COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13336

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13336

Items Outperform Adjectives in a Computational Model of Binary Semantic Classification

Evgeniia Diachek,[a] Sarah Brown-Schmidt,[a] Sean M. Polyn[b]

[a] *Department of Psychology and Human Development, Peabody College, Vanderbilt University*
[b] *Department of Psychology, College of Arts and Sciences, Vanderbilt University*

Received 11 November 2022; received in revised form 3 August 2023; accepted 23 August 2023

Abstract

Semantic memory encompasses one's knowledge about the world. Distributional semantic models, which construct vector spaces with embedded words, are a proposed framework for understanding the representational structure of human semantic knowledge. Unlike some classic semantic models, distributional semantic models lack a mechanism for specifying the properties of concepts, which raises questions regarding their utility for a general theory of semantic knowledge. Here, we develop a computational model of a binary semantic classification task, in which participants judged target words for the referent's size or animacy. We created a family of models, evaluating multiple distributional semantic models, and mechanisms for performing the classification. The most successful model constructed two composite representations for each extreme of the decision axis (e.g., one averaging together representations of characteristically big things and another of characteristically small things). Next, the target item was compared to each composite representation, allowing the model to classify more than 1,500 words with human-range performance and to predict response times. We propose that when making a decision on a binary semantic classification task, humans use task prompts to retrieve instances representative of the extremes on that semantic dimension and compare the probe to those instances. This proposal is consistent with the principles of the instance theory of semantic memory.

Keywords: Binary semantic judgment; Semantic projection; Distributional semantic models; Linear ballistic accumulator

Correspondence should be sent to Evgeniia Diachek, Department of Psychology and Human Development, Peabody College, Vanderbilt University, 230 Appleton Place, Hobbs 302, Nashville, TN 37212, USA. E-mail: evgeniia.diachek@gmail.com

1. Introduction

Human semantic memory is complex, encompassing one's knowledge about the world and the things in it. Characterizing the computations underlying the access and manipulation of semantic memory has been a central question in the field of cognitive science for decades (Barsalou, 2008; Binder, Desai, Graves, & Conant, 2009; Cree & McRae, 2003; Collins & Loftus, 1975; Collins & Quillian, 1969; Fodor, 1998; Jackendoff, 1992; Lambon Ralph & Patterson, 2008; Laurence & Margolis, 1999; Mahon & Caramazza, 2003; McRae, De Sa, & Seidenberg, 1997; Osgood, 1952; Saffran & Schwartz, 1994; Tulving, 1972;). Technological advances in the late 1990s offered a new powerful computational method of quantifying the meaning of words through ample text corpus data, creating a new class of distributional semantic models (DSMs). DSMs, such as latent semantic analysis (LSA; Landauer & Dumais, 1997), word2vec (Mikolov, Chen, Corrado, & Dean, 2013), and Global Vectors (GloVe; Pennington, Socher, & Manning, 2014) conceptualize semantic representations as vectors residing in a high-dimensional vector space. These models are based in part on the assumption that the meaning of a word is reflected in the pattern of its usage, namely, that words with similar or related meanings tend to occur in similar contexts (Harris, 1954; Firth, 1957). According to this idea, words like *virus*, *mask*, and *vaccine* tend to occur in proximity to each other (e.g., in the same sentence, paragraph, or document) because the meanings denoted by the words are semantically associated. In contrast, the words *virus* and *flowers* do not tend to occur in similar contexts, suggesting that the meanings associated with the words have little to no semantic association or similarity.

DSMs have been incorporated into a variety of cognitive models of semantic memory, predicting human performance on a variety of tasks including the Test of English as a Foreign Language (TOEFL) synonym task (Landauer & Dumais, 1997), word analogies (Mikolov et al., 2013), concept naming (Pennington et al., 2014), free recall (Morton & Polyn, 2016), feature generation (Cutler, Duff, & Polyn, 2019), the remote associates test (Smith, Huber, & Vul, 2013), the preferential decision-making task (Bhatia, 2019; Bhatia, Richie, & Zou, 2019), semantic fluency (Hills, Jones, & Todd, 2012), and binary semantic classification (Grand, Blank, Pereira, & Fedorenko, 2022). To model behavior in any one of these tasks, the semantic representational structure captured by the DSMs must be integrated with cognitive mechanisms that make use of it. For example, on the TOEFL synonyms task, participants are presented with a target word and several choices. The participant's task is to identify the synonym among the alternatives. In this example, the model's cognitive machinery is relatively simple—the algorithm calculates the cosine similarity of the target word to each choice word and picks the word with the greatest similarity to the target among the alternatives (Landauer & Dumais, 1997).

Despite the success of DSMs, challenges arise in broadly incorporating them into cognitive models of semantic tasks. While many of the tasks considered above involve evaluating words in terms of their similarity, problems arise with tasks involving the evaluation of specific properties of the words in question, since the dimensions of the semantic space are not necessarily meaningful. In other words, the proximity of the words within the representational space indicates semantic relatedness but not the nature of the relation (see Hill et al., 2015, for a deeper exploration of the distinction between relatedness and similarity). For example, many DSMs would perform well on identifying the oddball among the words *flower*, *garden*, and *vehicle*. However, it is unclear how they could identify which properties of a vehicle make it the oddball. These limitations are not true of all models of semantic memory. For example, the graph-theoretic semantic models of Collins, Quillian, and Loftus (Collins & Loftus, 1975; Collins & Quillian, 1969) overcome this issue by incorporating labeled links that specify the relationship between the properties of concepts. For example, the node *canary* is linked to other nodes *animal*, *yellow*, *beak*, and *fly* by the respective links *isa*, *is, hasa*, and *can*. Similarly, Rumelhart, McClelland, and the Parallel Distributed Processing (PDP) group (1986) developed connectionist models of semantic knowledge that are explicitly trained to store and retrieve item properties, for example, if *bird* and *hasa* are activated, *beak* is retrieved (McClelland & Rogers, 2003). Finally, Smith et al.'s (1974) featural model specifies that concepts have an associated list of features that can be queried to determine properties of the concept. While these classic models offer information about the properties of items, and the relationships between concepts, these relations have been experimenter coded, and we are unaware of any current technology that can generally automate this process. DSMs, on the other hand, offer a substantial advantage in terms of their scale (e.g., millions of words in word2vec vs. hundreds of concepts in a norming study by McRae, Cree, Seidenberg, & McNorgan, 2005), but lack specificity regarding the nature of the relations between concepts.

In a recent study, Grand et al. (2022) addressed this problem. Using a method similar to Osgood's semantic differential technique (Osgood, 1952; Osgood et al., 1957), Grand et al. (2022) collected human ratings evaluating words in terms of a variety of semantic dimensions (e.g., size, danger, gender, intelligence). For example, to evaluate the target word *elephant* on the size dimension, the words *small* and *large* were linked to the extremes of a 5-point scale, and the participant selected which number best went with the target word. They proposed a computational model which used distributional semantic representations to simulate these simple binary decisions about the characteristics of real-world objects on the semantic dimensions examined with the human ratings. Their model uses an average of three synonymous adjective labels assigned to each of the two extremes of the semantic dimension to construct a semantic axis in the representational space of the DSM. In other words, to make a *size* judgment, the vector representations of {*large, huge, big*} and {*small, little, tiny*} are retrieved and averaged together to create two semantic composite representations. By subtracting one of these semantic composites from the other, a difference vector is created, and this can be treated as a semantic axis in the representational space. A judgment is made by projecting a given word vector onto the semantic axis and calculating which extreme it is closer to (Fig. 1). We refer to this as an adjective-composite model of binary semantic classification. Grand et al. (2022) demonstrated the utility and flexibility of this semantic projection model, which was able to capture approximately 0.37 of variability in human ratings on a set of semantic classification tasks.

Overall, Grand et al. (2022) established that detailed, context-dependent conceptual knowledge can be flexibly extracted from the representational space of a DSM. They demonstrated the cognitive utility of the adjective-composite model but did not specifically propose it as a cognitive model of human performance in the binary semantic classification task. Rather, they

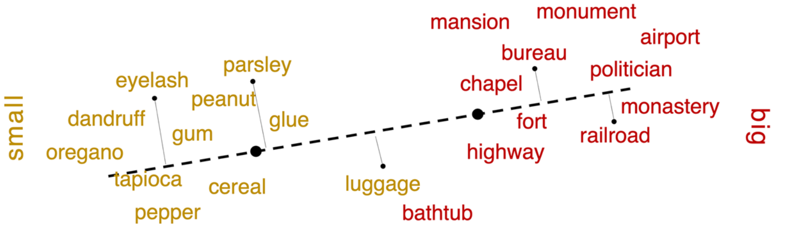

Fig. 1. Schematic depiction of a semantic projection. Each semantic model in this paper constructs a semantic axis as a difference vector (dashed line) by subtracting one semantic composite from another. Here we depict a semantic axis for a size judgment, using the item-composite semantic construction procedure. The two extremes of the *big–small* axis are computed as averages of the words in the dataset unanimously judged as *big* or *small* by all participants. The position of a projected word on the semantic axis is calculated with a dot product operation. The resulting value (a *dot-product score*) is used as evidence to determine the predicted probability for each response (*big* or *small*).

established that word embedding is constructed based on co-occurrence statistics that contain rich information capable of guiding flexible semantic classification judgments.

