OXFORD

Different computational relations in language are captured by distinct brain systems

Ze Fu [iD][1,2], Xiaosha Wang[1,2], Xiaoying Wang[1,2], Huichao Yang[1,2,3], Jiahuan Wang[1,2], Tao Wei[1,2], Xuhong Liao[3], Zhiyuan Liu[4],

Huimin Chen[5], Yanchao Bi[1,2,6,*]

[1]State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, China,
[2]Beijing Key Laboratory of Brain Imaging and Connectomics, Beijing Normal University, Beijing 100875, China,
[3]School of Systems Science, Beijing Normal University, Beijing 100875, China,
[4]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China,
[5]School of Journalism and Communication, Tsinghua University, Beijing 100084, China,
[6]Chinese Institute for Brain Research, Beijing 102206, China
*Corresponding author: State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University,
Beijing 100875, China. Email: ybi@bnu.edu.cn

A critical way for humans to acquire information is through language, yet whether and how language experience drives specific neural semantic representations is still poorly understood. We considered statistical properties captured by 3 different computational principles of language (simple co-occurrence, network-(graph)-topological relations, and neural-network-vector-embedding relations) and tested the extent to which they can explain the neural patterns of semantic representations, measured by 2 functional magnetic resonance imaging experiments that shared common semantic processes. Distinct graph-topological word relations, and not simple co-occurrence or neural-network-vector-embedding relations, had unique explanatory power for the neural patterns in the anterior temporal lobe (capturing graph-common-neighbors), inferior frontal gyrus, and posterior middle/inferior temporal gyrus (capturing graph-shortest-path). These results were relatively specific to language: they were not explained by sensory-motor similarities and the same computational relations of visual objects (based on visual image database) showed effects in the visual cortex in the picture naming experiment. That is, different topological properties within language and the same topological computations (common-neighbors) for language and visual inputs are captured by different brain regions. These findings reveal the specific neural semantic representations along graph-topological properties of language, highlighting the information type-specific and statistical property-specific manner of semantic representations in the human brain.

Key words: semantics; language computation models; fMRI; graph theory; neural network model.

Introduction

A typical adult human brain stores the meanings (semantics) of approximately tens of thousands of words (42,000 for typical English-speaking Americans) (Brysbaert et al. 2016), which allows naming objects and actions, understanding and producing sentences, and contributing to various kinds of reasoning. Decades of neuroimaging and neuropsychological literature have studied the cognitive neural representations of semantic knowledge, reaching the consensus that they are derived from, and grounded in, sensory experiences distributed across multiple sensory association cortices (Martin et al. 1995; Miceli et al. 2001; Fernandino et al. 2016), with those sharing physical properties (sensory/motor experiences) represented more closely in corresponding brain regions (Peelen et al. 2014; Aflalo et al. 2020; Wang et al. 2020). Even though rich knowledge can be acquired through language (e.g. reading or hearing the sentence "roses are red"), it has been commonly assumed that such language-derived knowledge is still grounded in sensory-derived representations (the visual perception system for seeing the color red) (Patterson et al. 2007; Barsalou 2008; Binder and Desai 2011; Barsalou 2016; Binder 2016; Binder et al. 2016; Martin 2016).

The specific role of language in deriving neural semantic representations beyond sensory-motor experiences has only recently been highlighted, motivated by the empirical findings that word meanings can be constructed in the complete absence of sensory experience (e.g. color and other visual concepts in congenital blindness) (Saysani et al. 2018; Bedny et al. 2019; Kim et al. 2019; Wang et al. 2020). The neural correlates of such nonsensory, presumably language-derived, knowledge have been identified in the dorsal part of the anterior temporal lobe (ATL) (Striem-Amit et al. 2018; Wang et al. 2020), which is part of a language

network and functionally connects with the other language-sensitive frontal–temporal cortices (such as the inferior frontal gyrus (IFG), the posterior part of the middle temporal gyrus (pMTG)) (Fedorenko et al. 2010). This set of regions, especially ATL, IFG, and pMTG, also tend to show stronger sensitivity to abstract words than concrete words (Noppeney and Price 2004; Binder et al. 2009; Wang et al. 2010; Hoffman et al. 2015), corroborating their potential functionality of supporting language-derived semantic representations.

What are the mechanisms by which language experience drives semantic representations in the human brain? Language is a highly rich faculty with a myriad of distinct processes. One parsimonious candidate is statistical properties. Using behavioral studies combined with computational modeling, ample evidence has shown that humans, during both development and adulthood, are sensitive to two main types of statistical patterns in language, including local statistical regularities, and global network patterns, which provide powerful computational mechanisms for various types of semantic relations to form (Romberg and Saffran 2010; Aslin and Newport 2014; Karuza et al. 2016; Lynn and Bassett 2020; Unger and Fisher 2021). Classical statistical learning studies show that humans can detect variations in *local statistical regularities* such as (first-order) simple co-occurrence (Fig. 1a and b, left panel) and/or transitional probability (Saffran et al. 1996; Schapiro et al. 2012), which is particularly salient in language acquisition (Saffran et al. 2001; Conway and Christiansen 2005). Recent studies further highlighted the effects of (higher-order) *global network topological patterns* in human language and/or knowledge learning. In the framework of network sciences (Cong and Liu 2014; Jackson and Bolger 2014), language can be constructed as a complex network (graph), with words as nodes and their simple co-occurrences as edges (Fig. 1b). Once represented as a graph, rich topological properties (node and edge layout patterns) can be computed to capture the local and global communication patterns that have been shown to affect human knowledge and/or language learning. For instance, humans can implicitly infer shared co-occurrence patterns (graph-common-neighbors) and path distance (graph-shortest-path) in various structural learning tasks, such as visual event segmentation (Schapiro et al. 2013), motor sequence learning (Lynn et al. 2020), and object relation learning (Yermolayeva and Rakison 2016; Garvert et al. 2017). These types of global topological patterns from language inputs have also shown to effectively predict semantic similarity ratings (graph-common-neighbors; Jackson and Bolger 2014) or reaction times of word-pair relatedness judgment (graph-shortest-path; Kenett et al. 2017). In contrast to these two types of computational properties (local and network topological) that are mathematically transparent, recent natural language processing (NLP) models exploit the higher order statistical patterns of language by *neural network*

(NN) learning methods, representing words by *vector embedding* obtained by model fitting (e.g. word2vec, Mikolov et al. 2013; GloVe, Pennington et al. 2014) and word relations (commonly) by geometric distance (e.g. cosine distance) in the resulting vector-embedding space (Fig. 1c). Although this vector-embedding space has also been shown to be correlated with human behavioral patterns such as semantic similarity ratings (Baroni et al. 2014; Pereira et al. 2016), its interpretability is opaque (Levy et al. 2015; Lenci 2018; Kumar 2021).

More direct evidence about how the neural system derives semantic representations from computing statistical properties of language inputs comes from neuroimaging studies that examine whether the brain responses to word meanings respect their statistical patterns captured by a particular computation model of the language corpora. The rationale is similar to that of classical cognitive computational modeling. If the "simulation" pattern of a computation model fits the observed pattern better than control models, then it provides a stronger candidate for explaining the computational mechanisms of the system of interest (here, a particular neural unit). A series of studies have shown that word relations computed from co-occurrence-based or NN-derived vector embedding correlate with word relations computed from brain activity patterns: with neural activity in distributed brain regions (count model, Huth et al. 2016; Mitchell et al. 2008); (GloVe model, Pereira et al. 2018; Anderson et al. 2019); and with the brain activity patterns of more specialized language processing regions/networks (word2vec, Wang et al. 2018; Carota et al. 2021). These models examined in these imaging studies aggregate multiple types of statistical regularities, with word relational structures constructed from various models correlating with each other and with nonlinguistic sensory properties (e.g. "cat" and "dog" are closely related in both verbal and visual relational patterns) (see Baroni et al. 2014; Lewis et al. 2019; Utsumi 2020). It is unknown whether the mathematically transparent statistical regularities, simple co-occurrence and network-(graph)-topological properties of language, which have been highlighted in explaining human behavioral patterns (Saffran et al. 1996; Jackson and Bolger 2014; Yermolayeva and Rakison 2016; Kenett et al. 2017), are captured by specific neural systems.

To test what types of computation of language statistical properties are better candidates to model the principles of the human brain's semantic representations, we systematically compared the effects of three types of language computational principles, theoretically and/or empirically motivated by developmental and adult behavioral and functional magnetic resonance imaging (fMRI) studies, in fitting brain activity patterns: simple co-occurrence, network-graph-topological (graph-common-neighbors and graph-shortest-path), and NN-derived vector-embedding (Fig. 1). For NN-derived vector-embedding models, word2vec (cosine distance) was tested because of its popularity and good performance
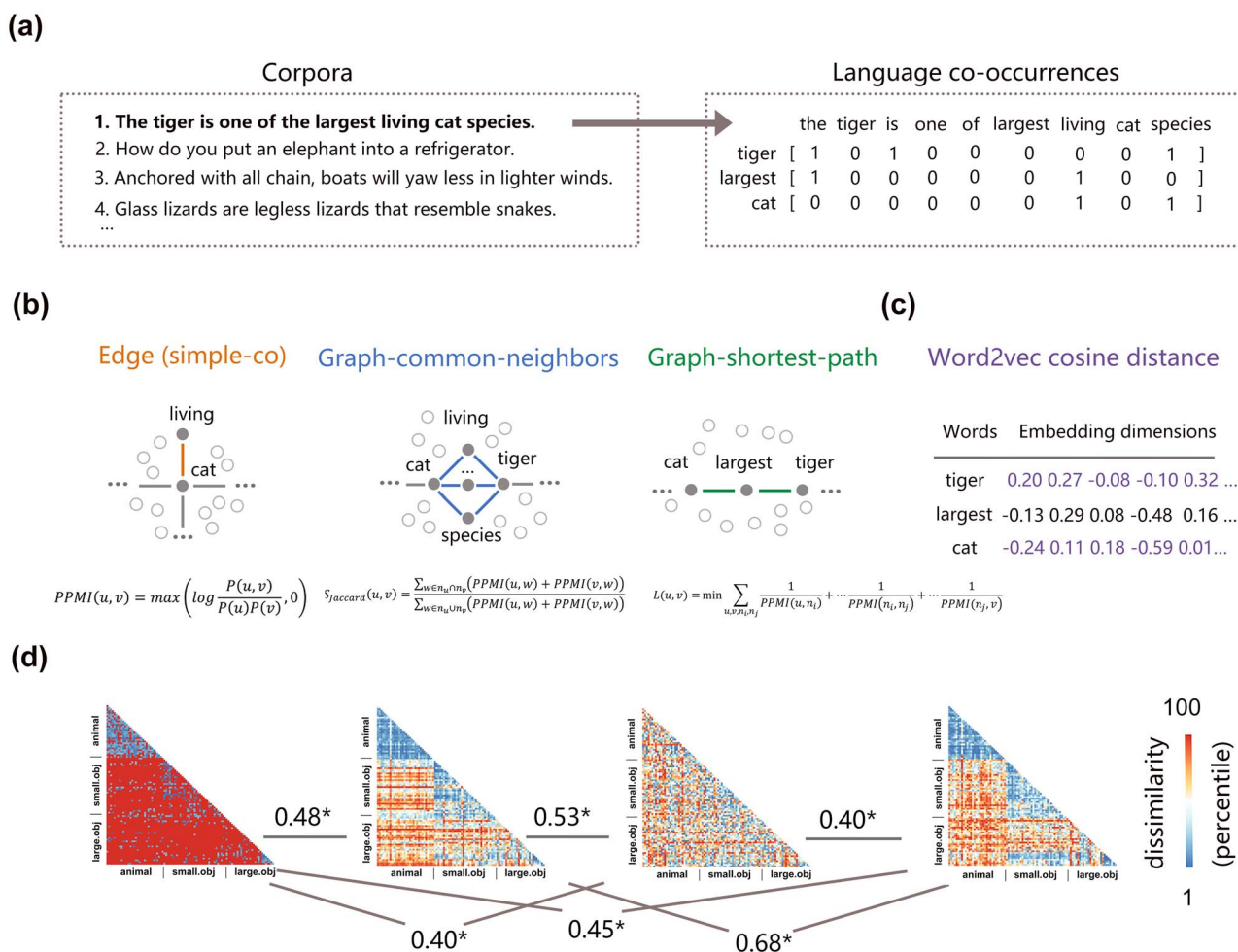
(a)

