Title: Decoding the Information Structure Underlying the Neural Representation of Concepts

Authors: Leonardo Fernandino[1,2]*, Lisa L. Conant[1], Colin J. Humphries[1], Jeffrey R. Binder[1,3]

Affiliations:

[1] Department of Neurology, Medical College of Wisconsin.

[2] Department of Biomedical Engineering, Medical College of Wisconsin

[3] Department of Biophysics, Medical College of Wisconsin.

*Corresponding author: Leonardo Fernandino, Medical College of Wisconsin, 8701 W. Watertown Plank Rd., Milwaukee, WI, USA. Email: lfernandino@mcw.edu. Phone: 414-955-7388.

Keywords:

Semantic memory; concept representation; embodied cognition; fMRI; representational similarity analysis.

Abstract:

The nature of the neural code underlying conceptual knowledge remains a major unsolved problem in cognitive neuroscience. Three main types of information have been proposed as candidates for the neural representations of lexical concepts: taxonomic (i.e., information about category membership and inter-category relations), distributional (i.e., information about patterns of word co-occurrence in natural language use), and experiential (i.e., information about sensory-motor, affective, and other features of phenomenal experience engaged during concept acquisition). In two experiments, we investigated the extent to which these three types of information are encoded in the neural activation patterns associated with hundreds of English nouns from a wide variety of conceptual categories. Participants made familiarity judgments on the meaning of written nouns while undergoing functional MRI. A high-resolution, whole-brain activation map was generated for each noun in each participant's native space. These word-specific activation maps were used to evaluate different representational spaces corresponding to the three types of information described above. In both studies, we found a striking advantage for experience-based models in most brain areas previously associated with concept representation. Partial correlation analyses revealed that only experiential information successfully predicted concept similarity structure when inter-model correlations were taken into account. This pattern of results was found independently for object concepts and event concepts. Our findings indicate that the neural representation of conceptual knowledge primarily encodes information about features of experience, and that - to the extent that it is represented in the brain - taxonomic and distributional information may rely on such an experience-based code.

Significance Statement:

The ability to identify individual objects or events as members of a kind (e.g., “knife”, “dog”, “party”) is a fundamental aspect of human cognition. It allows us to quickly access a wealth of information pertaining to a newly encountered object or event and use it to guide our behavior. How is this information represented in the brain? We used functional MRI to analyze patterns of brain activity corresponding to hundreds of familiar concepts, and quantitatively characterized the informational structure of these patterns. Our results indicate that, throughout the brain, conceptual knowledge is stored as patterns of neural activity that encode primarily experiential information grounded in modality-specific neural systems, challenging the idea that concept representations are independent of sensory-motor experience.

Main Text:

The capacity for conceptual knowledge is arguably one of the most defining properties of human cognition, yet it is still unclear how concepts are represented in the brain. Our tendency to organize objects, events and experiences into discrete categories has led most authors – dating back at least to Plato (1) – to take categorical structure as the central property of conceptual knowledge (2). Information about category membership and inter-category relations – i.e., taxonomic information – is traditionally seen as a hierarchically structured network, with basic level categories (e.g. “apple”, “orange”) grouped into superordinate categories (e.g., “fruit”, “food”), and subdivided into subordinate categories (e.g., “Gala apple”, “tangerine”) (3). A prominent account in cognitive science maintains that such categories are represented in the mind/brain as purely symbolic entities, whose semantic content and usefulness derive primarily from how they relate to each other (4, 5). Such representations are seen as qualitatively distinct from the sensory-motor processes through which we interact with the world, much like the distinction between software and hardware in digital computers.

An "experiential" representational system, on the other hand, encodes information about the experiences that led to the formation of particular concepts. It is motivated by a view, often referred to as embodied, grounded, or situated semantics, in which concepts arise primarily from generalization over particular experiences, as information originating from the various modality-specific systems (e.g., visual, auditory, tactile, motor, affective, etc.) is combined and re-encoded into progressively more schematic representations that are stored in memory. Since, in this view, there is a degree of continuity between conceptual and modality-specific systems, concept representations are thought to reflect the structure of the perceptual, affective and motor processes involved in those experiences (6–9).

A third type of information that can be encoded in concept representations, known as "distributional" information, pertains to statistical patterns of co-occurrence between lexical concepts (i.e., concepts that are widely shared within a population and denoted by a single word) in natural language usage. As is now widely appreciated, these co-occurrence patterns encode a substantial amount of information about word meaning (10–12). Although word co-occurrence patterns primarily encode contextual associations, such as those connecting the words "cow", "barn", and "farmer", semantic similarity information is indirectly encoded, since words with similar meanings tend to appear in similar contexts (e.g., "cow" and "horse", "pencil" and "pen"). This has led some authors to propose that concepts may be represented in the brain, at least in part, in terms of distributional information (10, 13).

Whether, and to what extent, each of these types of information plays a role in the neural representation of conceptual knowledge is a topic of intense research and debate. A large body of evidence has emerged from behavioral studies, functional neuroimaging experiments, and neuropsychological assessment of patients with semantic deficits, with results typically interpreted in terms of taxonomic (14–19), experiential (8, 20–25), or distributional (26–30) representations. However, the relative role of these representational systems in the neural representation of concepts remains controversial, in part because the evidence available is consistent with all three (18, 31, 32). Although these systems are qualitatively distinct, their representations of common lexical concepts are strongly inter-correlated: patterns of word co-occurrence in natural language are driven in part by taxonomic and experiential similarities between the concepts to which they refer, and the taxonomy of natural categories is systematically related to the experiential attributes of the exemplars (33, 34). Consequently, the empirical evidence currently available is unable to discriminate between these theoretical accounts.

We addressed this issue by deriving quantitative predictions from each of these representational systems for the similarity structure of a large set of concepts and then using representational similarity analysis (RSA) with high-resolution functional MRI (fMRI) to evaluate those predictions. Unlike the more typical cognitive subtraction technique, RSA focuses on the information structure of the pattern of neural responses to a set of stimuli (35). For a given stimulus set (e.g., words), RSA assesses how well the representational similarity structure predicted by a model matches the neural similarity structure observed from fMRI activation maps (Figure 1). This allowed us to directly compare, in quantitative terms, predictions derived from the three representational systems.

Results

In two experiments, participants made familiarity judgments on a large number of lexical concepts, selected from a broad range of taxonomic categories, while undergoing fMRI (see Materials and Methods section for details). Written nouns were presented, one at a time, on a computer screen, and participants rated each one according to how often they encountered the corresponding entity or event in their daily lives, on a scale from 1 ("rarely or never") to 3 ("often"). RSAs were conducted for a set of cortical areas previously associated with concept representation (36–38): angular gyrus (AG), supramarginal gyrus (SMG), superior temporal gyrus (STG), middle temporal gyrus (MTG), inferior temporal gyrus (ITG), anterior temporal lobe (ATL, including temporal cortical areas anterior to y = -2), parahippocampal gyrus (PHG), inferior frontal gyrus (IFG), caudal middle frontal gyrus (cMFG), superior frontal gyrus (SFG), precuneus (PCN), posterior cingulate gyrus (PCG), rostral anterior cingulate gyrus (rACG), and fusiform gyrus (FusG). These regions were anatomically defined following the Desikan-Killiany parcellation map (39). We also conducted RSA for a distributed, functionally defined region-of-interest (ROI) based on the voxel-based meta-analysis by Binder et al. (36) (Figure 1D; heretofore referred to as "semantic network ROI"). The neural similarity structure of concept representations in this ROI reflected not only information encoded in the activation patterns within each cortical area, but also in the pattern of activations *across* different areas.

