Psychological Review

Free Association in a Neural Network

Russell Richie, Ada Aka, and Sudeep Bhatia

Online First Publication, October 6, 2022. http://dx.doi.org/10.1037/rev0000396

CITATION

Richie, R., Aka, A., & Bhatia, S. (2022, October 6). Free Association in a Neural Network. *Psychological Review*. Advance online publication. http://dx.doi.org/10.1037/rev0000396

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review

© 2022 American Psychological Association
ISSN: 0033-295X

https://doi.org/10.1037/rev0000396

THEORETICAL NOTE

Free Association in a Neural Network

Russell Richie[1, 2], Ada Aka[1, 3], and Sudeep Bhatia[1, 3]
[1] Department of Psychology, University of Pennsylvania
[2] Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States
[3] Department of Marketing, The Wharton School, University of Pennsylvania

Free association among words is a fundamental and ubiquitous memory task. Although distributed semantics (DS) models can predict the association between pairs of words, and semantic network (SN) models can describe transition probabilities in free association data, there have been few attempts to apply established cognitive process models of memory search to free association data. Thus, researchers are currently unable to explain the dynamics of free association using memory mechanisms known to be at play in other retrieval tasks, such as free recall from lists. We address this issue using a popular neural network model of free recall, the context maintenance and retrieval (CMR) model, which we fit using stochastic gradient descent on a large data set of free association norms. Special cases of CMR mimic existing DS and SN models of free association, and we find that CMR outperforms these models on out-of-sample free association data. We also show that training CMR on free association data generates improved predictions for free recall from lists, demonstrating the value of free association for the study of many different types of memory phenomena. Overall, our analysis provides a new account of the dynamics of free association, predicts free association with increased accuracy, integrates theories of free association with established models of memory, and shows how large data sets and neural network training methods can be used to model complex cognitive processes that operate over thousands of representations.

Keywords: free association, neural networks, distributional semantics, memory modeling, machine learning

Supplemental materials: https://doi.org/10.1037/rev0000396.supp

In free association, subjects are asked to generate one or more words that come to mind in response to a cue word. For example, given the cue *stork*, a subject might respond *baby*, *egg*, and *mother*, in that order. Due to its simplicity, free association has been one of the most popular tasks in psychology for over a century. Francis Galton used this task as early as 1880, to better understand the nature of his own mental associations (Galton, 1880). William Wundt and his collaborators built on Galton's ideas to study response times and underlying cognitive operations in experimental participants (Cattell, 1887). Of course, free association is itself most associated with Sigmund Freud, who used the task to examine patients' thoughts free of censorship (Freud, 1913/1958). More recently, free association has emerged as a leading method in cognitive psychology and cognitive science, particularly in the study of mental representation and memory, and its relationship with language (Clark, 1970; Deese, 1959, 1962). In the 21st century, psychologists have collected large data sets of lexical free association norms (De Deyne et al., 2019; Nelson et al., 2004) and have used these norms to study phenomena such semantic organization (Steyvers & Tenenbaum, 2005), lexical access (De Deyne et al., 2013), similarity judgment (De Deyne et al., 2016), semantic memory search (Abbott et al., 2015), cued recall (Nelson et al., 1997), recognition memory (Nelson et al., 1998), visual word recognition (Balota et al., 2004), creativity (Kenett et al., 2014), cognitive development and aging (Dubossarsky et al., 2017), and much more.

While free association data have been useful in understanding various cognitive phenomena, free association is itself fundamental and ubiquitous, arguably underlying much of cognition and behavior in the wild. For this reason, free association deserves an explanation of its own (see Nelson et al., 2000, for a compelling argument). That is, free association needs a model, ideally specific enough to be implemented computationally and trained on free association data, and general enough to predict participant responses to novel cues (i.e., cues for which there are no training data). However, modeling free association presents some difficult challenges, as free association involves complex memory processes operating over thousands of semantically rich words and concepts. Thus, a computational model of free association requires specifying (a) representations for the many words and concepts that people may know about and (b) retrieval processes which operate over these representations to generate responses in free association tasks.

Sudeep Bhatia https://orcid.org/0000-0001-6068-684X

Funding was received from the National Science Foundation Grant SES-1847794. Versions of this article were presented at the Psychonomic Society Conference and the Context in Episodic Memory Symposium. Code and trained weights can be downloaded from https://osf.io/as26x/. This study was not preregistered.

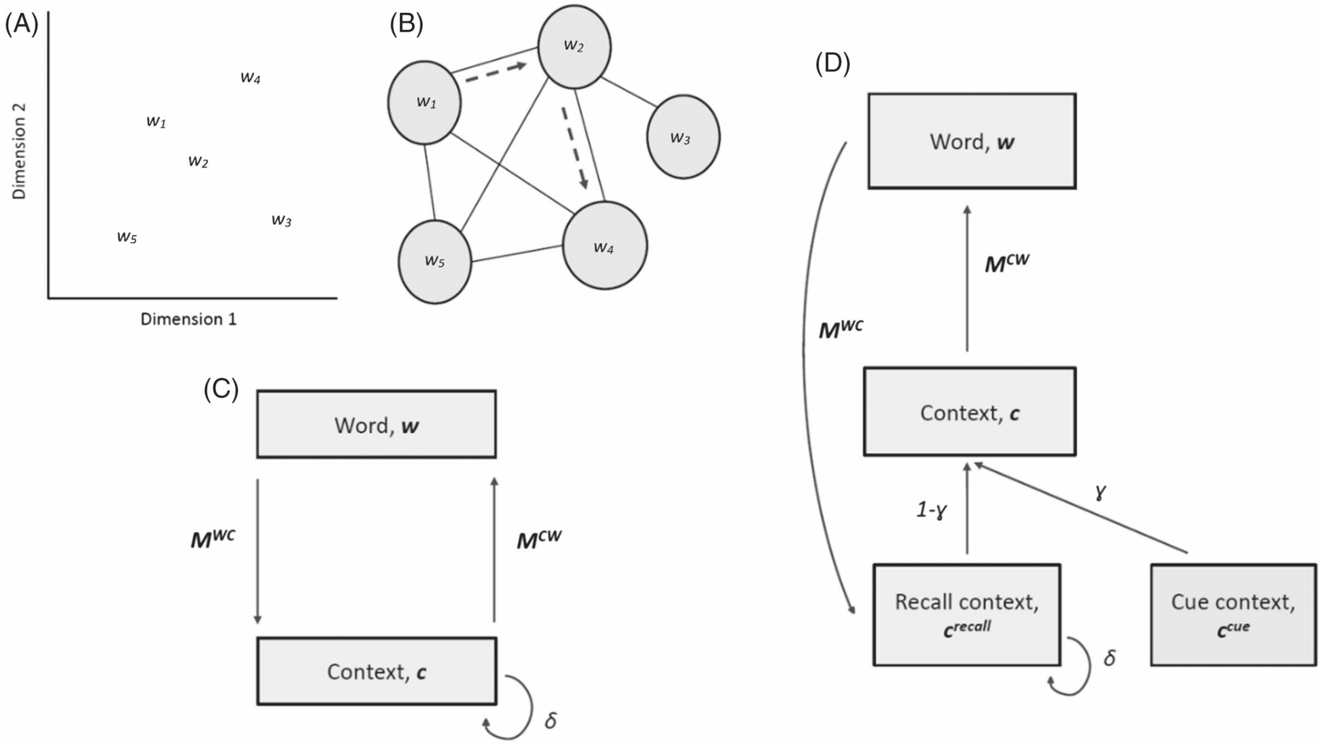
Correspondence concerning this article should be addressed to Sudeep Bhatia, Department of Psychology, University of Pennsylvania, 3720 Walnut Street, Philadelphia, PA 19104, United States. Email: bhatiasu@sas.upenn.edu

Researchers have made progress on (a) by using distributed semantics (DS) models (illustrated in Figure 1A), which exploit statistics of word use in large collections of text to derive semantic representations of words in the form of real-valued vectors (Howard et al., 2011; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mikolov et al., 2013; Pennington et al., 2014; for reviews, see Bhatia et al., 2019; Lenci, 2018; or Günther et al., 2019). These, and related models based on corpus statistics (e.g., Chaudhari et al., 2011; Ji et al., 2008; Matusevych & Stevenson, 2018; Peirsman & Geeraerts, 2009), can predict associations between pairs of words (e.g., *stork* and *baby*) by using the co-occurrence frequencies and the absolute frequencies of the words in language (Griffiths et al., 2007; Jones et al., 2018; Nematzadeh et al., 2017; Pereira et al., 2016). However, the combination of these representations with cognitive process models of memory—that is, requirement (b) from above—has been limited. This is why such models are unable to describe continued-response dynamics in free association and cannot easily predict how later responses (e.g., *mother*) depend on the combined effect of the cue (e.g., *stork*) and responses recalled earlier in the trial (e.g., *baby* and *egg*).

