Received: 1 June 2023 | Revised: 10 November 2023 | Accepted: 15 November 2023
DOI: 10.1002/hbm.26546

RESEARCH ARTICLE

WILEY

Unraveling lexical semantics in the brain: Comparing internal, external, and hybrid language models

Yang Yang[1,2,3] | Luan Li[1,2,3] | Simon de Deyne[4] | Bing Li[5] | Jing Wang[1,2,3] | Qing Cai[1,2,3]

[1]Shanghai Key Laboratory of Brain Functional Genomics (Ministry of Education), Affiliated Mental Health Center (ECNU), Institute of Brain and Education Innovation, School of Psychology and Cognitive Science, East China Normal University, Shanghai, China

[2]Shanghai Changning Mental Health Center, Shanghai, China

[3]Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, China

[4]School of Psychological Sciences, University of Melbourne, Melbourne, Victoria, Australia

[5]UMR 9193—SCALab—Sciences Cognitives et Sciences Affectives, Université de Lille, CNRS, Lille, France

Correspondence
Jing Wang and Qing Cai, School of Psychology and Cognitive Science, East China Normal University, 3663 North Zhongshan Road, Shanghai, 200062, China.
Email: wangjing@psy.ecnu.edu.cn and qcai@psy.ecnu.edu.cn

Funding information
Basic Research Project of Shanghai Science and Technology Commission, Grant/Award Number: 19JC1410101; Fundamental Research Funds for the Central Universities; National Natural Science Foundation of China, Grant/Award Numbers: 31970987, 32100857

Abstract

To explain how the human brain represents and organizes meaning, many theoretical and computational language models have been proposed over the years, varying in their underlying computational principles and in the language samples based on which they are built. However, how well they capture the neural encoding of lexical semantics remains elusive. We used representational similarity analysis (RSA) to evaluate to what extent three models of different types explained neural responses elicited by word stimuli: an External corpus-based word2vec model, an Internal free word association model, and a Hybrid ConceptNet model. Semantic networks were constructed using word relations computed in the three models and experimental stimuli were selected through a community detection procedure. The similarity patterns between language models and neural responses were compared at the community, exemplar, and word node levels to probe the potential hierarchical semantic structure. We found that semantic relations computed with the Internal model provided the closest approximation to the patterns of neural activation, whereas the External model did not capture neural responses as well. Compared with the exemplar and the node levels, community-level RSA demonstrated the broadest involvement of brain regions, engaging areas critical for semantic processing, including the angular gyrus, superior frontal gyrus and a large portion of the anterior temporal lobe. The findings highlight the multidimensional semantic organization in the brain which is better captured by Internal models sensitive to multiple modalities such as word association compared with External models trained on text corpora.

KEYWORDS
fMRI, free word association, lexical semantics, representational similarity analysis, semantic model

1 | INTRODUCTION

Our ability to navigate within a network of concepts and retrieve corresponding linguistic labels allows us to think, communicate, and make sense of the world. How the human brain represents and organizes

Yang Yang and Luan Li contributed equally to this work.

https://doi.org/10.1002/hbm.26546

meaning has therefore always been at the forefront of scientific explorations. Recent advancements in natural language processing and large human-generated language norms have made it possible to develop language models and hypotheses about how lexical-semantic representations are organized. These models have shown great success in performing semantic tasks and predicting human behavior. Yet to reveal the psychological reality of these models, it is critical we evaluate how well they relate to encoding of lexical meaning in the brain, which is still poorly understood.

One way to model semantic representations is the corpus-based distributional semantics approach. It holds that word meanings are derived from words' statistical co-occurrence patterns in language use, which could be extracted from large text corpora. Advancement in computing power and access to larger and better corpora in recent decades has spurred many word embedding models built within this approach. They represent a word's meaning in a high-dimensional semantic space by extracting its co-occurrence probabilities with other words from the text (e.g., Latent Semantic Analysis, Landauer & Dumais, 1997; word2vec, Mikolov et al., 2013). The relationships between two words can thus be measured as the overlap between their high-dimensional vectors that encode the context in which they are used, which is typically measured as a cosine similarity.

These models have successfully predicted several aspects of semantic cognition (for a recent review, see Lenci, 2018). Word vector similarities derived from these models were found to correlate with similarities computed from brain activity patterns (Anderson et al., 2019; Carota et al., 2017; Carota et al., 2021; Fu et al., 2022; Pereira et al., 2018; Xu et al., 2018). However, they have also been criticized for making no connections to the phenomenal experience of us human beings (Barsalou, 2008, 2016), particularly given the accumulating evidence for sensory-derived neural representations in which conceptual knowledge is grounded (Binder et al., 2016; Binder & Desai, 2011; Fernandino et al., 2022; Martin, 2016; Patterson et al., 2007). Due to the lack of symbol grounding, distributional models also failed to represent word meanings in novel situations (Glenberg & Robertson, 2000), further questioning the psychological plausibility of the mechanisms by which they could build up semantic representations. To reconcile, there have been some attempts at incorporating experience-related information into corpora-derived features in computational models (Bruni et al., 2014; Chen et al., 2017; Hoffman et al., 2018; Johns & Jones, 2012). Some opted for the addition of sensory-perceptual data into distributional data to form hybrid multimodal models (Johns & Jones, 2012). Some showed that semantic representations, including sensorimotor properties (Hoffman et al., 2018) and taxonomic structures (Chen et al., 2017), could emerge from sequences of co-occurrences under the distributional principle, suggesting mechanisms by which distributional models might be able to acquire interpretable associations for lexical concepts given additional information. Still, text-based distributional models themselves inadequately account for the full dimensions of semantic knowledge.

A related distributional view of meaning that does not use latent dimensions is the semantic network view. Like the distributional view, the relatedness between two words is captured by the distribution of direct and indirect paths connecting two nodes. In contrast to the distributional view, localist representations are used in which words are represented as individual nodes in a network connected through edges. This view has a long tradition rooted in the hierarchical taxonomic models (Collins & Quillian, 1969) and the spreading activation framework (Collins & Loftus, 1975; Quillian, 1967). In contrast to previous theoretical work, modern network-based approaches exploit recent large-scale empirical datasets of human-generated features and word relations to build networks that allow probing the relationships between nodes as well as the topological structure of the network (De Deyne et al., 2019; De Deyne & Storms, 2008). Among them, the free association task has been gaining increasing popularity in recent years. In this task, participants are asked to provide the first words that come to mind when seeing or hearing a cue word. Because it is unconstrained, compared with other tasks such as feature listing or taxonomic judgment, it has the potential to uncover the full breadth of conceptual representations. As such, word relations derived this way are considered to best approximate our inner conceptual representations (Jorge-Botana et al., 2018), including multidimensional semantic knowledge, encompassing distributed, grounded, and aggregated world experience. Several large-scale word-free association norms have been developed in different languages thanks to online crowdsourcing data collection. Associative models derived thereof could also reliably predict response latencies in lexical decision and word naming tasks, as well as semantic relatedness judgments (De Deyne et al., 2019; De Deyne & Storms, 2008).

The distributional and the localist network-based approaches diverge on the language samples based on which lexical semantic representations are computed. Distributional models derive semantic spaces for words from external language stimuli (i.e., text) and can thus be seen as external language models. Here, we refer to free association models as internal models since they are obtained directly from human judgments, which reflect both linguistic and non-linguistic experiences. Although both the external and internal models reflect aspects of lexical semantic dimensions, measures that transform words' association strength and consider the large-scale network structure in internal models better-captured relatedness and similarity judgments than the word similarity measures from external models (De Deyne et al., 2016; De Deyne et al., 2019); and when both models were supplemented with additional visual and/or affective features, internal models continued to predict behavioral performance better than external models (De Deyne et al., 2021).