From the fMRI data, we generated brain activation maps for each concept, reflecting the unique spatial pattern of neural activity for that concept (Figure 1). From these maps, a neural representational dissimilarity matrix (RDM) was generated for each ROI. RDMs consisted of all pairwise dissimilarities (1 – correlation) between the vectorized activation patterns across voxels. For each representational model investigated, a model-based RDM was computed, and its similarity to the neural RDM was evaluated via Spearman correlation. We evaluated six different representational models: two based on taxonomic information, two based on experiential information, and two based on distributional information (see details below). The resulting profile of relative model performances provided an assessment of the degree to which each type of information is reflected in the neural activation patterns corresponding to different concepts. Because the RDMs from different models were partly correlated with each other (Table S1), we also conducted partial correlation RSAs to evaluate the unique contribution of each model to neural concept representation. We stress that the representational models investigated here are models of *information content* (i.e., taxonomic, experiential, or distributional); they are not meant to model the neural architecture through which information is encoded in the brain (see Supplementary Materials for extended model descriptions).

Taxonomic models. *WordNet* is the most influential representational model based on taxonomic information, having been used in several neuroimaging studies to model semantic content (17, 28, 40, 41). It is organized as a knowledge graph in which words are grouped into sets of synonyms ("synsets"), each expressing a distinct concept. Synsets are richly interconnected according to taxonomic relations, resulting in a hierarchically structured network encompassing all noun concepts. Concept similarity is computed based on the shortest path connecting the two concepts in this network.

In contrast to WordNet, which provides a comprehensive taxonomy of lexical concepts, the *Categorical* model was customized to encode the particular taxonomic structure of the concept set in each study, based on a set of *a priori* categories. Therefore, the Categorical model ignores concept categories that were not included in the study, such as "furniture" or "clothing". To reduce the level of subjectivity involved in the assignment of items to categories and in the evaluation of inter-category similarities, we tested 18 *a priori* versions of the Categorical model, with different numbers of categories and different levels of hierarchical structure, for the concepts in Study 1 (Figure S1). Each version was tested via RSA against the fMRI data, and the best performing version was selected for comparisons against other types of models. The selected model for Study 1 (model N in Figure S1) consisted of 19 hierarchically structured categories: Abstract (Mental Abstract, Social Abstract, Social Event, Other Abstract), Event (Social Event, Concrete Event), Animate (Animal, Human, Body Part), Inanimate (Artifact [Musical Instrument, Vehicle, Other Artifact], Food, Other Inanimate), and Place.

In Study 2, the Categorical model consisted of 2 higher-level categories – Object and Event – each consisting of 4 sub-categories (Animal, Plant/Food, Tool, Vehicle; Sound Event, Social Event, Communication Event, Negative Event).

Experiential models. The *Exp48* model consists of 48 dimensions corresponding to distinct aspects of phenomenal experience, such as color, shape, manipulability, pleasantness, etc. (Table S2). This model is based on the experiential salience norms of Binder et al. (34), in which each dimension encodes the relative importance of an experiential attribute according to crowd-sourced ratings. Exp48 encompasses all perceptual, motor, spatial, temporal, causal, valuation, and valence dimensions present in those norms.

The *SM8* model consists of a subset of the Exp48 dimensions, focusing exclusively on sensory-motor information. These dimensions represent the relevance of each sensory modality (vision, audition, touch, taste, and smell) and of actions performed with each motor effector (hand, foot, and mouth) to the concept. The concept “apple”, for instance, has high values for vision, touch, taste, mouth actions, and hand actions, and low values for the other dimensions.

Distributional models. Distributional information was modeled with two of the most prominent distributional models available: *word2vec* (42) uses a deep neural network trained to predict a word based on a context window of a few words preceding and following the target word. In contrast, *GloVe* (43) is based on the *ratio* of co-occurrence probabilities between pairs of words across the entire corpus. In a comparative evaluation of distributional semantic models (12), word2vec and GloVe emerged as the two top performing models in predicting human behavior across a variety of semantic tasks.

Study 1 evaluated these six models with a set of 300 concepts spanning a wide variety of semantic categories, including animate objects (e.g., “elephant”, “student”), places (e.g., “kitchen”, “beach”), artifacts (e.g., “comb”, “rocket”) and other inanimate objects (e.g., “cloud”, “ice”), events (e.g., “hurricane”, “election”), and highly abstract concepts (e.g., “fate”, “hygiene”) (Figure 2A, Tables S3-S4).