Researchers have also made progress on (b) by using semantic network (SN) models (Abbott et al., 2015; De Deyne & Storms, 2008; De Deyne et al., 2013, 2016, 2019; Dubossarsky et al., 2017; Kenett et al., 2014; Kumar et al., 2019; Steyvers & Tenenbaum, 2005; illustrated in Figure 1B). Typically, SN models of free association measure the connection strengths between pairs of words using response frequencies in empirical free association data. Random walks on these networks (with transition probabilities proportional to connection strength) are subsequently used to describe retrieval processes, and sometimes describe asymmetry, clustering, and response sequences in free association data. Again, however, SN models leave some gaps in our understanding of free association. First, since their transition probabilities are based entirely on observed free association data, these networks cannot be easily used to make out-of-sample predictions (e.g., predictions for trials with new cues). In this way, these networks fail to solve requirement (a) outlined above, that is they do not provide (a priori) representations for the thousands of words and concepts over which free association processes can operate. Second, random walks on semantic networks typically satisfy the Markov property and assume that the retrieval probability of a word depends only on the previously recalled word. Although Markov models can generate long sequences of semantically related words (*stork* can activate *baby* which can activate *egg*, which can activate *mother*), the Markov property implies that once we know the word retrieved at time $t - 1$ (in our example, *egg*) no further improvements in our predictions of the word retrieved at t (*mother*) can be made by knowing the word at $t - 2$ (*baby*) or the cue word (*stork*). This property has been shown to be violated in human memory, as the cue and the previously retrieved

Figure 1

Models of Free Association



Note. Free association modeled using distributed semantics (A) or as a random walk on a semantic network (B). Architecture of the conventional CMR model for free recall from lists (C) and our modified CMR model for free association (D). M^{WC} and M^{CW} describe word-to-context and context-to-word connection weights, respectively, δ is a context feedback term, and γ captures the persistent effect of the cue. Our modified CMR model mimics distributed semantics models when $M^{WC} = M^{CW}$ and $\gamma = 1$, random walk models when $\gamma = \delta = 0$, and the CMR model for free recall from lists when $\gamma = 0$. CMR = context maintenance and retrieval.

responses have a significant effect on subsequent recall (Kahana & Caplan, 2002; Lohnas & Kahana, 2014; Posnansky, 1972). Note that some researchers have combined DS and SN models by equipping semantic networks with measures of association strength obtained from distributed semantic representations (Gruenenfelder et al., 2016; Rotaru et al., 2018; Utsumi, 2014). Random walks on these networks can predict out-of-sample participant responses in free association (Kumar et al., 2020), though they cannot capture violations of the Markov property in response chaining.

To help address these issues, we draw inspiration from models developed for a task closely related to free association: free recall from lists. In this task, subjects are given a list of words to study, and then asked to recall as many words from the list as they can, in any order that comes to mind. Due to its simplicity, the free recall from lists task is widely used to study memory retrieval and search (e.g., Atkinson & Shiffrin, 1968; Bousfield, 1953; Deese, 1959; Murdock, 1962; Tulving, 1962). Subjects in this task consistently demonstrate a range of effects, the most relevant for us being semantic clustering (adjacent recalled words tend to be semantically related; Bousfield, 1953; Romney et al., 1993), asymmetric response probabilities (words are more likely to cue the retrieval of following, rather than preceding, words in the list; Kahana, 1996), and compound cueing (retrieval probabilities of words depend on words recalled much earlier, a violation of the Markov property; Lohnas & Kahana, 2014).

A host of computational models have been developed to account for these dynamics (see Kahana, 2020, for review), but we take special interest in models based on retrieved context theory. This class of models, best exemplified by the temporal context model (TCM; Howard & Kahana, 1999) and context maintenance and retrieval model (CMR; Polyn et al., 2009), are recurrent neural networks with a layer for words and a layer for an evolving "context" representation (see Figure 1C). Recalling a word activates its corresponding node on the word layer, which sends activation to the context layer (i.e., retrieves the state of context at the time the word was studied), and the context layer in turn sends activation back to the word layer, cuing the next word for recall. Because the context-to-word and word-to-context weights are initialized with semantic similarity measurements from DS models, this process tends to produce a chain of semantically related words during retrieval. Further, the context-to-word weights can be different to the word-to-context weights, allowing for both temporal and semantic asymmetries in recall. Finally, the context encodes a (potentially decayed) representation of previously retrieved words as well as task-related variables such as cue words, allowing the CMR model to capture dynamics in free recall from lists, including cue-dependence and violations of the Markov property.

In this article, we develop a variant of the CMR model (illustrated in Figure 1D), which uses existing DS representations to specify word-to-context weights, but fits context-to-word weights flexibly to free association data. We train our network using stochastic gradient descent on a large free association data set (De Deyne et al., 2019) and test its ability to predict out-of-sample response sequences in free association.

Theoretical Background

Cognitive Process Models of Recall

Many influential cognitive models of memory search have been developed using the free recall from lists paradigm (see Kahana, 2020, for a review). In this task, participants study a list of words and later attempt to remember as many words as they can. There are numerous reliable empirical patterns in free recall from lists. These include serial position effects such as the primacy and recency effect (Murdock, 1962), as well as temporal contiguity effects, in which subjects cluster recalls based on their temporal encoding order (Kahana, 1996). Temporal contiguity effects involve asymmetric retrieval probabilities, as words usually cue the retrieval of following, rather than preceding, words in the presented list. Semantic similarity has also been shown to influence the output order and latency of recalled words (Bousfield, 1953; Romney et al., 1993). Words that are semantically similar to one another are more likely to be recalled in neighboring positions, generating semantic clustering in recall data. Finally, free recall from lists involves long-term retrieval dynamics, with retrieved words influencing many subsequent recalls (Kahana & Caplan, 2002; Lohnas & Kahana, 2014; Posnansky, 1972).

The earliest models of free recall assumed that participants associated words with neighboring positions in the list, resulting in a network representation of the list. Recall in these "associative chain" models took the form of a random walk over the network (Ladd & Woodworth, 1911). Associative chain models of free recall are closely related to the semantic network models discussed below, but have fallen out of favor due to their inability to account for empirical findings, such as error and intrusion patterns in serial recall (Kahana, 2020; Osth & Hurlstone, 2022). Instead, modelers have favored more complex models that make nuanced assumptions about storage, retrieval, as well as semantic and episodic context. For example, dual-store memory search models distinguish between two separate memory stores, the limited-capacity short-term store (STS) and the long-term store (LTS; Atkinson & Shiffrin, 1968; Raaijmakers & Shiffrin, 1981). According to the search of associative memory (SAM) model, a leading dual-store memory model, words in the STS are readily available for recall, whereas LTS is searched with the cue, and words are probabilistically sampled based on the strength of association between the cue and the target word.

Retrieved context theory (Howard & Kahana, 1999, 2002a) is another approach to modeling memory, that is built on ideas initially introduced by dual-store models. According to retrieved context theory, remembering a word brings back its encoding context, which in turn serves as a retrieval cue for subsequent recalls. Here, we will concentrate on a particular model, the CMR (Polyn et al., 2009), a generalized version of earlier retrieved context models. The CMR model includes two representational structures: a word layer and a context layer (Figure 1C). Two sets of weights connect these layers and characterize the strengths of the word-to-context and the context-to-word associations. If a new word is activated, the activation on the word layer is used to update the context layer and update the existing information in context. The new context activation is a linear combination of the previous context activation and the input from the word layer, which depends on the word-to-context connection weights. During the retrieval phase, context serves as a retrieval cue, so that the context activation determines the new word activation based on the context-to-word connection weights.

The connection weights are updated through a Hebbian learning process, the details of which are too complex to be described here, and are largely irrelevant to our own implementation of CMR, which models only associative retrieval, and not learning. However, it is important to note that this updating process allows for (pre-experimental) semantic similarity relationships to influence associations. Typically, this is done by initializing the connection weights with word vector representations from latent semantic analysis (Landauer & Dumais, 1997) or other similar distributed semantics models. This allows CMR to explain semantic clustering effects in free recall from lists, as words in context cue the retrieval of other semantically related words (Healey & Uitvlugt, 2019; Howard & Kahana, 2002b; Morton & Polyn, 2016; Polyn et al., 2009; Socher et al., 2009). Additionally, the Hebbian learning process responsible for updating the connection weights causes items presented in nearby positions during learning to be more strongly associated with each other, thus explaining temporal contiguity effects. Importantly, these effects are asymmetric, with recall transitions in the forward direction being more likely compared to those in the backward direction. In other words, the degree to which word *i* cues word *j* is not the same as the degree to which word *j* cues word *i* (see Howard & Kahana, 2002a; Kahana, 1996; and Polyn et al., 2009, for more details).

The retrieval mechanisms of CMR also allow for complex response dynamics. As discussed above, activating a word updates context. This updating combines the previous context activation with the new word activation. In this way, context evolves as new words are activated, and the effect of previously recalled words decays over time. However, this decay is partial, implying that words activated or recalled at t can influence context at $t + 2$ or later, violating the Markov property. Indeed, observed recall dynamics in free recall from lists have been shown to violate this property in the manner predicted by the CMR model (Kahana & Caplan, 2002; Lohnas & Kahana, 2014; Posnansky, 1972).

Overall, the CMR model successfully captures many patterns in free recall from lists. Thus, it is possible that this model could also be useful for describing free association. The tasks of course share an essential similarity—in both tasks subjects provide unstructured sequences of lexical responses drawn from memory and cued by one or more words presented or recalled earlier in the trial. Models of free association may therefore benefit from elements of free recall models, like CMR's asymmetry between word-to-context and context-to-word connection weights, and in the case of continued free responding, an evolving context layer reflecting recent and older responses.

However, while it is easy to see the similarities between the free recall from lists and free association, the two tasks do differ in key aspects. First, the responses participants give in free association tasks are much less constrained, being drawn from more or less the entire lexicon, while in free recall from lists participants attempt to remember a small set of studied items. Thus, unlike previous applications of CMR, in which the set of recallable items is limited to an experimenter-specified list, a model of free association must allow for nearly arbitrary responses. This poses several technical challenges, due to the complexity in the underlying structure of CMR, as well as the complex dynamics known to be at play in free association. Second, the continued free association task directs subjects to continue responding to a cue, whereas the free recall task asks participants to recall as many words as they can, in any order that comes to mind. In free association then, early responses and later responses may both be semantically related to the cue in some way. This necessitates modifications to CMR's context representation that can allow for a persistent effect of the cue word (in addition to the effect of the cue word at the start of retrieval).

