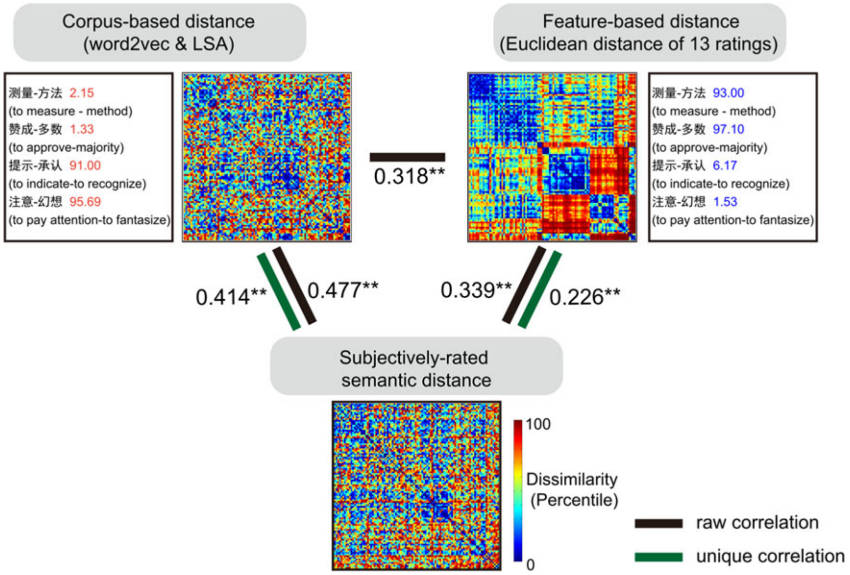


Figure 3. Correlations between representational models and subjective-rated semantic distance of abstract words. Four sample word pairs with semantic distances ranking top 10 percentile in one measure and bottom 10 percentile in the other are shown; numbers indicate the percentile ranks of semantic distance for each measure. The subjectively rated distance matrix, considered as the ground truth distance, was based on explicit rating of pairwise semantic distance of 100 randomly selected abstract words. Correspondence among these distance matrices were evaluated using Spearman correlation, with raw correlation referring to partial correlation when familiarity as a nuisance variable was controlled for and unique correlation to partial correlation when both familiarity and the other aspect of abstract concepts were controlled for. Asterisks indicate significance levels of the correlations relative to zero (2-tailed); $^{**}P < 10^{-10}$.

maximize condition (word) numbers, and conducted the RSA based on the group mean neural RDM (Kriegeskorte, Mur, Ruff, et al. 2008). In an additional leave-one-subject-out analysis we validated whether the main results were driven by specific subject outliers. In this validation we obtained the group-level neural RDMs in all but one subjects. The results were largely stable except for the regional-level RSA results of the corpus-based RDM (see the following sections for details).

Region-level RSA results—voxel-wise whole-brain Searchlight

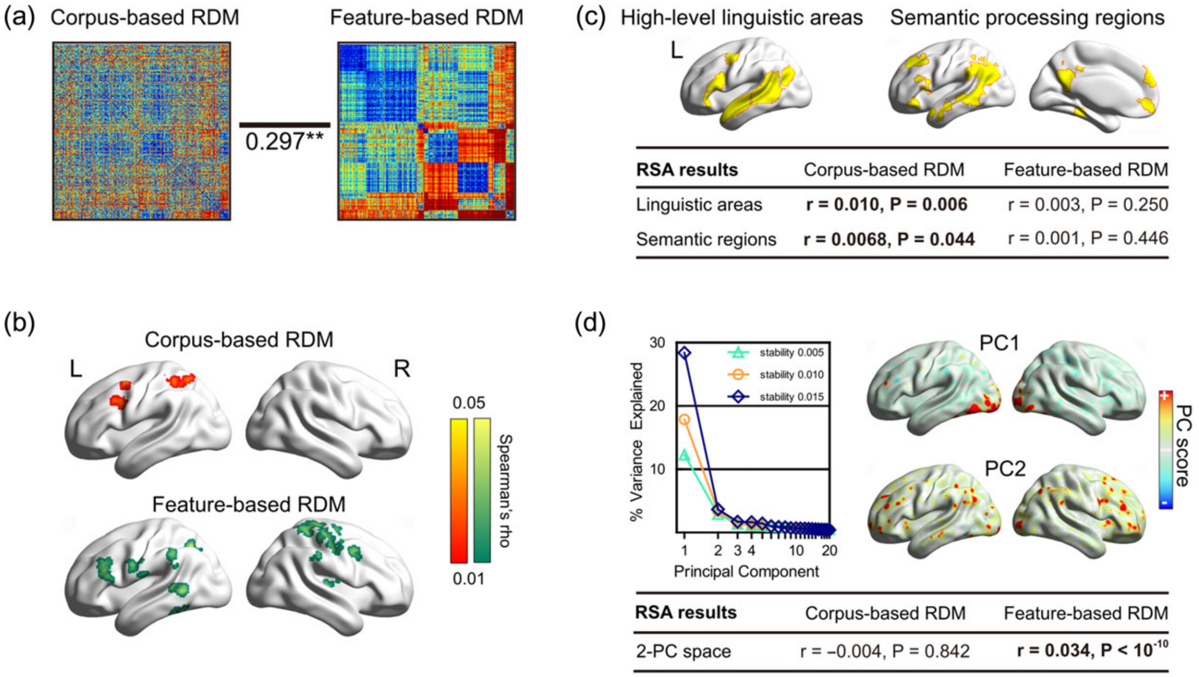
We performed whole-brain searchlight analyses (Kriegeskorte et al. 2006), in which neural RDMs in a given sphere that was centered on each voxel of the brain (radius 6 mm; corresponding to 123 voxels) were computed and their relationships with various cognitive RDMs were examined. Given that the stimuli were visual words, we first conducted a validation analysis for low-level visual effects—a whole-brain searchlight using the word stimuli pixel dissimilarity matrix. The results showed significant effects in early visual areas (peak MNI coordinate: 22, −102, 0; Supplementary Fig. S3), replicating previous findings showing the localization of low-level visual features of stimuli in the occipital cortex (Peelen and Caramazza 2012; Clarke and Tyler 2014) and validating our preprocessing and analysis procedures with the current fMRI data.