Corpora

1. **The tiger is one of the largest living cat species.**
2. How do you put an elephant into a refrigerator.
3. Anchored with all chain, boats will yaw less in lighter winds.
4. Glass lizards are legless lizards that resemble snakes.
...

Language co-occurrences

	the	tiger	is	one	of	largest	living	cat	species
tiger [	1	0	1	0	0	0	0	0	1]
largest [	1	0	0	0	0	0	1	0	0]
cat [	0	0	0	0	0	0	1	0	1]

(b) **(c)**

Edge (simple-co) Graph-common-neighbors Graph-shortest-path   Word2vec cosine distance

Words	Embedding dimensions
tiger	0.20 0.27 -0.08 -0.10 0.32 ...
largest	-0.13 0.29 0.08 -0.48 0.16 ...
cat	-0.24 0.11 0.18 -0.59 0.01...

$$PPMI(u,v) = max\left(log\frac{P(u,v)}{P(u)P(v)}, 0\right) \quad S_{Jaccard}(u,v) = \frac{\sum_{w \in n_u \cap n_v}(PPMI(u,w) + PPMI(v,w))}{\sum_{w \in n_u \cup n_v}(PPMI(u,w) + PPMI(v,w))} \quad L(u,v) = \min \sum_{u,v,n_i,n_j} \frac{1}{PPMI(u,n_i)} + \cdots \frac{1}{PPMI(n_i,n_j)} + \cdots \frac{1}{PPMI(n_j,v)}$$

(d)

Fig. 1. Construction of various language computation models for word (meaning) representations. (a) Word co-occurrence regularities, extracted from the Chinese web Google n-gram corpora (Liu et al. 2010; window size = 2). (b) Network-(graph)-topological relations. A word graph is constructed by having 83,007 unique Chinese words in the corpora as nodes and normalized raw word co-occurrence values in the corpora as edges (34,586,840 edges with positive PMI values). The edge values reflect the local proximity between two words in a graph space; common neighbors and shortest paths are then computed, reflecting words' commonality in a local community (graph-common-neighbors) and the closest transitional distance between 2 words (graph-shortest-path), respectively. Mathematical formulas for the computations are shown below (see Methods for detailed information). (c) Neural-network-vector-embedding relations. In the word2vec model, words are represented as 300-dimensional vectors and word relations are computed as cosine distances between these vectors. A pretrained and open-access word2vec model was adopted (Li et al. 2018). (d) RDMs of the 95 words used in the fMRI experiments, in which each cell represents the distance measure for a given word pair based on language computation models in (b) and (c). Numbers indicate the Spearman's correlation coefficients among these four RDMs; *, $P < 0.05$, Bonferroni corrected.

in fitting both behavioral and neural data (Baroni et al. 2014; Pereira et al. 2016; Wang et al. 2018). For 95 words, we computed word-pair distances along these four measures based on large-scale language corpora. To measure brain activity supporting word semantics, we conducted 2 fMRI experiments with varying input/output modalities but shared semantic processing of these 95 words and looked at the conjunction effects: a word production fMRI experiment (oral picture naming, entailing visual object recognition, semantic access, and phonological encoding/oral output) and a word recognition fMRI experiment (word familiarity judgment, entailing visual word recognition, semantic access, and button press). Representation similarity analysis (RSA) (Kriegeskorte, Mur and Bandettini 2008) was used to locate the neural circuits that are organized by specific language statistical properties, controlling for multiple types of potential confounding variables (e.g.

task peripheral, sensory-motor). Finally, to examine the universality/specificity of the observed word neural computations (language vs. nonverbal vision), the effects of graph-topological properties of visual object co-occurrence statistics derived from a large visual image database were evaluated.

Materials and methods
Participants

Twenty-nine participants (19 females; median age, 20 years; range, 18–32 years) were recruited in our study and were scanned in 2 fMRI experiments on separate days. All participants were right-handed native Mandarin Chinese speakers with normal or corrected-to-normal vision and had no history of neurological or language disorders. One hundred online participants (71 females; median age 21 years; range, 18–26 years)

were recruited in the behavioral experiments of sensory-motor attribute similarity judgment. They provided written informed consent. This study was approved by the institutional review board of the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University (ICBIR_A_0040_008), adhering to the Declaration of Helsinki for research involving human subjects. For the fMRI experiments, 3 participants in the oral picture naming experiment and 3 participants in the word familiarity judgment experiment were excluded from the data analysis due to successive head motions (>3 mm/3°). Another 6 participants in the word familiarity judgment experiment were excluded because of relatively poor behavioral responses (>20% nonresponsive trials in more than one run). Therefore, the fMRI data of 26 participants in the oral picture naming experiment and 20 participants in the word familiarity experiment were analyzed.

Stimuli

Ninety-five objects were selected in our fMRI experiments, including 3 common categories (32 animals, 35 small manipulable artifacts, and 28 large nonmanipulable artifacts). In 2 separate fMRI experiments, objects were shown as words and colored pictures (see Supplementary Fig. S1 and Table S1 for details). Pictures were 400 × 400 pixel images with the representative exemplar of the object presented against a white background (10.55° × 10.55° of visual angle). Words were presented in white FANG SONG font against a black background and subtended approximately 7.92° × 2.64° of the visual angle.

Computation of language models

Three major types of language computational principles were adopted to extract 4 kinds of statistical patterns between words: word simple co-occurrence relations, network-(graph)-topological (graph-common-neighbors and graph-shortest-path) relations and word2vec-derived vector-embedding relations. Pairwise distances between 95 objects were derived from these 4 measures to construct the theoretical representational dissimilarity matrices (RDMs) for the subsequent RSA computation with the neural data.

Construction of the PPMI-normalized simple co-occurrence matrix

Raw word co-occurrence counts were first collected in the Chinese Web Google n-gram corpora (https://catalog.ldc.upenn.edu/LDC2010T06) (Liu et al. 2010). The corpora included publicly accessible documents (a total of 882,996,532,572 tokens, 1,616,150 unique tokens including 864,629 unique Chinese words) on the Chinese internet by the end of 2008. In this dataset, n-grams within a context window ranging from 1 to 5 were extracted by the original developer using an auto segment parser. Bigram word co-occurrence counts were used in the main results. We also calculated our measures based on

tri-gram, four-gram, and five-gram word co-occurrence counts in the validation analyses. To reduce the computational loads and filter low-frequency and meaningless words, we further selected the keywords based on a human annotated Chinese Knowledge Database (Open-HowNet, 127,266 unique Chinese words by the end of January 2019) (Qi et al. 2019), resulting in 83,007 unique Chinese words. In the validation analysis, we also adopted different keyword selection methods based on different word frequency ranges and calculated the co-occurrence counts using the top 15%, 20%, 25%, and 50% most frequent words of a total of 864,629 Chinese word samples. Validation analyses of window size choice, matrix size, and keyword selection methods are presented in detail in the Supplementary Materials.

Importantly, pointwise mutual information (PMI)-normalized word co-occurrence counts between two words, u and v, were adopted to construct the 83,007 × 83,007 simple co-occurrence matrix, which reflects the direct proximity between 2 words in long-term language exposure (Church and Hanks 1990). There were 34,586,840 edges with positive PMI values (PPMI; see (1)) and the values of weak links (i.e. those with negative PMI values) were set to zero, i.e. having no connection, given that the negative PMI values might introduce "uninformative" noise (Levy and Goldberg 2014).

$$\text{PPMI}(u,v) = \max\left(\log\frac{P(u,v)}{P(u)P(v)}, 0\right) \quad (1)$$

Construction of the graph topological relations

To calculate graph-related distances, 83,007 words in the word co-occurrence PPMI matrix were taken as nodes, and word-pair PPMI values were taken as edges to construct the word graph space (see above). Considering that between-word relations in the graph space are remarkably rich, we mainly considered 2 graph-related measures based on current knowledge and algorithms from graph theory and semantic network practice (Newman 2001; Liben-Nowell and Kleinberg 2007; Lü and Zhou 2011; Jackson and Bolger 2014): the Jaccard similarity coefficient of common neighbors (graph-common-neighbors) and the shortest path distance (graph-shortest-path). For a given word pair, the Jaccard similarity coefficient of common neighbors was calculated as the summed PPMI weights of their shared edges of neighbors divided by the summed PPMI weights of union edges of neighbors connected with two words, which reflects the second-order proximity between the 2 words (Jackson and Bolger 2014). Precisely:

$$S_{\text{Jaccard}}(u,v) = \frac{\sum_{w \in n_u \cap n_v}(\text{PPMI}(u,w) + \text{PPMI}(v,w))}{\sum_{w \in n_u \cup n_v}(\text{PPMI}(u,w) + \text{PPMI}(v,w))} \quad (2)$$

where n_u is the set of nodes that are neighbors of node u.

For the calculation of weighted shortest path distance (Newman 2001), we summed weights along the shortest path between 2 words in a graph-defined space with inverted PPMI using Dijkstra's algorithm in Neo4j (http://neo4j.com/). The measure was precisely calculated as follows:

$$L(u,v) = \min \sum_{u,v,n_i,n_j} \frac{1}{PPMI(u,n_i)} + \cdots \frac{1}{PPMI(n_i,n_j)}$$

$$+ \cdots \frac{1}{PPMI(n_j,v)} \qquad (3)$$

Notably, both first-order edges and common neighbors measure the similarity between 2 words, while the weighted shortest paths measure the distance between two words. In the following analysis, we converted the former 2 into dissimilarity representations using a 1 minus calculation to obtain the corresponding RDMs.