Previously, Howard et al. (2011; also see Shankar et al., 2009) have attempted to address the first challenge by fitting a variant of the CMR model, the predictive temporal context model (pTCM), on natural language data. They have shown that pTCM can recover word-to-context and context-to-word associations for thousands of words that exceed the predictive accuracy of some distributed semantics models (discussed in the section below) for a variety of tasks, including free association. Another closely related model, the syntagmatic paradigmatic (SP) model, has been put forth by Dennis (2005). The SP model is a general account of verbal cognition based on associations that reflect both direct word co-occurrence as well as relational similarity. Crucially, it is capable of being trained on natural language data and has been shown to account for several findings in free recall. To our knowledge, the pTCM and SP models have not been used to explain nuanced dynamics of free association, such as asymmetric retrieval probabilities and response chaining effects. However, the results of Howard et al. (2011) and Dennis (2005) do demonstrate the promise of combining memory models (like CMR) with word representations obtained from language, for modeling free association. In this way, these models serve as theoretical steppingstones for our own work. In the discussion section of this article, we examine how the assumptions of these models can be used to extend our framework to more realistically capture the processes at play in word learning, sentence processing, and reasoning.

Distributed Semantics Models

While computational models of free recall from lists provide inspiration for a model of the operations or processes at play in free association, we still require a model of the representations over which those operations work. For this, we turn to distributed semantics (DS) models, which are a class of models that derive vector representations for the meanings of words, based on statistics of word–document or word–word co-occurrence in large collections of texts (Griffiths et al., 2007; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mikolov et al., 2013; Pennington et al., 2014; for reviews, see Lenci, 2018; Bhatia et al., 2019; or Günther et al., 2019). Perhaps the most well-known method in this class (at least to psychologists) is latent semantic analysis (LSA), in which a word–document co-occurrence matrix is built, and then reduced with singular value decomposition (Landauer & Dumais, 1997). Words with similar distributions in text end up with similar vectors through LSA, and thus similarity or relatedness judgments between words can be captured reasonably well through distance metrics like cosine between pairs of word vectors. Other methods for building word vectors include BEAGLE (Jones & Mewhort, 2007), GloVe (Pennington et al., 2014), Word2Vec (Mikolov et al., 2013), as well as the predictive TCM (Howard et al., 2011) described above. Whatever the approach, such vectors have proven quite successful in psychological applications beyond just similarity and relatedness judgments. Distance metrics like cosine predict strength of semantic priming in, for example, the lexical decision task, as measured by reaction times (Jones et al., 2006; Mandera et al., 2017), and have been used in models like CMR to account for semantic clustering in free recall from lists (Howard & Kahana, 2002b; Polyn et al., 2009).

Semantic judgments about words (e.g., the tastiness of a food) can also be approximated by calculating the relative vector similarity of a judgment target (e.g., *apple*) to words high (e.g., *delicious*, *tasty*) and low (e.g., *disgusting*) on a judgment dimension (Grand et al., 2022; Richie et al., 2019). However, even better (out-of-sample) approximations of semantic judgments can be made by directly regressing human ratings for a semantic dimension onto the vectors for judgment targets (Bhatia, 2019; Bhatia et al., in press; Gandhi et al., 2022; Hollis et al., 2017; Richie et al., 2019; Utsumi, 2020; Zou & Bhatia, 2021; see Snefjella & Blank, 2020, for a comprehensive list of "semantic norm extrapolation" studies, as well as caveats thereof). The advantage of this approach is that it allows for human data to directly supervise the setting of flexible weights on the attributes of the target representation. These weights, in a sense, provide a better model of the relationship between the judgment target and the judgment dimension than do hand-selected words for the judgment dimension. We will suggest taking a similar approach in the next section when building models of free association.

Further, because DS models exploit co-occurrence information in text, they can proxy the associations between words. For this reason, DS models have been used to study associations in probability judgment (Bhatia, 2017a) and social judgment (Bhatia, 2017b; Bhatia et al., 2018; Caliskan et al., 2017). Finally, as mentioned in the opening of this article, free association can and has been modeled with DS models. Perhaps the most straightforward way to model free association would be to simply correlate the probability of recalling one word (R) given another (C), $p(R|C)$, with a measure of similarity, like cosine similarity, between the word vectors for the response and the cue (Figure 1A). However, as many have pointed out over the years (e.g., Griffiths et al., 2007; Jones et al., 2018; Tversky, 1977; but also see Kintsch, 2014), cosine and related distance metrics are inherently symmetric, and thus cannot account for asymmetric associations, which by some estimates constitute over 85% of all pairs of associated words in norms (Griffiths et al., 2007). Instead, researchers have evaluated DS models on free association data in various ways that can handle asymmetries. For example, Pereira et al. (2016), who evaluated a variety of DS models ranging from LSA to Word2Vec on the well-known University of South Florida (USF) free association norms (Nelson et al., 2004), calculated the percentage overlap between the top 50 most frequent responses to a cue, and the top 50 most similar words to the cue by cosine similarity. Such an approach can handle asymmetry between words when they have neighborhoods of varying semantic density. For example, *baby* could be closer to *stork* than most other words, but the opposite may not be true if words such as *infant* and *mother* are in the immediate neighborhood of *baby*. Thus $p(baby|stork)$ can be higher than $p(stork|baby)$.

Note that Pereira et al. (2016) were not building a cognitive model of free association (nor did they claim to). This is important, because as Jones et al. (2018, p. 2) put it, "a cosine is not what people do in a task." In other words, cosine is not a process model of how people use representations (modeled with word vectors) to respond in a task like free association. To build such a process model, Jones et al. (2018) combined cosine similarity of word vectors with a version of the Luce (1959) choice rule, to simulate the probability of a target response to a cue. Crucially, Jones et al. assumed a minimum similarity threshold parameter, such that only words exceeding this minimum (i.e., within a certain radius of the cue vector) are considered during response generation. This model of choice can also account for asymmetries in recall when words have semantic neighborhoods of varying density. Consider again the case of s*tork-baby*. When *baby* is the cue, there are many competitors to *stork* whose similarity is higher than the minimum threshold (e.g., words like *infant* or *mother*), leading to a low $p(stork|baby)$. On the other hand, when *stork* is the cue, there are far fewer words whose similarity is higher than the minimum threshold, leading to a high value of $p(baby|stork)$.

While Jones et al. (2018) did not directly assess their model of $p(R|C)$ on empirical data, they did evaluate the ability of the choice rule to correctly predict the direction of asymmetry in word pairs in the USF norms (Nelson et al., 2004) whose asymmetry ratio $p(R|C)/p(C|R)$ was greater than 10 or less than .1. Selecting the similarity threshold leading to the best performance, they found that the choice rule applied to BEAGLE vectors could predict 80% of the asymmetries.

A similar approach was taken by Nematzadeh et al. (2017), work we pay special attention to as our models can be viewed as more general cases of theirs, and some of our evaluation metrics are based on theirs. Following work that provides a probabilistic interpretation of Word2Vec (Arora et al., 2016; Levy & Goldberg, 2014), Nematzadeh et al. (2017) computed $p(R|C)$ with a softmax transformation of the dot products of the cue and response word vectors. Thus, responses with a higher dot product with the cue, relative to other responses, were more likely to be recalled. As with the approaches in the previous two paragraphs, Nematzadeh et al.'s model can generate asymmetric response probabilities for words with neighborhoods of varying semantic density, as the calculation of relative exponentiated dot products in the softmax transformation leads to a nonlinear penalty for relatively dissimilar responses. Again, consider the case of s*tork-baby*. There are many competitors to *stork* which have high dot products with *baby*, but few competitors to *baby* which have high dot products with *stork*, leading to a higher $p(baby|stork)$ than $p(stork|baby)$.

Nematzadeh et al. (2017) combined this model with Word2Vec or GloVe embeddings trained on one of three corpora of varying size and used three metrics to evaluate performance of this model. First, they simply calculated the Spearman correlation between their model's prediction of $p(R|C)$ with the empirical $p(R|C)$ in the USF norms, finding that GloVe trained on the largest available corpus achieved the highest correlation, of $r = .27$. Second, for each cue, they calculated the median rank of its most common response in a choice model's predicted top responses (and likewise for the second- and third-most common responses). Once again, GloVe trained on the largest corpus outperformed other models, with median ranks of 11, 25, and 40.5 for the first, second, and third associates (which is impressive given that the number of possible responses they considered was 3,951). Finally, to assess their model's ability to capture asymmetries in association, they first calculated the empirical asymmetry ratio $p(R|C)/p(C|R)$ and their model's predicted asymmetry ratio, for each (cue, response) pair in the USF norms. They then computed the correlation between empirical and predicted asymmetry ratios across all (cue, response) pairs. GloVe trained on the largest corpus again performed well with $r = .48$. Following Nematzadeh et al. (2017), we will hereafter refer to these metrics as "p(R|C) correlation," "median rank of true associates," and "asymmetry ratio correlation." Note that Nematzadeh et al. (2017)

also evaluated topic models of free association (Griffiths et al., 2007). They found that topic models were outperformed by GloVe on the first two metrics, though they did marginally better than GloVe on the asymmetry ratio correlation metric ($r = .49$). Topic models are not as related to our CMR-based modeling framework, which is why we do not discuss them in detail here. However, we do introduce these models and analyze their predictions in our Supplemental Materials.

Semantic Network Models

There is another type of model that is relevant to the study of free association: the semantic network (Collins & Loftus, 1975; Collins & Quillian, 1969). In free association, nodes in semantic networks correspond to words and (directed or undirected) edges represent association strength (Figure 1B). As discussed above, similar "associative chaining" models were popular early theories of free recall from lists (Ladd & Woodworth, 1911), though they were later supplanted by more complex process models that were better able to describe nuanced retrieval dynamics observed in experimental research. Despite (or perhaps because of) their simplicity, semantic network models have remained popular in the study of free association, where they are often used to formally characterize response patterns in the empirical data (De Deyne et al., 2019). Semantic networks fit on free association data have also been used to study semantic organization (De Deyne & Storms, 2008; Steyvers & Tenenbaum, 2005) as well as differences in semantic organization as a function of development (Dubossarsky et al., 2017). Other researchers have used distance measures on variants of these networks (trained on free association data) to predict semantic similarity and lexical processing (De Deyne et al., 2013, 2016, 2019; Kenett et al., 2017; Kumar et al., 2019). More relevant to this article is work that uses random walks on semantic networks to describe memory search dynamics in semantic fluency tasks (Abbott et al., 2015). This approach derives transition probabilities between words from free association data, and uses these probabilities to make predictions for recall sequences composed of multiple words (see also Hills et al., 2012; Zemla & Austerweil, 2018; Zemla et al., 2020, for related applications).