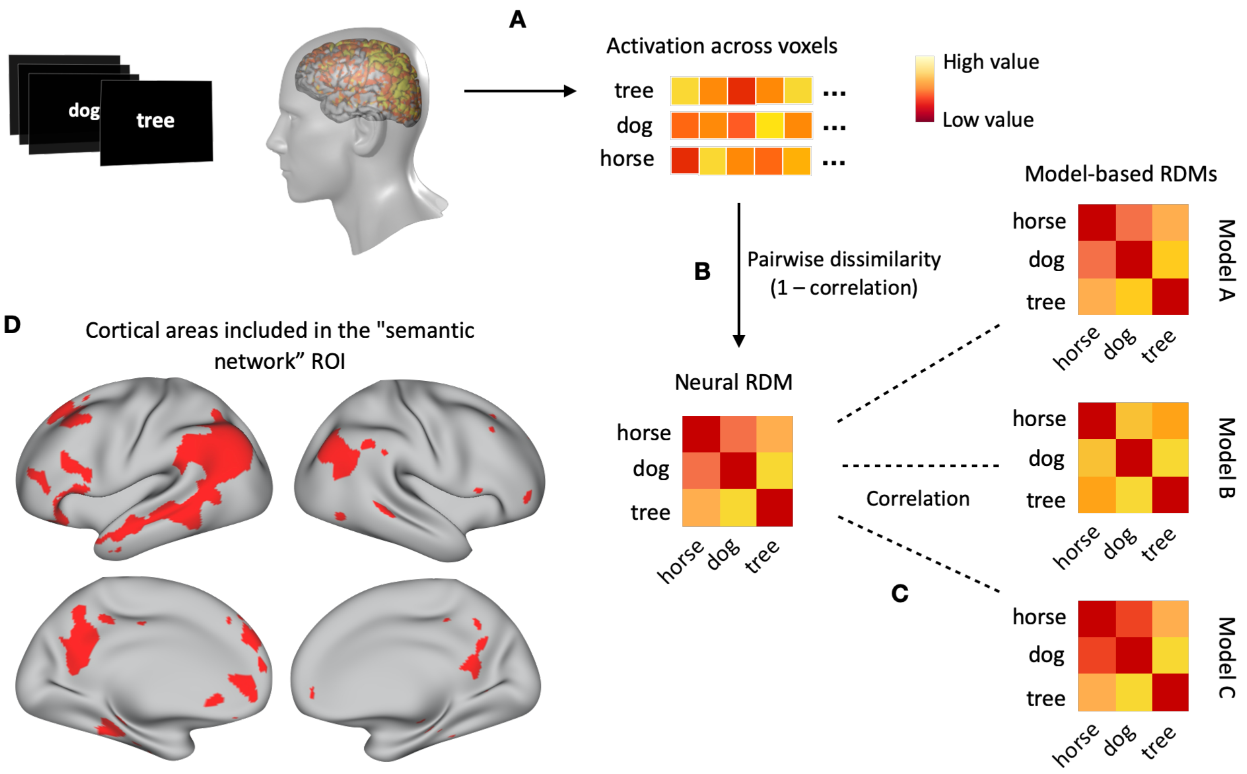

Figure 1: Representational similarity analysis. A. An fMRI activation map was generated for each concept presented in the study, and the activation across voxels was reshaped as a vector. B. The neural representational dissimilarity matrix (RDM) for the stimulus set was generated by computing the dissimilarity between these vectors (1 – correlation) for every pair of concepts. C. A model-based RDM was computed from each model, and the similarity between each model's RDM and the neural RDM was evaluated via Spearman correlation. D. Cortical areas included in the functionally defined semantic network ROI (36) (red).

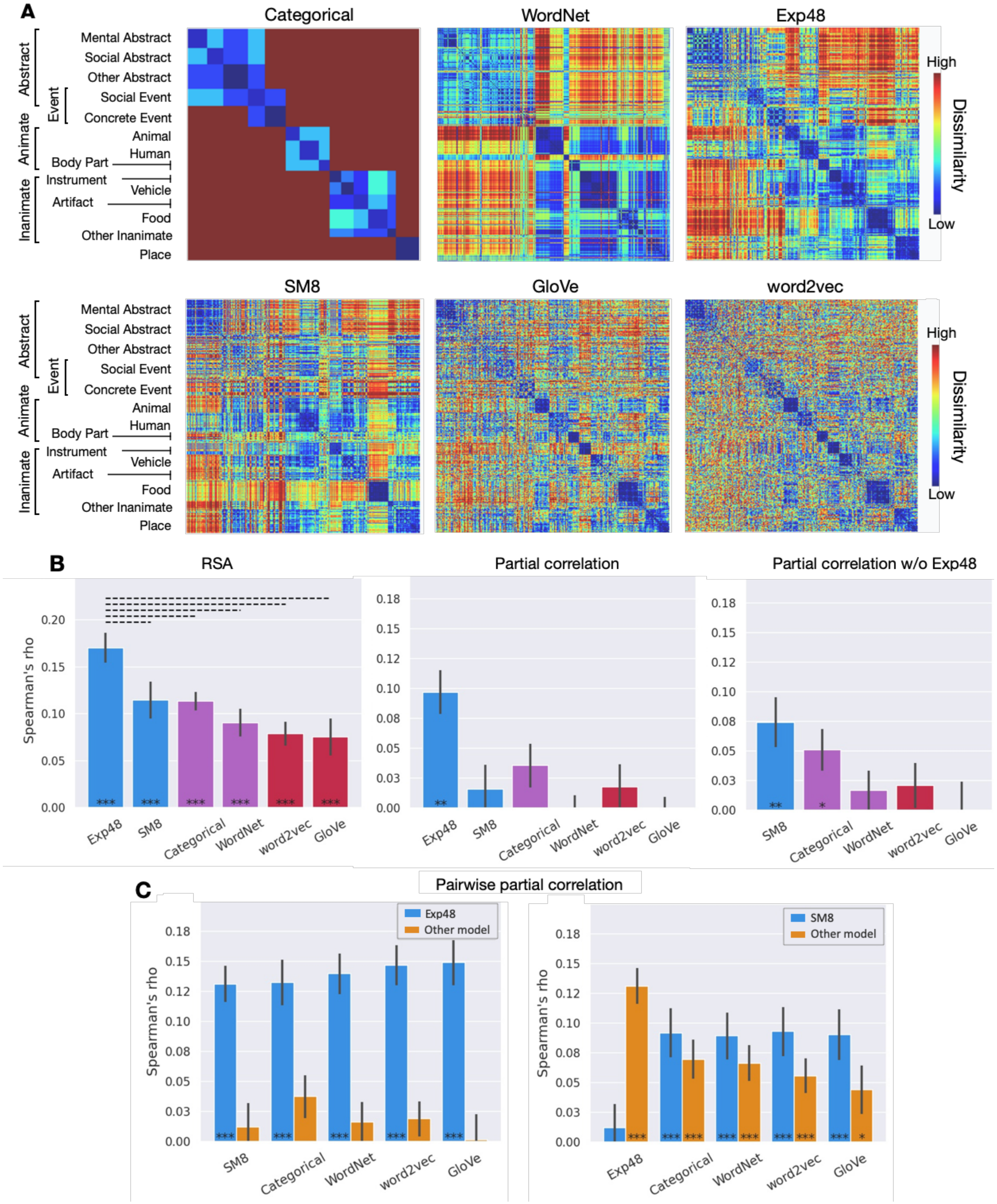
A
Categorical
WordNet
Exp48
SM8
GloVe
word2vec
Abstract
Event
Animate
Inanimate
Mental Abstract
Social Abstract
Other Abstract
Social Event
Concrete Event
Animal
Human
Body Part
Instrument
Vehicle
Artifact
Food
Other Inanimate
Place
High
Low
Dissimilarity
B
RSA
Partial correlation
Partial correlation w/o Exp48
Spearman's rho
0.00
0.05
0.10
0.15
0.20
0.03
0.08
0.12
0.18
Exp48
SM8
Categorical
WordNet
word2vec
GloVe
***
**
*
C
Pairwise partial correlation
Exp48
Other model
SM8
Other model

Figure 2: Study 1. A. Dissimilarity matrices for the representational spaces tested. Rows and columns represent each of the 300 concepts used in the study, grouped according to the Categorical model to reveal taxonomic structure. B. RSA results for the semantic network ROI. Experiential (blue), taxonomic (purple), and distributional (red) models; Left: Correlations between the group-averaged neural RDM and each model-based RDM. Center: Partial correlation results for each model while controlling for its similarity with all other models. Right: Partial correlation results when Exp48 was excluded from the analysis. C. Pairwise partial correlations for the semantic network ROI, with blue bars representing Exp48 (left) or SM8 (right) while controlling for its similarity to each of the other model-based RDMs; yellow bars correspond to each of the other model-based RDMs while controlling for their similarity to the model represented in blue. *** p < .0005, ** p < .005, * p < .05, Mantel test; solid bar: p < .001; dashed bar p < .05, permutation test. All p values are FDR-corrected for multiple comparisons (q = .05). Error bars represent the standard error.