Another proximity measure between words is Katzβ (communicability), which is calculated from ensembles of all paths with a damping factor β to discount the path weights. We did not consider this measure for the following reasons: (1) it is computationally expensive to calculate the global sums over the collection of all paths given the high degree of interconnection in the language graph; (2) the selection of dumping factor β is usually arbitrary and unexplored; and (3) the measure is similar to common neighbors when β is very small (Liben-Nowell and Kleinberg 2007).

Note that the graph-related measures we adopted here were based on the weighted version of graph space (i.e. the edge in the graph was a PPMI-normalized word co-occurrence value) to preserve as much statistical information as possible. In a validation analysis (see Supplementary Materials), we also calculated graph-based measures with the binary version (the edges with PMI greater than 0 were coded as 1, and other edges were coded as 0).

Construction of the word2vec vector-embedding relation

Another way to extract language statistical information is to project the word co-occurrence statistics in a hidden layer space using word embedding techniques. In this study, we adopted a pretrained, open-accessed word2vec vector-embedding dataset that achieved state-of-the-art performance on analogical reasoning of Chinese semantic relations (Li et al. 2018). Large mixed corpora (4037 million words) were selected from Baidu Encyclopedia, Chinese Wikipedia, People's Daily News, Sogou News, Financial News, Zhihu_QA, Weibo, and 8599 modern Chinese literature books. The basic parameter settings of this model were as follows: dynamic window size = 5, subsampling rate = 10^{-5}, negative sample number = 5, iteration = 5, low-frequency word = 10, dimension = 300. The skip-gram architecture (Mikolov et al. 2013) was adopted to predict the surrounding context words given an input target word. After intensive training and prediction, each word was represented as a 300-dimensional vector, and word relations were calculated as the cosine distance of these vectors. In addition to this pretrained w2v model, in a validation procedure we also repeated the analyses on a word2vec model trained on the same corpora that were used for calculating the edge and graph-related measures (see Supplementary Materials for details).

Computation of visual models

To examine whether our findings of language computation models were specific to language inputs or reflect domain-general computations for any kind of statistical co-occurrence pattern, we considered another kind of co-occurrence—visual co-occurrence statistics—which was collected based on an image database, VisualGenome (http://visualgenome.org/) (Krishna et al. 2017). The image database consisted of 108,077 images; objects in each image were annotated by human observers, and the labels were further mapped to Wordnet synsets. We first extracted the unique labels across the database (82,494 objects in total). The co-occurrence counts between objects were first obtained for each image and then summed over all images to obtain the raw object co-occurrence matrix. Similar to the language graph representation, we further constructed visual models using the same measures: visual edge (PPMI value, yielding 3,920,082 positive visual co-occurrence relationships), visual graph-common-neighbors and visual graph-shortest-path. RDMs were constructed accordingly.

Control RDMs

We constructed the following control RDMs:

Low-level stimulus & response properties: (1) *low level visual similarity*: To control for the low-level visual similarity effects between pictures/words, we calculated the pixel and gist dissimilarity of image pairs and the pixel dissimilarity of word pairs separately. The pixel dissimilarity was computed by Pearson's correlation distance between the grayscale values of images. The gist dissimilarity was computed by the Euclidean distance of 32 visual features of images (4 kinds of frequency, 8 kinds of orientation) (Oliva and Torralba 2001). (2) *Word phonological similarity*: Even though the word familiarity experiment did not explicitly require phonological output, there might still be automatic phonological access. To control for the effects of phonological similarity, we calculated the phonological dissimilarity for each word pair. Chinese is a syllabic language, with most words containing 2 syllables, and each syllable comprises an onset (consonant) and a rhyme (simple or complex vowel). We constructed the phonological RDM by calculating 1 minus the proportion of shared subsyllabic units (onset or rhyme) between each word pair (Fang et al. 2018). (3) *Response similarity*: A button-press RDM was constructed to control for the effects of motor responses in the word familiarity judgment task, computed as the absolute difference between the group-averaged button press responses (1 for left

index finger, familiar; 0 for left middle finger, unfamiliar) collected during scanning.

Nonlinguistic (sensory-motor) knowledge properties: Given the correlation between language and nonlinguistic structures ("cat" and "dog" tend to co-occur in language and in visual scenes and share sensory similarities), we carefully considered the effects of nonlinguistic sensory-motor properties, semantic domain membership, and visual object co-occurrence statistics: (1) For *sensory-motor attribute RDMs*, 5 sensory-motor attributes, including shape, motion, color, sound, and manipulation, were chosen based on the classical sensory-motor accounts of semantic neural representations (Binder et al. 2016; Fernandino et al. 2016). The dissimilarity structures of the 95 words on each of these attributes were collected from 100 college students (20 subjects per attribute) using the multiarrangement method (Kriegeskorte and Mur 2012), in which subjects were asked to arrange 95 word stimuli on a computer screen according to their distances along each attribute. The group-mean RDM was obtained by averaging across individual subjects' RDMs to serve as the control RDMs. (2) For *the semantic domain membership effects*, we constructed a binary RDM indicating whether the words belong to the same semantic domain, including animals, small manipulable objects, and large nonmanipulable objects; 0 means they are in the same semantic domain, 1 means they are not in the same semantic domain. (3) For *nonverbal visual object co-occurrence statistical properties*, the visual graph RDMs computed in the previous section were used.

fMRI experimental design

We adopted a condition-rich event-related fMRI experimental design to estimate the hemodynamic responses for each item (Kriegeskorte, Mur and Bandettini 2008). Two experiments were carried out: an oral picture naming experiment and a word familiarity judgment experiment. In the oral picture naming experiment, participants were instructed to overtly name the objects in colored pictures as precisely and quickly as possible. In the word familiarity judgment task, participants were asked to judge whether the presented written name of objects was familiar according to their personal experience by pressing the corresponding buttons (familiar: left index finger; unfamiliar: left middle finger).

The 2 experiments were conducted separately on 2 days for each participant, and the word familiarity judgment experiment was always carried out before the oral picture naming experiment to avoid eliciting the visual imagery of the object in the word experiment. Each experiment had 6 runs, with each word/picture repeated 6 times. Each run (8 min 48 s) consisted of 95 trials, and each word/picture was presented exactly once. In each trial, there was 0.5-s fixation and 0.8-s stimulus presentation, followed by the intertrial interval (ITI) ranging from 2.7 to 4.7 s. For the order of stimuli and length of ITI, we first determined the sequence of the 3 categories

using the optseq2 optimization algorithm (http://surfer.nmr.mgh.harvard.edu/optseq/) (Dale 1999) and further randomized the order of items in each category. Each run began and ended with a 10-s fixation period. The presentation and timing of stimuli was implemented using E-prime 2 (Psychology Software Tools) (Schneider et al. 2002).

Image acquisition

Functional and structural MRI images of 2 experiments were collected for each participant using a 3 T Siemens Trio Tim Scanner at the Beijing Normal University MRI Center. A high-resolution 3D structural image was collected with a 3D-MPRAGE sequence in the sagittal plane (144 slices, TR = 2530 ms, TE = 3.39 ms, flip angle = 7°, matrix size = 256 x 256, voxel size = 1.33 × 1 × 1.33 mm). Functional images were acquired with an echo-planar imaging (EPI) sequence (33 axial slices, TR = 2000 ms, TE = 30 ms, flip angle = 90°, matrix size = 64 × 64, voxel size = 3 × 3 × 3.5 mm with a gap of 0.7 mm).

Image data analysis
Preprocessing
Task-fMRI data were preprocessed and analyzed for each experiment using Statistical Parametric Mapping (SPM12; http://www.fil.ion.ucl.ac.uk/spm). For each individual participant, the first 5 volumes (10 s) of each run were discarded for signal equilibrium. The preprocessing of functional images included slice timing correction and head motion correction, and the resulting unnormalized and unsmoothed images were entered into general linear models (GLMs). The structural images were segmented into different tissue types; the resulting gray matter probabilistic images were coregistered to the mean functional image in the native space, resliced to the spatial resolution of functional images, and thresholded at one-third to obtain the gray mask of each subject. The forward and inverse deformation fields of each subject's native space to the Montreal Neurological Institute (MNI) space were also obtained at this step.

GLM
For the functional images in the native space in each subject, GLM was built to obtain object-level neural activation patterns. The GLM contained onset regressors for each of 95 items, 6 regressors of no interest corresponding to the 6 motion parameters, and a constant regressor for each run. All trials were included, as we did not record the behavioral responses in the picture naming task due to technical limitations, and only a small proportion of trials were omitted in the word familiarity judgment experiment (mean = 1.47%, SD = 0.94%). Each object regressor was convolved with a canonical HRF, and a high-pass filter cut-off was set as 128 s. Additionally, to ensure the maximal coverage of regions with a low ratio of signal-to-noise (e.g. ATL), the SPM implicit mask threshold was set to 10% of the mean of the global signal

(compared with the default threshold of 80%) (Devereux et al. 2013). For each experiment, the *t* value images (each condition relative to baseline) were obtained to capture the neural activation patterns.

RSA
Whole-brain searchlight analysis
To identify the brain regions that may represent different language computation models, we carried out RSA using a searchlight procedure (Kriegeskorte et al. 2006; Kriegeskorte, Mur and Bandettini 2008). For each voxel in the gray matter mask in the native space, the *t* values of 95 objects within a sphere (radium = 10 mm) centered at that voxel were extracted and correlated across object pairs to create the 95 × 95 neural RDM using Pearson's correlation distance. The neural RDM was then compared with language-computation-model-based RDMs using partial Spearman's rank correlation to obtain the "raw effects" of each language-model RDM (controlling for low-level control RDMs: pixel RDM, gist RDM, and phonological RDM in the oral picture naming experiment; pixel RDM, phonological RDM, and button-press RDM in the word familiarity judgment experiment), and further controlling for other language statistical models to obtain the "unique effects" of each model. The resulting *r* values were assigned to the center voxel of the sphere, and the searchlight procedure across each gray matter voxel produced a gray matter *r* map for each participant. These individual *r* maps were Fisher-z transformed, normalized into the MNI space, and spatially smoothed using a 6 mm full-width half-maximum Gaussian kernel.

For group-level random-effects analysis, one-sample *t* tests were performed across the individual *r* maps using permutation-based statistical nonparametric mapping (SnPM13; https://go.warwick.ac.uk/tenichols/snpm). No variance smoothing was used, and 10,000 permutations were performed. To localize the effects of theoretical models in each task, the RSA maps were thresholded at a conventional cluster extent-based inference threshold (voxelwise $P < 0.001$, FWE corrected cluster-level $P < 0.05$) unless explicitly stated. To demonstrate the task-invariant effects of language computation models, we performed conjunction analyses over two experiments. The overlapping regions were considered to show significant positive correlations between neural and theoretical RDMs in both experiments. As the conjunction method we adopted here is relatively conservative, which requires significant regions to be found in both experiments (Nichols et al. 2005; Caria et al. 2012; Kragel et al. 2018), we set the threshold in each experiment to uncorrected voxelwise $P < 0.005$, cluster size > 20 voxels. The brain results were projected onto the MNI brain surface using BrainNet Viewer (https://www.nitrc.org/projects/bnv) (Xia et al. 2013).