Connections in semantic networks can also be obtained from DS models, using, for example, the cue–response probability formulas proposed by Jones et al. (2018) and Nematzadeh et al. (2017). A number of papers have used this approach to better understand the implications of distributed semantic models for semantic organization (Gruenenfelder et al., 2016; Rotaru et al., 2018; Utsumi, 2014). Most recently, Kumar et al. (2020) have applied these types of semantic networks to free association and have shown how the combination of DS and SN modeling allows for both accuracy and generalizability in predicting cue–response probabilities. We will be adopting a similar approach below, though we will embed word vectors in a richer cognitive process model of memory and additionally modify word vector representations based on free association data.

Modeling Framework

Overview

Our goal is to use a leading model of memory search in free recall, the CMR model (Howard & Kahana, 1999, 2002a; Polyn et al., 2009), to characterize retrieval dynamics in free association. There are two reasons why we believe this is necessary. First, as we discuss in the introduction, existing distributed semantics and semantic network models are unable to describe complex dynamics of free association, particularly in continued free association. A model of memory search that has been shown to be successful in capturing the dynamics of free recall from lists may be able to address these limitations. Second, we believe that it is desirable to have a single general model of memory search that can be applied across domains (i.e., to both free recall from lists and to free association). The development of such a model can facilitate theoretical cohesion and potentially lead to new insights regarding the cognitive processes at play in memory search. This has the potential to improve our understanding of free association, as well as our understanding of free recall from lists. In fact, we will also be testing our model of free association on free recall data to evaluate its capacity for cross-domain predictions.

To develop a single general model, however, we will have to introduce a generalization of CMR that, under certain parameterizations, can handle idiosyncrasies of free association, particularly the possibility of a persistent, nondecaying effect of the cue on recall. In particular, in conventional CMR (Figure 1C), there are two directly interacting layers: a word layer, w, and a context layer, c. In our modified CMR (Figure 1D), however, w does not directly influence c. Instead, c depends on earlier layers c^{cue} and c^{recall}, which represent the cue and the previously recalled responses respectively. It is c^{recall} that is directly influenced by w. We focus below on describing this version of CMR and refer readers to Polyn et al. (2009) for more in-depth description of conventional CMR.

Recall Dynamics

We first assume a word layer, w, which represents words using individual nodes (i.e., implements localist or one-hot coding). In a setting with N recallable words, the word layer thus consists of N nodes. We write the $N \times 1$ dimensional activation vector of the word layer at time t as w_t with $w_{i,t}$ describing the activation of word i at t. By contrast, we assume that the context layer implements a distributed representation, with dimensionality M. We write the $M \times 1$ dimensional activation vector in the context layer at time t as c_t.

The context layer determines activation on the word layer through an $N \times M$ dimensional matrix of connection weights M^{CW}. We assume that activation is linear, so that:

$$w_t = M^{\text{CW}} \cdot c_t. \tag{1}$$

Activation, in turn, drives recall. We assume, for simplicity, that this happens via a softmax transformation of activation. Thus, the probability of recalling word i at time t is:

$$p_{i,t} = \frac{\exp(w_{i,t})}{\sum_{j=1}^{N} \exp(w_{j,t})}. \tag{2}$$

We specify the specific response recalled at time t using b_t^{recall}, which is an Nx1 dimensional binary vector with $b_{i,t}^{\text{recall}} = 1$ if word i is recalled at t, and $b_{i,t}^{\text{recall}} = 0$ if not. In this way, b_t^{recall} is a random variable drawn from the categorial distribution with probabilities p_t.

Activation on the context layer at time t, c_t, depends on both the cue, as well as the set of items recalled prior to t. For simplicity, we separate these two sets of variables into two context vectors, c^{cue} and

c_t^{recall}, which are combined additively, with parameter γ to determine c_t. Thus, we have the following:

$$c_t = \gamma \cdot c^{\text{cue}} + (1-\gamma) \cdot c_t^{\text{recall}}. \tag{3}$$

c_t^{recall} is a dynamically evolving context vector that updates after each recall. Following CMR, we assume that this update is a linear combination of the context at $t-1$ and the input generated by the word recalled at $t-1$ via the $M \times N$ dimensional matrix of connection weights M^{WC}. The weight used in this linear combination is proportional to self-feedback δ. Thus:

$$c_t^{\text{recall}} = \delta \cdot c_{t-1}^{\text{recall}} + M^{\text{WC}} \cdot b_{t-1}^{\text{recall}}. \tag{4}$$

c^{cue} captures the persistent effect of the cue on context. We assume, for simplicity, that c^{cue} also depends on the connection weights M^{WC}. We can write the cue in a given trial as b^{cue}, which is an $N \times 1$ dimensional binary vector with $b_i^{\text{cue}} = 1$ if word i is the cue and $b_i^{\text{cue}} = 0$ if not. This implies that:

$$c^{\text{cue}} = M^{\text{WC}} \cdot b^{\text{cue}}. \tag{5}$$

Finally, so that we can distinguish a persistent effect of the cue from an initial, possibly decaying effect of the cue, we assume that b_0^{recall} (the implicit zeroth recall) is simply b^{cue}, which implies that the context vector at the start of the trial is equal to the cue vector, that is $c_1 = c^{\text{cue}}$. In other words, the first recall is determined only by the cue (though subsequent recalls do depend on retrieved words).

Connection Weights

The above equations provide us with a model of recall dynamics in free association, in which the cue initiates recall, and combines with retrieved words to guide subsequent recall. The words that are recalled in a given trial depend on the matrix of connection weights, M^{WC} and M^{CW}. In the application of CMR to free recall from lists, these connection weights are initialized using word vector models. Thus, for example, a word vector model with 300-dimensional representations for words would involve an $M = 300$ dimensional context layer, with the rows of M^{WC} and columns of M^{CW} based on the word vector representations of the N recallable words. Of course, to accommodate list-specific effects on recall, these matrices are further updated based on presented words in each list learning trial. The only flexible parameters in this approach (besides the parameters at play in recall) are those that determine the effect of the presented list on the connection weights (e.g., learning rates in Hebbian updating; see Polyn et al., 2009, for further details).

Our application of CMR to free association adopts a similar approach, by initializing M^{WC} and M^{CW} using word vectors, but in the absence of a presented list, updates M^{CW} based on observed free association data. In this way, it considers the M^{CW} weights to be free parameters. Of course, unlike the parameters of the original CMR model, successfully updating the tens of thousands of entries in M^{CW} requires a very large data set. Fortunately, the Small World of Words project (De Deyne et al., 2019) offers such data set. It contains over one million trials in which participants provide continued responses to over 10,000 cues. We fit M^{CW} to these trials using stochastic gradient decent. Note that such a modeling exercise involves considerable flexibility, and for this reason, we evaluate our fits using cross validation. In particular, we train M^{CW} on a subset of cues in the SWOW data set, and test it on the held out cues. Doing so allows us to evaluate the generalizability of our best-fit M^{CW} and ensure that high accuracy rates are not a product of our model's flexibility. Of course, readers should be cautious in extrapolating our model to other tasks and data sets, as our best-fit model could reflect the peculiarities of the free association task, and the SWOW data set in particular. It is also worth noting here that despite our use of cross validation, our model is more flexible than previous distributed semantics approaches (e.g., those in Nematzadeh et al., 2017), which do not use any free association data to specify and constrain their representations.

Although it is theoretically possible for free association data to fully supervise the learning of weights and biases in our model (i.e., update M^{CW} as well as M^{WC}), this would require an extraordinary amount of data (and computation time), given the large number of parameters in this model. Thus, as a practical matter, it is useful to initialize the weights from the cue layer to the response layer with pretrained word vector representations and then let free association data guide departures from these already useful initial estimates. Additionally, although we choose to update M^{CW}, and keep M^{WC} fixed during our training exercise, we could do the opposite, with little effect on model performance.

Implementational Details

We use GloVe vectors to initialize M^{WC} and M^{CW}. GloVe builds a global word–word co-occurrence matrix whose entries encode the frequency with which words co-occur with one another in a given corpus. Vectors are then learned such that their dot product equals the logarithm of words' probability of co-occurrence. The GloVe vectors used in our analysis were pretrained on 6 billion tokens of a 2014 dump of Wikipedia as well as the GigaWord corpus (Pennington et al., 2014) and were shown to perform well in Nematzadeh et al.'s (2017) study of free association.

We fit two variants of the above model to data. Our first variant predicts the first word generated in response to a cue. In this setting, there are no previous recalls to influence context, and thus response probabilities are given entirely by the cue. The output of this model is a softmax transformation of activations $w_1 = M^{\text{CW}} \cdot c^{\text{cue}} \cdot c^{\text{cue}}$ is given by the initialized M^{WC} and is simply the GloVe vector for the cue. M^{CW} is initialized using GloVe and is further updated based on observed cue–response pairs. This updating process is implemented using gradient descent.