It is nonetheless unclear whether the relational information computed from internal models based on subjective, behavioral tasks maps onto the neural activation in the brain—the true internal architecture—better than that from external models. One way to address this open question is to construct some words' semantic representations using different language models, obtain representational dissimilarity matrices (RDMs) of the word pairs, and conduct representational similarity analysis (RSA) between the model RDMs and neural RDMs of brain responses to the words (Kriegeskorte et al., 2008). Previous neuroimaging research has not reached a consensus on this matter. Using the RSA method, Fernandino et al. (2022) compared RDMs of words built using a range of internal and external language models. They found that all models predicted similar structure of meaning-related neural activation patterns across the brain's language network with particularly strong activations in left angular gyrus (AG), precuneus (PCun), superior frontal gyrus (SFG) and inferior frontal gyrus (IFG), but only the model based on aspects of nonlinguistic experiential information, a type of internal model, independently accounted for variance in the neural data after controlling for the predictions of other models (for similar results, see Tong et al., 2022). On the other hand, it is also possible that external and internal models capture distinct semantic computations and therefore their neural correlates might be found in different brain regions. Carota et al. (2021) compared word2vec and an internal taxonomic model and found that they correlated with different patterns of neural activations. Specifically, word2vec had significantly higher correlations with activations in left IFG and AG, while the taxonomic model had a higher correlation in left posterior middle temporal gyrus (pMTG). Xu et al. (2018) also found distinct neural correlates of external distributional associations in the left temporoparietal junction (TPJ) and internal taxonomic relations in the anterior temporal lobe (ATL).

Note that in the previous work, word representations computed using the internal models were based on selected semantic properties such as taxonomic relations (Carota et al., 2021; Xu et al., 2018) and experiential attributes (Fernandino et al., 2022; Tong et al., 2022). The representational spaces used to model word similarity information in these studies were relatively coarse, encompassing only partial dimensions. Whereas neural representations of meaning should encode much finer-grained information about all aspects of lexical concepts. In this respect, these models might not be able to provide a full match on the mental semantic structure and approximate the neural response patterns. We aim to address this issue by using free word association data to compute the internal language models.

In this study, we investigate which types of theoretical language models better respect lexical semantics represented in the human brain. Three relational lexical networks were constructed based on free word associations, word2vec, and ConceptNet, leading to internal, external, and hybrid language models. The word associations were acquired from a new dataset taken from the Chinese Small World of Words (SWOW-ZH) project (Li et al., under review). ConceptNet was used as an example of hybrid models in this study. In ConceptNet, words are labeled with 36 different types of relations by humans, such as “capable of” and “is used for.” Importantly, a retrofitting procedure was used to combine these human-annotated relations with distributional information pre-trained with word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). Thus, ConceptNet represents an endeavor to reconcile localist and distributional approaches to semantic representations. ConceptNet indeed outperformed other distributional models in semantic tasks including relatedness judgments, sentence, and analogy completion (Speer et al., 2017). Therefore, we use it as a representative hybrid model here. RSA and functional MRI were used to evaluate how well the representational similarity predicted by each language model aligns with the neural similarity from fMRI activation patterns.

We additionally probe how the hierarchical semantic structure that emerges in language models maps onto the brain. Unlike previous work that derived semantic categories from human-generated taxonomic relations, we used a community detection algorithm to partition word nodes that are more densely connected to each other than to other words into groups to identify shared taxonomic structures across the three language models. This way we were able to obtain groups of concepts with minimal subjectivity and maximal dependence on the relationships from respective language samples. We also explore a prediction from the exemplar theory which stipulates that taxonomic representations are developed through repeated exposure to the exemplars of a category (Ashby & Rosedahl, 2017; Nosofsky, 1986), whereby categorical structures might be represented by the subordinate exemplars rather than an averaged summary of the category. The latter would be predicted by the competing prototype theory (Rosch & Mervis, 1975). To this aim, we also examined the similarity structures of a set of representative exemplars in the communities, which were used as the word stimuli in the fMRI experiment. Thus, RSAs were computed at three levels: community, exemplar, and node (word).

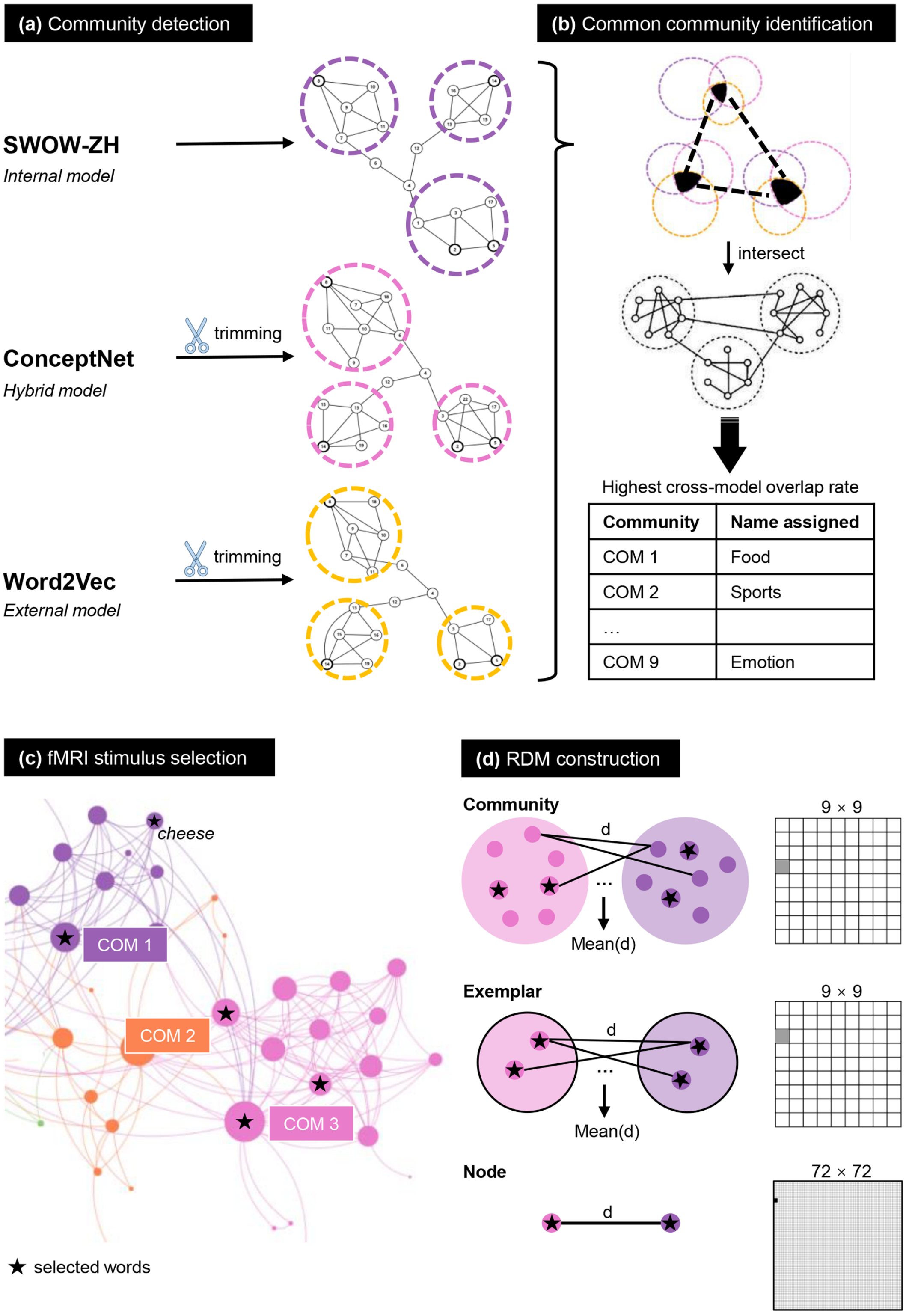
2 | METHODS

2.1 | Methods overview

We investigated how different language models account for the neural organization of concepts. Three semantic networks were constructed based on word association-based (SWOW-ZH), corpus-based (word2vec), and hybrid (ConceptNet) datasets. Community detection was performed on each network to obtain the natural groupings of concepts. fMRI data of word representation were collected on words in the common communities across the three theoretical models. The inter-stimulus RDMs were built for each model and the neural responses. Representational similarity analyses were performed by comparing the theoretical and neural RDMs within individual regions (Figure 1). The results revealed which language models were able to characterize the neural responses in specific brain areas, and which brain areas represented finer- or coarser-grained semantic information.