Validation analyses
We tested if the regions showing the task-invariant, unique effects of language computation models identified by the RSA searchlight analysis above (i.e. overlapping regions) could be explained by potential confounding variables or were robust across different graph construction methods: (1) We tested and controlled for (using partial correlation for RSA) the effects of *nonverbal visual object co-occurrence statistical properties* (see the "Construction of visual models" section); (2) We tested and controlled for *nonlinguistic sensory-motor attribute similarity structures*, including 5 sensory-motor attributes and a semantic domain model (see the "Control RDMs" section). (3) We employed various *graph construction methods*, including different graph types, window sizes, graph sizes, keyword selection methods, and corpus selection (see Supplementary Materials for details).

Language-ROI analyses
To more specifically test the effects of interest in those regions that have been consistently identified as being language sensitive, we further performed a region of interest (ROI) analysis, adopting a commonly used language mask (contrasting intact sentences to nonword lists; Fedorenko et al. 2010). The Fisher-transformed correlation values were averaged across voxels within each ROI for each subject. One-sample *t* tests across subjects were then conducted to test whether the RSA results of the theoretical models were significantly above zero. Multiple comparisons across ROIs and experiments were corrected using the Bonferroni method.

Results
Construction of language-computation-model-based RDMs
Three types of language computational principles were implemented. Simple co-occurrence counts were derived from large-scale language corpora (Chinese Web n-gram Corpora, consisting of approximately 883 billion words) and were PPMI-normalized to represent first-order proximity between 2 words (Fig. 1a). This normalized word co-occurrence of the 95 experimental stimuli (Supplementary Fig. S1 and Table S1) was used to construct the 95 × 95 simple co-occurrence RDM. Beyond the simple co-occurrence (i.e. edge RDM in the graph space), two types of network-(graph)-topological measures reflecting different aspects of statistical properties (graph-common-neighbors, graph-CN, and graph-shortest-path, graph-SP) were computed in a downsampled graph space (83,007 unique Chinese word samples as nodes, 34,586,840 PPMI-normalized simple co-occurrences as edges) to yield a graph-CN RDM and a graph-SP RDM (Fig. 1b). A word2vec RDM was constructed based on the cosine distance in a state-of-the-art pretrained word vector-embedding dataset (Li et al. 2018) (Fig. 1c). Visualizations of these four RDMs are presented in Figure 1d. These RDMs were moderately

to highly intercorrelated (edge RDM with CN RDM, Spearman's $r = 0.48$; edge RDM with SP RDM, Spearman's $r = 0.40$; edge RDM with word2vec RDM, Spearman's $r = 0.45$; CN RDM with SP RDM, Spearman's $r = 0.53$; CN RDM with word2vec RDM, Spearman's $r = 0.68$; SP RDM with word2vec RDM, Spearman's $r = 0.40$).

RSA searchlight results: relationship between language models and brain activity patterns

Neural RDMs of the 95 items were obtained and fit with language model RDMs for each fMRI experiment in an iterative sphere (10 mm) of each individual native space (an individually defined gray matter mask), following the procedure of whole-brain searchlight RSA (Kriegeskorte et al. 2006) (Fig. 2a). In each experiment, sanity check analyses were carried out for stimulus peripheral variables: pixel RDM, gist RDM, and phonological RDM in the oral picture naming experiment; pixel RDM, phonological RDM, and button-press RDM in the word familiarity judgment experiment. The RSA results of these control models were highly consistent with the previous literature (Kriegeskorte, Mur, Ruff, et al. 2008; Devereux et al. 2013; Carota et al. 2021) (Supplementary Fig. S2). In the main analyses of the language computation models, we first looked at the RSA results of each model independently ("raw" effects), with the peripheral RDMs in each experiment mentioned above regressed out. Given that these RDMs of language computation models are correlated (Fig. 1d), we further carried out a "unique effect" RSA for each language computation model, where the effects of the other language models were further controlled for (all using partial Spearman's rank correlations). The results for each fMRI experiment are shown in Supplementary Materials (voxelwise $P < 0.001$, FWE-corrected cluster-level $P < 0.05$), with positive results across both experiments, i.e. the shared cognitive components (word meanings) across experimental inputs/outputs, presented in detail (Table 1). As this conjunction method is relatively conservative (Nichols et al. 2005; Caria et al. 2012; Kragel et al. 2018), we reported clusters that survived the threshold of uncorrected voxelwise $P < 0.005$, cluster size > 20 voxels, across both experiments.

Language model-brain RSA raw results

The maps of group-level whole-brain searchlight RSA results from each language computation model in each experiment are shown in Figure 2b (see Supplementary Fig. S3 for the medial views of each hemisphere; see Supplementary Fig. S4a and Table S2 for more detailed results of the oral picture naming experiment, and Supplementary Fig. S4b and Table S3 for results of the word familiarity judgment experiment).

First-order-edge (simple co-occurrence) distance: In the oral picture naming experiment, the edge RDM correlated significantly with neural RDMs throughout the bilateral occipital–temporal cortex, with peak effects in the

lateral occipital cortex (LOC), extending into the early visual cortex, pMTG, and the posterior division of the temporal fusiform gyrus (pFG) and bilateral ATL, including the right temporal pole (TP), anterior division of the temporal fusiform gyrus (aFG) and parahippocampal gyrus (aPHG). In the word familiarity judgment experiment, the neural effects of edge RDM were confined to the bilateral ATL, including the TP, aPHG and left anterior division of the MTG (aMTG), the dorsal part of the medial frontal cortex (medPFC), orbital frontal cortex (OFC), right precuneus, precentral and postcentral gyrus, cingulate gyrus, insula and putamen. The overlap analysis showed that the bilateral ATL, especially the ventral part, was sensitive to the edge RDM in a task-invariant manner.

Graph-common-neighbors distance: In the oral picture naming experiment, neural effects of graph-CN RDM were found in the occipital and temporal regions, including the bilateral LOC, pMTG, pFG, and ATL. Clusters in the bilateral ATL encompassed the TP, aPHG, aFG, aMTG, and inferior temporal gyrus (aITG). In the word familiarity judgment experiment, similar patterns were found in bilateral ATL. Clusters extended into the medial temporal fusiform gyrus (medFG), orbital frontal cortex (OFC), and subcortical regions, including the hippocampus, amygdala, caudate, and thalamus. More dorsally, clusters were found in the medPFC, precuneous cortex, cingulate gyrus, precentral gyrus and postcentral gyrus, right AG, posterior part of the STG, bilateral insular cortex and primary auditory cortex. The overlap analysis showed the task-invariant neural representation of the graph-CN RDM in the bilateral ATL, including the TP, aFG, aPHG, aMTG and aITG, the OFC and the subcortical regions, including the hippocampus and amygdala.

Graph-shortest-path distance: In the oral picture naming experiment, the neural activity patterns in the frontal-temporal cortex were significantly associated with the graph-SP RDM, including the bilateral LOC, pFG, pMTG, left insula, and the pars triangularis part of the left IFG. More robust results were found in the word familiarity judgment experiment, which spread across the distributed brain regions, including bilateral temporal regions (with peak effects located in the left aSTG, aMTG, left amygdala), frontal regions (with peak effects located in the bilateral OFC, right caudate, and right medPFC) and widespread clusters located in the parietal cortex. The overlap analysis revealed that the task-invariant representation of the graph-SP RDM was in the bilateral LOC, bilateral pMTG, left pFG, ventral part of the left AG and pars triangularis part of the left IFG, as well as the OFC and insula.

Word2vec cosine distance: In the oral picture naming experiment, the neural patterns of bilateral LOC extending into the pMTG were found to be significantly correlated with word2vec RDM. In the word familiarity judgment experiment, significant mapping between
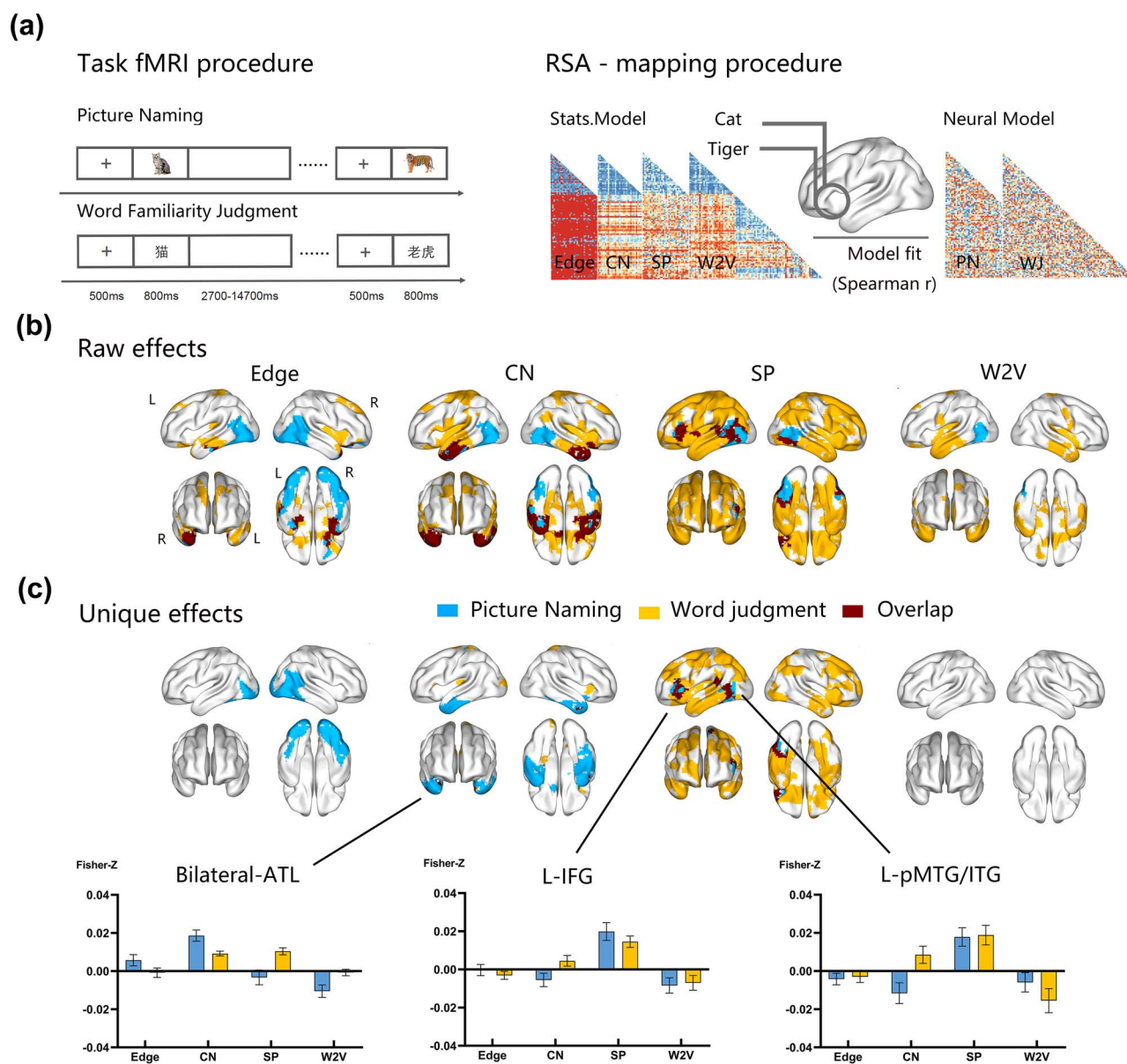
Fig. 2. Whole-brain searchlight RSA results of language computation models across 2 types of word processing fMRI experiments. (a) Task fMRI experimental design and RSA mapping procedure. Neural responses of 95 items were collected in 2 word processing fMRI experiments (oral picture naming, PN, and word familiarity judgment, WJ), which varied by input and output modalities but shared common semantic access. Neural RDMs were constructed as the pairwise correlation distances of neural activity patterns in a searching sphere (10 mm) across the 95 words. Second-order correlations between neural RDMs and each language computation RDM (simple co-occurrence (edge), graph-CN, graph-SP, and word2vec (w2v)) were calculated across the whole brain to evaluate where words' neural representational structures encode a specific computational principle of language experience. (b) The raw effects of each language computation model (controlling for low-level stimulus and response control variables). (c) The unique effects of each language computation model (i.e. additionally regressing out the effects of the other 3 models). Statistical maps were thresholded at voxelwise $P < 0.005$, cluster size > 20 voxels in each fMRI experiment (see Supplementary Fig. S3 for the medial views of brain results; see Supplementary Fig. S4, Table S2, and Table S3 for clusters surviving the conventional thresholds in each experiment). The bar plots exhibit the Fisher-Z transformed *r* values of the 4 models in the brain regions showing the overlap of unique effects across the 2 fMRI experiments, including graph-common-neighbors in the bilateral ATL and graph-shortest-path in left IFG and pMTG/ITG.