Our second variant predicts a K-length sequence of words generated in response to a cue. In a trial with $K = 3$ (as with the data set introduced below), there are three outputs generated by the model. The first output is the probability vector for the first response, contingent on the cue, and (as in the above paragraph) is a softmax transformation of activations $w_1 = M^{\text{CW}} \cdot c^{\text{cue}}$ generated by the cue. The second output is the probability vector for the second response, contingent on the cue and the observed first response. This is a softmax transformation of activations $w_2 = M^{\text{CW}} \cdot c_2$, with c_2 given by $c_2 = \gamma \cdot c^{\text{cue}} + (1-\gamma) \cdot c_2^{\text{recall}}$ and c_2^{recall} given by $M^{\text{WC}} \cdot b_1^{\text{recall}}$. Here b_1^{recall} is not a random variable, but rather the known outcome of the first recall. Finally, the third output is the probability vector for the third response, contingent on the cue and the observed first and second responses. This is a softmax transformation of activations $w_3 = M^{\text{CW}} \cdot c_3$, with c_3 given by $c_3 = \gamma \cdot c^{\text{cue}} + (1-\gamma) \cdot c_3^{\text{recall}}$, c_3^{recall} given by $(1-\delta) \cdot c_2^{\text{recall}} + M^{\text{WC}} \cdot b_2^{\text{recall}}$, and c_2^{recall} given by $M^{\text{WC}} \cdot$

b_1^{recall}. Again b_1^{recall} and b_2^{recall} are not random variables, but rather the known outcomes of the first two recalls. To ensure context c maintains constant magnitude over time, we L2 normalize c^{cue} and c_t^{recall} after each time step, and L2 normalize all GloVe vectors before training (see Polyn et al., 2009, for a more complex normalization scheme in CMR). Note that we do not apply any L2 normalization in the first-response variant, as it does not involve a changing context representation over time.

M^{CW} is initialized using GloVe and is further updated using stochastic gradient descent on observed data. Note that before updating, the data is transformed by separating the three observed responses in each trial into three separate observations: $(R_1|C)$, $(R_2|C, R_1)$ and $(R_3|C, R_1, R_2)$, so that the total number of observations used to train the model is $K{\cdot}T$ where T is the total number of trials (and K is, again, the number of responses generated to each trial). Also note that this updating process also involves flexible parameters γ and δ, which interact with M^{CW}. To ensure that gradient descent operates on a linear model, we manually search through values of γ and δ in the set $\{0, .1 \ldots .9, 1\}$, and fit M^{CW} separately for each γ and δ combination. Thus, for example, we start by setting $\gamma = 0$ and $\delta = 0$ and use gradient descent to find the best-fitting M^{CW} for these values of γ and δ. We then repeat this with $\gamma = 0$ and $\delta = .1$, $\gamma = 0$ and $\delta = .2$, and so forth, until we have evaluated all 121 of the possible combinations of γ and δ.

The above steps are repeated for a constrained model that sets M^{CW} to be identical to (the transpose of) M^{WC}, and whose weights do not need to be trained. We refer henceforth to the unconstrained model as the model with asymmetric weights and the constrained model as the model with symmetric weights. Both the first-response and continued-response models (and their constrained counterparts) are evaluated on held out cues, that is, cues that are completely absent from the training data. Finally, all models are trained in Keras (Chollet, 2015), using used stochastic gradient descent with a categorical cross-entropy loss function, and with a learning rate of .01, decay of 10^{-6}, and (Nesterov) momentum of .9. We do not fine-tune any hyperparameters, and it is likely that model performance will improve with alternate training procedures. Training is subject to early stopping if 10 epochs pass with no improvement. Because participants only very rarely respond to a cue with itself, we disallow self-cuing in all our models. To do this, after all models have been trained and predictions generated, we set the probability of responding to a cue with the same cue to 0 (and rescale the remaining response probabilities to sum to one).

Training and Test Data

We trained and tested our models on free association data from the English Small World of Words project (SWOW-EN, De Deyne et al., 2019). This is the largest set of English free association norms collected to date, with 12,292 unique cues and 100 trials per cue, from over 90,000 participants varying in age ($M = 36$ years, $SD = 16$), gender (62% female), level of education (81% with bachelor's degree), and country of origin (e.g., 58% from the U.S., 13% from the U.K., 8% from Canada). In contrast to the currently most-cited free association norms from Nelson et al. (2004, USF Norms), De Deyne et al. (2019) asked participants to provide three responses to a given cue, so that weaker associations could be detected.

With over one million trials, SWOW-EN is a massive data set by the standards of cognitive modeling. Importantly, the number of unique responses generated by participants in this data set (over 130,000) is extremely large. Most of these responses are generated very infrequently, making the recall data quite sparse. We wish to predict participant-generated recall sequences by fitting 300 dimensional connection weights for each response word. To make achieving this goal computationally feasible, we subsetted the data for training and evaluating our models. To select data for the first-response model variant, we excluded all trials whose cues or first responses were not in our GloVe model's vocabulary and excluded all first responses that occurred in fewer than 100 trials across all of SWOW-EN. This meant that about 750,000 trials were retained. From these trials, we sampled a training set of about 100,000 trials and a test set of about 10,000 trials. To ensure that train and test sets were disjoint in terms of the sets of cues, we randomly selected cues and all their associated trials (of cue–response pairs) until ~100,000 training trials were selected. From the remaining cues, we used the same procedure to select ~10,000 test trials. In the general discussion section, we consider ways of using our trained model to make predictions for highly infrequent responses (on which our model has not been trained).

To select data for the continued-response model, we excluded all trials for which the cue or any of the three generated responses were missing from our GloVe model's vocabulary and excluded all responses that occurred in fewer than 300 trials (in any response slot) across all of SWOW-EN. This left us with about 250,000 trials. We used the same procedure to construct train and test sets as above, except now we only selected 30,000 trials for the training set, and 3,000 trials for the test set, since each "trial" from SWOW-EN contains three responses. Each trial was "unpacked" into three new observations, as discussed in the previous section. Thus, after unpacking these trials, we were left with 90,000 training observations and 9,000 test observations.

To calculate response asymmetries $p(R|C)/p(C|R)$, every training or test cue must also be an allowable response in our models. Thus, the above data led to a first-response model with $N = 3{,}616$ nodes in the word layer (2,221 response words and 1,395 cue words) and a continued-response model with $N = 2{,}154$ nodes in the word layer (2,030 response words and 124 cue words). These values of N also characterize the dimensionality of the output probability vectors of these models. The response layer is bigger in the single response model as, for this model, we ensured all training and test cues (total 1,395 cues) were in the response layer, so we could calculate cue–response asymmetry ratios for both the training set and the test set. But in the continued-response model, we only added the test set cues (total 124 cues) in the response layer, as calculating asymmetry ratios for the training set was not part of our analysis plan. Finally, both models had context layers with $M = 300$ nodes (corresponding to the dimensionality of the GloVe model). Code and trained weights can be downloaded from https://osf.io/as26x/. This study was not preregistered.

Free Recall Data

Although our primary interest in this article lies in free association, it is also possible that one or more of the above variants trained on free association data generalize effectively to free recall data, owing to the similarities in the tasks. If this turned out to be the case, it would possibly provide the basis for a unified cognitive modeling framework for these (and maybe other) related tasks in memory and semantic cognition. Therefore, after training on models on free association data from SWOW-EN, we evaluated them on the Penn Electrophysiology of Encoding and Retrieval Study (PEERS) data set of free recall from lists. PEERS is a publicly available, large-scale, multisession memory study that aims to understand the correlates of memory encoding and retrieval. While the PEERS data set consists of multiple experiments with slightly varying paradigms, we concentrate on free recall data from Experiment 4, which has been previously reported in Kahana et al. (2018) and Aka et al. (2021). In this experiment, participants performed a delayed free recall task consisting of 23 experimental sessions. In each of the sessions, participants completed 24 trials, with each trial containing a list of 24 words. For our analyses, we concatenated all subjects and lists into a single data set. Every single word in Experiment 4 of PEERS had an embedding in our GloVe model, but only 249,523 of the 618,859 responses generated by participants were in the set of SWOW-EN responses we retained for modeling first responses, as explained above. We used only these responses for our modeling purposes. Similar to our training and testing on free association data, we simply attempted to predict the next recalled word given the previously recalled word in the PEERS data. To keep things simple, we did not model list-presentation effects (e.g., primacy, recency, and temporal contiguity).

Relationship With Existing Models

Before continuing to our results section, it is useful to briefly note the similarities between our model and prior work. Our model is of course most closely related to CMR, from which it is derived. Indeed, we retain all of the retrieval dynamics of CMR, with two minor exceptions. First, our model allows for a persistent cue effect on context, parameterized with γ. Setting $\gamma = 0$ removes this effect, and results in a model whose context representation is largely identical to that of CMR, as it has been previously applied to free recall from lists. Second, we assume that recall probabilities depend on a softmax transformation of activations, rather than an accumulation process. This assumption is necessary to keep our modeling tractable. We do not expect it to alter any of the core properties of our model.

Our model is also related to other process models of free recall, like the search of associative memory (SAM) model (Raaijmakers & Shiffrin, 1981; see also Kahana, 2020). Most of the free association effects we wish to describe would have likely emerged if we had chosen to implement a variant of SAM. We decided to use CMR's architecture because of its neural network interpretation, which allows for the easy application of deep learning libraries (like Keras) to fit free association data. We also used CMR because its separate context-to-word and word-to-context weight matrices potentially allow for a better account of asymmetric response probabilities in free association. (See also Asr et al., 2018, for related evidence that modeling free association benefits from using both word and context vectors).

Relatedly, our approach to initializing connection weights using the GloVe model, and then updating a subset of these weights based on empirical free association data is essentially the "pre-train, then fine-tune" paradigm (also known as transfer learning) that has driven many of the recent successes in natural language processing (e.g., Devlin et al., 2019; Pan & Yang, 2009). In this paradigm, a (language) model is trained on a very generic objective on a large data set (e.g., word or sentence prediction on a corpus of billions of words), and then (some of) its parameters are fine tuned in a supervised fashion on a more specific task that has less labeled data available. The advantage of this approach is that it has the best of both worlds: the scale that comes from optimizing a generic objective on a large data set as well as the specificity and task appropriateness that comes from optimizing on a particular objective on a potentially smaller data set. In prior work, we have used this approach to predict human judgments and decisions (Bhatia, 2019; Bhatia et al., in press; Gandhi et al., 2022; Richie & Bhatia, 2021; Richie et al., 2019; Zou & Bhatia, 2021). We believe this approach is also suitable for the present application to free association, for which we have labeled data for only a few hundred thousand trials.