2.2 | The internal model: SWOW

The internal model was based on word associations from the newly released Chinese SWOW dataset (Li et al., under review). A semantic network was constructed where nodes corresponded to cues in the word association data, and edges represented the strength (i.e., the number of cued responses) between a cue and response word derived from the word association task. As part of the Small World of Words project, participants were recruited and performed the task online



Community	Name assigned
COM 1	Food
COM 2	Sports
…	
COM 9	Emotion

FIGURE 1 Method overview. (a) Community detection was performed in each language model to obtain a group of words that were densely connected. The ConceptNet and word2vec models were first trimmed so that the network density matched that of SWOW-ZH. (b) The top 9 communities with the highest overlap of words across language models were selected. (c) fMRI data were collected on 72 two-character words, 8 from each of the 9 common communities. The communities were matched on word frequency, visual complexity, and node in-degree in the SWOW network. (d) Representational dissimilarity matrices of semantic distance were constructed at three levels. The community-level matrix represented the inter-community distance based on all the cross-model common words in the community. The exemplar-level matrix computed the inter-community distance based on the experimental word stimuli. The node-level matrix represented the inter-word distance.

(https://smallworldofwords.org/zh). In each of the 18 trials, participants were asked to type in three unique words that first came to their mind when a cue word was presented (De Deyne et al., 2013). The data were cleaned by removing non-Chinese characters, excluding strings of more than seven characters, and converting the traditional Chinese into simplified Chinese characters (github.com/BYVoid/OpenCC).

A network of concepts was then constructed based on the associations between word nodes. Two words were connected if one was the cue and the other was the response word to it. This network (referred to as SWOW in the following text) is the first Chinese word association network and by far the largest Chinese semantic network resource. The data used in the presented study were collected between years 2016 and 2018, including 543,952 response entries from 15,870 participants. To control for the sparsity of the network, only the first response word in each association trial was included. The total number of unique word concepts in the current network was 9399.

2.3 | The external model: word2vec

We utilized a word embedding model pre-trained using word2vec (Mikolov et al., 2013) to characterize the semantic relations of words according to external resources. This model used the Baidu Encyclopedia corpus that contained approximately one billion tokens (https://github.com/Embedding/Chinese-Word-Vectors). The vocabulary consisted of the 24,922 most frequent words. A skip-gram word2vec model was trained to construct a 300-dimensional feature space based on word co-occurrence statistics (window size $=$ 5, sub-sampling rate $=$ .0001, negative sample number $=$ 5, learning rate $=$ .025).

2.4 | The hybrid model: ConceptNet

To examine the joint effect of external and internal representation of concepts in accounting for the neural responses, we constructed a theoretical RDM using ConceptNet Numberbatch word vectors v17.06, a hybrid of knowledge graph (ConceptNet) with annotated word relations and multiple pre-trained word embedding models, namely word2vec, GloVe, and OpenSubtitles 2016 (https://github.com/commonsense/conceptnet-numberbatch). By incorporating graph-structured knowledge into fully corpus-based distributional semantics, ConceptNet Numberbatch outperformed many other systems in various evaluations of word relatedness (Speer et al., 2017). ConceptNet Numberbatch is multilingual, meaning that words in different languages share the same semantic space. The dimension of the vector representations in this model was 300. We note that the Hybrid model in this study was not a combination of the current Internal and External models. Rather, it is an example of hybrid models in the sense that it draws from both corpus-based distributional information and annotated relations produced by humans.

2.5 | Community detection within each language model

A community is a group of nodes that are more densely connected with each other than with other nodes. We performed community detection of the semantic networks derived from the three language models to obtain natural groupings of word concepts following a data-driven approach.

The vector representations from the word2vec and ConceptNet models could be viewed as fully connected networks. They have large community sizes with far more nodes and edges than the SWOW network. To allow for direct comparisons in terms of the network structure, we trimmed the word2vec and ConceptNet networks so that their network density, that is, the proportion of the number of actual edges to the number of all the possible edges, was comparable to that of SWOW. The trimming procedure was done step-by-step, removing the edges with the smallest cosine similarity until the density approximated that of SWOW. Admittedly, the final networks in our analysis still differed in size (Table 1). Yet, by aligning the network density, we were able to focus on the communities that have a similar level of connectivity and are thus more comparable.

Community detection was then performed using the Louvain method (Blondel et al., 2008), a heuristic greedy algorithm that optimized the Newman-Girvan modularity, implemented in the Python toolkit NetworkX (https://networkx.org/; Hagberg et al., 2008). The detections yielded modularity scores of 0.34, 0.63, and 0.57 for the SWOW, word2vec, and ConceptNet networks, respectively, suggesting good partitions of the three networks (Table 1).

2.6 | Identifying common communities across models

To investigate how different language models account for neural representations of semantic concepts, we focused on a subset of communities that met two criteria: (1) they were consistently present in all three semantic networks, and (2) each formed a distinct semantic category that could be interpreted by humans.

The number of overlapping nodes over all the possible triplets of communities (one community per network) was first computed. For each triplet, we then calculated the proportion of common nodes over all the unique words in the triplet and selected the top nine triplets with the highest overlap rates. Three native speakers were independently presented with the overlapping words and asked to name the topic for each triplet. The names provided by them were highly consistent across individuals (Table 2; Table S1), indicating that the identified words covered a diverse range of topics. These communities were then used to sample stimulus words for the fMRI task.

2.7 | Stimuli of fMRI task

We sampled 72 two-character words, with eight words from each of the nine common communities. Three native speakers were independently

TABLE 1 Properties of the inter-word semantic networks.

Model	SWOW	word2vec	ConceptNet Numberbatch
Number of nodes	6525	6771	7981
Number of edges	127,983	131,115	167,404
Number of communities	39	539	97
Average community size	167.3	12.7	82.9
Max community size	1365	936	2194
Minimum community size	4	2	2
Density	0.006	0.005	0.006
Modularity	0.342	0.625	0.571

TABLE 2 Nine most consistent communities across language models.

Community	Number of all nodes in the triplets	Number of overlapping nodes	Overlap rate (%)
Food	817	304	37.2
Sports	154	41	26.6
Music	200	53	26.5
Economics	702	131	18.7
Astronomy	118	18	15.3
Crime	222	32	14.4
Social interaction	1456	203	13.9
Animal	405	51	12.6
Emotion	1363	164	12.0

asked to select 82-character words from each community that they considered "typical exemplars of that community." The resulting words were further balanced across communities based on: (1) visual complexity, namely the number of strokes of the Chinese characters in a word ($F_{(8,63)} = .73$); (2) in-degree of the node (word) in SWOW, which has been found to affect the behavioral performances on semantic tasks (De Deyne & Storms, 2008; $F_{(8,63)} = .26$); and (3) word frequency in SUBLEX-CH database (Cai & Brysbaert, 2010; $F_{(8,63)} = .90$), all *ps* >.05. In addition, a set of eight proper nouns referring to places were generated as the probes for the semantic detection task (described below).

2.8 | Participants

Twenty-one adults (13 female; mean age 22.5 years, age ranged from 18 to 29 years) were recruited in the fMRI experiment. All the participants were native Chinese speakers, right-handed, had normal or corrected-to-normal vision, and had no self-reported history of neurological diseases. All the participants gave written informed consent prior to the experiment. The procedure was approved by the East China Normal University Committee on Human Research Protection.