In the current work, we evaluate the adjective-composite model as a potential cognitive mechanism involved in binary semantic classification and compare it with an alternative item-composite model. This alternative model proposes that while performing a binary semantic classification task, participants use each adjective label of the judgment (i.e., *big* and *small)* to retrieve a set of items representative of each extreme. In other words, the cognitive system retrieves a set of vectors corresponding to big things and another set of vectors of small things. Each of these sets is blended to create a composite semantic representation of that extreme of the semantic decision axis (see Fig. 1). As with the adjective-composite model, the judgment is made by projecting a given word vector onto this semantic decision axis and calculating which extreme it is closer to.

In order to compare these two models, we present a likelihood-based computational modeling framework for the binary semantic classification task. The framework allows us to contrast different semantic projection mechanisms in terms of their ability to predict human task performance. The framework consists of three parts: a DSM to define the representational space (word2vec or GloVe), a semantic evaluation algorithm (adjective-composite or item-composite), and a decision mechanism (a logistic decision rule or linear ballistic accumulators [LBA]). We examine two DSMs to establish that the simulation results are robust and are not dependent on the exact distributional model used. Each DSM constructs a set of word vectors from word co-occurrence statistics in a large natural language corpus. The word2vec model uses a neural network algorithm to construct its vectors (Mikolov et al., 2013), and the GloVe model extracts the vectors more directly from the global corpus statistics (Pennington et al., 2014). Both decision mechanisms allow us to evaluate model performance in terms of response probabilities, and the LBA model additionally allows us to examine response latencies. These models are evaluated with respect to how well they can predict human performance in a large dataset with two semantic classification tasks (size and animacy), 42 participants, 1,650 unique target words, and 47,520 unique responses. Results from a second large dataset are presented in the Supporting Information and are consistent with all of the results presented in the main paper.

The cognitive mechanisms proposed for the semantic evaluation algorithms can be described in terms of prominent instance-theoretic cognitive models (Jamieson, Avery, Johns, & Jones, 2018). The judgment labels associated with the classification task (i.e., *big*, *small*) can be thought of as retrieval cues that prompt the retrieval of semantic representations used to guide task performance. The Grand et al. mechanism retrieves composites of the adjective labels of the decision axes, while our novel mechanism retrieves composites of the *items* associated with those labels. This is reminiscent of the memory retrieval mechanisms proposed in Hintzman's seminal work with MINERVA 2 (Hintzman, 1984, 1986, 1988), in which a memory store composed of many instance-based traces of past experience can be flexibly probed to reactivate composite representations. Such machinery was used to great effect in the recently proposed instance theory of semantic memory (ITS; Jamieson et al., 2018), which treats multiple instances of a word's usage in natural language as independent traces in memory. This allows ITS to, among other things, interpret homonyms correctly by flexibly constructing a representation of the word's meaning on the basis of the word's current context.

The novel item-composite semantic mechanism retrieves representative members of a given judgment class (i.e., big things) on the basis of the properties of those items. We note that the representational structure of a DSM does not support direct targeting of vectors on the basis of specific properties. As such, we take inspiration from the property-based semantic models described above and consider the possibility that a secondary system allows the participant to directly target item representations known to possess that characteristic. This allows the model to retrieve big things, small things, living things, and non-living things directly when it constructs its composite item representations, and then to evaluate new items in terms of their similarity to these composites. We demonstrate that these item composites allow the model to successfully predict performance for many items that are not part of the composite and that the model is successful even when only a few items are used to construct each composite. In the discussion, we revisit the question of whether a secondary property-knowledge system is strictly necessary, or whether a modified DSM could potentially account for human performance.

2. Methods

The behavioral data, modeling scripts, and supplemental materials are available on the project's associated OSF page: https://osf.io/mwvx3/

2.1. Behavioral data

The data used to create and evaluate the computational models were collected for a study described in Polyn, Norman, and Kahana (2009). We provide relevant methodological details here, and the original paper may be consulted for additional detail. Forty-two individuals (28 female, 14 male) from the University of Pennsylvania community received payment in

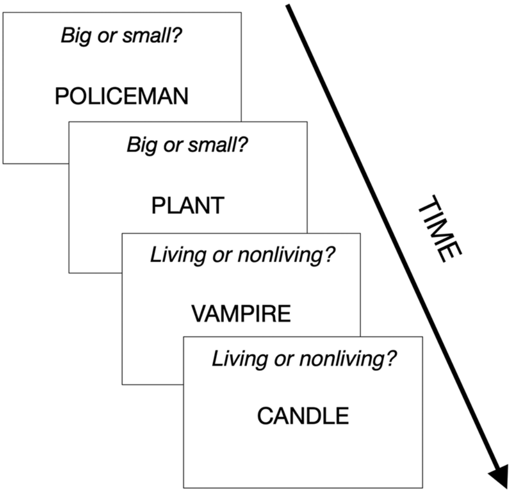

Fig. 2. Schematic representation of the binary semantic classification task. Target words were presented one at a time underneath a task cue, and participants indicated their response via keypress. (See the text for details.)

accordance with the University's IRB guidelines (for more demographic information, see Polyn et al., 2009). Participants were presented with a series of target words one at a time along with a task cue on a computer screen (Fig. 2). On size trials, participants indicated whether the referent was big or small (compared to a shoebox) with a keypress. On animacy trials, participants indicated whether the referent was living or nonliving with a keypress. During the initial instructions, participants were told that for some words there would not be an unambiguous correct answer and that they should respond according to their first reaction to the word. For the size task, the word *dog* was given as an example: A small dog could fit into a shoebox, but a large dog could not. For the animacy task, the word *dinosaur* was given as an example. A fossilized dinosaur would be nonliving, but a dinosaur from *Jurassic Park* would be living.

Each target word was presented in the middle of the screen for 3 s. If the participant did not make a response within 3 s, a warning message was displayed, and they advanced to the next trial automatically. Each set of 24 target words was followed by a 90-s free-recall period which is not examined here. The words were presented in two experimental conditions: *task shift* and *no shift*. In the *no shift* condition, the 24 target words were all associated with the same task (e.g., *size* or *animacy*). In the *task shift* condition, participants alternated judging the target words on either their size or animacy. Target words were drawn pseudorandomly from a word pool of 1,650 unique words. Each participant only saw a given word once, and the words were chosen such that no two words in a set of 24 were highly semantically related (this was done for the benefit of the free-recall task, which is not examined here). Each participant completed four experimental sessions, each with 12 sets of 24 target words, for a total of 1,152 word judgment trials. The final dataset contained 47,520 unique responses (after excluding missed trials).

A preliminary analysis of the binary semantic classification data was used to identify a set of representative words for each of the four response categories (*big*, *small*, *living*, and *nonliving*). For each of the 1,650 words, we calculated two proportions: one indicating the outcome of all big versus small responses for that word and the other indicating the outcome of all living versus nonliving responses for that word. Each proportion was an average across all participants who saw that word, but we note that a given participant would only see a given word in the context of one of the two judgment tasks. These proportions were used to identify words unanimously judged the same way by every participant. These unanimously judged words are treated as representative of that response category and are used to construct the item composite representations in the item-composite model.

The number of unanimous items for each of the four semantic response categories varied somewhat: *big* ($n = 402$, 24% of all unique words), *small* ($n = 222$, 13%), *living* ($n = 203$, 12%), and *nonliving* ($n = 569$, 34%). This variability did not seem to systematically affect model performance in any obvious way. This analysis yielded a set of representative unanimous words (624 words for big/small, 772 words for living/nonliving) that were used to construct the item-composite model, as described below. The non-unanimous words (1,026 words for big/small, 878 words for living/nonliving) were used to evaluate the performance of the model, as described below. The full list of unambiguous words used in the construction of the composite semantic axis is provided in the Appendix.

2.2. Modeling

We created a likelihood-based modeling framework to simulate and predict human performance on the binary semantic classification task. Each individual model was defined in terms of the following subcomponents: a DSM that was used to retrieve semantic vectors (GloVe or word2vec); a *semantic evaluation model* that was used to produce an evidence estimate for each choice alternative (single-adjective, adjective-composite, or item-composite); and a *decision model* that was used to convert the evidence into a decision likelihood for each choice alternative, and for the second decision model calculate response latency likelihood (logistic or LBA). This yielded a family of 12 models, which were evaluated against each other.

2.2.1. Distributional semantic models

We used two different word embeddings (word2vec and GloVe) for the semantic vectors for the target words and adjective labels. For word2vec, we used the Continuous Skip-gram version (Mikolov et al., 2013), trained on the English CoNLL17 corpus (Conference on Computational Natural Language Learning, English language subcategory, approximately 9 billion tokens; Zeman et al., 2017), producing 100-dimensional vectors. For GloVe, we used a version trained on a combination of Wikipedia 2014 and Giga-word 5 (6 billion tokens), producing 300-dimensional vectors (Pennington et al., 2014).

2.2.2. Semantic evaluation algorithm

Three semantic evaluation algorithms were created to calculate evidence (referred to here as dot-product scores) for each choice alternative. Each evaluation algorithm constructs a decision axis in the semantic space for each judgment task (size or animacy). To do this, the algorithm selects one or more representational vectors for each extreme of the continuum. These vectors are used to construct the decision axis as described below. In each case, a difference vector is constructed by subtracting the vectors associated with one extreme of the semantic axis from those associated with the other extreme. This difference vector is then used as the semantic axis against which words are judged (as described below).