the word2vec RDM and neural RDMs was found in the bilateral ATL, extending into the medFG, OFC and subcortical regions, including the hippocampus, putamen, and right thalamus. More dorsally, clusters were found in the precentral and postcentral gyrus, as well as the insula and primary auditory cortex. No voxels survived in the overlap analysis when investigating the task-invariant neural representation of word2vec RDM.

Language model-brain RSA unique results

The unique RSA effects of each language computation model, with the effects of other language models (and the peripheral variables) partially removed, reveal the relative specificity of the target model in explaining the neural activity in a particular brain region (see above). The results surviving the conventional cluster extent-based inference threshold were presented in each

Table 1. Overlap results of language computation model RSA across two types of word processing experiments.

Brain regions		Cluster size (Voxels)	MNI Coordinates			Brodmann
			X	Y	Z	
Overlapping regions of raw effects						
First-order edge:						
L	aMTG/aITG	60	−59	−23	−25	20/21
L	aPHG/HIP/AMYG	235	−26	−8	−26	36/20
R	TP/aPHG/aFG/OFC/HIP/AMYG	612	30	5	−32	38/36/35/28/11
Common neighbor:						
L	TP/aPHG/aMTG/aITG/aFG/OFC/HIP/AMYG	993	−37	−2	−30	38/36/35/28/25/11/20/21
R	TP/aPHG/aMTG/aITG/aFG/OFC/HIP/AMYG	1,147	37	0	−25	
Shortest path:						
L	IFGtriang/OFC/INS	352	−43	23	3	45/47/48
L	LOC/pMTG/AG/pFG	805	−49	−53	2	37/20/21/39/22
R	LOC/ITG/pMTG	225	49	−65	−10	37/19
W2V: Null results						
Overlapping regions of unique effects						
First-order edge: Null results						
Common neighbor:						
L	HIP	46	−28	−13	−23	36
L	TP	24	−24	7	−35	38
R	TP/aMTG	97	42	5	−28	38/20
Shortest path:						
L	IFGtriang/OFC/INS	319	−43	24	2	45/47/48
L	SFG	39	−7	34	31	32
L	SMA	60	−10	8	64	6
L	LOC/pMTG/pFG	427	−51	−52	−1	37/20/21/42
W2V: Null results						

Note: Effects in clusters with extent < 20 voxels are not shown. Regions are labeled according to the Harvard–Oxford cortical and subcortical atlas. AG, Angular gyrus; AMYG, amygdala; FG, temporal fusiform cortex; HIP, hippocampus; IFGtriang, inferior frontal gyrus, triangular part; INS, insula; ITG, inferior temporal gyrus; LOC, lateral occipital cortex; MTG, middle temporal gyrus; OFC, orbital frontal cortex; PHG, parahippocampal gyrus; SFG, superior frontal gyrus; SMA, supplemental motor area; TP, temporal pole; a, anterior; p, posterior.

experiment separately (oral picture naming in Supplementary Fig. S4a and Table S2; word familiarity judgment in Supplementary Fig. S4b and Table S3). The overlap results across the two experiments are shown in Figure 2c (see Supplementary Fig. S3 for medial views) and Table 1.

First-order-edge (simple co-occurrence) distance: In the oral picture naming experiment, the unique neural effects of edge RDM were in the bilateral LOC, pFG, and early visual cortex. In the word-judgment experiment, no regions showed significant effects.

Graph-common-neighbors distance: The significant mappings between the graph-CN RDM and neural RDMs in the bilateral ATL were preserved after regressing out 3 other RDMs in the oral picture naming experiment. The overlap analysis further confirmed that the task-invariant neural representation of the graph-CN RDM was confined in ATL, including the right TP and aMTG, as well as the left hippocampus.

Graph-shortest-path distance: In the oral picture naming experiment, RSA mappings between graph-SP RDM and neural RDMs revealed significant clusters in the pars triangularis of the left IFG, pMTG, and pFG after regressing out 3 other RDMs. In the word familiarity judgment experiment, the significant clusters in multiple frontal–temporal regions were preserved, including the medFG, OFC, SMA, IFG, and pMTG, as well as widespread clusters located in the parietal cortex. The overlap analysis revealed the task-invariant neural representation of graph-SP RDM in the pars triangularis part of the left IFG, left SMA and left pMTG/ITG.

Word2vec cosine distance: No clusters survived the convention cluster extent-based inference threshold in either experiment.

In summary, the raw effects (across fMRI experiments) of the different language computation models were observed in both overlapping and different brain regions. The unique effects revealed interesting dissociations: language graph-CN exhibited unique, task-invariant

effects in the bilateral ATL, and language graph-SP exhibited unique effects in the left IFG and left pMTG/ITG (see Fig. 2c for bar plots of the unique effects). Edge (simple co-occurrence) and word2vec did not show overlapping regions of unique effects, i.e. those that cannot be explained by other models. Note that the left SMA showed the unique effects of language graph-SP but did not exhibit significant raw effects, which may result from complicated intercorrelations between these language computation RDMs, and it was not included in the following ROI analyses.

Validation of language specificity 1: visual object co-occurrence statistical patterns

To investigate whether the observed neural effects of the language computation models were specific to computing language-derived statistical information or reflecting certain domain-general computations for information from any type of input, we constructed the same kind of graph representation using visual co-occurrence statistics from large image corpora (a visual graph with 82,494 nodes and 3,920,082 visual nonzero co-occurrence edges, which was derived from a large image dataset—VisualGenome with 108,077 images) (Krishna et al. 2017) (Fig. 3a). The same graph-related measures were calculated, including visual edge, visual graph-CN, and visual graph-SP.

The visual models were intermediately correlated with each other (visual edge RDM with visual CN RDM, Spearman's $r = 0.38$; visual edge RDM with visual SP RDM, Spearman's $r = 0.46$; visual CN RDM with visual SP RDM, Spearman's $r = 0.52$) and were weakly yet significantly correlated with the language models (visual edge RDM with language edge RDM, Spearman's $r = 0.15$; visual CN RDM with language CN RDM, Spearman's $r = 0.23$; visual SP RDM with language SP RDM, Spearman's $r = 0.12$; $Ps < 0.001$) (Fig. 3b).

In the whole-brain searchlight RSA of the visual RDMs (Fig. 3c; see Supplementary Fig. S5 for more complete visualization), the only significant cluster was the visual graph-CN in the right transverse occipital sulcus (TOS) (peak $t_{(25)} = 4.96$, peak MNI, $x = 18$, $y = -99$, $z = 21$, cluster size = 113 voxels) in the oral picture naming experiment under the convention cluster-extent-based inference threshold (voxelwise $P < 0.001$, FWE-corrected cluster-level $P < 0.05$), with a cluster located in the left parahippocampal place area (PPA) visible at a lower threshold (voxelwise $P < 0.005$, cluster size > 10 voxels).

We focused on the ROIs showing language-model RSA unique effects (Fig. 2c)—bilateral ATL (language graph-CN effect), left IFG and left pMTG/ITG (language graph-SP effect). None of the visual RDMs had any effects in these regions ($Ps > 0.12$, uncorrected). Furthermore, the positive RSA result patterns of the language models in these regions were preserved after regressing the corresponding visual RDMs ($Ps < 0.05$, Bonferroni corrected, number of correction = 6) (Fig. 3d and Supplementary Table S4).

Validation of language specificity 2: controlling for sensory-motor attribute and semantic domain similarity structures

To examine the possibility that the observed neural effects of the language computation models were attributed to the sensory-motor similarities of objects, we constructed 5 sensory-motor attribute RDMs (shape, motion, color, sound, manipulation) based on the mean of group responses (20 subjects for each attribute) in a multiarrangement task. A binary semantic domain model was also constructed as a comprehensive means to control for potential categorically based sensory-motor similarities. These nonlinguistic RDMs were intermediately correlated with the language models (Spearman's $rs = 0.09$–0.57) (Supplementary Fig. S6a). Critically, for the ROIs showing task-invariant, unique effects of language graph models (Fig. 2c), the RSA results of the language models remained robust after regressing out these sensory-motor attributes and semantic domain similarity structures (Supplementary Fig. S6b and Table S4; $Ps < 0.05$, Bonferroni corrected, number of correction = 6).

Validation of graph construction methods

To test the robustness of the language graph representation, validation analyses were performed to address the following concerns: (1) Are the results affected by the graph type (weighted-graph vs. binary-graph)? (2) Are the results affected by the specific window size choice in calculating the simple co-occurrence? (3) Are the results affected by the graph size and the method to select which words are included in the graph representation? (4) Are the results affected by specific language corpora being used? We performed analyses with different graph types (weighted vs. binary), with different window sizes (2–5 words), a wide range of graph sizes (words with different frequency ranges—top 15%, 20%, 25%, and 50%), and with identical language corpora for all 4 models. The main results were robust across these different graph construction methods (Fig. 4; for details, see Supplementary Materials, Supplementary Figs. S7 and S8 and Table S4).