Study 1

Study1 revealed that all six models predicted the similarity structure of concept-related neural activation patterns in left AG, PCN, SFG, and IFG (Figures S2 and S3). In all of these areas, as well as in left and right PCG, left PHG, right IFG, and right SFG, the Exp48 model achieved the strongest correlation. SM8 did not perform as well as Exp48, and this difference reached significance in several ROIs, particularly in the left IFG. The same areas also showed a significant advantage for Exp48 over WordNet. The difference between Exp48 and the Categorical model reached significance in left AG, PCN, PC, and PHG, and in right IFG and SFG. There were no significant differences between Categorical and WordNet, although there was a trend toward higher correlations for Categorical in most areas. Relative to other models, the Categorical model performed best in the left IFG, STG, MTG, and SMG, and in the right AG. Several areas showed a trend toward lower performance for word2vec and GloVe relative to the other models, and in most areas these two models performed similarly, with two exceptions: in left IFG, GloVe significantly outperformed word2vec, while the inverse effect was found in right PCN. Intriguingly, in none of the ROIs was Exp48 significantly outperformed by another model, although Categorical was the only model to reach significance in the left STG and the right cMFG.

The functionally defined semantic network ROI reflected the overall pattern observed across the other ROIs, with overall stronger correlations (Figure 2B, left). Exp48 performed significantly better than any other model, with no other differences between models reaching significance. To investigate the unique contribution of each type of information to the similarity structure of neural activation patterns, we conducted partial correlation RSAs. These analyses assessed how well each model explained the neural data after controlling for its similarity to other models. We conducted RSAs with the RDM of each model after regressing out the RDMs of all other models. As shown in Figure 2B (center), only Exp48 explained significant variance in the neural data after controlling for the other models. To verify whether Exp48 individually explains all of the variance accounted for by each of the other models, we performed pairwise partial correlation RSAs (Figure 2C, left). These analyses showed that no other model accounted for any measurable additional variance after controlling for their similarity to Exp48. In other words, Exp48 was the only model that made unique contributions to explaining the representational structure of lexical concepts in the semantic network ROI, while the other models only predicted the data to the extent that their similarity structure correlated with that of Exp48.

We then investigated whether information about the relative importance of eight sensory-motor modalities, by itself, successfully predicted the neural similarity structure of concepts after controlling for taxonomic and distributional information. The results showed that, when Exp48 was not included in the partial correlation analysis, SM8 remained highly significantly correlated with the data after controlling for all other models, while WordNet, word2vec and GloVe did not (Figure 2B, right). Categorical was the only other model displaying a significant partial correlation, although numerically lower and less significant than that of SM8. Pairwise partial correlation RSAs showed that SM8 did not explain all the variance accounted for by WordNet, word2vec, or Glove; however, there was a trend toward stronger correlations for SM8 than for any of the non-experiential models (Figure 2C, right). Together, these results suggest that experiential information – including but not restricted to sensory-motor information – plays a fundamental role in the neural representation of conceptual knowledge in heteromodal cortical areas, while taxonomic and distributional representational spaces may contribute little, if any, independent information.

Study 2

Study 2 was designed to further investigate the information content of concept representations with a different set of concepts, a larger participant sample size, and matched subsets of object and event concepts. This study examined whether the pattern of model performances found in Study 1 would be observed independently for objects and events, or – given their markedly distinct ontological status – whether these two types of concepts would differ in the degree to which they encode experiential, taxonomic, and distributional information. The 320 concepts included in the study were selected to be typical exemplars of four categories of objects (animals, tools, plants/foods, and vehicles) and four categories of events (sounds, negative events, social events, and communication events). The larger sample size (36 participants) allowed for statistical testing across participants as well as across stimuli (i.e., words).

In all ROIs, RSAs based on the entire concept set (i.e., all categories included) revealed an advantage for experiential models relative to taxonomic and distributional models (Figures S4 and S5). As in Study 1, correlations were particularly strong in the left IFG, AG, and PCN, with the IFG showing substantially stronger correlations than other ROIs. For all models, correlations were stronger in left hemisphere ROIs than in their right hemisphere homologs, consistent with left lateralization of lexical semantic representations. In almost all ROIs examined (28 out of 30), correlations were significantly stronger for the experiential models than for the taxonomic and distributional models. Nowhere did a taxonomic or distributional model perform significantly better than SM8 or Exp48. Exp48 significantly outperformed SM8 in left ATL, STG, PCG, AG, MTG, IFG, FusG, cMFG, and SFG; and in right PCG, AG, SMG, cMFG, and SFG. No significant differences between the two experiential models were found in the other ROIs.

In the semantic network ROI, Exp48 outperformed all other models when tested across participants and across words (Figures 3B, left and S6, left), and, in both analyses, it was the only model that retained explanatory power after controlling for the predictions of all other models (Figure 3B, center, and S6, center). SM8 also performed significantly better than all non-experiential models, and Categorical outperformed WordNet, word2vec, and GloVe.

Partial correlation RSAs revealed that Exp48 accounted for all the variance explained by the other models (Figure 3B, center and 3C, left). When Exp48 was left out of the analysis, SM8 and Categorical together accounted for all the variance explained (Figure 3B, right), with SM8 performing significantly better than Categorical. Confirming the main findings of Study 1, these results indicate that lexical concepts are encoded in high-level heteromodal cortical areas primarily in terms of experiential information.