The *single-adjective model* used a difference vector that was constructed by subtracting the vector for the adjective label *small* from the vector for *big* for the size trials and subtracting *inanimate* from *animate* for the animacy trials. Following the method of Grand et al. (2022), the *adjective-composite model* used a difference vector that was constructed by taking the difference between the two averages of two or three synonyms, that is, the difference of {*huge*, b*ig*, *large*} and {*small*, *little*, *tiny*} for the size trials, and the difference of {*animate*, *living*} and {*inanimate*, *nonliving*} for the animacy trials. Finally, for the *item-composite model*, we took the full set of vectors for the representative words (as described in Section 2.1) for each extreme of the semantic axis and averaged them together to make two item-composite representations (i.e., a *big* composite, and a *small* composite). The difference vector was constructed by subtracting one of these item-composite representations from the other.

Because the item-composite model is constructed using unanimously judged words representative of each semantic category, these unanimous words were excluded from our evaluation of the model. The exclusion of the unanimously judged words from model evaluation has a secondary benefit: The remaining words, by definition, show more variability in responses and as such provide a stronger test of the model's ability to capture the responses associated with potentially ambiguous words.

To derive the evidence for each individual trial, we calculated the dot product between the semantic vector for the target word and the difference vector resulting in one value, which we refer to as a dot-product score. It is important to note that the difference vector was only created when evaluating the logistic decision model, not the LBA model. The LBA model requires two competing accumulators for each response alternative. Whereas the logistic decision model produces a single evidence value by combining the two semantic composites into a single semantic decision axis, the LBA model uses the same semantic composites but does not combine them. Rather, it calculates a separate evidence score for each extreme of the classification by calculating the dot product of the target word representation with that extreme's semantic composite. Each of these evidence scores is used to drive one of the accumulators (as depicted schematically in Fig. 3).

2.2.3. Decision models

The two decision models—logistic transformation and LBA—convert evidence values into a decision likelihood for each choice alternative.

Logistic transformation. In the logistic version of the decision model, we generated the predicted responses for a given word using the logistic function. The probability for a given response was calculated using the logistic function using the following equation:

$$f(x) = \frac{1}{1 + e^{-k(x)}},$$

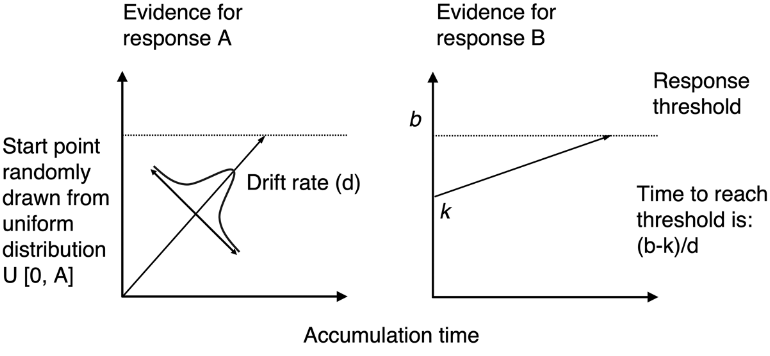

Fig. 3. Two-choice version of the LBA. The left panel shows the evidence for Response A, and the right panel shows the evidence for Response B. Starting values *k* are randomly drawn from a uniform random distribution. The drift rate *d* is an additive combination of evidence (calculated for the target word on that trial) and noise (drawn from a random normal distribution with standard deviation *s*). A response is made when the first accumulator reaches the threshold *b* (adapted from Brown & Heathcote, 2008).

where e is the natural logarithm base and x is the evidence (dot-product score) for a given target word. The free parameter k controls the steepness of the logistic curve.

Linear ballistic accumulator. In addition to the logistic decision model, we implemented an LBA model (Brown & Heathcote, 2008), which allowed us to evaluate model performance not only in terms of the predicted responses but also response latencies. While the logistic decision model simply produces a probability for each response option, LBA creates a probability distribution across the set of possible response latencies. LBA is a simple model of decision and response time that assumes multiple independent accumulators racing towards a certain decision threshold in a linear and deterministic manner until the decision is made. Across-trial variability in response latencies arises from noise added to the starting point of the accumulation process and from noise affecting that trial's drift rate. Each evidence accumulator begins with a certain amount of evidence reflected as a starting point k. Accumulated evidence increases at a speed determined by the drift rate d until it reaches the response threshold b (Fig. 3). The first accumulator to reach the threshold determines the response and the time to reach the threshold (the response latency) is calculated as $(b - k)/d$.

2.3. Model evaluation

We evaluated the fitness of each model variant using a maximum likelihood estimation technique. The probabilistic nature of the model allows it to predict the likelihood of each semantic classification response on a trial-by-trial basis, on the basis of the identity of the target work, and the identity of the classification task. To the extent that a given model tends to assign a higher probability to the observed response, that model will perform better in the model comparison analyses described below.

2.3.1. Maximum likelihood estimation

For both the logistic decision model and LBA, we calculated the likelihood of each model given the observed data by summing the log-transformed probability values for the model's trial-level predictions (this is equivalent to taking the product of the probability values associated with these trials). Each model produces a probability for the observed response on a given trial. The overall probability of a given dataset is the product of the estimated probabilities of each of these trial events. This overall probability was log-transformed into a log-likelihood value for model comparison statistics. We used the log-likelihood value to calculate Akaike information criterion (AIC) scores using the finite-sample correction algorithm described by Wagenmakers and Farrell (2004). These AIC scores were converted into weighted Akaike's information criterion (wAIC) scores (again following Wagenmakers and Farrell, 2004) to evaluate the fitness of multiple models relative to each other using the formula

$$w_i\,(AIC) = \frac{\exp\left\{-\frac{1}{2}\ \Delta_i\,(AIC)\right\}}{\sum_{k=1}^{K} \exp\left\{-\frac{1}{2}\Delta_k\,(AIC)\right\}},$$

where $\Delta_i(AIC)$ is the difference in AIC scores between each candidate model and the best candidate model. These Akaike weights w_i (AIC) sum to 1, with each weight indicating the conditional probability that the corresponding model M_i is the best model given the data and the set of candidate models (Wagenmakers & Farrell, 2004).

We additionally evaluated each model using a linear correlation analysis and a pairwise order consistency analysis. Grand et al. (2022) reported these analyses but not AIC, as their framework did not explicitly incorporate a decision model to produce response probabilities.

2.3.2. Linear correlation

For each word in the dataset not used in the construction of the item-composite vector (the restricted dataset), we first calculated the evidence score and the mean judgment value separately for each task (with a mean of 0 indicating all participants judged the word as small or inanimate, and 1 indicating all participants judged it as big or animate) averaged across the participants. Then we calculated a Pearson correlation between the evidence scores and mean judgment values for the words in the restricted dataset.

2.3.3. Pairwise order consistency

Following the method of Grand et al. (2022), we calculated the proportion of two-word combinations in the restricted dataset for which the difference between the human judgment and the dot-product scores was in the same direction, out of all possible two-word combinations. For example, if the word *elephant* was judged on average as larger than the word *mouse* and the dot-product score for *elephant* was larger than for *mouse*, then the *elephant–mouse* word pair would get a score of 1 and 0 otherwise. We repeated this procedure for each possible two-word combination, resulting in $1{,}650^2$ possible word combinations and scores (0 and 1). The final score is the proportion of 1s across all possible two-word combinations.

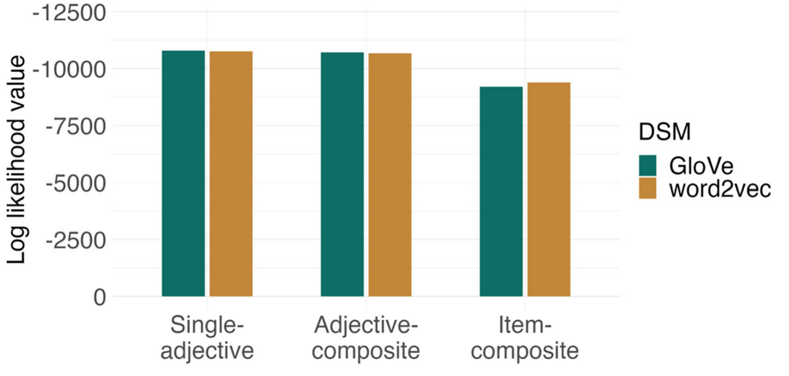

Fig. 4. Log-likelihood values for the logistic decision model combined across the size and animacy tasks. The two colors indicate the two DSMs. Values closer to zero correspond to better predictive power.

3. Results

3.1. *Maximum likelihood estimation: Logistic*

In terms of overall ability to predict behavioral responses, model variants containing the item-composite evaluation mechanism perform substantially better than model variants containing the other two semantic evaluation mechanisms. The log-likelihood fitness values for each model variant (Fig. 4) indicate that the item-composite semantic evaluation model was most likely to have generated the observed data (see Table SI-3b for raw values in the Supporting Information). The weighted AIC scores show that the advantage of the item-composite model variants is substantial (wAIC for item-composite: 1.0, for adjective-composite and single-adjective: 0.0 each). The results additionally indicate that when summed across all models and tasks, the models that use GloVe outperform the models that use word2vec (wAIC for GloVe: 1.0, word2vec: 0.0). As mentioned above, these model comparisons were done using a restricted set of more ambiguous words, which exclude any trial where the target word was used to construct the item-composite representations. In the Supporting Information, we report the results of this and other analyses on the full set of trials, which show similar results (though, as expected, some of the performance scores are inflated for the item-composite model) (see Tables SI-1c, SI-2c, SI-3c, SI-4c in the Supporting Information).