Classical language-area ROI results

To more specifically test the effects of interest in those regions that have been consistently identified as being language sensitive, we further performed an ROI analysis, adopting a commonly used language mask (contrasting intact sentences to nonword lists) (Fedorenko et al. 2010), including left ATL, IFG, IFGorb, pMTG, AG, and MFG (Fig. 5, left panel). We carried out ROI-based RSA analyses across these 6 language-processing regions for each of the 4 language models after regressing out the other 3 models (i.e. testing unique effects) and the range of control models (visual object co-occurrence, sensory-motor attribute similarity, semantic domain similarity, and peripheral stimulus properties). Converging with the whole-brain searchlight results, the effects of edge RDM and word2vec
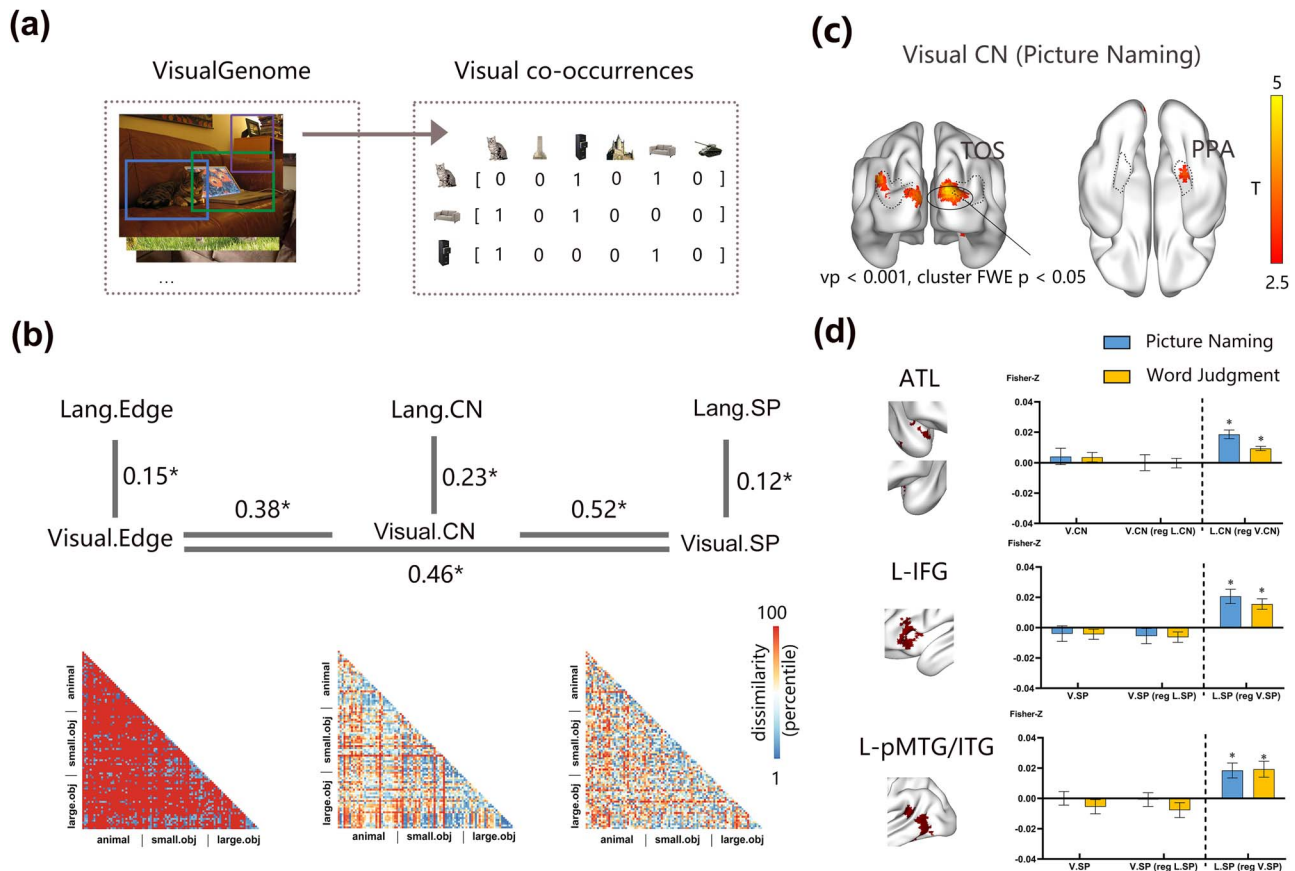
Fig. 3. Neural effects of language graph-topological models were not explained by visual co-occurrence statistics. (a) Visual co-occurrence regularities were derived from a large-scale, human-annotated image database, VisualGenome (approximately 108,077 images, 82,494 objects). The raw co-occurrence counts in a given image "context" were calculated and normalized into PPMI values. Similar to the computation of language inputs, simple co-occurrence and graph-related measures were adopted for the visual statistics, including visual object simple co-occurrence (edge), visual graph-CN, and visual graph-SP. (b) Visual graph RDMs of the 95 words and their relationships with language graph RDMs. Correlations among these visual RDMs and language RDMs were obtained using Spearman's rank correlations. (c) Whole-brain searchlight results of the unique effects of visual object graph-CN in the picture naming task, thresholded at voxelwise $P < 0.005$, cluster size > 20 voxels (see Supplementary Fig. S5 for the raw and unique effects of other visual models). The cluster in the black circle survived the threshold of voxelwise $P < 0.001$, FWE-corrected cluster-level $P < 0.05$. The dotted lines show the classical place processing regions (transverse occipital sulcus, TOS; parahippocampal place area, PPA), localized by contrasting large-place-related object pictures with animal/face pictures in an independent block-designed localizer from 21 healthy subjects. (d) the effects of the visual graph RDMs in the brain regions showing task-invariant, unique effects of the language graph measures. Bar plots on the left side of the dotted line show the RSA results of visual graph RDMs, which did not approach significance when they were examined alone or when language graph RDMs were regressed out ($Ps > 0.12$, uncorrected). Bar plots on the right side of the dotted line show the unique RSA results of language graph RDMs, when visual graph RDMs were further included as covariates. Asterisks in (b) and (d) indicate statistical significance surviving $P < 0.05$, Bonferroni corrected.

RDM were not significant in any of the ROIs ($Ps > 0.121$, uncorrected, see Fig. 5 and Supplementary Table S4 for details); the language graph-CN RDM showed effect trend only in the ATL ($Ps < 0.036$, uncorrected); the graph-SP RDM showed significant results in the IFG, IFGorb, and pMTG ($Ps < 0.05$, Bonferroni corrected, number of correction = 12). Additionally, AG and MFG showed a tendency for the graph-SP effects ($Ps < 0.033$, uncorrected) in this ROI analysis, which were not visible in the whole-brain searchlight.

Discussion

To understand whether there are neural structures representing word semantics derived from statistical properties of language and the types of properties being captured, we fitted brain activity patterns with word relations obtained using 3 computational principles: simple co-occurrence, 2 network topological relations (graph-common-neighbors and graph-shortest-path), and NN-derived vector-embedding relations. In 2 fMRI experiments sharing common semantic representations, word relations computed from all 4 models correlated with word brain activity patterns across broadly distributed brain regions. Importantly, the word relations derived from the network graph space, and not the other 2 types, have unique explanatory power for the neural activity patterns in brain regions associated with language processing, including ATL, IFG, pMTG/ITG. Intriguingly, different graph relations were respected by these regions, with ATL based on the proportions of common neighbors in a graph (i.e. number of shared co-occurrence) and IFG and pMTG/ITG based on the shortest path distances (with trends in AG and MFG).
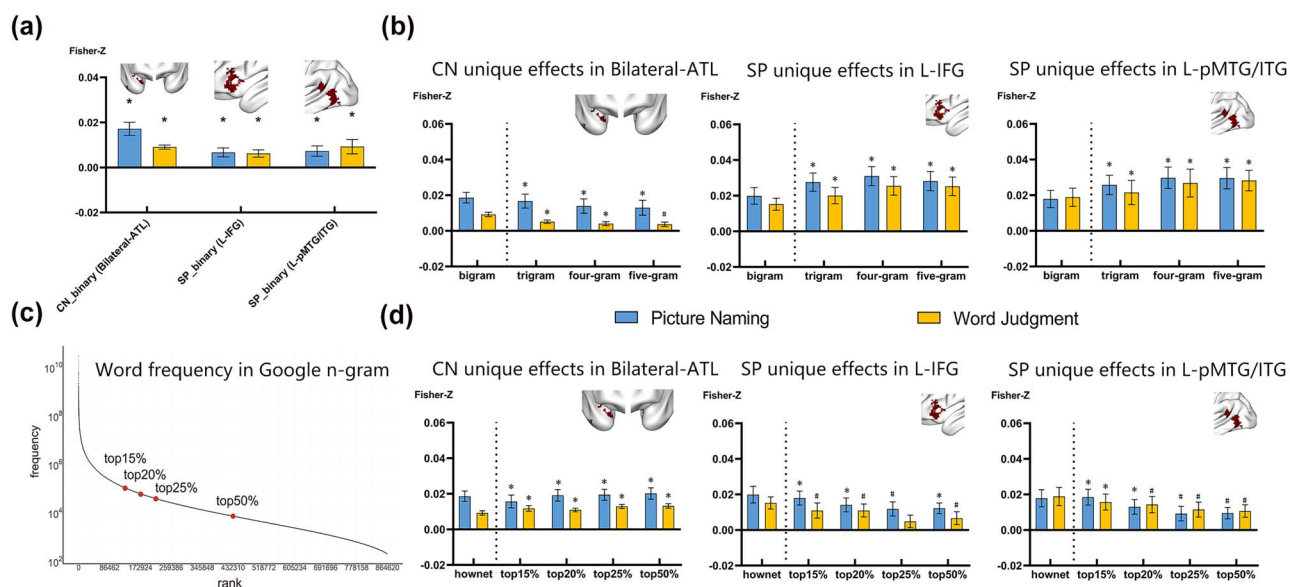
Fig. 4. Validation of different graph construction methods in the brain regions showing task-invariant, unique effects of language graph-relation models. (a) ROI results using a different graph type (binary version). (b) ROI results using different window sizes ranging from bigram to five-gram. (c) Log-transformed frequency distribution of Chinese words in the Chinese Google n-gram corpora. (d) ROI results using graphs of different sizes, with keywords selected based on word frequency (top 15%, top 20%, top 25%, top 50%) shown in (c). Bar plots on the left side of the dashed line in (b) and (d) show the main results obtained in the bigram word co-occurrence graph with keywords selected based on OpenHowNet for the purpose of visualization without additional statistical inference. *, $P < 0.05$, Bonferroni corrected, numbers of corrections were 6, 18, and 24 in (a), (b), and (d), respectively; #, $P < 0.05$, uncorrected.
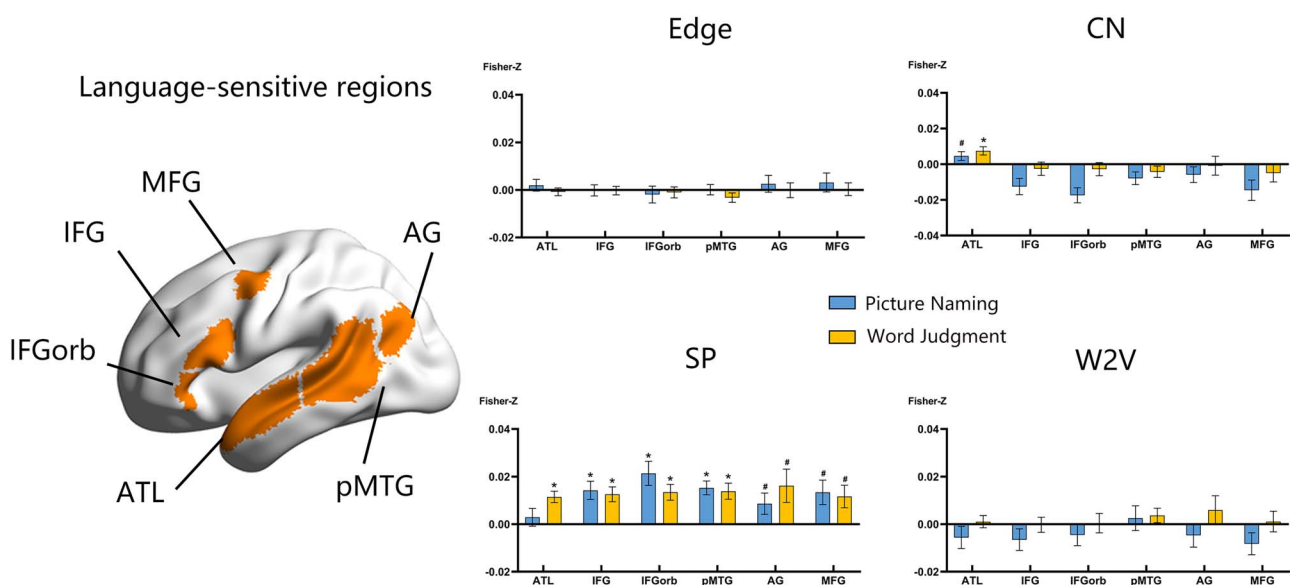
Fig. 5. ROI RSA results of language computation models in classical language areas. The left panel shows the language-sensitive ROIs (Fedorenko et al. 2010), including left ATL, IFG, IFGorb, pMTG, AG, and MFG. Bar plots show the Fisher-transformed Spearman's partial correlations between neural RDMs and each language computation model in each ROI after regressing out the other 3 language models and the full range of control models (visual object co-occurrence, sensory-motor attribute similarity, semantic domain similarity, and peripheral stimulus properties). *, $P < 0.05$, Bonferroni corrected, number of corrections for each model = 12; #, $P < 0.05$, uncorrected.