Finally, and most importantly, our model has similarities to existing distributed semantics and semantic network models of free association. In particular, our first-response model reduces to the dot product model of Nematzadeh et al. (2017) when we have symmetric weights, that is, when we set M^{CW} equal to (the transpose of) M^{WC}. In this case, the activation of a word given a cue is simply the dot product of the word vector and the cue vector, and the retrieval probability is the softmax transformation of the vector of dot products. This model also emerges in the continued-response case when we set $\gamma = 1$. Here, retrieval probabilities depend only on the cue (and not on previously retrieved responses) and are thus independently and identically distributed over time, with probabilities given by the Nematzadeh et al. dot product model. Conversely, values of $\gamma = \delta = 0$ lead to a Markov random walk over a semantic network, in which the cue initiates memory search and each jump leads to a new recall. The transition probabilities in this model depend on the connection weights and are closely related (though not strictly identical to) prior applications of semantic networks that specify node connections using distributed semantics models (Gruenenfelder et al., 2016; Kumar et al., 2020; Rotaru et al., 2018; Utsumi, 2014). Below, we will be comparing our most flexible model against various constrained models to measure the predictive and explanatory gains offered by asymmetric connection weights, persistent cue effects, and recurrent context representations.

Results

First Responses

We first examined our models' abilities to predict the first response (R_1) generated for each cue (C). Table 1 contains results on the training and test sets for negative log likelihood (NLL), the three metrics from Nematzadeh et al. (2017), and accuracy in predicting direction of asymmetry, for our main model (asymmetric weights) and the constrained model that constrains M^{CW} to be equal to (the transpose of) M^{WC} (symmetric weights). It also contains the results Nematzadeh et al. (2017) for their best-performing (largest) GloVe model. As discussed above, the symmetric weights model is identical to the dot product model of Nematzadeh et al. (2017). Note that although the test set has ~10,000 trials, the three Nematzadeh et al. metrics are calculated at the level of unique cues or unique cue–response pairs in the test set.

A few results are worth highlighting. First, the model with symmetric weights performed comparably to what Nematzadeh et al. (2017) report (although our median ranks are a bit worse), but our CMR model, with asymmetric weights, outperformed this constrained model, for all metrics and on both the training and test

Table 1
Model Performance

			Training set		Test set		
Data set	Evaluation metric		Asymmetric weights	Symmetric weights	Asymmetric weights	Symmetric weights	Nematzadeh GloVe
SWOW-EN	Total negative log-likelihood		417,031	1,126,676	86,747	115,248	n/a
	Average trial-level negative log-likelihood		4.17	11.26	8.64	11.48	n/a
	$p(R_1 \mid C)$ correlation		.54 (.53, .54)	.27 (.26, .28)	.37 (.34, .40)	.29 (.26, .33)	.27
	Median rank of true associates in model predictions	1st associate	1	20	5	23	11
		2nd associate	2	47	11	31	25
		3rd associate	4	70.5	46	86	40.5
	Asymmetry ratio correlation		.83 (.82, .83)	.45 (.44, .46)	.59 (.58, .61)	.38 (.36, .40)	.48
	Asymmetry direction accuracy (in %)		97 (97, 97)	71 (71, 72)	79 (78, 80)	67 (66, 68)	n/a
PEERS list recall	Total negative log-likelihood		n/a	n/a	7,385,121	8,556,284	n/a
	Average trial-level negative log-likelihood		n/a	n/a	25.01	28.97	n/a

Note. Performance of both CMR first response model variants on all model evaluation metrics, on our training and test sets selected from SWOW-EN, and on PEERS set of list recall. CMR = context maintenance and retrieval; SWOW-EN = English Small World of Words project; PEERS = Penn Electrophysiology of Encoding and Retrieval Study. 95% confidence intervals are in parentheses. In the last column, we present analogous statistics from the best-performing GloVe model in Nematzadeh et al. (2017; note that the results of this model are from a different test set and are not directly comparable to those presented in the other columns).

sets. It also did better than Nematzadeh et al.'s best performing GloVe model on all metrics (except for the median rank of the 3rd associate). In the test set, our model achieved a correlation between empirical and model $p(R_1|C)$ of .37, and the median ranks of the 1st, 2nd, and 3rd strongest associates in the model's predicted top associates were 5, 11, and 46, respectively. Finally, this model's predicted asymmetry ratios correlated with empirical asymmetry ratios at $r = .59$ and predicted the direction of asymmetry with 79% accuracy.

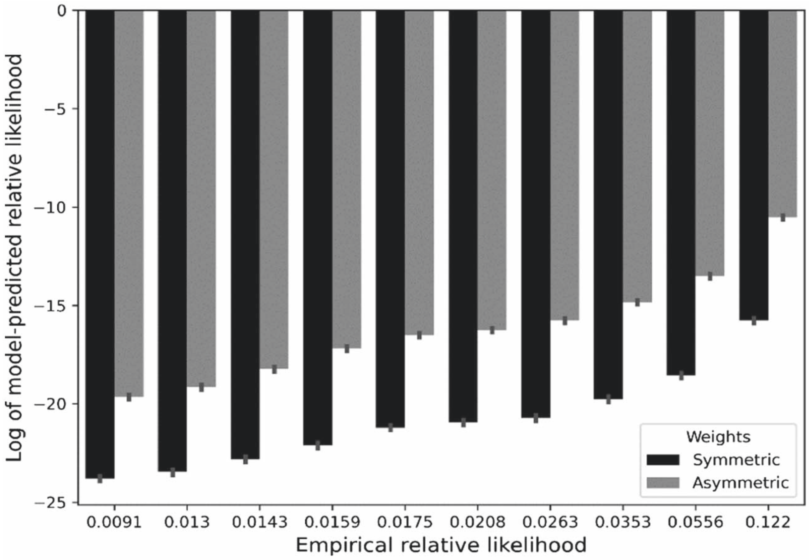
As a more detailed analysis of these models' abilities to capture free association, we conducted the following analysis, visualized in Figure 2. We binned all the empirically observed cue and first response (C, R_1) pairs from all of SWOW-EN into 10 equally sized sets of increasing relative frequency, $f(R_1|C)/f(R_1|{\sim}C)$, and also computed the log relative likelihood, $\log[p(R_1|C)/p(R_1|{\sim}C)]$, for each (C, R_1) pair according to our main model with asymmetric weights, and according to the constrained model with symmetric weights. Figure 2 shows that both models' log relative likelihoods tracked empirical relative frequency. In other words, both models predicted which responses are relatively more likely given a cue. However, relative likelihoods were always higher in the model with asymmetric weights, consistent with this model's lower negative log likelihood on the training and test sets (see Table 1). Interestingly, this advantage persisted for cue–response pairs with high empirical frequencies, as well as cue–response pairs with low empirical frequencies, indicating that allowing for asymmetric weights improves predictions uniformly for strong and weak associations.

Finally, we also examined the performance of these two models on Experiment 4 of the PEERS data set of free recall from lists. In particular, we evaluated each model's ability to predict the next recalled word, given only the previous recalled word. Table 1 shows the resulting negative log likelihoods of both models. Our model once again outperformed the constrained model, showing that fine-tuning the CMR model on free association data improves predictions for free recall from lists.

In Table 2, we present the results of a learning curve analysis in which we test the asymmetric model's performance on the above metrics after training on a random sample of 75%, 50%, and 25% of the training data set. As expected, model performance drops monotonically as the size of the training data set is reduced. Additionally, performance is bounded on the upper-end by the full asymmetric model (trained on 100% of the training data set) and on the lower-end by the symmetric weights model (which is implicitly the asymmetric model trained on 0% of the training data set). Somewhat surprisingly, reducing the size of the training data set does not have a large effect on metrics like $p(R_1|C)$ correlation, asymmetry ratio correlation, or asymmetry direction accuracy. Thus, it seems that even a small amount of training data is enough to achieve good results for these behavioral patterns (of course, to achieve the highest log-likelihood, it is necessary to train the model on the full data set).

Note that we have also applied topic models (Griffiths et al., 2007; Nematzadeh et al., 2017) to the analysis in this section. These models are important theories of free association, however, we found that our fits of these models were quite poor. These models are also less related to the CMR and its variants (whose connection weights correspond to vector representations for words and not probabilistic distributions over topics). For this reason, we present the results of our topic model analysis in the Supplemental Materials.

Figure 2
Model Predictions of First Responses



Note. Log of model-predicted relative likelihood $p(R_1|C)/p(R_1|{\sim}C)$ for each (C, R_1) pair, according to the models with symmetric weights (black) or asymmetric weights (gray), as a function of empirical relative likelihood $f(R_1|C)/f(R_1|{\sim}C)$. *X*-ticks represent the left edge of each bin of empirical relative likelihood. Error bars represent 95% confidence intervals.

Continued Responses

We also tested whether our models were able to describe the (3-length) sequences of words generated in response to each cue. For this purpose, we not only allowed for models with asymmetric weights, but also models with varying values of δ and γ (which control the effect of the cue and previously recalled words on subsequent recall). The underlined rows of Table 3 show the values of δ and γ which gave the symmetric and asymmetric model variants the best performance on the training set and the test set, as well as the negative log-likelihoods of those best-fitting values of δ and γ on the training set and test set. As can be seen, the model with asymmetric weights improved over the model with symmetric weights, on both the training set and on the test set. Thus, the benefits of fine-tuning CMR on free association extend to continued-response modeling.

How does performance of each model vary as a function of γ and δ? That is, what balance of a persistent effect of c^{cue} and previously recalled words on c^{recall} (controlled by γ), and how much self-feedback in c^{recall} (controlled by δ), are needed for modeling continued free association? Figure 3 contains heatmaps of the negative log likelihoods of both models on the training set and the test set, as a function of γ and δ. Most apparent in this figure are the downward sloping dark bands, representing a trade-off between γ and δ, which implies that models must find some way of representing the cue in the context layer, *c*. This can happen with either a high γ (reflecting strong, persistent contribution of c^{cue} to context *c*), or with a high δ (reflecting strong maintenance of the cue's initial impact on c^{recall} and therefore on context *c*). However, careful examination of these bands also shows that darker cells, that is, lower negative log likelihoods, tend to be in the bottom right, suggesting the best models had low values of γ and high values of δ, reflecting small persistent effects of the cue and strong influence of all previously recalled responses (including the cue).