3.2. *Correlation analysis*

Using the logistic decision-making model, we carried out a correlation analysis to characterize the degree of correspondence between each model's predicted responses and the observed responses (Fig. 5) (see Table SI-1b for raw correlation coefficients in the Supporting Information). The item-composite model yields numerically the largest and consistently reliable (all $ps < .05$) correlations between the predicted and observed responses across the two embedding spaces. When averaged across both tasks and both embeddings, the item-composite model has the highest correlation of .63, followed by the adjective-composite

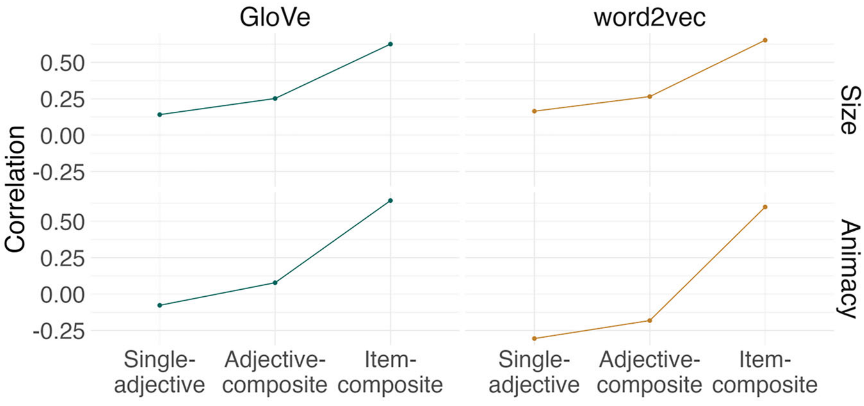

Fig. 5. Pearson r coefficients between the predicted responses and the mean human judgments for the size task (upper panels) and animacy task (lower panels). Left panels indicate results from models using the GloVe DSM and right panels the word2vec DSM.

model with a correlation of .10, and the single-adjective model with a correlation of $-.02$. All of the pairwise comparisons between these Fisher-transformed correlation coefficients are significant, with all zs > 3.53, and all ps $< .0004$). When averaged across all models and embeddings, the correlation coefficients for the size and animacy tasks were significantly different with the means of .35 and .13 for size and animacy tasks, respectively (the difference between Fisher-transformed correlation coefficients $z = 6.88$, $p < 10^{-11}$). When averaged across all models and tasks, the GloVe and word2vec embeddings produced numerically slightly different results with the means of .28 and .20, respectively (the difference between Fisher-transformed correlation coefficients $z = 2.38$, $p = .02$).

We sought to determine whether the substantial advantage of the item-composite model was due to the larger number of word vectors used to construct the semantic composites relative to the other two models. A follow-up analysis suggests that the item-composite model performs at a superior level even when the number of words used to make the composites is matched across the different model types. This analysis involves a specialized permutation analysis on the trials using the size task. For each permutation, we randomly selected three words each from the sets of unanimous big and small words used to construct three-item semantic composite representations for the item-composite model. We used these words to construct a new difference axis and reran the correlation analysis reported above. We repeated this procedure 100 times (for both the GloVe and word2vec model variants) to obtain a distribution of correlation values. The mean correlation coefficient across the 100 permutations was .48 for GloVe and .52 for word2vec (Fig. 6). While these correlation values were numerically smaller than for the original item-composite model (means of .69 and .72 for GloVe and word2vec, respectively), they were reliably larger than the correlation values associated with adjective-composite model (.25 and .26 for GloVe and word2vec, respectively). In other words, the three-item semantic composite model showed a better correspondence to human responses than the adjective-composite model, for 99 out of the

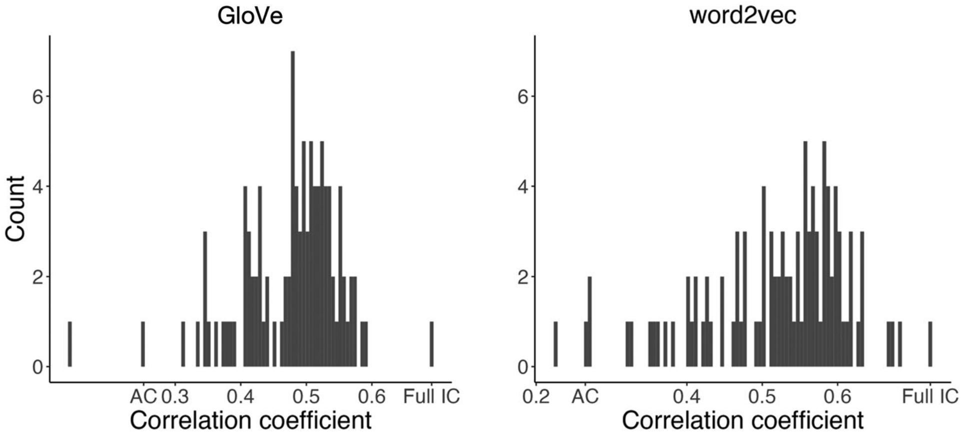

Fig. 6. A distribution of Pearson *r* correlation coefficients for 100 three-item variants of the item-composite models, constructed using a permutation procedure. Correlation coefficients reflect the correspondence between the model's predicted responses and the mean human judgments for the size task using GloVe (left panel) and word2vec (right panel). *AC* indicates the correlation coefficient calculated for the adjective-composite model, and *Full IC* indicates the correlation coefficient calculated for the item-composite model with all unanimously judged items included in the semantic composite.

100 permutation-based models. This was true for both GloVe and word2vec. This indicates that the predictive advantage of the item-composite model is due to the semantic identities of the words used to construct the semantic model, rather than the quantity of words.

3.3. Pairwise order consistency

A pairwise order consistency analysis (following Grand et al., 2022) also demonstrated the superiority of the item-composite model to the other models in terms of the degree of correspondence between each model's predicted responses and the observed responses (Fig. 7) (see Table SI-2b for raw pairwise order consistency coefficients in the Supporting Information). A pairwise order consistency score of 100% indicates perfect correspondence between model and observed behavior, and 50% indicates chance-level performance. We assessed statistical significance using a permutation analysis with 10,000 random shuffles of the model-produced evidence values. This allowed us to construct a null distribution and calculate p-values for responses from each judgment task within each distributional model, yielding four pairwise order consistency statistics for GloVe-size, GloVe-animacy, word2vec-size, and word2vec-animacy.

On average, the single-adjective model performed at chance levels, with mean pairwise order consistency values: GloVe-size = 55% ($p = .98$), GloVe-animacy = 48% ($p = .90$), word2vec-size = 56% ($p = < .0001$) and word2vec-animacy = 40% ($p = .99$). The adjective-composite model performed better by a few percentage points, which caused three of the pairwise order statistics to rise above the permuted distribution, but word2vec-animacy remained at chance levels, GloVe-size = 58% ($p < .0001$), GloVe-animacy = 54% ($p < .01$),

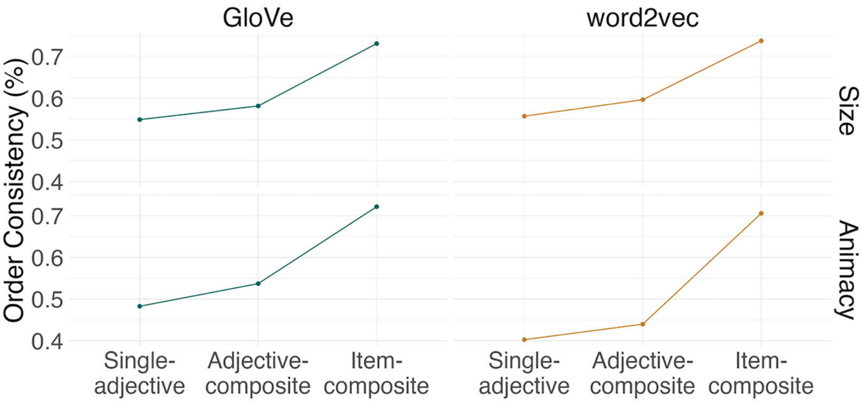

Fig. 7. Pairwise order consistency values for the size task (upper panel) and animacy task (lower panel). Different colors represent two DSMs. Left panels indicate results from models using the GloVe DSM and right panels the word2vec DSM.