These neural results of language graph representation were relatively specific to language, as they were not associated with relational structures derived from visual object co-occurrence statistics when using the same computation methods or explained by (nonlinguistic) sensory-motor similarities.

Our results describing correlations between all 4 language model patterns and broadly distributed brain regions are consistent with the literature findings (Huth et al. 2016; Carota et al. 2017; Pereira et al. 2018; Wang et al. 2018; Anderson et al. 2019; Carota et al. 2021). In these previous studies, it was not clear what specific kinds of computations drove these effects, given the medium-to-high correlations among different types of language computation models and the lack of computational transparency with the NN-based models tested.

The current study, comparing the effects of 3 different kinds of computational principles that capture specific aspects of the statistical properties of language (local statistical regularity, 2 specific types of global-topological properties, and global-NN-learning based), allows for the inference that global-network-topological properties explain the neural activity patterns better than simple co-occurrence and word2vec, and beyond sensory-motor similarities, in ATL, IFG, and pMTG/ITG. Such results provide positive evidence for language-derived semantic representations in these regions (Bi 2021).

What do these network-graph-topological relations reflect, how are they different from word2vec cosine distances, and what are the implications for the representations in these brain regions? The graph representation retained the original dimensions given the size of the word co-occurrence matrix (Jackson and Bolger 2014), with the extraction of statistical information achieved through relationships between highly informative neighbors and paths. In this way, more "historical information" can be preserved. Furthermore, one can select subgraphs or prune the edges based on word frequency without dramatically changing the network structures, especially the interconnected neighborhoods, indicating the robustness of this type of representational structure (Eom 2018; see also our own validation results). By comparison, word embedding techniques project the word co-occurrence statistics into a dense vector space, resulting in a holistic representation of word meanings. In the case of word2vec, statistical information was compressed into a fixed, arbitrary, usually 300-dimensional vector space through error-driven training and hyperparameter optimizations. It remains controversial whether these hyperparameter tuning processes are psychologically meaningful and what information is retained and lost after dimension reductions (Kumar 2021). The robustness and the "historical faithfulness" principle of the graph representation may explain its advantage in fitting the neural representation over NN spaces such as w2v, providing a neural basis for the behavioral effects observed (see introduction and reviews in Karuza et al. 2016).

Network graph representation also has the advantage of unpacking different relationships in the same framework, which provides novel computational insights into the functionality of distributed "language regions," including ATL, IFG, pMTG, along with AG and MFG (Fedorenko et al. 2010). These regions have been shown to manifest complex functional profiles, with ATL and AG showing different representational preferences for different semantic relations, such as taxonomic and thematic relations (Xu et al. 2018). These findings, while informative of parsing semantic relations into multiple facets, invite further questions about how exactly the neural system computes such psychological dimensions. Our current findings provide a parsimonious mechanistic account of how the neural system gives rise to these rich semantic structures: these brain structures

organize neural representations along specific graph-based statistical properties of language inputs. ATL was found to track graph-common-neighbors and IFG/pMTG (and AG/MFG) graph-shortest-path. The shortest path captures long-range dependency in graph space—it measures association strength between 2 nonadjacent nodes, while the common neighbors capture structural similarity based on second-order proximity. These two types of statistical properties tend to be associated with different types of meaning relations, with CN with taxonomy (semantic categorical) similarity and edge with associative relations (Jackson and Bolger 2014). Here, we performed an ad hoc analysis on a word set where taxonomic and thematic relations were dissociated based on subjective ratings (Xu et al. 2018) and found such correspondence: CN was more strongly correlated with taxonomic relations (Spearman's $r = 0.46$; Spearman's r with thematic relations $= 0.31$), and SP with thematic relations (Spearman's $r = 0.46$; Spearman's r with taxonomic relations $= 0.20$). That is, the graph-topological effects we observed provide a parsimonious computational mechanistic explanation for the previous findings that ATL represents word taxonomic similarity (Martin et al. 2018; Xu et al. 2018), IFG and/or pMTG/ITG contribute to the retrieval of infrequent word associations, i.e. longer path length distance (Badre et al. 2005; Whitney et al. 2011), and AG thematic relations (Xu et al. 2018). Finally, these findings are consistent with the topological structural observations of the intrinsic functional semantic network, in which these regions were identified as connector hubs that bind together different brain subnetworks: ATL binds the perisylvian language network and multimodal experiential network, and pMTG binds the perisylvian language network and frontoparietal control network (Xu et al. 2016, 2017).

Are these brain regions sensitive to graph-topological structures from all stimulus modalities/domains? On the one hand, several sequential learning experiments have shown that changes in neural representations in medial temporal, anterior temporal, and frontal regions, after a short training session involving sequence exposure (Schapiro et al. 2012; Schapiro et al. 2013; Garvert et al. 2017), follow graph-based statistical regularities (simple edge, community structure, path distance) of arbitrary visual stimuli sequences during the formation of episodic memory. On the other hand, we did not observe any sensitivity to common neighbors or shortest path measures from the large visual object co-occurrence database in these language-sensitive regions. The statistical patterns computed from the graph relations based on the visual object corpora instead showed significant associations with the visual cortex activity pattern in the picture naming experiment, indicating an input modality-specific representation of similar computational structures. This visual object result is aligned with recent findings of co-occurrence measures based on visual object2vec (Bonner and Epstein 2021),

while our study further showed that it is the graph common structures shared between two objects that are represented in the "place areas" (TOS and PPA). More generally, a wave of recent studies has highlighted the explanatory power of "spatial relationship" or "grid-like" structures in representing conceptual knowledge and information in memory in general (Constantinescu et al. 2016; Theves et al. 2019; Theves et al. 2020). Here, we showed that the topological distances in a graph space are better predictors than the cosine distance in the vector-embedding space in explaining word representations. The breadth of the application of the observed computational structures in representing information in these regions warrants further testing (Peer et al. 2020).

A few methodological caveats need to be considered. First, in our current investigation, we used large-scale language corpora as the proximity of collective language experience on a group level. It may not be an accurate reflection of specific language inputs at the individual level. Computation modeling of word meaning representations in the future could benefit from collecting and estimating individual participants' language inputs with the help of personalized big data techniques. Second, we specifically considered relatively simple models that are fully data driven, without prior knowledge such as grammatical information and attentional allocation mechanisms. In recent years, there has been a surge of computation models with improved performances in various language tasks, such as recurrent neural-network models (ELMo) (Peters et al. 2018) and attention neural-network models (BERT) (Devlin et al. 2018); whether their computational architecture is relevant to brain computations in ATL/IFG/pMTG awaits further study.

To conclude, combining two fMRI experiments investigating word meaning representations, we found that the human brain is sensitive to specific graph-topological properties of language, providing positive evidence for language-derived semantic representations. Graph-based topological models of language had unique explanatory power on words' neural activity patterns beyond simple co-occurrence and vector-embedding models, showing effects in the ATL (capturing graph-common-neighbors), IFG, and pMTG/ITG (capturing graph-shortest-path), in contrast to the visual cortex, which shows sensitivity to graph-common-neighbors computations of visual experiences. Together, the distributed neural semantic representations across different brain regions exhibit both information type-specific (language vs. visual) and computation-specific (graph-common-neighbors vs. graph-shortest-path) patterns of organization.

Authors' contributions

YB conceived and designed the study; ZF performed the study; XSW, XYW, and HY contributed to critical discussions; JW computed the visual relation measures; TW acquired fMRI data; XL contributed to graph measure computation; ZL and HC contributed to natural language computation modeling; and YB and ZF wrote the paper.

Supplementary material

Supplementary material is available at *Cerebral Cortex Journal* online.

References

Aflalo T, Zhang C, Rosario E, Pouratian N, Orban G, Andersen R. A shared neural substrate for action verbs and observed actions in human posterior parietal cortex. *Sci Adv*. 2020:6:eabb3984.