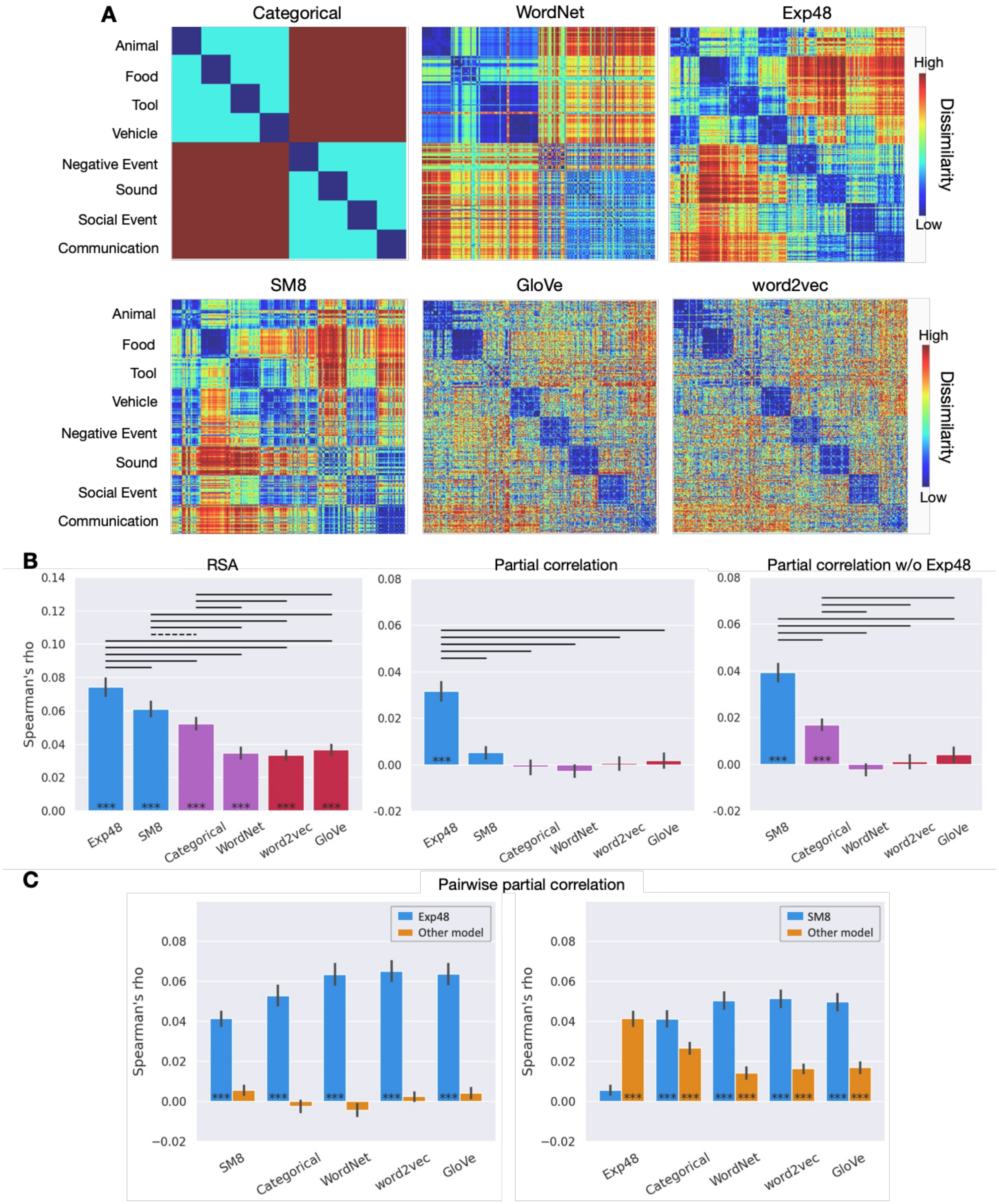
A
Categorical
WordNet
Exp48
Animal
Food
Tool
Vehicle
Negative Event
Sound
Social Event
Communication
High
Dissimilarity
Low
SM8
GloVe
word2vec
Animal
Food
Tool
Vehicle
Negative Event
Sound
Social Event
Communication
High
Dissimilarity
Low
B
RSA
Partial correlation
Partial correlation w/o Exp48
Spearman's rho
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
***
Exp48
SM8
Categorical
WordNet
word2vec
GloVe
0.08
0.06
0.04
0.02
0.00
-0.02
Exp48
SM8
Categorical
WordNet
word2vec
GloVe
0.08
0.06
0.04
0.02
0.00
-0.02
SM8
Categorical
WordNet
word2vec
GloVe
C
Pairwise partial correlation
Exp48
Other model
SM8
Other model
Spearman's rho
0.08
0.06
0.04
0.02
0.00
−0.02
SM8
Categorical
WordNet
word2vec
GloVe
Exp48
Categorical
WordNet
word2vec
GloVe

Figure 3. Study 2. A. Dissimilarity matrices for the representational spaces tested. Rows and columns have been grouped according to the Categorical model to reveal taxonomic structure. B. RSA results across participants for the semantic network ROI. Group mean values of the correlations between each participant's neural RDM and the model-based RDMs. Full RSA correlations (left) and partial correlations with Exp48 included (center) and excluded (right). C. Pairwise partial correlations, with blue bars representing Exp48 (left) or SM8 (right) while controlling for its similarity to each of the other models; yellow bars correspond to each of the other models while controlling for their similarity to the model represented in blue. *** p < .0005, ** p < .005, * p < .05; solid bar: p < .001; dashed bar p < .05; Wilcoxon signed-rank tests. All p values are FDR-corrected for multiple comparisons (q = .05). Error bars represent the standard error of the mean.

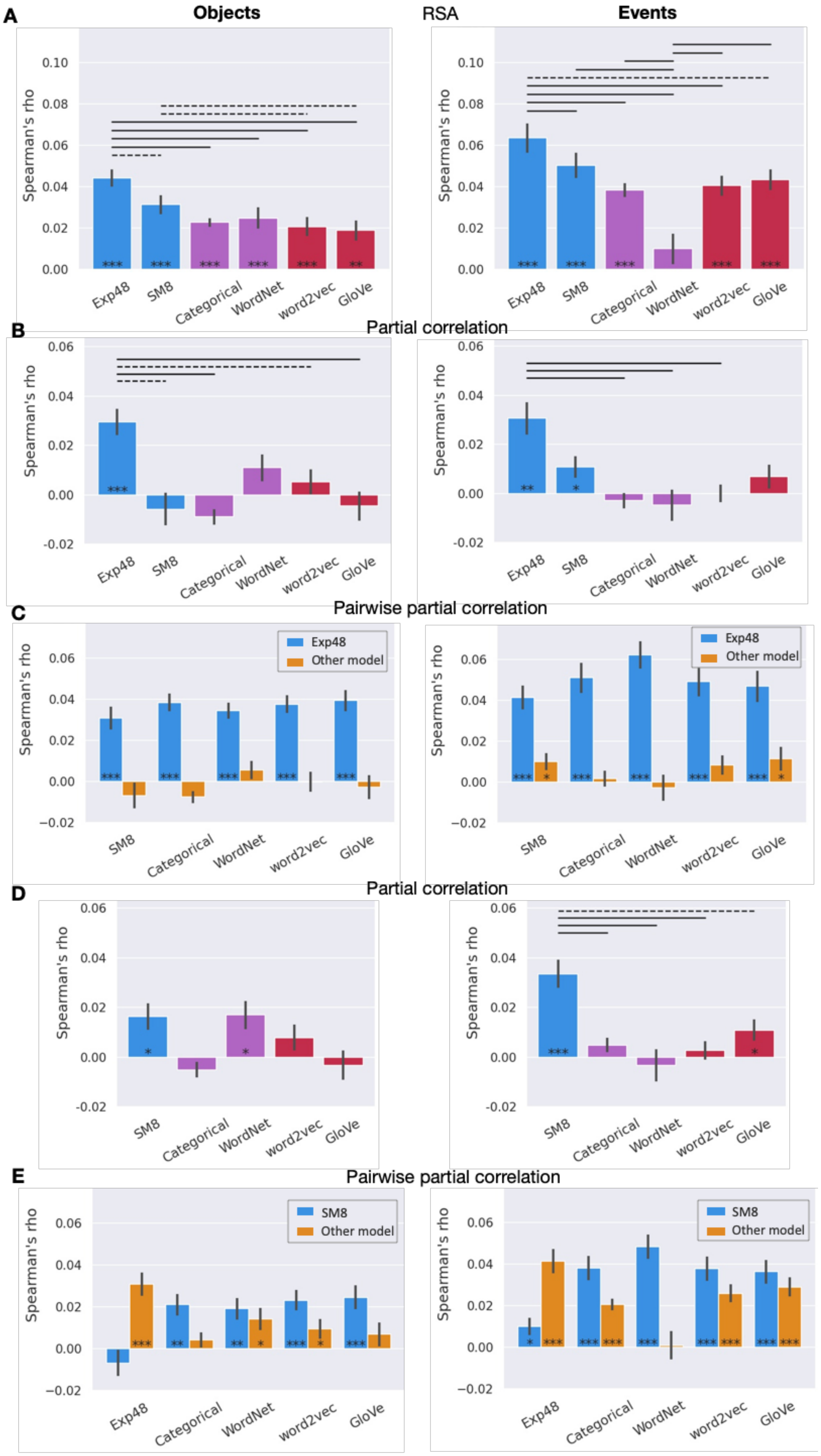
A
Objects
RSA
Events
Spearman's rho
0.10
0.08
0.06
0.04
0.02
0.00
Exp48
SM8
Categorical
WordNet
word2vec
GloVe
***
**
B
Partial correlation
0.06
0.04
0.02
0.00
-0.02
*
C
Pairwise partial correlation
Exp48
Other model
−0.02
D
Partial correlation
E
Pairwise partial correlation
SM8
Other model