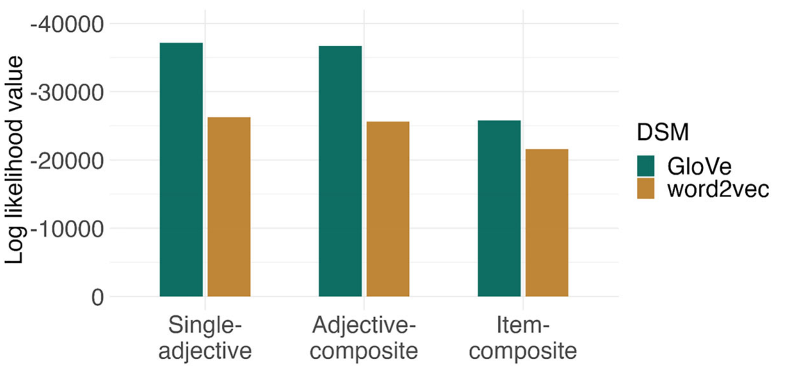

Fig. 8. Log likelihood values for the LBA decision model combined across the size and animacy tasks. The different colors indicate the two DSMs. Values closer to zero indicate a better fit.

word2vec-size = 60% (p < .0001) and word2vec-animacy = 44% (p = .99). The item-composite model performed substantially better than the other two models, with all pairwise order statistics substantially above chance, GloVe-size = 73% (p < .0001), GloVe-animacy = 72% (p < .0001), word2vec-size = 74% (p < .0001) and word2vec-animacy = 71% (p < .0001). The GloVe and word2vec models performed similarly well when averaged across model variants, with mean pairwise order consistency of 59.96% and 57.33%, respectively.

3.4. Maximum likelihood estimation: Linear ballistic accumulator

Broadly speaking, simulations using the LBA decision rule also demonstrated the superiority of the item-composite semantic evaluation algorithm to the other algorithms, in terms of overall predictive power of the models (Fig. 8) (see Table SI-4b for raw values in the Supporting Information). This demonstrates that the item-composite algorithm can be integrated

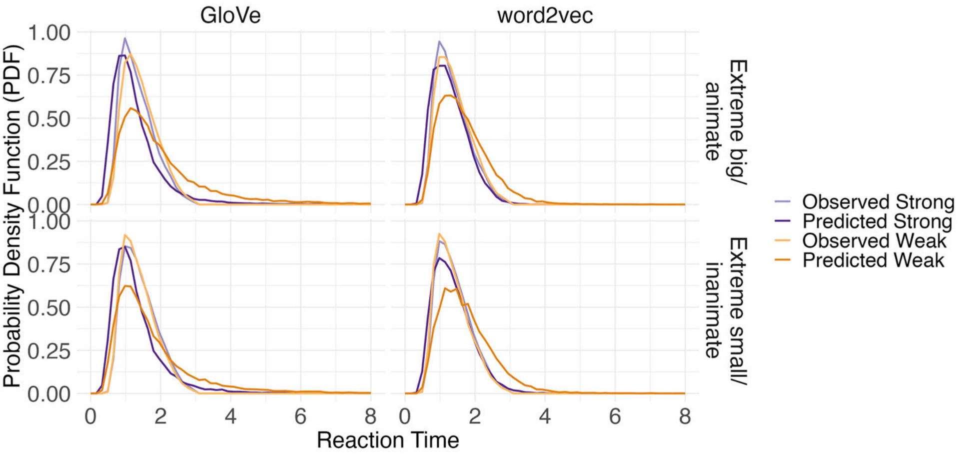

Fig. 9. Probability density functions for observed and predicted reaction times using the item-composite LBA model with GloVe (left panels) and word2vec (right panels) embeddings. Top panels correspond to response times for *big* and *animate* responses. Bottom panels correspond to response times for *small* and *inanimate* responses. The best-fitting item-composite model predicts a larger response time difference between strong and weak words than is observed.

into a framework that predicts response times, though some of the following analyses reveal substantial room for improvement in this regard. As with the logistic decision model, model variants including the item-composite mechanism were substantially more likely to have generated the observed data (wAIC for item-composite: 1.0, for adjective-composite and single-adjective: 0.0 each). With the logistic decision model, model variants including GloVe provided better fits to the observed data. Here, model variants including word2vec yielded a better fit to the observed data (wAIC for GloVe: 0.0, word2vec: 1.0). We return to this point in the discussion.

A closer examination of trial-level predictions indicated that all models produced qualitatively poor fits to certain aspects of the observed response latencies. Fig. 9 shows the probability density for the observed and predicted reaction times for correct responses using the item-composite LBA model. For this analysis, we partitioned the trial events based on the model's estimate of evidence strength (the *evidence scores* described in Section 2.2.2) for a given word being judged in a given task. The top 50% of evidence scores were treated as strong evidence, and the bottom 50% of evidence scores were treated as weak evidence.

This analysis reveals that the item-composite LBA model predicts a much larger difference in response times between strong-evidence and weak-evidence trials than is seen in the observed data. In the observed data, participants are reliably faster to respond for words labeled as having strong evidence than weak evidence. For this analysis, we aggregated across the two judgment tasks. For big and animate responses, trials with strong evidence were 64 ms faster than trials with weak evidence ($t_{(41)} = -21.73, p < 10^{-15}$). For small and inanimate responses, trials with strong evidence were 30 ms faster than trials with weak evidence ($t_{(41)} = 8.51, p < 10^{-9}$). The model correctly predicts that trials with strong evidence will be faster than trials with weak evidence, but the model gets the magnitude of the effect wrong. For simulated big and animate responses, trials with strong evidence were 648 ms faster than trials with weak evidence ($t_{(41)} = -58.65$, $p < 10^{-15}$). For simulated small and inanimate responses, trials with strong evidence were 521 ms faster than trials with weak evidence ($t_{(41)} = -58.66$, $p < 10^{-15}$).

The adjective-composite LBA model has a similar problem, though the problematic overprediction is even more pronounced. For simulated big and animate responses, trials with strong evidence were 1,111 ms faster than trials with weak evidence ($t_{(41)} = -76.46$, $p < 10^{-15}$). For simulated small and inanimate responses, trials with strong evidence were 1,085 ms faster than trials with weak evidence ($t_{(41)} = -66.54$, $p < 10^{-15}$).

4. Discussion

Semantic memory stores information about the world and the things in it. DSMs offer insight into the nature of human semantic memory and have been used both as a tool to understand behavioral data and as theories of the cognitive representation of semantic knowledge (Landauer & Dumais, 1997; Lund & Burgess, 1996; Jones & Mewhort, 2007). These models provide an automated way to construct semantic spaces and can be combined with the cognitive mechanisms of decision-making to characterize human semantic categorization behavior.

In the current paper, we combined the principles of the multiple-trace theory of memory (MINERVA 2; Hintzman, 1986), the instance theory of semantic knowledge (ITS; Jamieson et al., 2018), and the methods from Grand et al. (2022) to build a computational model of binary semantic classification. ITS proposes that encounters with words are stored as individual traces in episodic memory and that the semantic meaning of a word can be constructed on the fly by retrieving a blend of many memory traces containing independent instances of usage of that word from episodic memory. Jamieson et al. (2018) demonstrate that ITS does a good job inferring the meaning of homonyms from the local linguistic context and can capture the taxonomic structure of sets of words from distinct categories.

In our simulation of the binary semantic classification task, we compare two instance-inspired cognitive mechanisms. In the case of the adjective-composite algorithm, the semantic identity of a set of adjective labels is retrieved (as in Grand et al., 2022). In the case of the item-composite algorithm, the semantic identities of representative items are retrieved. We paired these semantic evaluation algorithms with two cognitive models of decision-making. The first model uses a logistic function to simulate the likelihood of each choice decision. The second model incorporates LBAs (Brown & Heathcote, 2008) to simulate both responses and response latencies as a race between accumulators representing the two extremes of a decision axis.

Our findings demonstrate that the model variants containing the item-composite semantic evaluation algorithm provide a better account of human classification responses and response times in the binary semantic classification task, relative to two other semantic evaluation algorithms. The item-composite algorithm constructs its item composites using the vector representations of words judged unanimously by all participants for each response category. To fairly evaluate these models, we only examined judgments for the non-unanimous words not used in the construction of the item-composite model. Examining the set of ambiguous words provides a challenging test for the models.

As mentioned earlier, model comparisons using the full dataset (including all trials) are presented in the Supporting Information. These results show a consistent pattern of results to those on the restricted dataset for all three main analyses, though item-composite model performance is generally inflated, as expected. This inflation is particularly noticeable in the results of the correlation analysis. The Supporting Information also evaluates the models on an independent dataset using the same semantic classification tasks, finding similar results in all regards (see Tables SI-1a, SI-2a, SI-3a, SI-4a in the Supporting Information).

The item-composite model uses more item representations to construct its composite representations than the other models. However, this difference does not seem to explain the difference in performance of the two models. Using a permutation analysis, we subsampled the words used to construct the semantic decision axis in the item-composite model, matching the number of items used in the adjective-composite model. The item-composite model retained its predictive advantage in 99% of these permutations, suggesting it is the quality of the words used to construct the difference axis, not the quantity, that drove the observed pattern of results. Together, these findings suggest that a cognitive mechanism involving the retrieval and blending of items that are representative of the extremes of the semantic decision axis is more promising than a mechanism using the adjective labels directly.

We used two DSMs, GloVe and word2vec, in the modeling framework, primarily to demonstrate that the advantage of the item-composite model is not dependent on the particular DSM used to construct the word embeddings. A comparison of the two embedding spaces to one another was generally inconclusive regarding their relative utility for cognitive modeling. The two DSMs performed similarly well overall. While GloVe outperformed word2vec using the logistic model, word2vec outperformed GloVe using the LBA model. It is not clear what aspects of the DSMs are responsible for these differences. Word2vec is a predictive model with hidden layers that learn representations of words through prediction and self-correction, and GloVe is a latent semantic abstraction model which lacks this predictive component. However, both models use co-occurrence information in similar ways, and they were not matched in terms of secondary characteristics, such as the specific text corpus used for training. Other groups have also found that these two models have similar utility in cognitive model development. For example, Pereira, Gershman, Ritter, and Botvinick (2016) found that word2vec and GloVe produced comparable results in a large study comparing various DSMs on word association, synonyms and analogy problems, and similarity and relatedness judgments.