2.9 | Experimental paradigm and procedure

The fMRI task adopted a rapid event-related design (Kriegeskorte et al., 2008). A list of 72 concepts and eight place names was presented for 12 iterations (Figure 2). The word order in each presentation was pseudo-randomized. The 12 presentations were separated into six runs, two presentations per run. A run started with a 10-s fixation period and lasted about 410 s. In each stimulus trial, a two-character word was displayed at the center of the screen for 500 ms, followed by a fixation cross. The duration of the fixation was 3500 ms for 1/4 of the trials and 1500 ms for 3/4 of the trials, the distribution of which was pseudo-randomized within each presentation. Participants were asked to pay attention to the words and press a button

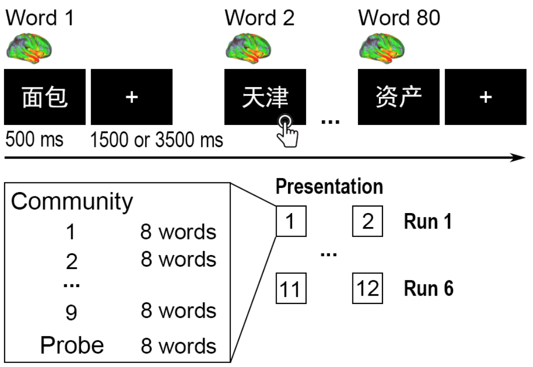


FIGURE 2 Experimental paradigm. In each presentation, 80 words were presented once in pseudorandomized order. Each word was followed by a fixation cross. Participants were asked to press a button when a place name was presented. The scan was composed of six runs, two presentations per run.

each time they saw a probe: a proper noun denoting the name of a place.

2.10 | MRI acquisition

Participants were scanned in a 3 T MRI scanner (Siemens Prisma; Siemens, Erlangen, Germany) using a 64-channel head coil. Functional images were acquired using a single-shot T2*-weighted gradient-echo echo-planar imaging pulse sequence (TR = 1000 ms, TE = 32 ms, flip angle [FA] = 55°, each volume comprising 72 axial slices, matrix = 96 × 96, field of view [FoV] read = 192 mm, voxel size = 2 × 2 × 2 mm^3). T1-weighted anatomical image was acquired using a multi-echo MPRAGE sequence, TR = 2300 ms, TE = 3.32 ms, FA = 8°, matrix = 256 × 256, FoV read = 240 mm, slice thickness = .9 mm).

2.11 | Image preprocessing

Functional images were preprocessed using fMRIPrep pipeline 20.1.1 (Esteban et al., 2019). Head motion correction was performed using MCFLIRT (FSL 5.0.9). Slice timing correction was performed using the AFNI 3dTshift. Images were normalized to the MNI152 template. AFNI 3dDeconcolve was used to fit a general linear model for each voxel, where all the trials of interest were modeled, and the six rigid-body head motion parameters were included as nuisance regressors. The time series were then high-pass filtered with a cutoff of 128 s.

2.12 | Constructing theoretical representational dissimilarity matrices

To represent different hypotheses of semantic organization, RDMs of the pairwise distance between stimulus words were constructed. The inter-concept distance in the sparse graph of SWOW was measured by the weighted shortest path length. The distance between words i and j was calculated as $D_{ij} = \sum_{k=1}^{n} \frac{1}{e_k}$, where n is the total number of steps in the shortest path from i to j, and e_k is the cue-response frequency of the pair of words on step k. The minimal value of the distance from word i to j and the distance from j to i was used as the distance between i and j to construct a symmetrical RDM. The inter-concept distance in the word2vec and ConceptNet Numberbatch was measured using two metrics: (a) the 1—cosine similarity of the vector representations of words, and (b) the weighted shortest path in the sparse graphs derived from the two embedding models, which used the same metric as the SWOW-based RDM. Including two metrics allowed for a more comprehensive comparison between the internal, hybrid, and external models. Thus, in total, five 72 × 72 inter-word RDMs were constructed, and we use the following acronyms for them: SWOW_g, CON_g, CON_v, W2V_g, and W2V_v, where "_v" and "_g" refer to the two metrics used to compute inter-concept distance.

To investigate the neural organization at the categorical level, we also constructed two types of 9 × 9 inter-community RDMs. The mean distance between all pairs of words from a pair of communities was computed to represent the inter-community distance. The exemplar RDMs were computed based on the stimulus words, whereas the community RDMs used all the words in the corresponding behavioral or corpus dataset in each community to characterize the community (Figure 1). The community RDMs included information about concepts that were not in the fMRI experiment but provided a characterization of the communities that were not biased by the specific stimulus set.

2.13 | Representational similarity analysis

Participant-specific responses to each trial of the target word were estimated using general linear models implemented with the AFNI 3dLSS function (https://afni.nimh.nih.gov/pub/dist/doc/program_help/3dLSS.html), in which a single target trial was estimated using one regressor and all the rest of the trials were estimated using another regressor (Mumford et al., 2012). The beta estimates of 12 presentations of the same word concept were then averaged, producing 72 estimates per voxel per participant.

RSA was performed on a regional basis using Nilearn (http://nilearn.github.io/) in Python. The Harvard-Oxford atlas (Desikan et al., 2006) was used to define a total of 96 regions of interests (ROIs) in both hemispheres. Pairwise Euclidean distance between words was calculated over all the voxels within each ROI, resulting in 92 72 × 72 dissimilarity matrices per participant. These neural RDMs were averaged across participants to represent how the neural response patterns associated with representing individual concepts were dissimilar from each other at the group level. The 9 × 9 neural RDMs at the exemplar level were constructed in the same way as the theoretical RDMs.

Representational similarity between the neural RDM and each theoretical RDM at the node level was measured using Spearman's correlation over the pairs of vectorized RDMs (upper triangles without diagonals) for each ROI. The community-level similarity was computed between the exemplar-level neural RDMs and the theoretical RDMs at the exemplar level or at the community level. The resulting similarity map was tested against a null-hypothesis distribution, in which the similarity was computed using randomly shuffled RDMs over 5000 iterations. The significance test was performed at a false discovery rate (FDR) q value of .05.

3 | RESULTS

3.1 | Comparison of the theoretical models

Representational dissimilarity matrices of concepts were constructed based on internal word association data (SWOW_g), multi-source knowledge graph (hybrid model, CON_g and CON_v), and external corpus-based word embeddings (W2V_g and W2V_v) at the community, exemplar, and node levels (Figure 3). At the community level, where the concepts within a community were collapsed, the similarity between SWOW and the other models ranged from 0.56 to 0.63 (Figure 4). The similarities between the word2vec- and ConceptNet-derived models ranged from 0.72 to 0.91, the mean of which was significantly larger than the mean similarity between SWOW and the other models (Mann–Whitney *U* test on Fisher's z-transformed correlations, $U = 0$, $n_1 = 6$, $n_2 = 4$, two-tailed $p = .0095$). At the exemplar and the node levels, word2vec and ConceptNet RDMs were also