Figure 4. Results for object concepts (left column) and event concepts (right column) for the semantic network ROI in Study 2 (across participants). A. RSA results. B. Partial correlations. C. Pairwise partial correlations for Exp48. D. Partial correlations with Exp48 excluded from the analysis. E. Pairwise partial correlations for SM8. Color and symbol conventions as in Figure 3.

Separate analyses for objects and events revealed two main differences between these two types of concepts (Figures 4, S7-S10). First, RSA correlations were substantially higher for events than for objects, across all three types of representational models, in almost all ROIs. This result reflects the higher variability of pairwise similarities for the neural representations of event concepts, as evidenced by the higher noise ceiling in this condition (mean noise ceiling across anatomical ROIs, lower bound = 0.17) relative to object concepts (0.14). Inspection of the neural RDM (Figure S11) reveals a more pronounced categorical structure for events than for objects, with particularly high pairwise similarities for communication events and social events.

Except for the left AG, WordNet appears to perform worse for events than for objects in all other ROIs, both in absolute terms and in relation to other models. This is likely due to taxonomic trees for events being relatively shallower in this model (mean tree depth for objects and events is 10.6 and 8.1, respectively), resulting in a narrower range of possible pairwise similarities. Apart from WordNet, the overall profile of relative model performances was similar across objects and events. For both types of concepts, experiential models achieved the strongest RSA correlations in most anatomical regions, with no ROIs showing significantly stronger correlations for any taxonomic or distributional model. Partial correlation RSAs in the semantic network ROI showed that, for object concepts, Exp48 accounted for all the variance explained by any other model (Figure 4, B and C). For events, SM8 and GloVe maintained relatively weak but statistically significant correlations after controlling for Exp48. SM8 was second to Exp48 in explaining the most variance for both objects and events (Figure 4, D and E).

Discussion

We conducted a quantitative assessment of the degree to which three different kinds of information are encoded in the neural representations of lexical concepts. We found that Exp48 –

a representational model based on 48 distinct experiential dimensions grounded in known neurocognitive systems – consistently outperformed taxonomic and distributional models in predicting the neural similarity structure of a large number of lexical concepts (524 across both studies), spanning a wide variety of semantic categories. Additionally, a model based on eight sensory-motor dimensions, reflecting the relative importance of neural systems dedicated to vision, hearing, touch, taste, and smell, as well as motor control of the hands, mouth, and feet, matched or exceeded the performance of the best performing taxonomic and distributional models in all ROIs associated with concept representation (except for the left IFG in Study 1, where the Categorical model performed better). These results provide compelling evidence that experiential information grounded in sensory, motor, spatial, temporal, affective, and reward neural systems plays a fundamental role in the neural representation of concepts – not only in sensory-motor cortical areas, as shown in previous studies – but also in high level association areas.

In both experiments, Exp48 was the only model whose predictions correlated with the neural similarity structure of the entire concept set when controlling for the predictions of all other models, and in Study 2 this was also true for object concepts in particular. For event concepts, SM8 was the only other model that independently accounted for variance in the neural data. The fact that taxonomic and distributional models did not predict the neural RDM when controlling for the predictions of Exp48 alone (Figures 2C and 3C, left panel) indicates that those models provide no measurable additional information about concept similarity structure. This suggests that taxonomic information, while seemingly central to the organization of conceptual knowledge, may be represented only indirectly, via its interdependency with experiential information (34).

One limitation of the present study is that the representational spaces used to model experiential information content are relatively coarse. Neural concept representations must encode detailed information about aspects of the phenomenal experience associated with different lexical concepts, such as particular colors, shapes, or motor schemas, but Exp48 and SM8 only encode the relative importance of each kind of experience. In other words, these models represent information about how much the neural system underlying a particular process (e.g., color perception, motor control of the hand) contributes to a lexical concept representation, but it contains no information about the particular representations contributed by each system (e.g., different hues or different motor schemas). This coarseness, which results from practical limitations in obtaining participant ratings about fine-grained experiential attributes, imposes limitations on our ability to detect experiential information in the neural data. Nevertheless, it is still surprising that these models performed so well compared to some of the most sophisticated taxonomic and distributional models available, pointing to the development of more detailed experiential models as a promising direction for future work.

Together, the present results indicate that concept representations in heteromodal cortical areas primarily encode information about features of phenomenal experience grounded in sensory-motor, spatiotemporal, affective, and reward systems. While other studies have shown that areas involved in sensory perception and motor control are activated during concept retrieval (8, 22, 24), our results imply that information pertaining to these systems is directly encoded in high-level association areas during concept retrieval. The idea that concepts are neurally represented as amodal symbols whose representational code is independent of modality-specific systems has a long history (18, 31); however, the finding that taxonomic and distributional information did not make any measurable independent contribution to the neural representation of concepts in high-level cortical areas challenges that view. Instead, it supports a view in which concept representations emerge from the multimodal integration of signals originating in modality-specific systems during concept acquisition (9, 24). In this framework, concept retrieval consists of the transient activation of a neuronal cell assembly distributed across the heteromodal cortical hubs that make up the semantic network, with possible downstream activation of the relevant modality-specific assemblies depending on contextual demands. Further research is required to determine the extent to which different experiential features contribute to concept representation in different parts of the semantic network, and how their relative importance varies across ontological categories.

In addition to their implications for the nature of the neural code underlying conceptual knowledge, the present results are also relevant to computational approaches to natural language processing in the field of artificial intelligence. They suggest that incorporating experiential information into computational models of word meaning could lead to more human-like performance. These findings can also inform the development of brain-machine interface systems by demonstrating that information about the experiential content of conceptual thought can be decoded from the spatial distribution of neural activity throughout the cortex.

Materials and Methods

Study 1

Participants: 8 native speakers of English (4 women), ages 19-37 (mean = 28.5) took part in Study 1. Participants were all right-handed according to the Edinburgh Handedness Scale (44), had at least a high school education, and had no history of neurological or psychiatric conditions. All participants provided written informed consent following procedures approved by the Medical College of Wisconsin Institutional Review Board.

Stimuli: Stimuli included 300 English nouns of various categories, including concrete and abstract concepts. Nouns were relatively familiar, with mean log-transformed HAL frequency of 8.7 (range 4.0–12.7) and mean age of acquisition of 6.7 years (range 2.7–13.8; English Lexicon Project (https://elexicon.wustl.edu) (45); Tables S3-S4).