Our assumption that semantic reasoning is based on an on-the-fly retrieval of individual word instances is broadly consistent with a variety of findings from the study of real-time lexical processing, which show that word meanings are flexible in context, drawing on multiple possible meanings in a context-dependent manner (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Metzing & Brennan, 2003). For example, interpretation of referential expressions like "the girl" and "the peanut" is shaped by the properties of the overall discourse they are embedded in, including the referents and their properties. For example, in a context that illustrates the animacy of a cartoon peanut, a sentence like "*The peanut was in love*" is easily processed, but a locally coherent sentence like "*The peanut was salted*" results in confusion as indicated by an increased N400 effect (Nieuwland & Van Berkum, 2006; also see Nieuwland, Otten, & Van Berkum, 2007). Likewise, an instruction like "*Put the cube inside the can*" given a context with two differently sized cans causes momentary confusion if the cube is small enough to fit in either can, whereas this confusion is lifted if the cube is larger and only will fit in the larger of the two cans as indicated by the earlier eye fixations on the target object (Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002; also see Chambers, Tanenhaus, & Magnuson, 2004).

The idea that semantic classification involves comparing a target with other items is supported by findings from studies of language production. For example, adjectives like *small* and *large* tend not to be produced by speakers unless the immediate context contains items that contrast along the size dimension and the speaker has noticed them (Brown-Schmidt & Tanenhaus, 2006; Brown-Schmidt & Konopka 2008; Pechmann, 1989). For example, when naming a butterfly, if the speaker fails to notice a larger one in the scene, they are likely to simply say "butterfly," and if they do notice the larger butterfly the timing of when the adjective is produced is strongly predicted by the latency of the eye-fixation to the size-contrasting item, with early looks producing prenominal modifiers (e.g., "the small butterfly"), and later looks producing late modifiers (e.g., "the butterfly, uh small one"), unless the speaker is using a language that affords postnominal modification (e.g. "la mariposa pequeña"; Brown-Schmidt & Konopka, 2008).

The mechanism comparing a target word with representative items from a response category potentially provides insight into how a relevant comparison class shapes semantic judgment. While we do not address this question in the present work, the set of extreme exemplars that are retrieved may itself be a contextually dependent process; if so, this may explain some of the contextual dependency in how certain linguistic expressions are *interpreted* in rich contexts. In our study, when making a size judgment, participants had a reference point as they were asked to judge a size of an object compared to a shoebox. As such, it was not necessary for participants to alter the set of comparison items from trial to trial. However, the flexible nature of the retrieval process described here opens up the possibility of using this model to make more flexible classification judgments.

The item-composite model allows for the reference point to shift in different contexts by altering the set of retrieved representative examples for each semantic category. For example, the model could be used to judge the relative size of things in a cellular environment by using the descriptive phrase *cellular environment* alongside the category label (*big* or *small*) to retrieve representative items. In this context, *ribosomes* could be judged relative to small items like *virus* and *RNA*, and large items like *nucleus* and *endoplasmic reticulum*. In contrast, the model could use the phrase *geographical entities* to judge *Texas* relative to small things like *Galapagos Islands* and *Switzerland* and large things like *Spain* and *Africa*.

Indeed, it is well-established in the referential processing literature that the real-time interpretation of phrases like *the small glass* is driven by the relevant comparison set in the immediate context (Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Sedivy, 2003): The adjective "small" evokes a 4-cm tall glass when the context contains a 4-cm and an 8-cm glass, but "small" evokes the 8-cm glass when it is paired with a 12-cm one. Further, these comparison classes are created on the fly, based on multiple cues in the local context. In a context where a listener views three drinking glasses (4 cm, 8 cm, 12 cm tall), and the speaker says "Pick up the small glass," this sentence is typically interpreted as referring to the smallest glass that the speaker can see: If the 4-cm glass is obscured from the speaker's view, the listener interprets "the small glass" to be the 8-cm tall one, rather than the "small" 4-cm glass that the speaker cannot see (Heller, Grodner, & Tanenhaus, 2008; Heller, Parisien, & Stevenson, 2016; Ryskin, Benjamin, Tullis, & Brown-Schmidt, 2015). Findings like these might be captured by a semantic model that can sculpt a retrieved set of representative exemplars on the basis of properties of the local context.

This account could also provide theoretical leverage regarding the flexibility of human semantic knowledge. Previous studies indicate that humans are capable of rapidly and flexibly reconfiguring their semantic knowledge to meet various task demands. A good example of such conceptual flexibility is ad hoc categories (Barsalou, 1983), such as *things to sell at a garage sale* or *things that can fall on one's head*. While these categories are unlikely to be part of a person's core semantic knowledge, participants can nevertheless perform such a classification rapidly, suggesting that they can quickly construct a representation of a category they have never encountered before. An attempt to simulate performance on such a task could begin with the retrieval of a few representative items from the category, though there might be a theoretical challenge in determining what items should be used to define the other end of the semantic decision axis (i.e., *things inappropriate to sell at a garage sale*).

Many natural language models have grappled with the importance of contextual information. For example, the probabilistic Topics model (Griffiths et al., 2007) uses principles similar to LSA (Landauer & Dumais, 1997), and in addition incorporates the idea that certain words tend to be distributed over certain discourse topics (e.g., nature, education, health). Thus, the Topic model partially includes contextual information in the representation of words. As a result, the Topic model can produce better fits to free association data than LSA and was able to account for homonym, disambiguation, word prediction, and discourse effects can be problematic for cognitive models incorporating LSA (Griffiths et al., 2007).

The importance of contextual information has become evident with the advent of a new class of DSMs: transformer models such as BERT (Devlin et al., 2018), ELMo (https://allenai.org/allennlp/software/elmo), and GPT-2 (Radford et al., 2019). The key difference between this novel class of models and older models is that it integrates contextual information within the representation of each word, significantly improving performance on a variety of semantic tasks, including tasks involving the production of coherent language (Bhatia and Richie, 2021). While transformer models outperform many other types of computational models on semantic tasks, work remains to be done to determine their cognitive plausibility. These models process each word in a phrase or sentence in parallel, but evidence from sentence processing literature suggests that sentence processing is linear and incremental (Kamide et al., 2003).

Our study leaves open a number of questions for future work. The item-composite model uses representations constructed from large linguistic corpora. However, these semantic vectors do not have easily interpretable semantic dimensions, which makes it unclear how the relevant words, used to construct the axes, are retrieved from memory. One possibility is that some perceptual features of concepts can be recovered through linguistic co-occurrence statistics. Previous research has shown that individuals who lack certain sensory experiences—for example, congenitally blind individuals—possess detailed semantic knowledge about perceptual features of various objects. For example, van Paridon, Liu, and Lupyan (2021) demonstrated that congenitally blind people, despite the lack of visual perceptual experience, formed associations between colors and adjectives (e.g., blue is cold, red is hot) that were similar to the intuitions of sighted people. Similarly, Kim, Elli, and Bedny (2019) compared blind and sighted people's knowledge of the appearance of common animals. The authors found that individuals who were blind inferred features of animal appearance from taxonomy and habitat properties (e.g., because sharks live in the water, they must have scaley skin like other fish). These results indicate that knowledge of animal appearance (even if incorrect) can be acquired through inference from language, rather than through memorization of facts directly specifying those properties. An alternative explanation is in line with a computational model of language processing described by Johns and Jones (2015), which relates to both usage-based theories of language learning and the instance theory of semantic knowledge. According to this proposal, during language processing, linguistic information (e.g., *flowers bloom in the spring*) is encoded along with referential information (i.e., perceptual information experienced during language comprehension, e.g., flower color, size, etc.). Later, when the linguistic memory trace is retrieved, the attached experiential referential information is retrieved with it, making it possible to judge flowers on various perceptual properties.

Finally, our study does not answer the question of whether the semantic decision axes used in this work are part of an individual's existing representational knowledge or if they are constructed on the fly to meet specific task demands. Instance-based theories of semantic knowledge describe how a representation of word meaning can be constructed on the fly in a highly parallel, probe-driven retrieval process (Jamieson et al., 2018). Following Jamieson et al. (2018), we speculate that the composite representations used in our models might be constructed during task performance and not necessarily constitute a part of the participant's core semantic knowledge.

Open Research Badges

This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/mwvx3/?view_only=22173ead608b4405be9fef330e4b512e

References

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*(3), 211–227.

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

Bhatia, S. (2019). Semantic processes in preferential decision-making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(4), 627.

Bhatia, S., & Richie, R. (2022). Transformer networks of human conceptual knowledge. *Psychological Review*. Advance online publication https://psycnet.apa.org/doi/10.1037/rev0000319

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, *29*, 31–36.

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.

Brown-Schmidt, S., & Konopka, A. E. (2008). Little houses and casas pequeñas: Message formulation and syntactic form in unscripted speech with speakers of English and Spanish. *Cognition*, *109*(2), 274–280.