The RSA searchlight mapping (Supplementary Fig. S2*b*) for the language-corpus-based distance of abstract concepts yielded one significant cluster in the left posterior inferior and middle frontal gyri (IFG/MFG) (FDR corrected $P < 0.01$, cluster size > 800 mm^3). Regions containing information about the feature-based conceptual space were found to be widely distributed in both hemispheres, including the bilateral posterior IFG/MFG, left triangular IFG, left posterior intraparietal sulcus (IPS), left posterior middle temporal gyrus (MTG), left inferior temporal gyrus (ITG), left amygdala, right postcentral and precentral gyri and adjacent parietal and frontal cortices, and a cluster in the right insula. These 2 sets of results overlapped in the left posterior IFG/MFG (68 voxels).

We further tested the unique effects of each cognitive dimension by using partial correlation to control for the effects of the other dimension while carrying out RSA with 1 dimension (Fig. 4*b*). The effects of the language-corpus-based distance remained significant in the left posterior IFG/MFG, which extended to the left precentral gyrus. Additional correlations were found in the left inferior parietal lobule. The unique effects of semantic features were similar to the raw correlation effects except that the left posterior IFG/MFG did not approach significance and that additional correlations were found in the bilateral supramarginal gyrus, left precentral gyrus and right amygdala. No overlap was observed.

Detailed information about the correlation peak coordinates and cluster sizes is provided in Supplementary Table S3. Changing the size of searchlight spheres (from 6 mm radius to 8 mm radius, corresponding to 2056 mm^3, 257 voxels) produced similar results (Supplementary Fig. S2*c*). In the leave-one-subject-out validation analyses (Supplementary Table S4), all the feature-based clusters remained significant in all iterations. The 2 corpus-based clusters were significant in all but 1 iteration (when subject 6 was excluded), indicating that these 2 clusters may be driven by 1 particular subject rather than being stable across subjects.

Note that the current whole-brain searchlight did not yield clusters in the left anterior temporal lobe, a region that has been found to be more strongly activated by abstract than concrete words (Binder et al. 2009; Wang et al. 2010). We thus performed an RSA in this region, defined from the abstract versus concrete contrast in a synonym judgment task (Hoffman et al. 2015). We first calculated the temporal signal-to-noise ratio (TSNR) of this region by dividing the mean of motion-corrected, unsmoothed time series across the whole run by its standard deviation and averaging the TSNR across all the runs for each subject. The TSNR of this region ranged from 31 to 36 in our



RSA results	Corpus-based RDM	Feature-based RDM
Linguistic areas	**r = 0.010, P = 0.006**	r = 0.003, P = 0.250
Semantic regions	**r = 0.0068, P = 0.044**	r = 0.001, P = 0.446

RSA results	Corpus-based RDM	Feature-based RDM
2-PC space	r = –0.004, P = 0.842	**r = 0.034, P < 10^{-10}**

Figure 4. RSA results of abstract words. (*a*) The relationship between the corpus-based and feature-based representational dissimilarity matrices (RDMs) of abstract concepts. (*b*–*d*) RSA at 3 neural scales. Only unique correlations are shown. See Supplementary Figure S2*b* for raw correlation results. (*b*) Regional-level RSA (i.e., voxel-wise searchlight with 6-mm-radius spheres) revealed regions whose activation patterns had significantly positive correlation with the corpus-based distance and the feature-based distance, respectively. The significant levels of correlations greater than zero (1-tailed) were determined at FDR corrected $P < 0.01$, combined with a minimum cluster size of 800 mm^3 (100 voxels). (*c*) For system-level RSA, neural RDMs were derived from dissimilarity of activation patterns across all voxels in brain regions implicated in high-level linguistic processing (Crutch et al. 2013) and semantic processing (Binder et al. 2009). The table below shows the correlations between neural RDMs and 2 representational models of abstract words (1-tailed). (*d*) For whole-brain level RSA, PCA was performed to extract the PCs from whole-brain activation patterns of 360 abstract words. The scree plot showing the variance explained by each of the top 20 PCs at various stability thresholds identifies the first 2 PCs as the important dimensions. Cortical maps of PC scores in one representative subject are shown. See Supplementary Figure S4*a* for cortical maps in the remaining subjects. The neural RDM was computed as the Euclidean distance between word pairs based on the 2 PCs. Visual properties of words were removed from PC1 loadings to exclude potential contamination of visual effects. The table below shows the correlations between the neural RDM and 2 representational models of abstract words (1-tailed). Numbers in bold indicate $P < 0.05$.

subjects, indicating acceptable signal levels (Murphy et al. 2007). We still did not find any significant effects of either the corpus- or feature-based aspects of the abstract semantic space here (rs < –0.001, Ps > 0.576).