Anderson AJ, Binder JR, Fernandino L, Humphries CJ, Conant LL, Raizada RD, Lin F, Lalor EC. An integrated neural decoder of linguistic and experiential meaning. *J Neurosci*. 2019:39:8969–8987.

Aslin RN, Newport EL. Distributional language learning: mechanisms and models of category formation. *Lang Learn*. 2014:64:86–105.

Badre D, Poldrack RA, Paré-Blagoev EJ, Insler RZ, Wagner AD. Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex. *Neuron*. 2005:47:907–918.

Baroni M, Dinu G, Kruszewski G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; 2014. p. 238–247.

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

Barsalou LW. On staying grounded and avoiding quixotic dead ends. *Psychon Bull Rev*. 2016:23:1122–1142.

Bedny M, Koster-Hale J, Elli G, Yazzolino L, Saxe R. There's more to "sparkle" than meets the eye: knowledge of vision and light verbs among congenitally blind and sighted individuals. *Cognition*. 2019:189:105–115.

Bi Y. Dual coding of knowledge in the human brain. *Trends Cogn Sci*. 2021:25:883–895.

Binder JR. In defense of abstract conceptual representations. *Psychon Bull Rev*. 2016:23:1096–1108.

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

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

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

Bonner MF, Epstein RA. Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nat Commun*. 2021:12:1–16.

Brysbaert M, Stevens M, Mandera P, Keuleers E. How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Front Psychol*. 2016:7:1116.

Caria A, de Falco S, Venuti P, Lee S, Esposito G, Rigo P, Birbaumer N, Bornstein MH. Species-specific response to human infant faces in the premotor cortex. *NeuroImage*. 2012:60:884–893.

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

Carota F, Nili H, Pulvermüller F, Kriegeskorte N. Distinct fronto-temporal substrates of distributional and taxonomic similarity among words: evidence from RSA of BOLD signals. *NeuroImage*. 2021:224:117408.

Church K, Hanks P. Word association norms, mutual information, and lexicography. *Comput Linguist*. 1990:16:22–29.

Cong J, Liu H. Approaching human language with complex networks. *Phys Life Rev*. 2014:11:598–618.

Constantinescu AO, O'Reilly JX, Behrens TE. Organizing conceptual knowledge in humans with a gridlike code. *Science*. 2016:352: 1464–1468.

Conway CM, Christiansen MH. Modality-constrained statistical learning of tactile, visual, and auditory sequences. *J Exp Psychol Learn Mem Cogn*. 2005:31:24–39.

Dale AM. Optimal experimental design for event-related fMRI. *Hum Brain Mapp*. 1999:8:109–114.

Devereux BJ, Clarke A, Marouchos A, Tyler LK. Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *J Neurosci*. 2013:33: 18906–18916.

Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 2019. p. 4171–4186.

Eom Y-H. Resilience of networks to environmental stress: from regular to random networks. *Phys Rev E*. 2018:97:042313.

Fang Y, Wang X, Zhong S, Song L, Han Z, Gong G, Bi Y. Semantic representation in the white matter pathway. *PLoS Biol*. 2018:16:e2003993.

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

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

Garvert MM, Dolan RJ, Behrens TE. A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *elife*. 2017:6:e17086.

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

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

Jackson AF, Bolger DJ. Using a high-dimensional graph of semantic space to model relationships among words. *Front Psychol*. 2014:5:385.

Karuza EA, Thompson-Schill SL, Bassett DS. Local patterns to global architectures: influences of network topology on human learning. *Trends Cogn Sci*. 2016:20:629–640.

Kenett YN, Levi E, Anaki D, Faust M. The semantic distance task: quantifying semantic distance with semantic network path length. *J Exp Psychol Learn Mem Cogn*. 2017:43:1470–1489.

Kim JS, Elli GV, Bedny M. Knowledge of animal appearance among sighted and blind adults. *Proc Natl Acad Sci U S A*. 2019:116: 11213–11222.

Kragel PA, Kano M, van Oudenhove L, Ly HG, Dupont P, Rubio A, Delon-Martin C, Bonaz BL, Manuck SB, Gianaros PJ, et al. Generalizable representations of pain, cognitive control, and negative emotion in medial frontal cortex. *Nat Neurosci*. 2018:21: 283–289.

Kriegeskorte N, Mur M. Inverse MDS: inferring dissimilarity structure from multiple item arrangements. *Front Psychol*. 2012: 3:245.

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

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

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

Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L-J, Shamma DA. Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int J Comput Vis*. 2017:123:32–73.

Kumar AA. Semantic memory: a review of methods, models, and current challenges. *Psychon Bull Rev*. 2021:28:40–80.

Lenci A. Distributional models of word meaning. *Annu Rev Linguist*. 2018:4:151–171.

Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization. *Adv Neural Inf Process Syst*. 2014:27:2177–2185.

Levy O, Goldberg Y, Dagan I. Improving distributional similarity with lessons learned from word embeddings. *Trans Assoc Comput Linguist*. 2015:3:211–225.

Lewis M, Zettersten M, Lupyan G. Distributional semantics as a source of visual knowledge. *Proc Natl Acad Sci U S A*. 2019:116: 19237–19238.

Li S, Zhao Z, Hu R, Li W, Liu T, Du X. Analogical reasoning on Chinese morphological and semantic relations. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2018. p. 138–143.

Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. *J Assoc Inf Sci Technol*. 2007:58:1019–1031.

Liu F, Yang M, Lin D. *Chinese web 5-gram version 1*. Philadelphia: Linguistic Data Consortium; 2010

Lü L, Zhou T. Link prediction in complex networks: a survey. *Physica A*. 2011:390:1150–1170.

Lynn CW, Bassett DS. How humans learn and represent networks. *Proc Natl Acad Sci U S A*. 2020:117:29407–29415.

Lynn CW, Kahn AE, Nyema N, Bassett DS. Abstract representations of events arise from mental errors in learning and memory. *Nat Commun*. 2020:11:1–12.

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

Martin A, Haxby JV, Lalonde FM, Wiggs CL, Ungerleider LG. Discrete cortical regions associated with knowledge of color and knowledge of action. *Science*. 1995:270:102–105.

Martin CB, Douglas D, Newsome RN, Man LL, Barense MD. Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *elife*. 2018:7:e31873.

Miceli G, Fouch E, Capasso R, Shelton JR, Tomaiuolo F, Caramazza A. The dissociation of color from form and function knowledge. *Nat Neurosci*. 2001:4:662–667.

Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th International Conference on Neural Information Processings Systems (NIPS)*. 2013. p. 3111–3119.

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

Newman ME. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys Rev E*. 2001:64:016132.

Nichols T, Brett M, Andersson J, Wager T, Poline J-B. Valid conjunction inference with the minimum statistic. *NeuroImage*. 2005:25:653–660.

Noppeney U, Price CJ. Retrieval of abstract semantics. *NeuroImage*. 2004:22:164–170.

Oliva A, Torralba A. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int J Comput Vis*. 2001:42:145–175.

Patterson K, Nestor PJ, Rogers TT. Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat Rev Neurosci*. 2007:8:976–987.

Peelen MV, He C, Han Z, Caramazza A, Bi Y. Nonvisual and visual object shape representations in occipitotemporal cortex: evidence from congenitally blind and sighted adults. *J Neurosci*. 2014:34:163–170.

Peer M, Brunec IK, Newcombe NS, Epstein RA. Structuring knowledge with cognitive maps and cognitive graphs. *Trends Cogn Sci*. 2020:25:37–54.

Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2014. p. 1532-1543.

Pereira F, Gershman S, Ritter S, Botvinick M. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cogn Neuropsychol*. 2016:33:175–190.

Pereira F, Lou B, Pritchett B, Ritter S, Gershman SJ, Kanwisher N, Botvinick M, Fedorenko E. Toward a universal decoder of linguistic meaning from brain activation. *Nat Commun*. 2018:9:1–13.

Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 2018. p. 2227–2237.

Qi F, Yang C, Liu Z, Dong Q, Sun M, Dong Z. Openhownet: an open sememe-based lexical knowledge base. 2019: arXiv preprint arXiv:1901.09957.

Romberg AR, Saffran JR. Statistical learning and language acquisition. *Wiley Interdiscip Rev Cogn Sci*. 2010:1:906–914.

Saffran JR, Aslin RN, Newport EL. Statistical learning by 8-month-old infants. *Science*. 1996:274:1926–1928.

Saffran JR, Senghas A, Trueswell JC. The acquisition of language by children. *Proc Natl Acad Sci U S A*. 2001:98:12874–12875.

Saysani A, Corballis MC, Corballis PM. Colour envisioned: concepts of colour in the blind and sighted. *Vis Cogn*. 2018:26:382–392.

Schapiro AC, Kustner LV, Turk-Browne NB. Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Curr Biol*. 2012:22:1622–1627.

Schapiro AC, Rogers TT, Cordova NI, Turk-Browne NB, Botvinick MM. Neural representations of events arise from temporal community structure. *Nat Neurosci*. 2013:16:486–492.

Schneider W, Eschman A, Zuccolotto A. *E-prime reference guide*. Psychology Software Tools, Incorporated; 2002

Striem-Amit E, Wang X, Bi Y, Caramazza A. Neural representation of visual concepts in people born blind. *Nat Commun*. 2018:9:1–12.

Theves S, Fernandez G, Doeller CF. The hippocampus encodes distances in multidimensional feature space. *Curr Biol*. 2019:29:1226–1231.

Theves S, Fernández G, Doeller CF. The hippocampus maps concept space, not feature space. *J Neurosci*. 2020:40:7318–7325.

Unger L, Fisher AV. The emergence of richly organized semantic knowledge from simple statistics: a synthetic review. *Dev Rev*. 2021:60:100949.

Utsumi A. Exploring what is encoded in distributional word vectors: a neurobiologically motivated analysis. *Cogn Sci*. 2020:44:e12844.

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

Wang X, Wu W, Ling Z, Xu Y, Fang Y, Wang X, Binder JR, Men W, Gao J-H, Bi Y. Organizational principles of abstract words in the human brain. *Cereb Cortex*. 2018:28:4305–4318.

Wang X, Men W, Gao J, Caramazza A, Bi Y. Two forms of knowledge representations in the human brain. *Neuron*. 2020:107:383–393.

Whitney C, Kirk M, O'Sullivan J, Lambon Ralph MA, Jefferies E. The neural organization of semantic control: TMS evidence for a distributed network in left inferior frontal and posterior middle temporal gyrus. *Cereb Cortex*. 2011:21:1066–1075.

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

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

Xu Y, He Y, Bi Y. A tri-network model of human semantic processing. *Front Psychol*. 2017:8:1538.

Xu Y, Wang X, Wang X, Men W, Gao J-H, Bi Y. Doctor, teacher, and stethoscope: neural representation of different types of semantic relations. *J Neurosci*. 2018:38:3303–3317.

Yermolayeva Y, Rakison DH. Seeing the unseen: second-order correlation learning in 7- to 11-month-olds. *Cognition*. 2016:152:87–100.