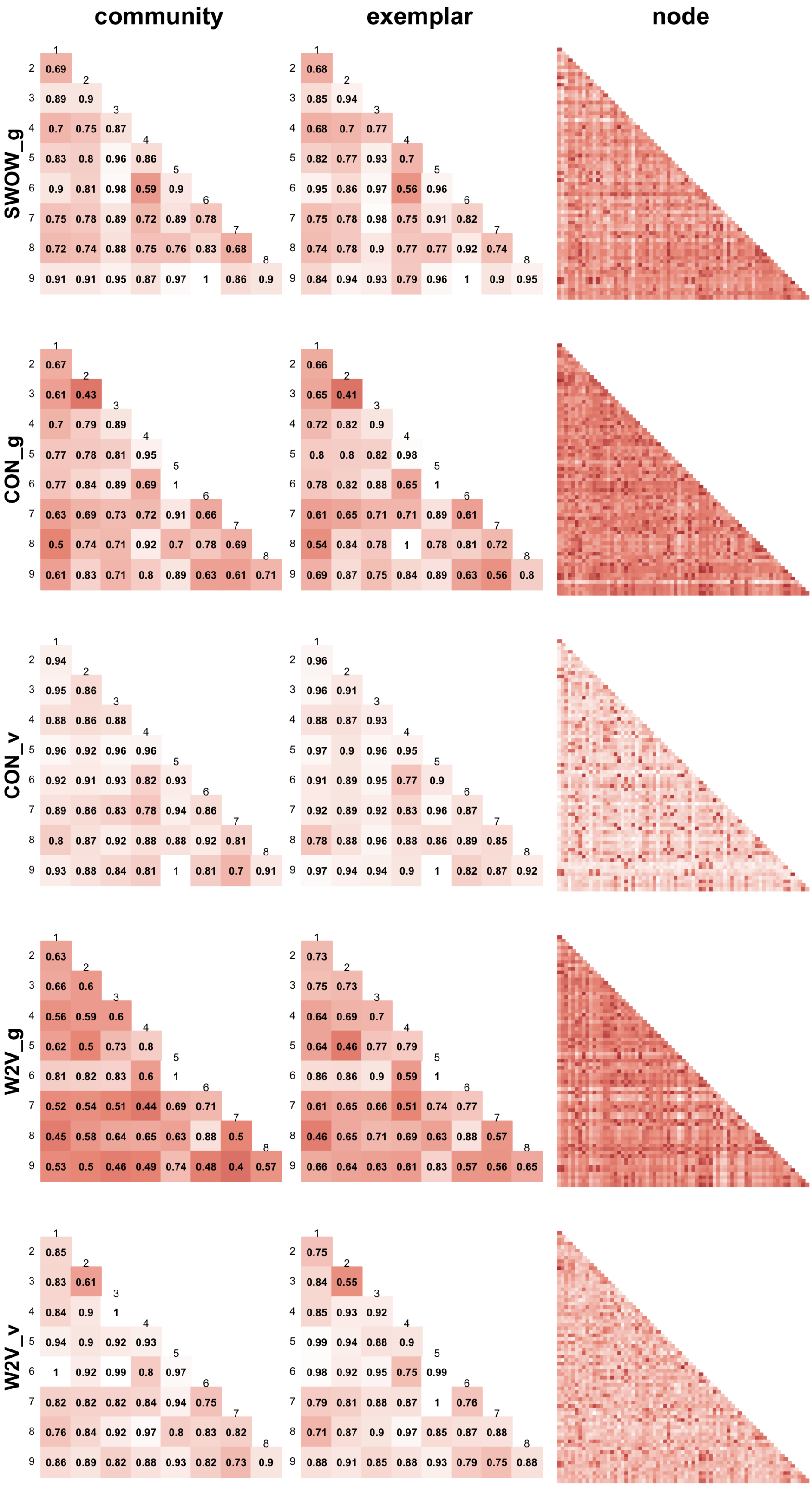


FIGURE 3 Representational dissimilarity matrices of concepts based on different hypotheses. The 9 × 9 matrices in Columns 1 and 2 used different methods to characterize inter-community distances. The 72 × 72 matrices in Column 3 characterized inter-word distances. Matrices in different rows represented the inter-concept relations computed in different language models. *SWOW_g*: RDM constructed using data of SWOW and graph-theory-based metrics of inter-concept distance, namely the weighted shortest path. *CON_g*: RDM using ConceptNet Numberbatch and graph-theory-based metrics. *CON_v*: RDM using ConceptNet Numberbatch and vector-based metrics of inter-concept distance, namely (1—cosine similarity). *W2V_g*: RDM using word2vec and graph-theory-based metrics. *W2V_v*: RDM using word2vec and vector-based metrics. The Communities 1 to 9 corresponded to Food, Sports, Music, Economics, Astronomy, Crime, Social interaction, Animal, and Emotion.

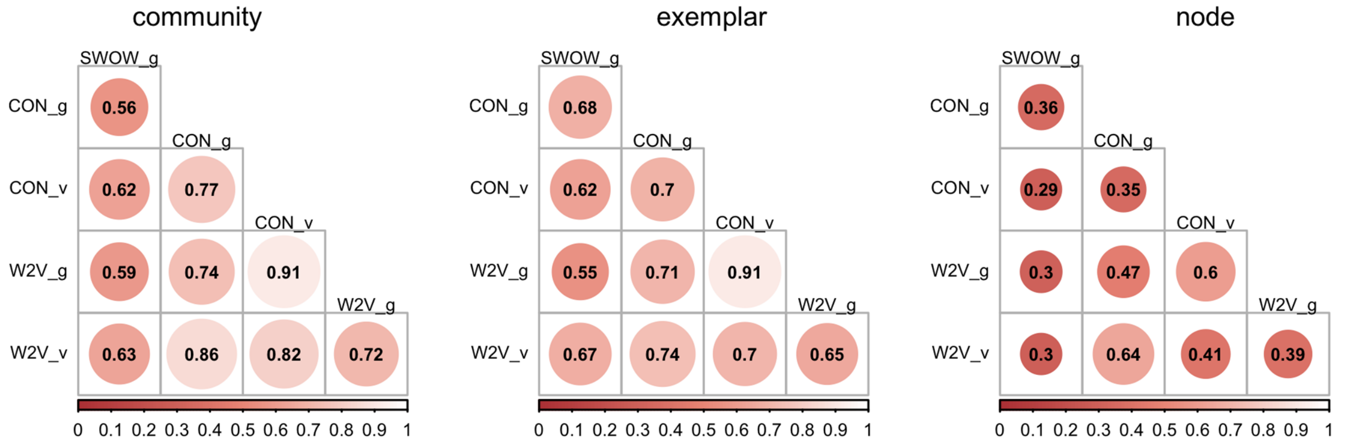


FIGURE 4 Pairwise correlations between five different RDMs at the community, exemplar, and node levels.

more similar to each other than to SWOW (exemplar: U = 2, *p* = .0286; node: U = 1, *p* = .0190). Thus, SWOW is consistently more distinct from word2vec or ConceptNet at all three levels of representations.

3.2 | Similarity between the neural representation and different theoretical models

The participants' performance in the detection task was at ceiling (overall: *M* accuracy = .99, *SD* = .004; detection words: *M* accuracy = .96, *SD* = .035). Therefore, all target trials were included in the analysis of the fMRI data.

The similarity between each theoretical RDM and the neural RDM of each anatomically defined region was computed and thresholded against the null distribution generated by a 5000-iteration random permutation. At the community level, SWOW was found to resemble the neural representations of concepts in multiple regions in the left temporal–parietal-occipital network, the superior frontal gyrus, and several of the right-hemisphere homologs (Figure 5a; Table 3). By contrast, only one region showed similarity with one of the ConceptNet RDMs (CON_v) at the community level, specifically the right frontal operculum (Figure 5d; Figure 5j). No region displayed similar organization principles with the word2vec RDMs (Figure 5g; Table 3).

At the exemplar level, neural representations showing significant correlations with the SWOW-based model were found in 12 regions, including the left superior temporal gyrus (STG), supramarginal gyrus, medial temporal lobe, bilateral anterior temporal lobes (ATLs), and bilateral superior frontal gyri (Figure 5b; Table 3). Representational similarity patterns in eight regions were significantly correlated with CON_v, including bilateral ATLs, right posterior temporal cortices, and right frontal areas (Figure 5e; Figure 5k). Very few or no region was found similar to word2vec or CON_g (Figure 5e; Figure 5h).

At the node level, 14 ROIs showed significant correlations with SWOW, including bilateral ATLs, middle and superior temporal cortices, and the fusiform gyrus (Figure 5c; Table 3). Only four regions showed a significant correlation with CON_v (Figure 5f; Figure 5l), and fewer regions showed similarity with the other models (Figure 5f; Figure 5i).

Within SWOW, the three types of characterizations of concept relations accounted for neural representation to different degrees. The correlation of neural representation with the community-level model was significantly greater than with the exemplar-level model across the ROIs (paired-sample t-test on Fisher's z-transformed correlation, $t = 7.06$, $p = 2.68 \times 10^{-10}$). The correlation of neural representation to the exemplar-level model was significantly greater than the node-level model across the ROIs (paired-sample t-test on Fisher's z-transformed correlation, $t = 10.11$, $p = 9.45 \times 10^{-17}$).