System-level RSA Results

Past fMRI studies have consistently revealed that language and semantic processing activate distributed regions in the left-lateralized temporal, frontal, and parietal regions. We selected 2 previously published maps as masks for the brain networks underlying these 2 processing domains. Brain regions for high-level linguistic processing were derived from a group-level localizer contrasting sentences to nonword lists in 220 participants (Fedorenko et al. 2010). Brain regions for semantic processing were taken from the results of a comprehensive meta-analysis across 120 functional neuroimaging studies involving various semantic contrasts (Binder et al. 2009). Different from voxel-wise searchlight analyses, the system-level RSA tested the possibility that semantic knowledge may be represented across multiple regions in the system as a whole (i.e., larger-scale population coding). We thus constructed system-level neural RDMs based on the dissimilarity of activation patterns across all of the voxels in these 2 masks.

The high-level linguistic mask (Fig. 4c) included regions in the left hemisphere covering lateral temporal regions, temporo-parietal junction, inferior frontal gyrus and precentral gyrus. Its neural RDM significantly correlated with the language-corpus-based RDM (r = 0.011, P = 0.002), not with the semantic feature RDM (r = 0.006, P = 0.062). The unique correlations between this neural RDM and each of the 2 cognitive RDMs showed similar patterns. Significant correlation was found with the language-corpus-based RDM (partial r = 0.010, P = 0.006), not with the semantic feature RDM (partial r = 0.003, P = 0.250). That is, the significant association between the neural RDM of the language mask and the language-corpus-based RDM was not explained by the potential confounding of semantic feature similarities. This effect was robust in the leave-one-subject-out analysis (Supplementary Table S4).

The semantic processing mask (Fig. 4c) included regions in the middle temporal gyrus, the fusiform and parahippocampal gyri, the inferior frontal gyrus, the dorsomedial prefrontal cortex and the ventromedial prefrontal cortex, the inferior parietal lobe and the posterior cingulate gyri. Its neural RDM showed significant correlation with the language-corpus-based RDM (r = 0.0073, P = 0.034), not with the semantic feature RDM (r = 0.003, P = 0.241). These results held when we considered the unique effects of the 2 cognitive RDMs, that is, excluding the effects of the other RDM (language-corpus-based: partial r = 0.0068, P = 0.044; semantic-feature-based: partial r = 0.001, P = 0.446). Leave-one-subject-out analyses revealed the similar tendency of the language-corpus-based RDM to approach significance (Supplementary Table S4). Interestingly, when we looked at sub-networks within this large semantic mask that were defined using graph-theoretical modularity analysis on the intrinsic functional connectivity patterns (Xu et al. 2016), both the raw and unique correlations with the language-corpus-based RDM remained significant in the neural RDM of the left perisylvian module (raw $r = 0.0071$, $P = 0.037$; partial $r = 0.0067$, $P = 0.046$), whose anatomical locations correspond with the high-level linguistic areas. These effects were not found in the default mode module (Ps > 0.153) and only the raw, but not the unique, correlation approached significance in the left fronto-parietal module (raw $r = 0.0073$, $P = 0.034$; partial $r = 0.0045$, $P = 0.130$). This indicates that the linguistic contextual information of abstract concepts in the semantic mask is primarily driven by the high-level linguistic areas, converging with the results using the high-level linguistic mask above.

Whole-brain level RSA Results—PCA of the whole-brain Patterns

Finally, we explored how the whole-brain activation patterns may represent linguistic contextual and feature aspects of abstract concepts. At this level, we considered those voxels whose responses for a given word were stable across multiple repetitions (Mitchell et al. 2008) (see Methods). The whole-brain neural RDMs constructed directly on these voxels at various stability thresholds did not correlate with either the language-corpus-based or semantic-feature-based RDM (Ps > 0.138), which may be due to the low signal-to-noise ratio in the whole-brain fMRI data.

Recent research has suggested that the PCs derived from whole-brain activity patterns are sensitive to major conceptual dimensions such as animacy or social interaction (Huth et al. 2012, 2016). We adopted this approach to obtain the neural dimensions along which the whole-brain activity patterns of abstract concepts are organized. Following previous studies (Huth et al. 2012), we applied PCA to the activation patterns of 360 abstract words in the pooled stable voxels of all the subjects to extract the neural PCs shared across individuals. As shown in Figure 4*d*, for various stability thresholds, the characteristic "elbow" points of the scree plots were always found at the third PC. The first 2 PCs were thus considered the most important dimensions. The following results were conducted with the stability threshold set to 0.01, and in additional analyses we found that varying the stability threshold did not strongly affect the neural PC interpretation (see below).