Table 3 also contains negative log likelihoods with one or both of γ and δ constrained to key values (bolded values represent flexible values of γ and δ). For example, the first row of this table reports the negative log likelihood of the model with symmetric weights and with $\delta = \gamma = 0$, which is essentially a Markov random walk model, where responses are purely driven by the most recent response. The seventh row likewise reports the performance of a symmetric weight model in which $\gamma = 1$ (and δ is subsequently inconsequential). This model is identical to that proposed by Nematzadeh et al. (2017). Both the Markov random walk and the Nematzadeh et al. models are outperformed by an asymmetric weight model with flexible γ and δ. Overall, Table 3 shows that models with flexible γ and δ are always superior over constrained counterparts. All of the differences in log likelihoods are significant (*p*s < .05 by likelihood ratio rests), except for the symmetric weights model on the test set when comparing the fully flexible model to one with δ constrained to 1.

That the best models had (in addition to asymmetric weights) both $\delta > 0$ and $\gamma > 0$, suggests that continued free association in response to a cue is described with the combination of (a) representations of previously recalled words in c_t^{recall}, but also (b) a persistent effect of the cue from c^{cue}. To intuitively understand this result, we examined, for each observation in our test data, the prediction of our best-fitting asymmetric weight model with flexible γ and δ, as well as the predictions of a Markov random walk model with $\delta = \gamma = 0$ and a cue-only model with $\gamma = 1$. In Table 4, we illustrate the five observations in our test data for which the ratio of predicted probability (equivalent to the difference in the log likelihood) of our flexible model and our Markov random walk model is the highest. Here, we see that the flexible model can make better predictions

Table 2
Learning Curve Analysis of Our Asymmetric Weights Model

Data set	Evaluation metric		Proportion of training data used				
			100%	75%	50%	25%	0%
SWOW-EN	Total negative log-likelihood		86,747	90,336	95,721	101,895	115,248
	Average trial-level negative log-likelihood		8.64	9.00	9.54	10.15	11.48
	$p(R_1\|C)$ correlation		.37 (.34, .40)	.37 (.33, .4)	.36 (.33, .39)	.35 (.32, .38)	.29 (.26, .33)
	Median rank of true associates in model predictions	1st associate	5	5	5	6	23
		2nd associate	11	12	13	12	31
		3rd associate	46	43	41	44	86
	Asymmetry ratio correlation		.59 (.58, .61)	.59 (.57, .60)	.57 (.55, .59)	.56 (.54, .58)	.38 (.36, .40)
	Asymmetry direction accuracy (in %)		79 (78, 80)	79 (78, 80)	78 (77, 79)	77 (75, 78)	67 (66, 68)
PEERS list recall	Total negative log-likelihood		7,385,121	7,649,163	7,968,176	8,391,467	8,556,284
	Average trial-level negative log-likelihood		25.01	25.90	26.98	28.41	28.97

Note. SWOW-EN = English Small World of Words project; PEERS = Penn Electrophysiology of Encoding and Retrieval Study. Here, we present model evaluation metrics on the test set for the model trained on 100% of the training data set, as well as on a random sample of 75%, 50%, and 25% of the training data set. We also present the results for our symmetric weights model, which is implicitly the asymmetric weights model trained on 0% of the data set. 95% confidence intervals are in parentheses.

because it keeps track of the cue. Thus, for example, the flexible model predicts a high probability of R_2 = *pee* when R_1 = *small* and C = *wee*. The Markov random walk model, by contrast, predicts a low probability of R_2 = *pee* following R_1 = *small*, as it does not remember the cue (i.e., it sets $\delta = \gamma = 0$). Table 4 also shows five observations in our test data for which the ratio of predicted probability of our flexible model and our cue-only model is the highest. In this setting, the flexible model can make better predictions because it keeps track of the previously recalled responses. Thus, for example, the flexible model predicts a high probability of R_3 = *shoes* when R_2 = *underwear*, R_1 = *jeans* and C = *recognition*. The cue-only model, by contrast, predicts a low probability of R_3 = *shoes* in response to C = *recognition*, as it does not remember the intervening responses (i.e., it sets $\gamma = 1$).

Finally, the best value $\gamma = .3$ on the training set suggests that representations of previously recalled words (and the cue) in c^{recall} dominates the persistent effect of the cue from c^{cue}. Likewise, the best-fitting value of $\delta = 1$ suggests that decay on c^{recall} is minimal, that is, that later responses are driven by more recent responses just as much as by older responses (including the cue as it initially influences c_t^{recall}). This could be due to the relatively small number of responses generated in our free association data (three) versus more than twenty in standard free recall from lists tasks, that is, decay could be minimal (or perhaps imperceptible) with such short response sequences (see also Oberauer et al., 2016, for a discussion of the possible sources of working memory limits).

Response Chaining Effects

To further analyze the properties of our best-fitting models, we used a binning analysis like that for the first response model. First, we examined response chaining effects by binning all the empirically observed (C, R_1, R_2) triples into five equally sized sets of increasing relative frequency, $f(R_2|C, R_1)/f(R_2|C, \sim R_1)$. Intuitively, this measure captures the frequency of R_2 in response to the cue when it is preceded by R_1 relative to when it is not preceded by R_1. Triplets with high values on this measure involve R_2s that are especially likely to occur after the R_1s (keeping the effect of the Cs constant).

We also computed the predicted log relative likelihood, $\log[p(R_2|C,R_1)/p(R_2|C, \sim R_1)]$ for each triplet according to various models. These are our models' predictions for this relative frequency measure. We considered the best-performing asymmetric weights CMR model (with $\delta = 1$ and $\gamma = .3$) as well as constrained versions of this model with $\delta = \gamma = 0$ (corresponding to a Markov random walk model) and $\gamma = 1$ (corresponding to a persistent effect of the cue, and no effect of previously recalled responses). The (C, R_1, R_2) triplet with the highest predicted log relative likelihood by our CMR model is (*gorge*, *eat*, *food*), whereas the (empirically observed) triplet with the lowest predicted log relative likelihood by this model is (*cylinder*, *maths*, *water*). These two triplets show how CMR constrains its predictions of R_2 by attaching a positive weight to R_1. In the former case, CMR attaches a high probability to *food* when it is preceded by *eat* relative to when it is not (in which case, the model predicts an R_2 that is semantically related to *canyon* or *river*, reflecting the alternate meaning of *gorge* as a narrow valley between mountains). Likewise, in the latter case, the model expects *maths* (when cued by *cylinder*) to be followed by mathematical or geometrical concepts, rather than by *water* (which is more likely when

Table 3
Model Fits

Training set			Test set		
δ	γ	NLL	δ	γ	NLL
Symmetric					
0.0	0.0	675,945	0.0	0.0	68,061
1.0	0.0	672,509	**1.0**	0.0	67,725
1.0	0.0	672,509	1.0	0.0	67,725
0.0	**0.6**	672,494	0.0	**0.6**	67,733
1.0	***0.3***	*672,219*	***0.9***	***0.3***	*67,704*
1.0	**0.3**	672,219	1.0	**0.3**	67,704
n/a	1.0	673,855	n/a	1.0	67,897
Asymmetric					
0.0	0.0	634,292	0.0	0.0	64,898
1.0	0.0	617,793	**1.0**	0.0	63,718
1.0	0.0	617,793	1.0	0.0	63,718
0.0	**0.6**	617,745	0.0	**0.5**	63,899
1.0	***0.3***	*616,169*	***1.0***	***0.1***	*63,701*
1.0	**0.3**	616,169	1.0	**0.1**	63,701
n/a	1.0	624,007	n/a	1.0	64,921

Note. Total negative log likelihoods of the symmetric and asymmetric weights models with various constraints on γ and δ. NLL = negative log likelihood. Bolded numbers indicate parameters that were fit freely. Italicized rows identify best-performing parameter configurations.

R_1 is *can* or *cup*). The Markov random walk model makes similar predictions, whereas the cue-only model is unable to distinguish such cases (matching the patterns shown in Table 3).

Are our model's predictions accurate? As can be seen in Figure 4, the CMR model with flexible δ and γ, as well as the Markov model with $\delta = \gamma = 0$, generated higher predicted likelihoods for cue–response triplets (C, R_1, R_2) with higher empirical likelihoods. In other words, both models predicted which R_2 are relatively more likely given R_1 (keeping the effect of the cue constant). As expected, the model with the effect of only the cue ($\gamma = 1$) failed to capture the effect of R_1 on the recall probability of R_2 as it did not allow feedback from recalled words to context. Additionally, the likelihoods predicted by the Markov model, which did not allow for a persistent effect of the cue, were higher than those of the flexible CMR model. This was due to the fact that our analysis in this figure controlled for the effect of the cue and tried to model only the effect of R_1 on R_2. The Markov model is best suited for capturing this effect and thus makes very good predictions.

Persistent Cue Effects

If the Markov model was indeed the best-performing model, we would not expect to observe an effect of C on R_2 controlling for $R_{1.}$ To test for this type of persistent cue effect, we performed a third binning analysis in which we binned all the empirically observed (C, R_1, R_2) triples into five equally sized sets of increasing relative frequency, $f(R_2|C, R_1)/f(R_2|{\sim}C, R_1)$. Intuitively, this measure captures the frequency of R_2 following R_1 when C is the cue relative to when C is not the cue. Triplets with high values on this measure involve R_2s that are especially likely to occur in response to the Cs (keeping the effect of the R_1s constant).