Task: Participants rated each noun according to how often they encountered the corresponding entity or event in their daily lives, on a scale from 1 (“rarely or never”) to 3 (“often”). The task was designed to encourage semantic processing of the word stimuli without emphasizing any particular semantic features or dimensions. Participants indicated their response by pressing one of three buttons on a response pad with their right hand. On each trial, a noun was displayed in written form on the center of the screen for 500 ms, followed by a 3.5-second blank screen. Each trial was followed by a central fixation cross with variable duration between 1 and 4 s (mean = 2 s). The entire stimulus set was presented 6 times over the course of the study. In each presentation, the order of the stimuli was randomized. Each presentation was split into 3 runs of 100 trials each. The task was performed over the course of 3 scanning sessions on separate days, with 2 presentations (6 runs) per session. Each run started and ended with an 8-second fixation cross.

Equipment: Scanning was performed on a GE Healthcare Discovery MR750 3T MRI scanner at the Medical College of Wisconsin’s Center for Imaging Research. Stimulus presentation and response recording were performed via E-prime 2.0 software running on a Windows desktop computer and a Psychology Software Tools Serial Response Box. Stimuli were back-projected on a screen positioned behind the scanner bore and viewed through a mirror attached to the head coil.

Scanning protocol: MRI scanning was conducted over 3 separate sessions on different days. Each session consisted of a structural T1-weighted MPRAGE scan, a structural T2-weighted CUBE scan, 3 pairs of T2-weighted spin echo echo-planar scans (5 volumes each) acquired with opposing phase-encoding directions (for correction of geometrical distortions in the functional images), and 6 gradient echo echo-planar runs using simultaneous multi-slice acquisition (8x multiband, TR = 1200 ms, TE = 33.5 ms, 512 volumes, flip angle = 65, matrix = 104 x 104, slice thickness = 2.0 mm, axial acquisition, 72 slices, field-of-view = 208 mm, voxel size = 2 x 2 x 2 mm).

Data analysis: Functional MRI images were pre-processed (slice timing correction, motion correction, distortion correction, volume alignment) using AFNI (46). Statistical analyses were conducted in each participant's original coordinate space. Activation (beta) maps were generated for each noun relative to the mean signal across all other nouns using AFNI's 3dDeconvolve. Response time z scores and head motion parameters were included as regressors of no interest. Anatomically defined ROIs were created based on the probabilistic Desikan-Killiany parcellation atlas included in AFNI (TT_desai_dk_mpm). The ATL ROI was created by the union of STG, MTG, ITG, temporal pole, and entorhinal cortex ROIs, cropped to exclude voxels posterior to y = -2 (37). The functionally defined semantic network ROI corresponded to the 1% most consistently activated voxels in an activation likelihood estimate (ALE) meta-analysis of 120 neuroimaging studies of semantic language processing (36) (main text Figure 1D). All ROI masks were created in MNI coordinate space and non-linearly aligned to each participant's anatomical scan using AFNI's 3dQwarp. Neural RDMs were computed for each ROI as the Spearman correlation distance (1 – rho) for all pairs of concepts. A group-averaged neural RDM was computed by averaging the neural RDMs of all participants. A model-based RDM was computed from each model (except WordNet) as the cosine distance (1 – cosine) for all pairs of concept vectors. The WordNet RDM was based on the Wu & Palmer similarity metric (WPsim), which relies on the depth of the two synsets in the taxonomic tree and that of their Least Common Subsumer (LCS, i.e., their most specific common hypernym):

$$\text{WPsim} = 2 \times \frac{\text{depth}(\text{LCS}(\text{s1, s2}))}{(\text{depth}(\text{s1}) + \text{depth}(\text{s2}))}$$

We used the package Natural Language Toolkit (NLTK 3.4.5; https://www.nltk.org) to compute WPsim; WordNet dissimilarity was computed as 1 – WPsim.

The RSAs computed the Spearman correlation between the group-averaged neural RDMs and each model-based RDM. Statistical significance was tested using the Mantel test with 10,000 permutations. Differences in RSA correlations between models were assessed for significance via permutation tests. Significance was controlled for multiple comparisons via false discovery rate (FDR) with q = .05. Since the model-based RDMs can be strongly correlated between models (Table S1), these analyses were supplemented with partial correlation analyses to reveal the unique contribution of each model after controlling for other models.

Study 2

Participants: 36 native speakers of English (21 women), ages 20-41 (mean = 28.7) took part in Study 2. Participants were all right-handed according to the Edinburgh Handedness Scale (44), had at least a high school education, and had no history of neurological or psychiatric conditions. None of the participants took part in Study 1.

Stimuli: Stimuli included 160 object nouns (animals, foods, tools, and vehicles – 40 of each) and 160 event nouns (social events, verbal communication events, non-verbal sound events, negative events – 40 of each) (Tables S5-S6). Of the 320 concepts included in Study 2, 62 objects and 34 events were also used in Study 1.

Task: Task instructions were identical to Study 1. On each trial, a noun was displayed in written form on the center of the screen for 500 ms, followed by a 2.5-second blank screen. Each trial was followed by a central fixation cross with variable duration between 1 and 3 s (mean = 1.5 s). The entire stimulus set was presented 6 times over the course of the study in randomized order. Each presentation was split into 4 runs of 80 trials each. The task was performed over the course of 3 scanning sessions on separate days, with 2 presentations (8 runs) per session. Each run started and ended with an 8-second fixation cross.

Equipment: Scanning was performed on a GE Healthcare Premier 3T MRI scanner at the Medical College of Wisconsin's Center for Imaging Research. Stimulus presentation and response recording were performed with Psychopy 3 software (47) running on a Windows desktop computer and a Celeritas fiber optic response system (Psychology Software Tools, Inc.). Stimuli were displayed on an MRI-compatible LCD screen positioned behind the scanner bore and viewed through a mirror attached to the head coil.

Scanning protocol: MRI scanning was conducted on 3 separate visits. Each session consisted of a structural T1-weighted MPRAGE scan, a structural T2-weighted CUBE scan, 3 pairs of T2-weighted spin echo echo-planar scans (5 volumes each) acquired with opposing phase-encoding directions, and 8 gradient echo echo-planar functional scans (4x multiband, TR = 1500 ms, TE = 23 ms, 512 volumes, flip angle = 50, matrix = 104 x 104, slice thickness = 2.0 mm, axial acquisition, 72 slices, field-of-view = 208 mm, voxel size = 2 x 2 x 2 mm).

Data analysis: Data preprocessing and statistical analysis were performed as described in Study 1. RDMs for the models tested in Study 2 are depicted in Figure 3 (main text), and inter-model correlations are displayed in Table S1.

RSAs were conducted in two ways. In the analysis across stimuli (i.e., words), voxel-based RDMs from all participants were combined into a group-averaged neural RDM, and, for each model, a single correlation was computed between this RDM and the model-based RDM. These correlations were tested for significance via the Mantel test with 10,000 permutations. In the analysis across participants, a correlation was computed, for each model, between each participant's neural RDM and the model-based RDM. The Fisher Z-transformed correlation coefficients were then averaged across participants to compute the group mean rho, and significance tested via Wilcoxon's Signed Rank test.