The first 2 PCs explained 17.87% and 3.37% of the total variance, respectively. These results for abstract concepts were comparable to those reported with concrete concepts (Huth et al. 2012), which found that on average the first 4 PCs explained 19% of the total variance. Cortical maps of PC scores for 1 representative subject is shown in Figure 4*d* and for the remaining subjects in Supplementary Figure S4*a*. These maps reveal that the neural PC1 was positively associated with word-evoked activations in occipital, parietal and frontal regions and that the neural PC2 was evenly distributed in both hemispheres. Previous investigation of neural PC1 (Huth et al. 2012) and our observation of its association with the occipital regions indicate that PC1 may contain information about stimulus properties. Indeed, word loadings on PC1 were significantly associated with visual complexity of the word stimuli (numbers of strokes: $r = 0.277$, $P < 10^{-7}$; number of radicals: $r = 0.196$, $P < 0.001$), whereas PC2 loadings were not (Ps > 0.725). To rule out the potential contamination of visual effects, we regressed out the numbers of strokes and radicals from the PC1 loadings and used the residue PC loadings in the following analyses.

To evaluate the underlying principles for the whole-brain activity pattern of abstract words, we constructed a neural RDM using the Euclidean distance between word pairs in these 2 PCs (Fig. 4*d*). This neural RDM significantly correlated with the feature-based RDM ($r = 0.035$, $P < 10^{-10}$) and not with the language-corpus-based RDM ($r = 0.006$, $P = 0.055$). Significant correlation with the feature-based RDM was also found when the stability threshold was 0.005 ($r = 0.036$, $P < 10^{-10}$) and 0.015 ($r = 0.033$, $P < 10^{-10}$). That is, the relationship between word pairs in this neural RDM is associated with how related 2 abstract concepts are in terms of their semantic features, not in terms of how likely they would be to occur in similar linguistic contexts. RSA using partial correlations yielded the same results: the 2-dimensional neural RDM correlated with the semantic feature RDM when the language-corpus-based RDM was controlled for (stability 0.01: partial $r = 0.034$, $P < 10^{-10}$; stability 0.005: partial $r = 0.035$, $P < 10^{-10}$; stability 0.015: partial $r = 0.032$, $P < 10^{-10}$), and not with the language-corpus-based RDM when including the semantic feature RDM as a covariate (stability 0.01: partial $r = -0.004$, $P = 0.842$; stability 0.005: partial $r = -0.004$, $P = 0.862$; stability 0.015: partial $r = -0.005$, $P = 0.881$). This feature-related effect was stable in the leave-one-subject-out analysis (Supplementary Table S4).

We performed additional validation analyses using the neural RDM derived from raw PC loadings and controlling for visual complexity in the Spearman's partial correlation, as we did in the searchlight and system-level RSA. The results of cognitive RDMs represented in the whole-brain activation patterns were similar: this neural RDM significantly correlated with the semantic-feature-based RDM ($r = 0.020$, $P < 10^{-6}$), and not with the language-corpus-based RDM ($r = -0.002$, $P = 0.695$). Similar results were found with the unique effects (the semantic-feature-based RDM, partial $r = 0.022$, $P < 10^{-7}$; the language-corpus-based RDM, partial $r = -0.008$, $P = 0.982$).

To understand the potential semantic features that the neural PCs may encode, we examined with which semantic feature ratings the PC loadings were associated (Supplementary Fig. S4*b*). Words with top 10 loadings on each PC and their semantic feature ratings are shown in Supplementary Table S5. Correlation analyses showed that among 13 semantic features, the neural PC1 significantly correlated with valence ($r = -0.119$, $P = 0.025$). Similar correlations were found when the stability threshold was 0.005 ($r = -0.123$, $P = 0.020$) and 0.015 ($r = -0.104$, $P = 0.049$). Not regressing out visual complexity from the PC1 loadings first ($r = -0.088$, $P = 0.095$) and not controlling for familiarity ($r = -0.174$, $P = 0.001$) had minimal influence on this correlation. That is, the most important single semantic feature (dimension) that affects the whole-brain patterns of abstract concepts is whether the concept is deemed positive or negative. None of the correlations between PC2 and semantic features approached significance.

Discussion

To understand how abstract words are represented in the brain, we tested the neural correlates of 2 classical representational models: linguistic contexts and semantic feature decomposition. The distance spaces of 360 abstract words were constructed using computational language-corpus-based algorithms and the behavioral ratings of 13 semantic features, respectively. The neural response patterns of abstract words on 3 scales—regional, system, and whole-brain—were obtained and compared to these 2 spaces using RSA. The main findings were as follows: both types of organizational principles made unique contributions to the cognitive semantic space of abstract concepts, as defined by subjective ratings of semantic distance; both affected neural representations, but on different scales and involving different brain regions. The language-corpus-based co-occurrence patterns of abstract concepts were significantly associated with the neural patterns of the high-level linguistic system. In contrast, the high-dimensional semantic-feature-based space was represented in a more distributed manner: it was significantly associated with the whole-brain activation patterns and with the activation patterns of regions widely distributed across both hemispheres including the left triangular IFG, left ITG and posterior MTG, left IPS, bilateral supramarginal gyri and amygdala. Taken together, both linguistic contextual information and feature composition of abstract concepts are implemented in the brain, with dissociable neural correlates.