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*(4), 592–609.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, (3), 153–178.

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, (3), 687.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language*, *47*(1), 30–49.

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

Collins, A. M., & Quillian, R. M. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247.

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, *132*(2), 163–201. https://doi.org/10.1037/0096-3445.132.2.163

Cutler, R. A., Duff, M. C., & Polyn, S. M. (2019). Searching for semantic knowledge: A vector space semantic analysis of the feature generation task. *Frontiers in Human Neuroscience*, *13*, 341.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*, (6), 409–436.

Firth, J. R. (1957). *Papers in linguistics, 1934–1951*. London, England: Oxford University Press.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford, England: Oxford University Press.

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, *6*, 975–987.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211.

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*(3), 831–836.

Heller, D., Parisien, C., & Stevenson, S. (2016). Perspective-taking behavior as the probabilistic weighing of multiple domains. *Cognition*, *149*, 104–120.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological review*, *95*(4), 528.

Hintzman, D. L. (1986). " Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*(4), 411.

Jackendoff, R. S. (1992). *Semantic structures* (Vol. 18). MIT press.

Jamieson, R. K., Avery, J. E., Johns, B. T., & Jones, M. N. (2018). An instance theory of semantic memory. *Computational Brain & Behavior*, *1*, (2), 119–136.

Johns, B. T., & Jones, M. N. (2015). Generating structure from experience: A retrieval-based model of language processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *69*(3), 233.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37. https://doi.org/10.1037/0033-295X.114.1.1

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language*, *49*(1), 133-156.

Kim, J. S., Elli, G. V., & Bedny, M. (2019). Knowledge of animal appearance among sighted and blind adults. *Proceedings of the National Academy of Sciences*, *116*(23), 11213–11222.

Lambon Ralph, M. A., & Patterson, K. (2008). Generalization and differentiation in semantic memory: Insights from semantic dementia. *Annals of the New York Academy of Sciences*, *1124*(1), 61–76.

Landauer, T. K., & Dumais, S. T. (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

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. *Concepts: Core Readings*, *3*, 81.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, *28*, (2), 203–208. https://doi.org/10.3758/BF03204766

Mahon, B. Z., & Caramazza, A. (2003). Constraining questions about the organization and representation of conceptual knowledge. *Cognitive Neuropsychology*, *20*(3–6), 433–450. https://doi.org/10.1080/02643290342000014

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*, (4), 547–559.

McRae, K., De Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99–130. https://doi.org/10.1037/0096-3445.126.2.99

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201–213.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv. https://doi.org/10.48550/arXiv.1301.3781.

Morton, N. W., & Polyn, S. M. (2016). A predictive framework for evaluating models of semantic organization in free recall. *Journal of Memory and Language*, *86*, 119–140.

Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111.

Nieuwland, M. S., Otten, M., & Van Berkum, J. J. (2007). Who are you talking about? Tracking discourse-level referential processing with event-related brain potentials. *Journal of Cognitive Neuroscience*, *19*(2), 228–236.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, *49*(3), 197–237. https://doi.org/10.1037/h0021468

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning* (No. 47). University of Illinois press.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 89–110.

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). Stroudsburg, PA: Association for Computational Linguistics.

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3–4), 175–190. https://doi.org/10.1080/02643294.2016.1176907

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Richie, R., & Bhatia, S. (2021). Similarity judgment within and across categories: A comprehensive model comparison. *Cognitive Science*, *45*(8), e13030.

Rumelhart, D. E., & McClelland, J. L. & the PDP Research Group.(1986) *Parallel distributed processing: Explorations in the microstructure of cognition*, *Vol. 1: Foundations*. Cambridge, MA: MIT Press.

Ryskin, R. A., Benjamin, A. S., Tullis, J., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. *Journal of Experimental Psychology: General*, *144*(5), 898.

Saffran, E. M., & Schwartz, M. F. (1994). Of cabbages and things: Semantic memory from a neuropsychological perspective—A tutorial review. In C. Umiltà & M. Moscovitch (Eds.), *Attention and performance 15: Conscious and nonconscious information processing* (pp. 507–536). Cambridge, MA: MIT Press.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147.

Smith, K. A., Huber, D. E., & Vul, E. (2013). Multiply-constrained semantic search in the remote associates test. *Cognition*, *128*(1), 64–75.

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological review*, *81*(3), 214.

Tulving, E. (1972). Episodic and Semantic Memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). New York, NY: Academic Press.

van Paridon, J., Liu, Q., & Lupyan, G. (2021). How do blind people know that blue is cold? Distributional semantics encode color-adjective associations. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43). Austin, TX: Cognitive Science Society.

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic bulletin & review*, *11*, 192–196.

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinkova, S., Hajic, J., Hlavacova, J., Kettnerová, V., Uresova, Z., …, Li, J. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 1–19). Vancouver, Canada: Association for Computational Linguistics.

Supporting Information

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

Independent Dataset
Restricted Dataset (reported in the manuscript)
Full Dataset (the dataset used is the same as reported in the manuscript, but fit on the full set of items)

Appendix

List of unambiguous words used in the construction of the composite semantic evaluation model.

Big	Small	Animate	Inanimate
acrobat	almond	acrobat	acid
actor	ant	actress	aircraft
actress	apple	adolescent	airport
adolescent	aspirin	adult	album
adult	bacon	alligator	alley
africa	bait	antelope	ambulance
agent	bandage	ape	anchor
aircraft	bead	apple	antenna
airplane	bean	architect	apartment
airport	bee	artist	application
alley	berry	assistant	apron
ambulance	bible	astronaut	article
ancestor	bluejay	athlete	ashtray
antelope	bracelet	audience	atlas
antler	broccoli	author	attic
apartment	bruise	ballerina	automobile
ape	bubble	bartender	award
arena	buckle	bear	badge
army	bug	beaver	bag
artist	butter	beggar	balcony
asia	butterfly	biologist	ball
assistant	button	bird	balloon
astronaut	camera	boy	ballot
atmosphere	candle	boyfriend	bandage
attorney	card	brother	barn
audience	cardinal	bull	baseball
aunt	carrot	burglar	basement
author	cashew	butcher	basket
automobile	caterpillar	butler	basketball
baker	cent	butterfly	bassinet
ballerina	chalk	camel	bath
bandit	charcoal	canary	bathroom
bank	checkers	candidate	bathtub
banker	cheddar	captain	battery
barn	cheek	carpenter	bay
bartender	chemical	cat	beach
bay	cherry	cheerleader	bedroom
beach	chip	chef	beer
bedroom	chocolate	child	belt
beggar	cinnamon	chimpanzee	bench
bicycle	clove	climber	beverage
bike	coal	cobra	bicycle

(*Continued*)

Big	Small	Animate	Inanimate
biologist	cocktail	colonel	bill
bison	coin	comedian	biscuit
blackboard	coleslaw	companion	blackboard
blockade	collar	consumer	blanket
boat	compass	cousin	blockade
body	cookie	cow	blueprint
booth	cork	cowboy	board
boss	cotton	creature	boat
boy	cream	cricket	bolt
boyfriend	crumb	criminal	bomb
bridge	crystal	crocodile	book
brother	cue	crow	boot
brunette	cuff	customer	booth
buffalo	cup	dad	bottle
building	daisy	dancer	bouillon
bull	dandruff	deer	boulder
bully	diamond	dentist	boulevard
bureau	diaper	dictator	bowl
bus	dice	doctor	box
camel	dime	driver	bracelet
canoe	dollar	eagle	brake
canvas	doorbell	electrician	brandy
canyon	dough	elephant	brick
capital	drug	elk	bridge
captive	dust	emperor	brook
car	ear	employee	broom
caravan	earring	employer	brush
carnival	egg	farmer	buckle
carpenter	electron	father	buggy
carriage	envelope	fireman	building
cashier	eyelash	fish	bulb
castle	feather	flower	bulletin
cathedral	fig	friend	bun
cattle	finger	frog	bureau
ceiling	fingernail	gentleman	bus
cellar	fish	girl	button
champion	fist	goose	cabin
chapel	flask	gorilla	cafe
chauffeur	flea	grasshopper	cafeteria
cheerleader	flower	guest	cage
chef	fly	gymnast	cake
chemist	foot	hawk	calculator
chief	fragrance	hen	calendar
church	freckle	hornet	camera
citizen	fries	horse	can
clerk	frost	hostess	canal
cliff	garlic	hound	candle

(*Continued*)

Big	Small	Animate	Inanimate
climber	gem	husband	cane
closet	gene	infant	cannon
coach	germ	instructor	canoe
college	gin	inventor	canvas
colonel	glasses	kid	cap
comet	grape	lady	cape
commander	gum	leader	caravan
community	hand	lion	card
computer	heel	lover	carpet
concert	honey	mailman	cart
conductor	jar	man	carton
consumer	jello	manager	cash
contractor	jewel	mayor	casket
convent	key	miner	castle
cook	kitten	mob	cathedral
cooler	label	mongoose	cave
cop	lace	monk	cellar
copier	leaf	monkey	cello
corporation	lemon	moth	cemetery
couch	lens	mother	cent
country	lime	mouse	chain
county	lint	mule	chalk
cow	lipstick	navigator	chamber
cowboy	lizard	nephew	champagne
criminal	lock	niece	charcoal
critic	lollipop	nun	check
cupboard	loop	nurse	checkers
curtain	magnet	octopus	chime
cyclone	mascara	officer	chimney
dad	match	otter	chisel
dam	mint	outlaw	church
daughter	mitten	owl	cigar
dentist	molecule	ox	cigarette
department	money	oyster	cinnamon
designer	mosquito	parent	clay
detective	moss	parrot	cliff
dictator	moth	partner	clippers
dinosaur	mouse	patient	closet
dishwasher	mouth	pedestrian	clothes
diver	nail	pelican	coal
doctor	napkin	penguin	cobweb
dolphin	necklace	person	coffin
donkey	needle	philosopher	coin
door	nitrogen	pig	coleslaw
dorm	nose	pirate	cologne
dragon	note	plumber	column
driver	novel	poet	compass