Differences in RSA correlations between models were tested across words via permutation test (Mantel) and across participants via Wilcoxon's Signed-Rank test. Significance was controlled for multiple comparisons via false discovery rate (FDR) with q = .05.

Acknowledgments: The authors wish to thank Elizabeth Awe for assistance with data collection and management, and Jia-Qin Tong for assistance with preprocessing of fMRI data. **Funding:** This project was supported by grants NIH/NIDCD R01-DC016622 to J.R.B. and AHW Project #5520462 to Brian-Fred Fitzsimmons. **Author contributions:** L.F. conceived the study, coordinated data acquisition, analyzed the data, prepared figures and tables, and wrote the initial draft; C.J.H. contributed software tools for data acquisition; J.R.B. and L.L.C. developed the stimulus set and reviewed/edited the manuscript; J.R.B supervised the project and acquired financial support. **Competing interests:** Authors declare no competing interests.

References:

1. W. D. Ross, *Plato's Theory of Ideas* (Clarendon Press, 1951).
2. E. E. Smith, D. L. Medin, *Categories and Concepts* (Harvard University Press, 2013).
3. C. B. Mervis, E. Rosch, Categorization of natural objects. *Annu. Rev. Psychol.* **32**, 89–115 (1981).
4. A. M. Collins, E. F. Loftus, A spreading-activation theory of semantic processing. *Psychol. Rev.* **82**, 407–428 (1975).
5. J. R. Anderson, A spreading activation theory of memory. *J. Verbal Learning Verbal Behav.* **22**, 261–295 (1983).
6. A. R. Damasio, Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. *Cognition* **33**, 25–62 (1989).
7. A. M. Glenberg, What memory is for. *Behav Brain Sci* **20**, 1–55 (1997).
8. L. W. Barsalou, W. K. Simmons, A. Barbey, C. Wilson, Grounding conceptual knowledge in modality-specific systems. *Trends Cogn. Sci.* **7**, 84–91 (2003).
9. J. R. Binder, R. H. Desai, The neurobiology of semantic memory. *Trends Cogn. Sci.* **15**, 527–536 (2011).
10. T. K. Landauer, S. T. Dumais, A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* **104**, 211–240 (1997).
11. M. N. Jones, W. Kintsch, D. J. K. Mewhort, High-dimensional semantic space accounts of priming. *J. Mem. Lang.* **55**, 534–552 (2006).

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

13. M. Andrews, S. Frank, G. Vigliocco, Reconciling embodied and distributional accounts of meaning in language. *Top. Cogn. Sci.* **6**, 359–370 (2014).

14. E. M. Markman, Constraints children place on word meanings. *Cogn. Sci.* **14**, 57–77 (1990).

15. M. Lucas, Semantic priming without association: a meta-analytic review. *Psychon Bull Rev* **7**, 618–630 (2000).

16. A. Caramazza, J. R. Shelton, Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *J. Cogn. Neurosci.* **10**, 1–34 (1998).

17. C. E. Crangle, M. Perreau-Guimaraes, P. Suppes, Structural similarities between brain and linguistic data provide evidence of semantic relations in the brain. *PLoS One* **8**, e65366--16 (2013).

18. G. Handjaras, *et al.*, How concepts are encoded in the human brain: A modality independent, category-based cortical organization of semantic knowledge. *Neuroimage* **135**, 232–242 (2016).

19. Y. Xu, *et al.*, Doctor, Teacher, and Stethoscope: Neural Representation of Different Types of Semantic Relations. *J. Neurosci.* **38**, 3303–3317 (2018).

20. M. H. Fischer, R. a Zwaan, Embodied language: a review of the role of the motor system in language comprehension. *Q. J. Exp. Psychol. (Hove).* **61**, 825–850 (2008).

21. E. K. Warrington, R. A. McCarthy, Categories of knowledge: Further fractionations and

an attempted integration. *Brain* **110**, 1273–96 (1987).

22. M. Kiefer, F. Pulvermüller, Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex* **48**, 805–825 (2011).

23. L. Fernandino, *et al.*, Parkinson's disease disrupts both automatic and controlled processing of action verbs. *Brain Lang.* **127**, 1–10 (2013).

24. L. Fernandino, *et al.*, Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cereb. Cortex* **26**, 2018–2034 (2016).

25. L. Fernandino, *et al.*, Predicting brain activation patterns associated with individual lexical concepts based on five sensory-motor attributes. *Neuropsychologia* **76**, 17–26 (2015).

26. M. N. Jones, D. J. K. Mewhort, Representing word meaning and order information in a composite holographic lexicon. *Psychol. Rev.* **114**, 1–37 (2007).

27. F. Günther, C. Dudschig, B. Kaup, Predicting lexical priming effects from distributional semantic similarities: A replication with extension. *Front. Psychol.* **7**, 313–339 (2016).

28. A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, J. L. Gallant, Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).

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

30. F. Pereira, *et al.*, Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.*, 1–13 (2018).

31. B. Z. Mahon, A. Caramazza, A critical look at the embodied cognition hypothesis and a

new proposal for grounding conceptual content. *J Physiol Paris* **102**, 59–70 (2008).

32. R. A. Zwaan, Embodiment and language comprehension: Reframing the discussion. *Trends Cogn. Sci.* **18**, 229–234 (2014).

33. B. C. Malt, E. E. Smith, Correlated properties in natural categories. *J. Verbal Learning Verbal Behav.* (1984).

34. J. R. Binder, *et al.*, Toward a brain-based componential semantic representation. *Cogn. Neuropsychol.* **33**, 130–174 (2016).

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

36. J. R. Binder, R. H. Desai, W. W. Graves, L. L. Conant, Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* **19**, 2767–2796 (2009).

37. M. Visser, E. Jefferies, M. A. L. Ralph, Semantic processing in the anterior temporal lobes: a meta-analysis of the functional neuroimaging literature. *J. Cogn. Neurosci.* **22**, 1083–1094 (2010).

38. J. Wang, J. A. Conder, D. N. Blitzer, S. V Shinkareva, Neural representation of abstract and concrete concepts: a meta-analysis of neuroimaging studies. *Hum. Brain Mapp.* **31**, 1459–1468 (2010).

39. R. S. Desikan, *et al.*, An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968–980 (2006).

40. G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller, Introduction to WordNet: An on-line lexical database. *Int. J. Lexicogr.* **3**, 235–244 (1990).

41. A. J. Anderson, B. Murphy, M. Poesio, Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. *J. Cogn. Neurosci.* **26**, 658–681 (2014).

42. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space in *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, (2013).

43. J. Pennington, R. Socher, C. D. Manning, GloVe: Global vectors for word representation in *Empirical Methods in Natural Language Processing (EMNLP)*, (2014), pp. 1532–1543.

44. R. C. Oldfield, The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* **9**, 97–113 (1971).

45. D. A. Balota, *et al.*, The English Lexicon Project. *Behav. Res. Methods* **39**, 445–459 (2007).

46. R. W. Cox, AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162–173 (1996).

47. J. W. Peirce, PsychoPy-Psychophysics software in Python. *J. Neurosci. Methods* **162**, 8–13 (2007).