The 2 principles of meaning representation, especially for abstract words, is a classical debate. Current dominant neuroanatomical models embraced the feature theories, attempting to ground semantic representations in sensory, motor and affective systems of the brain (Barsalou 2008; Martin 2016). The distributional approaches of semantics in the natural language processing field (Landauer and Dumais 1997; Mikolov et al. 2013), however, represent words as vectors reflecting the co-occurrence patterns with other words in large language corpora and has recently made significant progress with the help of neural network learning models in capturing semantic similarities. These 2 principles largely correspond to the nonverbal and verbal systems proposed by the Dual Coding Theory (Paivio 1986), which have been incorporated in some unifying theories of semantic cognition, such as the hub-and-spoke/controlled semantic cognition model (Lambon Ralph et al. 2017), and also debated extensively in the representation of abstract concepts (Borghi et al. 2017). The current study is the first to explicitly investigate whether and how these 2 principles are respected by the brain, revealing their dissociations at various neural levels. Below we discuss the brain basis of each principle in detail.

Representing Abstract Words in Verbal Co-occurrence Patterns

To represent meaning in terms of linguistic context, we used word-word distances obtained from natural language processing models, that is, language-corpus-based word co-occurrence pattern analyses, as 1 potential means. RSA results showed that this space, not the semantic feature space, significantly correlated with the neural response pattern of the classical language system as a whole (the left lateral temporal, inferior parietal, and inferior frontal regions). Interestingly, no significant correlation was observed in the homologous right-hemisphere regions ($P = 0.46$), indicating that unlike natural speech that recruits both hemispheres for comprehension (Huth et al. 2016), the corpus-based space is mainly left lateralized. This system was defined by the contrast between sentences and nonword lists (Fedorenko et al. 2010) and has been found to show functional specificity for high-level linguistic processing, but not for a variety of non-linguistic functions including arithmetic, cognitive control, music and working memory (Fedorenko et al. 2011). The manner in which these regions are engaged in language processing, and the aspects of language with which they are involved, however, is not fully understood. Our results indicate 1 particular mechanism through which they may represent language—the neural response pattern of the whole system encodes the statistics-derived word distance patterns in natural language. We further examined each of the 6 subregions within the system and found significant correlations in the triangular and orbital portions of the inferior frontal gyrus ($Ps < 0.0034$), surviving corrections for multiple comparisons. This is consistent with the findings of a recent study (Carota et al. 2017), lending support to the sensitivity of the left inferior frontal gyrus to distributional semantics.

Given that the information contained by text-derived word distance could be very rich, including semantic, syntactic and/or lexical form regularities (Mikolov et al. 2013), the exact nature of representation here warrants further investigation. At least some of this representation likely includes semantic information, as these regions fall within the semantic processing network (Binder et al. 2009), distributed in a left-lateralized network comprised of temporal, inferior frontal and inferior parietal regions, which was obtained through meta-analyses of 120 neuroimaging studies of semantic memory using multiple types of semantic contrasts differentiating between semantic and linguistic surface form (phonology, orthography) processing. That is, regions in this linguistic mask are more deeply involved in the semantic aspect of language processing. The neural response pattern of the semantic processing mask also significantly associated with the language-corpus-based co-occurrence space. Intrinsically, this large semantic network is decomposed into 3 sub-networks (Xu et al. 2016). The subnetwork that corresponds to the language mask in the left perisylvian cortices was the only subnetwork showing a significantly unique correlation with the language-corpus-based distance of abstract concepts, providing supporting evidence for the sensitivity of this subnetwork to the co-occurrence of abstract concepts in language.

We considered 2 popular algorithms of extracting word co-occurrence statistics in natural language: LSA and word2vec. Both are not simple word co-occurrence counts but are some kind of second-order co-occurrence patterns that represent words as continuous vectors from large language corpora, with the assumption that words that are close in meaning will appear in similar contexts (Firth 1957). LSA is a count-based algorithm that constructs a matrix containing word counts per document (emphasizing global context) and obtains word vectors by retaining a few hundred dimensions after matrix factorization. The continuous skip-gram model of word2vec we used adopts the neural network approach by training word vectors to predict surrounding linguistic context given a single word (hence the local context). The main analyses in our study used the mean of these 2 spaces to capture both global and local contextual information. Analyses using these measures separately showed that the word2vec distance tended to be more strongly correlated with the subjectively rated and feature-based semantic space at the behavioral level and with the neural distance at the system level (Supplementary Table S1), suggesting that the main results using the mean of these 2 spaces may be primarily driven by the word2vec distance. This is consistent with the state-of-art performance of word vectors provided by word2vec in extracting semantic similarities (Mikolov et al. 2013). While the types of information embedded in the vector space in neural network models are unspecified, the current findings show that they are indeed neurobiologically realistic, especially within the language network mask. Future studies are also warranted to directly compare the neural correlates of different types of linguistic information extracted from various computational models.