(*Continued*)

Big	Small	Animate	Inanimate
dryer	nucleus	politician	computer
dungeon	nut	pony	cone
earth	ointment	preacher	contract
editor	olive	president	convent
egypt	ornament	priest	cookbook
electrician	peanut	prince	cookie
elephant	pear	princess	cooler
elk	pearl	prisoner	copier
emperor	pedal	producer	cord
empire	pen	professional	cottage
employee	penny	puppy	couch
employer	pill	quail	court
engineer	pimple	queen	cracker
escalator	pin	rabbit	crater
europe	plaque	raccoon	crayon
factory	pocket	referee	crevice
family	poison	reptile	crown
farm	popcorn	robber	crutch
farmer	proton	roommate	cube
father	prune	rooster	cuff
field	puck	rose	cup
fighter	quarter	runner	cupboard
fleet	raisin	sailor	curb
florida	rat	salesman	cushion
forest	razor	salmon	custard
fort	ribbon	scallop	cyclone
fountain	ring	secretary	cylinder
france	salt	sergeant	dagger
freeway	sand	serpent	dam
friend	sapphire	shark	dart
furniture	saucer	sheep	dashboard
galaxy	sausage	shepherd	deck
gang	screw	sibling	denim
gangster	seed	sister	deodorant
garage	shoe	snake	desk
garden	shoelace	son	dessert
general	shrimp	spider	detergent
gentleman	signature	spouse	diagram
giraffe	slime	stewardess	dial
girl	slug	stranger	diamond
gorilla	snack	student	diary
governor	soap	surgeon	dice
graduate	sock	swan	dime
grave	spice	swimmer	diner
groom	spider	teacher	dinner
guard	sponge	teenager	diploma

(*Continued*)

Big	Small	Animate	Inanimate
guardian	spool	termite	disc
gym	staple	thief	dish
gymnast	straw	toad	dock
haystack	strawberry	tortoise	doll
helicopter	string	tourist	dollar
herd	syringe	traitor	doorbell
hero	tack	turtle	dough
highway	tag	typist	drawer
hiker	tangerine	uncle	dress
horse	tart	victor	drink
hospital	tea	visitor	driveway
house	thermometer	waiter	drug
human	thimble	waitress	dryer
hurricane	thorn	walrus	dune
iceberg	thumb	warrior	dungeon
igloo	tick	whale	dustpan
inmate	ticket	winner	earring
instructor	toad	witness	elevator
inventor	toast	wolf	encyclopedia
island	toe	woman	engine
jeep	tomato	zebra	eraser
jet	toothbrush		escalator
judge	toothpaste		essay
jungle	trigger		explosion
jupiter	tulip		factory
kangaroo	turnip		feast
keeper	tweezers		feather
king	twig		fiddle
kitchen	virus		fireplace
lady	vitamin		flag
landscape	wallet		flannel
lawn	wasp		flashlight
lawyer	wax		flask
leader	wick		fleet
leopard	wire		floor
lieutenant	worm		flour
limousine	wound		fort
lion	wrench		fossil
lodge	wrist		fragrance
london	yolk		freeway
lounge			fudge
lover			funeral
magician			fur
man			furniture
manager			gallon
mansion			garage
mars			garbage

(*Continued*)

Big	Small	Animate	Inanimate
mattress			gauze
meteor			gavel
microwave			gin
military			glacier
mister			glass
moat			glasses
mob			glue
monster			gold
moon			gown
moose			grave
mother			gravel
motorcycle			grease
mountain			grill
museum			ground
neptune			hail
newsstand			hammer
nun			hammock
ocean			hamper
office			handbag
officer			handcuffs
opera			hanger
orchestra			hatchet
outdoors			haystack
owner			heater
painter			helmet
palace			hoe
parent			hood
paris			hook
partner			hoop
party			horizon
passenger			hospital
path			hurricane
patient			hut
patriot			igloo
pavement			incense
pedestrian			inn
people			iron
person			item
philosopher			jacket
piano			jar
picnic			jeans
pirate			jello
planet			jelly
playground			jewel
plumber			journal
pluto			jug
police			keg

(*Continued*)

Big	Small	Animate	Inanimate
politician			kettle
pony			key
pool			keyboard
pope			kitchen
prairie			kite
preacher			kleenex
predator			knapsack
president			knife
priest			knob
primate			knot
prince			labyrinth
prison			lace
producer			lamp
professor			lash
pub			letter
publisher			lightning
queen			linen
radiator			lint
raft			literature
railroad			lock
ram			lodge
rebel			lollipop
receptionist			lounge
referee			luggage
refrigerator			lunch
reindeer			macaroni
resort			magazine
restaurant			magnet
river			mailbox
road			mall
robber			marble
robot			marker
roof			mask
room			mat
roommate			match
runner			mattress
sailor			mayonnaise
salesman			medal
saturn			medication
scientist			medicine
seashore			meteor
senate			microphone

(*Continued*)

Big	Small	Animate	Inanimate
senator			microscope
servant			mirror
shark			missile
shed			mitten
sheep			monument
shelter			moon
shepherd			mop
sheriff			motel
ship			motor
shore			motorcycle
shrine			mug
sibling			nail
sister			napkin
skeleton			needle
slope			net
society			newspaper
soldier			newsstand
spouse			nickel
stable			nicotine
stairs			nightgown
stallion			nitrogen
statue			notebook
store			oboe
stranger			office
stream			ointment
street			ornament
student			outfit
submarine			oval
suburb			oven
sun			pad
supermarket			paddle
supervisor			pail
suspect			paint
sword			painting
tank			palace
tavern			pan
taxi			pants
teacher			paper
team			parcel
technician			passage
temple			pasta
territory			path
tiger			patio
toilet			pavement
tornado			pedal
tower			pen
town			pencil

(*Continued*)

Big	Small	Animate	Inanimate
tractor			penny
traitor			pepper
tree			perfume
tribe			periscope
tricycle			phone
trombone			pick
tunnel			pill
umpire			pipe
uncle			pistol
unicorn			pit
universe			pitchfork
university			plaid
van			plaster
vehicle			plate
venus			plaza
villain			pliers
visitor			pocket
volcano			pocketbook
volunteer			poison
waiter			polyester
waitress			pool
wall			port
walrus			portrait
warehouse			pot
warrior			pottery
waterfall			powder
well			pub
whale			puck
wife			pudding
winner			pump
wolf			puzzle
woman			quill
worker			racket
world			radiator
yacht			radio
yard			raft
zoo			rag
			railroad
			rake
			razor
			receipt
			recipe
			record
			refrigerator
			relish
			report
			restaurant

(*Continued*)

Big	Small	Animate	Inanimate
			rifle
			ring
			road
			robe
			rock
			rocket
			roof
			room
			roost
			ruby
			rum
			saddle
			saloon
			salt
			sand
			sandwich
			sapphire
			saturn
			saucer
			scale
			scalpel
			scissors
			scotch
			screen
			screw
			screwdriver
			scribble
			sculpture
			seat
			shack
			shampoo
			shears
			shed
			shelf
			ship
			shirt
			shoe
			shoelace
			shop
			shortcake
			shovel
			shutter
			sickle
			sidewalk
			siding
			sign
			signature

(*Continued*)

Big	Small	Animate	Inanimate
			sink
			sketch
			ski
			skyscraper
			slacks
			sleeve
			slime
			sliver
			slope
			snack
			snorkel
			soap
			sock
			sofa
			spatula
			spit
			spoon
			stage
			stairs
			stake
			stamp
			stapler
			step
			stereo
			stethoscope
			sticker
			stocking
			stone
			stool
			stove
			straw
			street
			string
			submarine
			suit
			suite
			sunrise
			sunset
			supermarket
			supper
			survey
			swing
			switch
			table
			tack
			tag
			tank

(*Continued*)

Big	Small	Animate	Inanimate
			tape
			taxi
			teapot
			telephone
			telescope
			temple
			thermometer
			thimble
			tie
			tile
			toilet
			tool
			toothbrush
			toothpaste
			torch
			towel
			toy
			tractor
			train
			trash
			tray
			tread
			treasure
			treat
			trench
			triangle
			tricycle
			trophy
			truck
			trumpet
			tub
			tunnel
			twine
			typewriter
			umbrella
			underwear
			uniform
			vacuum
			van
			vehicle
			velvet
			vent
			venus
			vinegar
			viola
			violin

(*Continued*)

Big	Small	Animate	Inanimate
			volleyball
			wagon
			wall
			wallet
			wand
			wardrobe
			wave
			wax
			well
			wheel
			whip
			whistle
			wick
			windshield
			xerox
			yacht
			yarn