4 | DISCUSSION

To investigate how concepts are represented and organized in the brain, we built three theoretical language models to compute the similarity between word nodes, communities defined by exemplars, and communities defined by all the members. We then investigated the neural correlates of these relationships using RSA. Three models that quantified concept relations based on different hypotheses were included: the Internal model, which was based on large-scale free word association data; the External word2vec model; and the Hybrid ConceptNet model, which incorporated experiential information into corpus-derived word embeddings. Our results showed that semantic relations computed with the Internal model provided the closest approximation on average to the similarity pattern found in the brain, followed by the Hybrid model. In contrast, semantic relations computed from the External model had little direct neural mapping. Further, among the three levels of semantic relationships computed from the Internal model, the coarsest community level demonstrated the broadest involvement of brain regions, compared with the exemplar and the node levels. These community-level relations engaged a large portion of the brain regions critical for semantic processing, including the AG, SFG, and a large portion of the anterior temporal lobe.

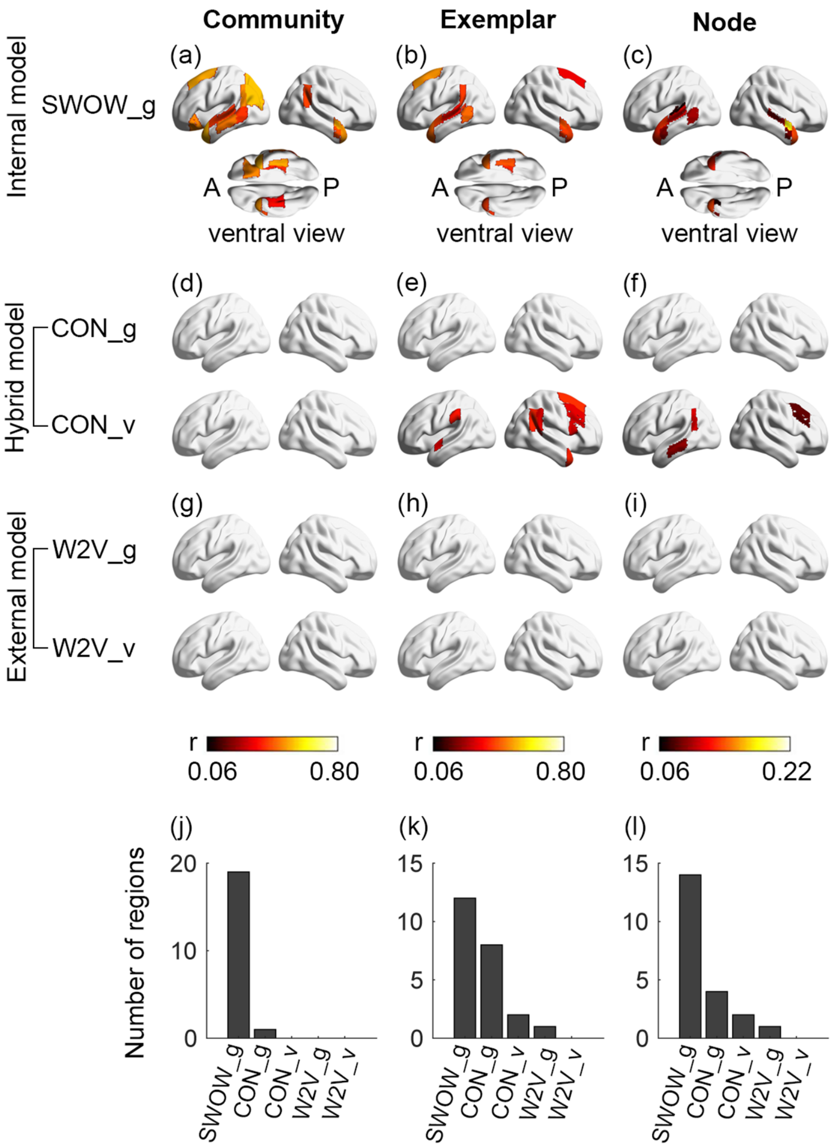


FIGURE 5 Results of representational similarity analyses. (a)–(i): Colored brain areas were the regions showing significant similarity with different theoretical models of inter-concept relations. (j)–(l): number of significant regions for different theoretical models.

Our results offer compelling evidence that the SWOW-based Internal language model surpasses the corpus-based distributional model in mapping with the neural similarity structure. Notably, even when distributional models were supplemented with additional human-labeled relations (i.e., the Hybrid ConceptNet model), they still could not match the explanatory power of SWOW. These results are consistent with prior research indicating that representational similarity structure computed by experience-based models, which we classify as another type of internal model, outperform external models like word2vec and GloVe in predicting the neural similarity structure (Fernandino et al., 2022; Tong et al., 2022). They also align with behavioral findings showing an advantage of SWOW over external models in predicting human performance in semantic tasks (De Deyne et al., 2016; De Deyne et al., 2019; De Deyne et al., 2021).

We found no significant correlation between word2vec and neural response patterns in any brain region. It is worth noting that previous research has reported finding conceptual similarities that correlated with neural similarities using external models like word2vec (Anderson et al., 2019; Carota et al., 2021; Fu et al., 2022; Xu et al., 2018). In those studies, either concrete words (Anderson et al., 2019; Carota et al., 2021) or abstract nouns (Wang et al., 2018) were used as stimuli, potentially tapping semantic dimensions necessary for computing similarity within the set of stimuli, which might have been captured by external models to some extent (Bi, 2021). However, the present study utilized noun stimuli of various categories, including concrete and abstract concepts. Moreover, the categories in this study were not arbitrarily chosen but obtained by community detection that found common taxonomic structures across language models. While the common communities were used for a fair comparison between different language models, this approach overlooked unique concept communities in each language model. It is likely that our designs allowed us to probe a more general similarity structure in the semantic space which corpus-based language models could not adequately explain.

Consider how corpus-based distributional models like word2vec are built: it compresses statistical information in a large corpus into a

TABLE 3 Representational similarity (Spearman's correlation) of the region of interest with each theoretical language model. Only 35 out of 96 regions that showed significantly similar organization of concepts with at least one model were listed.

	Community		Exemplar				Node			
	SWOW_g	CON_v	SWOW_g	CON_g	CON_v	W2V_g	SWOW_g	CON_g	CON_v	W2V_v
L ant STG	0.40		0.41		0.34		0.13			
L TP	0.54		0.45				0.13			
R TP	0.50		0.40		0.38		0.14			
L t-o MTG	0.39		0.45				0.10			
L pos STG	0.40		0.35				0.11			
L ant STG	0.52		0.43				0.18			
R ant MTG	0.43		0.40				0.09			
L SFG	0.52		0.50							
R ant PHG	0.34			0.35				0.10		0.06
R AG	0.41				0.38					
L pos MTG	0.47								0.10	
L pos PHG	0.39		0.38							
R MFG					0.31				0.08	
L pos TFC	0.52		0.43							
R SFG			0.32		0.39					
R pos TFC	0.35						0.06			
L AG	0.49								0.11	
R pos CG				0.44		0.43				
L pos ITG									0.05	
R pos PHG	0.32									
L pos SMG			0.40							
R pos SMG					0.31					
L FO										0.06
R ant TFC							0.07			
L ant SMG					0.36					
R FO		0.40								
L lat SOC	0.55									
L ant PHG	0.38									
L ant MTG							0.10			
R pos STG							0.10			
R PP							0.11			
R HG							0.08			
R oper IFG					0.30					
L PT							0.07			
L FOC	0.49									

Note: No regions showed significant representational similarity with models CON_g, W2V_g, W2V_v at the community level, W2V_v at the exemplar level, or W2V_g at the node level. Regions were defined according to the Harvard-Oxford atlas.
Abbreviations: AG, angular gyrus; ant, anterior division; CG, cingulate gyrus; FO, frontal operculum; FOC, frontal orbital cortex; HG, Heschl's gyrus; ITG, inferior temporal gyrus; L, left; lat, lateral; MFG, middle frontal gyrus; MTG, middle temporal gyrus; oper, pars opercularis; PHG, parahippocampal gyrus; pos, posterior division; PP, Planum Polare; PT, Planum Temporale; R, right; SFG, superior frontal gyrus; SMG, supramarginal gyrus; SOC, superior occipital cortex; STG, superior temporal gyrus; TFC, temporal fusiform cortex; t-o, temporooccipital part; TP, temporal pole.