Representing Abstract Words in a High-dimensional Semantic Feature Space

The dominant view of representing concepts in brain-based experiential features attempts to explain abstract meaning representation through a high-dimensional feature space (Crutch et al. 2013; Binder et al. 2016; Fernandino, Binder, et al. 2016). We constructed a 13-dimensional semantic feature space following previous studies establishing its cognitive relevance (Crutch et al. 2013; Primativo et al. 2016). In the search for the brain bases of this organizational principle at 3 scales, we found that the neural response pattern across the semantic processing mask as a whole (Binder et al. 2009) did not have any significant effects. This could reflect the fact that nearly all of the 120 studies used to define this mask involved concrete concepts, thus this system may primarily represent concrete sensory-motor aspects of meaning (Fernandino, Humphries, et al. 2016). Region-based and whole-brain analyses, however, generated positive results. The region-based analyses identified widely distributed regions that are sensitive to the semantic feature space. Some, including the left triangular IFG, IPS, ITG, posterior MTG, and supramarginal gyrus, fall well within the semantic processing mask (Binder et al. 2009), suggesting that these semantic subregions may represent meaning through feature composition, that is, by integrating and coordinating multiple semantic features represented in segregated neural systems. Intriguingly, the anatomical locations of these regions are also in line with the left frontoparietal module of the semantic system (Xu et al. 2016), suggesting the potential relationship between this feature space and semantic control. Additionally, the bilateral amygdala, right insula, right postcentral and precentral gyri and adjacent parietal and frontal cortices were found to be sensitive to the feature space. Some of these regions play various roles in emotion processing (Phillips et al. 2003; Lindquist et al. 2012), consistent with the proposal that emotion is a particularly important dimension for abstract concept processing (Kousta et al. 2011).

The highly distributed nature of feature representations is supported by the results on the whole-brain scale. The neural space based on the first 2 neural PCs of the whole-brain activation patterns correlated with the semantic space derived from semantic features, not with that from word co-occurrence. Intriguingly, this result accords well with the previous observation that semantic deficits in a patient with global aphasia were predicted by semantic similarity measured by feature distance and not LSA cosines (Crutch et al. 2013).

Our measure combined 13 semantic features of abstract concepts using the Euclidean distance; therefore the highly distributed regions may represent different aspects of this semantic feature space. In addition, it is important to note that many of these features are abstract by themselves (e.g., morality) and may not correspond to "primitive" experience-based brain systems (Crutch et al. 2013). They also are not entirely independent (i.e., they do not represent different aspects of the feature space orthogonally; consider emotion and valence). The neural correlates for these features remain to be understood. Our results did reveal that the first neural PC of the whole-brain activity pattern is associated with valence information, and not with other features such as emotion or social interaction. This is consistent with behavioral evidence for a central role of valence in the meaning of abstract words. For instance, it has been shown that valence could account for the abstractness effect observed in the lexical decision task after controlling for confounding variables such as imageability and context availability ratings (Kousta et al. 2011). Our neural-data-driven results show that valence properties indeed underlie the brain's response pattern to abstract words but that its effect goes beyond the effects of emotion or social interaction. The special status of valence may be because valence judgment—that is, whether and to what extent a given word is related to positive or negative feelings—lies at the core of human behaviors, exerting strong influence over a wide range of psychological phenomena, including attitudes, decision making, predicting the future, personality, even perception of spoken words and everyday objects (Barrett and Bliss-Moreau 2009). It is also possible that valence is particularly more important for abstract concepts. A post hoc analysis showed that in our stimuli valence is associated with more abstract words than other features: 34% of our abstract words contain valence information, using the criterion of rating scores greater than 5 (positive) or smaller than 3 (negative), whereas other semantic features are relevant for only 5% of words on average using a similar criterion (rating scores greater than 5, range: 1.1–16%). Whether the effects of valence on whole-brain neural activity are only the byproduct of having greater variance in the stimulus set or because it is an important distinctive feature for abstract concepts awaits further study.

Relationships Between the 2 Representations

Our study focused on the unique aspects of these 2 types of representation—linguistic contexts and semantic features—to elucidate the neural mechanisms underlying conceptual representation for abstract words. Similar principles may also be applied to concrete words. Indeed, evidence from several approaches highlights the complementary effects of these 2 aspects of meaning in concrete concepts. In development, the presence of both experience and verbal cues works better than when information of one type is present alone (Smith and Yu 2008; Andrews et al. 2009). Bayesian-based statistical incorporation of both types of information more closely resembles human conceptual cognitive performances than the use of only one type (Andrews et al. 2009). How these 2 types of information are orchestrated together in the brain to support rich conceptual representation remains to be explored.

Conclusion

Both linguistic contextual information and semantic features of abstract concepts are implemented in the brain, with dissociable neural correlates on different scales. The classical language regions, which overlap with one of the sub-networks of the established semantic system, represent abstract word meanings by verbal co-occurrence. The whole-brain patterns and regions widely distributed across both hemispheres including the left inferior frontal gyrus, posterior temporal cortex, bilateral supramarginal gyri and amygdala, represent abstract word meaning in a high-dimensional feature space. These results not only elucidate how multiple principles, including experience-based and language-based principles, serve to represent abstract word meanings in the brain, but also provide insight into the representation mechanisms of other types of meaning, including concrete concepts.

Supplementary Material

Supplementary data are available at *Cerebral Cortex* online.

Funding

This project is funded by the National Basic Research Program of China Program (2013CB837300 and 2014CB846100 to Y.B.), National Natural Science Foundation of China (31521063 to Y.B.), the China Postdoctoral Science Foundation (212400203 to X.W.;

212400205 to Y.F.) the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB02070006 to Z.L.), the Fundamental Research Funds for the Central Universities (WK2350000001 to Z.L.; 2017XTCX04 to Y.B.), Beijing Municipal Science & Technology Commission (Z161100000216152 to W.M.; Z16110100020000 & Z161100000216125 to Y.B.), and the Interdiscipline Research Funds of Beijing Normal University (to Y.B.).

Notes

We thank Paul Hoffman for generously sharing their semantic masks. We thank Kevin Shapiro, Yong He, Nan Lin, Tao Wei, and Gaolang Gong for comments on earlier versions of the manuscript. We are also grateful to all research participants. *Conflict of Interest*: None declared.

References

Andrews M, Vigliocco G, Vinson D. 2009. Integrating experiential and distributional data to learn semantic representations. Psychol Rev. 116:463–498.

Barrett LF, Bliss-Moreau E. 2009. Affect as a psychological primitive. Adv Exp Soc Psychol. 41:167–218.

Barsalou LW. 2008. Grounded cognition. Annu Rev Psychol. 59: 617–645.

Barsalou LW, Wiemer-Hastings K. 2005. Situating abstract concepts. New York: Cambridge University Press.

Binder JR, Conant LL, Humphries CJ, Fernandino L, Simons SB, Aguilar M, Desai RH. 2016. Toward a brain-based componential semantic representation. Cogn Neuropsychol. 33:130–174.

Binder JR, Desai RH. 2011. The neurobiology of semantic memory. Trends Cogn Sci. 15:527–536.

Binder JR, Desai RH, Graves WW, Conant LL. 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex. 19: 2767–2796.

Borghi AM, Binkofski F, Castelfranchi C, Cimatti F, Scorolli C, Tummolini L. 2017. The challenge of abstract concepts. Psychol Bull. 143:263–292.

Bradley MM, Lang PJ. 1999. Affective norms for English words (ANEW): stimuli, instruction manual and affective ratings (Technical Report No. C-1). Gainesville, FL: NIMH Center for Research in Psychophysiology, University of Florida.

Caramazza A, Shelton JR. 1998. Domain-specific knowledge systems in the brain the animate-inanimate distinction. J Cogn Neurosci. 10:1–34.

Carota F, Kriegeskorte N, Nili H, Pulvermuller F. 2017. Representational similarity mapping of distributional semantics in left inferior frontal, middle temporal, and motor cortex. Cereb Cortex. 27:294–309.

Chikazoe J, Lee DH, Kriegeskorte N, Anderson AK. 2014. Population coding of affect across stimuli, modalities and individuals. Nat Neurosci. 17:1114–1122.

Clark JM, Paivio A. 2004. Extensions of the Paivio, Yuille, and Madigan (1968) norms. Behav Res Methods Instrum Comput. 36:371–383.

Clarke A, Tyler LK. 2014. Object-specific semantic coding in human perirhinal cortex. J Neurosci. 34:4766–4775.

Crutch SJ, Troche J, Reilly J, Ridgway GR. 2013. Abstract conceptual feature ratings: the role of emotion, magnitude, and other cognitive domains in the organization of abstract conceptual knowledge. Front Hum Neurosci. 7:186.

Fairhall SL, Caramazza A. 2013. Brain regions that represent amodal conceptual knowledge. J Neurosci. 33:10552–10558.

Fedorenko E, Behr MK, Kanwisher N. 2011. Functional specificity for high-level linguistic processing in the human brain. Proc Natl Acad Sci USA. 108:16428–16433.

Fedorenko E, Hsieh PJ, Nieto-Castanon A, Whitfield-Gabrieli S, Kanwisher N. 2010. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. J Neurophysiol. 104:1177–1194.

Fernandino L, Binder JR, Desai RH, Pendl SL, Humphries CJ, Gross WL, Conant LL, Seidenberg MS. 2016. Concept representation reflects multimodal abstraction: a framework for embodied semantics. Cereb Cortex. 26:2018–2034.

Fernandino L, Humphries CJ, Conant LL, Seidenberg MS, Binder JR. 2016. Heteromodal cortical areas encode sensory-motor features of word meaning. J Neurosci. 36:9763–9769.

Firth JR. 1957. A synopsis of linguistic theory 1930–1955. Oxford, England: Blackwell Publishers.

Hoffman P. 2015. The meaning of 'life' and other abstract words: Insights from neuropsychology. J Neuropsychol. 10: 317–343.

Hoffman P, Binney RJ, Lambon Ralph MA. 2015. Differing contributions of inferior prefrontal and anterior temporal cortex to concrete and abstract conceptual knowledge. Cortex. 63: 250–266.