vector space, typically with 300 dimensions through error-driven training and hyperparameter optimizations, which involve adjusting various settings to improve performance. It is unclear what information is preserved or lost during these tuning processes (Kumar et al., 2021). It is possible that distributional models are able to account for the fine-grained neural representations of concepts, but in a high-dimensional space, all concepts are distinct and far away from each other, making any characterization of the inter-concept relations difficult to generate. In contrast, the human brain tends to use low dimensions to characterize the world and perform various tasks (Bottini & Doeller, 2020). Therefore, semantic relations are likely computed in the brain with some low-dimensional principles absent from distributional language models. The results of the community detection analysis lend some support to this idea. When density was matched across networks, the word2vec network had a larger number of communities and greater modularity than the SWOW network (Table 1). This might be due to the word2vec's ability to characterize every aspect of semantics associated with the word concepts that are present in the corpora, leading to the great ability to capture fine-grained distinctions among concepts and more detailed groupings. By contrast, when human beings perform word-to-word free association tasks, the semantic features are unlikely to be analyzed in a comprehensive or algorithmic way. The heuristic access to numerous semantic features during free association might lead to a bias toward a limited number of features each time, resulting in a more interconnected representation of concepts.

It should be noted that we used word2vec as a starting point for comparing SWOW with external models. However, we acknowledge that there are other computational models (e.g., Transformers) and methods (e.g., pointwise mutual information) available for computing distributed word relations from text corpora. They might capture different or additional aspects of semantic knowledge and might provide a better mapping of neural semantic representations. Future studies should consider incorporating these alternative models and methods to further investigate the relationships between language models and neural semantic representations.

Despite the debate between distributional and localist approaches to semantic memory, recent evidence indicates that multimodal language models incorporating both distributional architecture and semantic features better explain neural (Anderson et al., 2019) and behavioral responses (Johns & Jones, 2012) than unimodal models. They strongly support the neural dual coding framework, which integrates language-derived and sensory-derived knowledge representations in the brain (Bi, 2021). Our results show that word association similarities only correlated moderately with word2vec and ConceptNet similarities, which is consistent with previous findings (Nematzadeh et al., 2017). This suggests that they capture different aspects of meaning. Further, word associations likely encode multimodal representations containing both distributed and experiential information. This suggestion is preliminarily supported by research showing that adding sensory or affective information to corpus-based models could enhance their performance in semantic tasks, but the improvement was minimal for the free word association model (De Deyne et al., 2021), suggesting that word associations already encode sensorimotor and affective information. Nevertheless, further investigation comparing word association and experiential models is warranted which also hinges on a better understanding of the nature of the relationships encoded in word associations.

We have identified neural representations in widespread regions in the left temporal–parietal-occipital network, particularly in the extensive cortex in the ATL, which showed similar patterns to the SWOW model across all three levels of comparison. These regions resemble the network identified in previous neuroimaging studies on semantic cognition (e.g., Binder et al., 2009; Tong et al., 2022). In particular, the ATL has been widely recognized as a "hub" for cross-modal semantic processing, with intrinsic connectivity to modality-specific brain regions (Patterson et al., 2007; Ralph et al., 2017). Our findings build upon previous research that has established the ATL's role in encoding conceptual similarity based on experiential features (Anderson et al., 2019; Fernandino et al., 2022; Tong et al., 2022; Wang et al., 2018), rated categories (Carota et al., 2017, 2021; Devereux et al., 2013) as well as distributional information (Anderson et al., 2019; Pereira et al., 2018; Xu et al., 2018), and extend them by suggesting that it also encodes free associative relations. Notably, it encodes not only the relations between lexical concepts (node) but also the categorical information (community and exemplar) latently computed through word's relationships with other words.

Our results also highlight other brain regions that were previously implicated in both unimodal and multimodal integration of semantic processes. Specifically, significant correlations with the SWOW model were observed at the community and exemplar levels in left pSTG. At the community level, we also found significant correlations in AG, which is involved in the integration of semantic information from unimodal inputs (Binder et al., 2009; Binder & Desai, 2011; Fernandino et al., 2016; Price et al., 2015) and left SOC, which serves as a visual input source to the temporal cortex (Humphreys & Riddoch, 2006), bilateral PHG, which is involved in scene recall and visual environment associations (Bonner et al., 2016; Epstein et al., 1999) and the orbitofrontal cortex, which processes emotion and value judgments (Wikenheiser & Schoenbaum, 2016).

Our exploration of three types of conceptual relations yielded interesting findings. While both the community and the exemplar RDMs presumably characterized inter-community similarity, the former showed significantly higher correlations with neural similarity patterns across the brain. Notably, the community-based model used all the words in the behavioral or corpus dataset for each community to construct the representational dissimilarity matrix, whereas the exemplar-based model used the exact sets of words that were presented in the fMRI study. The present results suggested that the neural representation of inter-community relations derived from individual concepts was more similar to an unbiased characterization of the community than to a stimulus-wise matching representation. Although the exemplar theory of category learning is strongly supported by the literature (Ashby & Rosedahl, 2017; Liu et al., 2023), it likely characterizes the learning process of individual instances better than the representation of categories (Murphy, 2016). Despite the mixed results regarding whether the exemplar model or the prototype model provides a better fit for brain data (Bowman et al., 2020; Mack et al., 2013), recent research has shown that during the course of category learning, exemplar- and prototype-approaches engaged different brain systems and both types of representations emerged, while prototype correlates dominated by the end of learning (Bowman et al., 2020). Therefore, together with our finding, it is likely that multiple types of category representation co-exist in the brain, more so as a summary characterization of its members.

In conclusion, we have shown that conceptual similarities at levels of community, exemplar, and word node computed from an internal, word association model map well onto neural similarities in a brain network of semantic processing. Notably, it outperformed an external, corpus-based distributional model and a hybrid model with both distributional and human-labeled relations in multiple brain areas. These findings indicate that word associations may encode conceptual representations with a general similarity principle that is absent in language models based on text corpora. Additionally, community-level RSA showed significant effects in broader brain regions than exemplar- and node-level analyses. This provides some additional support to previous research on the prototype theory of category representation in the brain. We propose avenues in which word association models and experiential models might be compared in future research to disambiguate the nature of the conceptual relations in word associations and to better understand what information is encoded in the conceptual representation system.

ACKNOWLEDGEMENTS

This work was funded by the National Natural Science Foundation of China (31970987 to Qing Cai; 32100857 to Jing Wang), the Basic Research Project of Shanghai Science and Technology Commission (No. 19JC1410101), and the Fundamental Research Funds for the Central Universities. We are grateful to Ziyi Ding for helpful discussion and the anonymous reviewers for the constructive comments that helped improve the quality of this paper.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The word stimuli in the fMRI experiment are provided in Table S1. The MRI data are available upon reasonable request from the authors.

ORCID

Jing Wang https://orcid.org/0000-0001-6022-8240

REFERENCES

Anderson, A. J., Binder, J. R., Fernandino, X. L., Humphries, C. J., Conant, L. L., Raizada, R. D. S., Lin, F., & Lalor, X. E. C. (2019). An integrated neural decoder of linguistic and experiential meaning. *Journal of Neuroscience*, *39*, 8969–8987. https://doi.org/10.1523/JNEUROSCI.2575-18.2019

Ashby, F. G., & Rosedahl, L. (2017). A neural interpretation of exemplar theory. *Psychological Review*, *124*(4), 472–482. https://doi.org/10.1037/REV0000064

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Barsalou, L. W. (2016). On staying grounded and avoiding quixotic dead ends. *Psychonomic Bulletin and Review*, *23*(4), 1122–1142. https://doi.org/10.3758/S13423-016-1028-3

Bi, Y. (2021). Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences*, *25*(10), 883–895. https://doi.org/10.1016/J.TICS.2021.07.006

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

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, *33*(3–4), 130–174. https://doi.org/10.1080/02643294.2016.1147426

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *10*, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Bonner, M., Price, A., & Peelle, J. E. (2016). Semantics of the visual environment encoded in parahippocampal cortex. *Journal of Cognitive Neuroscience*, *28*(3), 361–378. https://doi.org/10.1162/jocn_a_00908

Bottini, R., & Doeller, C. F. (2020). Knowledge across reference frames: Cognitive maps and image spaces. *Trends in Cognitive Sciences*, *24*(8), 606–619. https://doi.org/10.1016/J.TICS.2020.05.008

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *eLife*, *9*, 1–47. https://doi.org/10.7554/ELIFE.59360

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1–47. https://doi.org/10.1613/jair.4135

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One*, *5*(6), e10729. https://doi.org/10.1371/journal.pone.0010729

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

Carota, F., Nili, H., Pulvermüller, F., & Kriegeskorte, N. (2021). Distinct fronto-temporal substrates of distributional and taxonomic similarity among words: Evidence from RSA of BOLD signals. *NeuroImage*, *224*, 117408. https://doi.org/10.1016/J.NEUROIMAGE.2020.117408

Chen, L., Lambon Ralph, M. A., & Rogers, T. T. (2017). A unified model of human semantic knowledge and its disorders. *Nature Human Behaviour*, *1*(3), 0039. https://doi.org/10.1038/s41562-016-0039

Collins, A., & Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428.

Collins, A., & Quillian, M. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *82*(2), 240–247.

De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2021). Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, *45*(1), e12922. https://doi.org/10.1111/cogs.12922

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987–1006. https://doi.org/10.3758/S13428-018-1115-7/TABLES/5

De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2016). Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General*, *145*(9), 1228–1254. https://doi.org/10.1037/xge0000192.supp

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, *45*(2), 480–498. https://doi.org/10.3758/S13428-012-0260-7

De Deyne, S., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior Research Methods*, *40*(1), 213–231. https://doi.org/10.3758/BRM

Desikan, R., Ségonne, F., Fischl, B., & Quinn, B. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*, 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Devereux, B. J., Clarke, A., Marouchos, A., & Tyler, L. K. (2013). Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *Journal of Neuroscience*, *33*(48), 18906–18916. https://doi.org/10.1523/JNEUROSCI.3809-13.2013

Epstein, R., Harris, A., Stanley, D., & Kanwisher, K. (1999). The parahippocampal place area: Recognition, navigation, or encoding? *Neuron*, *23*(1), 115–125.

Esteban, O., Markiewicz, C., Blair, R., & Moodie, C. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C. J., Gross, W. L., Conant, L. L., & Seidenberg, M. S. (2016). Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral Cortex*, *26*(5), 2018–2034. https://doi.org/10.1093/cercor/bhv020

Fernandino, L., Tong, J. Q., Conant, L. L., Humphries, C. J., & Binder, J. R. (2022). Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(6), e2108091119. https://doi.org/10.1073/PNAS.2108091119/SUPPL_FILE/PNAS.2108091119.SD03.XLSX

Fu, Z., Wang, X., Wang, X., Yang, H., Wang, J., Wei, T., Liao, X., Liu, Z., Chen, H., & Bi, Y. (2022). Different computational relations in language are captured by distinct brain systems. *Cerebral Cortex*, *1–17*, 997–1013. https://doi.org/10.1093/cercor/bhac117

Glenberg, A., & Robertson, D. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, *43*(3), 379–401. https://doi.org/10.1006/jmla.2000.2714

Hagberg, A., Swart, P., & Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th python in science conference (SciPy2008)* (pp. 11–15). Pasadena.

Hoffman, P., McClelland, J. L., & Lambon Ralph, M. A. (2018). Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. *Psychological Review*, *125*(3), 293–328. https://doi.org/10.1037/REV0000094

Humphreys, G., & Riddoch, M. (2006). Features, objects, action: The cognitive neuropsychology of visual object processing, 1984–2004. *Cognitive Neuropsychology*, *23*(1), 156–183. https://doi.org/10.1080/02643290542000030

Johns, B., & Jones, M. (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, *4*(1), 103–120. https://doi.org/10.1111/j.1756-8765.2011.01176.x

Jorge-Botana, G., Olmos, R., & Luzón, J. M. (2018). Word maturity indices with latent semantic analysis: Why, when, and where is Procrustes rotation applied? *Wiley Interdisciplinary Reviews: Cognitive Science*, *9*(1), 1457. https://doi.org/10.1002/wcs.1457

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(NOV), 1–28. https://doi.org/10.3389/neuro.06.004.2008

Kumar, A. A., Steyvers, M., & Balota, D. A. (2021). A critical review of network-based and distributional approaches to semantic memory structure and processes. *Topics in Cognitive Science*, *0*, 1–24. https://doi.org/10.1111/tops.12548

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*, 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Li, B., Ding, Z. Y., De Deyne, S., & Cai, Q. (under review). A large-scale database of mandarin Chinese word associations from the small world of words project.

Liu, Z., Liao, S., & Seger, C. A. (2023). Rule and exemplar-based transfer in category learning. *Journal of Cognitive Neuroscience*, *35*(4), 628–644. https://doi.org/10.1162/jocn_a_01963

Mack, M., Preston, A., & Love, B. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology*, *23*(20), 2023–2027. https://doi.org/10.1016/j.cub.2013.08.035

Martin, A. (2016). GRAPES—Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychonomic Bulletin and Review*, *23*(4), 979–990. https://doi.org/10.3758/S13423-015-0842-3

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

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

Murphy, G. L. (2016). Is there an exemplar theory of concepts? *Psychonomic Bulletin and Review*, *23*(4), 1035–1042. https://doi.org/10.3758/S13423-015-0834-3/METRICS

Nematzadeh, A., Meylan, S., & Griffiths, T. (2017). Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th annual conference of the cognitive science society* (pp. 859–864). Cognitive Science Society.

Nosofsky, R. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976–987. https://doi.org/10.1038/nrn2277

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*, 1–13. https://doi.org/10.1038/s41467-018-03068-4

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

Quillian, M. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, *12*(5), 410–430. https://doi.org/10.1002/bs.3830120511

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

Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Science*, *7*(4), 573–605.

Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 4444–4451). AAAI.

Tong, J., Binder, J. R., Humphries, C., Mazurchuk, S., Conant, L. L., & Fernandino, L. (2022). A distributed network for multimodal experiential representation of concepts. *Journal of Neuroscience*, *42*(37), 7121–7130. https://doi.org/10.1523/JNEUROSCI.1243-21.2022

Wang, X., Wu, W., Ling, Z., Xu, Y., Fang, Y., Wang, X., Binder, J. R., Men, W., Gao, J.-H., & Bi, Y. (2018). Organizational principles of abstract words in the human brain. *Cerebral Cortex*, *28*(12), 4305–4318. https://doi.org/10.1093/cercor/bhx283

Wikenheiser, A., & Schoenbaum, G. (2016). Over the river, through the woods: Cognitive maps in the hippocampus and orbitofrontal cortex. *Nature Reviews Neuroscience*, *17*(8), 513–523. https://doi.org/10.1038/nrn.2016.56

Xu, Y., Wang, X., Wang, X., Men, W., Gao, J.-H., & Bi, Y. (2018). Doctor, teacher, and stethoscope: Neural representation of different types of semantic relations. *The Journal of Neuroscience*, *38*(13), 3303–3317. https://doi.org/10.1523/JNEUROSCI.2562-17.2018

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Yang, Y., Li, L., de Deyne, S., Li, B., Wang, J., & Cai, Q. (2024). Unraveling lexical semantics in the brain: Comparing internal, external, and hybrid language models. *Human Brain Mapping*, *45*(1), e26546. https://doi.org/10.1002/hbm.26546