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature. 532:453–458.

Huth AG, Nishimoto S, Vu AT, Gallant JL. 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron. 76:1210–1224.

Kousta ST, Vigliocco G, Vinson DP, Andrews M, Del Campo E. 2011. The representation of abstract words: why emotion matters. J Exp Psychol Gen. 140:14–34.

Kriegeskorte N, Goebel R, Bandettini P. 2006. Information-based functional brain mapping. Proc Natl Acad Sci USA. 103: 3863–3868.

Kriegeskorte N, Kievit RA. 2013. Representational geometry: integrating cognition, computation, and the brain. Trends Cogn Sci. 17:401–412.

Kriegeskorte N, Mur M, Bandettini P. 2008. Representational similarity analysis—connecting the branches of systems neuroscience. Front Syst Neurosci. 2:4.

Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA. 2008. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron. 60:1126–1141.

Kuperman V, Estes Z, Brysbaert M, Warriner AB. 2014. Emotion and language: valence and arousal affect word recognition. J Exp Psychol Gen. 143:1065–1081.

Lambon Ralph MA, Jefferies E, Patterson K, Rogers TT. 2017. The neural and computational bases of semantic cognition. Nat Rev Neurosci. 18:42–55.

Lancaster JL, Tordesillas-Gutierrez D, Martinez M, Salinas F, Evans A, Zilles K, Mazziotta JC, Fox PT. 2007. Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template. Hum Brain Mapp. 28:1194–1205.

Landauer TK, Dumais S. 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol Rev. 104:211–240.

Lindquist KA, Wager TD, Kober H, Bliss-Moreau E. 2012. The brain basis of emotion: a meta-analytic review. Behav Brain Sci. 55:121–202.

Martin A. 2016. GRAPES-Grounding representations in action, perception, and emotion systems: how object properties and categories are represented in the human brain. Psychon Bull Rev. 23:979–990.

McRae K, Cree GS, Seidenberg MS, McNorgan C. 2005. Semantic feature production norms for a large set of living and nonliving things. Behav Res Methods. 37:547–559.

Meteyard L, Cuadrado SR, Bahrami B, Vigliocco G. 2012. Coming of age: a review of embodiment and the neuroscience of semantics. Cortex. 48:788–804.

Mikolov T, Chen K, Corrado G, Dean J. 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781v3 [cs.CL].

Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, Just MA. 2008. Predicting human brain activity associated with the meanings of nouns. Science. 320:1191–1195.

Mumford JA, Turner BO, Ashby FG, Poldrack RA. 2012. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. Neuroimage. 59: 2636–2643.

Murphy K, Bodurka J, Bandettini P. 2007. How long to scan? The relationship between fMRI temporal signal to noise ratio and necessary scan duration. Neuroimage. 34:565–574.

Paivio A. 1986. Mental representations: a dual coding approach. London: Oxford University Press.

Peelen MV, Caramazza A. 2012. Conceptual object representations in human anterior temporal cortex. J Neurosci. 32:15728–15736.

Phillips ML, Drevets WC, Rauch SL, Lane R. 2003. Neurobiology of emotion perception I: the neural basis of normal emotion perception. Biol Psychiatry. 54:504–514.

Primativo S, Reilly J, Crutch SJ. 2016. Abstract conceptual feature ratings predict gaze within written word arrays: evidence from a visual wor(l)d paradigm. Cogn Sci. 41:659–685.

Pulvermuller F. 2013. How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. Trends Cogn Sci. 17:458–470.

Recchia G, Jones MN. 2012. The semantic richness of abstract concepts. Front Hum Neurosci. 6:315.

Russell JA. 1980. A circumplex model of affect. J Pers Soc Psychol. 39:1161–1178.

Schwanenflugel PJ, Shoben EJ. 1983. Differential context effects in the comprehension of abstract and concrete verbal materials. J Exp Psychol Learn Mem Cogn. 9:82–102.

Smith L, Yu C. 2008. Infants rapidly learn word-referent mappings via cross-situational statistics. Cognition. 106: 1558–1568.

Sun HL, Huang JP, Sun DJ, Li DJ, Xing HB, editors. 1997. Introduction to language corpus system of modern Chinese study. City: Peking University Publisher.

Troche J, Crutch S, Reilly J. 2014. Clustering, hierarchical organization, and the topography of abstract and concrete nouns. Front Psychol. 5:360.

Vigliocco G, Kousta ST, Della Rosa PA, Vinson DP, Tettamanti M, Devlin JT, Cappa SF. 2014. The neural representation of abstract words: the role of emotion. Cereb Cortex. 24:1767–1777.

Wang J, Conder JA, Blitzer DN, Shinkareva SV. 2010. Neural representation of abstract and concrete concepts: a meta-analysis of neuroimaging studies. Hum Brain Mapp. 31: 1459–1468.

Xia M, Wang J, He Y. 2013. BrainNet Viewer: a network visualization tool for human brain connectomics. PLoS One. 8:e68910.

Xu Y, Lin Q, Han Z, He Y, Bi Y. 2016. Intrinsic functional network architecture of human semantic processing: modules and hubs. Neuroimage. 132:542–555.