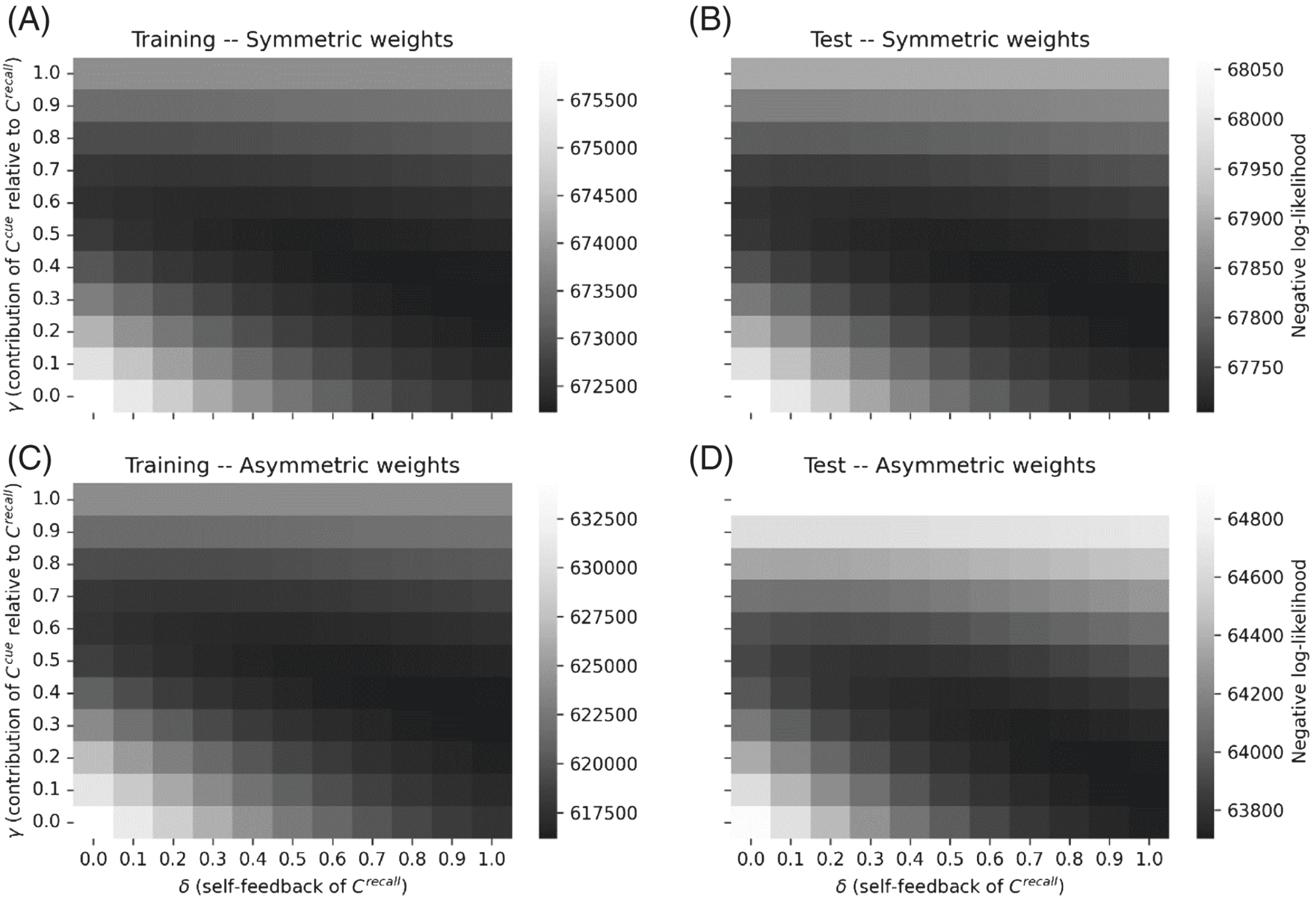
We also computed $\log[p(R_2|C, R_1)/p(R_2|{\sim}C, R_1)]$ for each triplet according to various models. Note that the denominator of this term cannot be calculated because our models do not define $p({\sim}C)$. We therefore approximated the denominator by simply averaging $p(R_2|C\text{'}, R_1)$ for all $C' \neq C$ (for consistency, we calculated the denominator of $f(R_2|C, R_1)/f(R_2|{\sim}C, R_1)$ in the same way). This is why we call this measure the model-predicted pseudorelative likelihood. Again, we considered the best-performing CMR model with flexible δ and γ, and versions with $\delta = \gamma = 0$ (corresponding to a Markov random walk model) and $\gamma = 1$ (corresponding to an effect of only the cue). The (C, R_1, R_2) triplet with the highest predicted pseudorelative likelihood by our CMR model is (*taxes*, *death*, *money*), whereas the (empirically observed) triplet with the lowest predicted likelihood by this model is (*cd*, *music*, *money*). These two triplets show how CMR constrains its predictions of R_2 by placing a positive weight on C. In the former case, CMR attaches a high probability of *death* being followed by *money*, when the cue is *taxes*, relative to when the cue is not *taxes* (in which case, the model predicts an R_2 that is more directly associated with *death*, like *murder* or *funeral*). Likewise, in the latter case, the model expects *music* (when cued by *cd*) to be followed by another music-related term, rather than by *money*. The cue-only model makes similar predictions, whereas the Markov model is unable to distinguish such cases (matching the patterns shown in Table 3).

Are our model's predictions accurate? Yes. Figure 5 shows that the models with flexible δ and γ, as well as the model with $\gamma = 1$, generated higher predicted likelihoods for cue–response triplets (C, R_1, R_2) with higher empirical likelihoods. In other words, both models predicted which R_2s are relatively more likely given C (controlling for the effect of R_1). The Markov model with $\delta = \gamma = 0$ failed to capture the effect of C on the relative likelihood of R_2 as it did not allow for a persistent effect of the cue on context. Finally, the likelihoods predicted by the cue-only model, which did not allow for feedback from recalled words to context, were often higher than those of the flexible CMR model. This was due to the fact that our analysis in this figure controlled for the effect of R_1 and tried to model only the effect of C on R_2. The cue-only model is best suited for capturing this effect and thus makes very good predictions, though of course it fails at capturing the effect of R_1 on R_2 shown in Figure 4. The only model that can capture both the response chaining effect shown in Figure 4 and the persistent cue effect shown in Figure 5 is the CMR model with flexible δ and γ.

Predicting Second and Third Responses

As another test to understand the necessity of the dynamic properties of our CMR variant for modeling continued responding, we performed likelihood ratio tests comparing asymmetric weight models with flexible γ and δ to asymmetric weight models with constrained γ and δ. In particular, we compared (a) models with γ and δ that optimized performance on only the first, second, or third response of each trial in the training or test set, to (b) a model with $\gamma = 1$ when only evaluating on the first, second, or third response of each trial in the training set or test set. For each model (a), which flexibly fits γ and δ to only first, second, or third responses on either the training or test set, we calculated the log likelihood of the data (first, second, or third responses) under that model. For (b) we calculated the log likelihood of the data under a comparable model with $\gamma = 1$. We then calculated the ratio of the two likelihoods, where a ratio above one indicates that the data are more likely under the flexible model. Statistical significance was calculated by conducting chi-squared tests on the likelihood ratio with two degrees of freedom (as the flexible model varies both γ and δ). Results are in Table 5.

Figure 3
Model Fits



Note. Heatmaps of negative log-likelihood of the models with symmetric (Panels A and B) or asymmetric (Panels C and D) weights, on the training set (Panels A and C) and the test set (Panels B and D), as a function of γ and δ.

Here, for example, the left side of the first row shows that the model with γ = .6, and δ = .9 had a significantly better log likelihood than a comparable model with γ = 1, when evaluated on the first response of every trial in the training set. Note that the value of δ is irrelevant when γ = 1, as the cue is the sole determinant of context and recall.

Our intention with this analysis was to show that the CMR-style dynamics, which arise when γ < 1 and δ > 0, are necessary to account for the second and third response but not the first response, since only second and third responses in a trial provide the opportunity for response chaining. Interestingly, this turned out not to be the case. Models with asymmetric weights always favored γ < 1 and were better than models with γ = 1 (all likelihood ratio tests have $p <$ .0001), even for the first response in the training or test set. Our explanation for this finding is that the asymmetric weights model with γ = 1 fits M^{CW} to predict not just the first response, but also the next two responses. However, the second and third responses tend to be less prepotent responses and less likely to appear in the first response slot (De Deyne et al., 2019), and so by optimizing M^{CW} to predict these less prepotent responses, the model with γ = 1 becomes worse at predicting the first, prepotent responses. This is why having flexible γ and δ leads to better fits for the first response (as well as the second and third responses). The results of this analysis are consistent with the idea that free association models benefit from the inclusion of CMR-style dynamics. This is the case even if the goal of the model is to predict only the first response.

The Effects of Fine Tuning

The above sections have shown that the asymmetric weight model, which fine tunes M^{CW} on free association data, consistently outperforms the symmetric weight model, which constrains M^{CW} to be identical to (the transpose of) M^{WC}. We have argued above that this is due to the benefits conferred by transfer learning, which allow pretrained models to be modified to better predict a given data set. Here, we wish to examine what it is about our free association data set that necessitates modifications to the distributed semantics representations obtained from text. For this purpose, we first examine the distribution of the dot products of the M^{CW} and M^{WC} vectors for our symmetric and asymmetric first response models, that is, the distribution of $M_i^{\mathrm{WC}} \cdot M_j^{\mathrm{CW}}$ for each i and j (where i indexes a column of M^{WC} and j indexes a row of M^{CW}). $M_i^{\mathrm{WC}} \cdot M_j^{\mathrm{CW}}$ measures the degree to which a word i cues word j. In the symmetric case (shown in Figure 6A), $M_i^{\mathrm{WC}} \cdot M_j^{\mathrm{CW}}$ is merely the dot product of one GloVe vector with another GloVe vector. In the asymmetric case

Table 4
Model Predictions of Continued Responses

Observation	Cue	Prev. Responses	Target Response	Prob. Ratio
		Flexible versus Markov random walk		
1	*wee*	*small*	*pee*	1.64
2	*swimmer*	*water*	*athlete*	1.62
3	*horrendous*	*awful, mean*	*terrible*	1.56
4	*loneliness*	*alone*	*sadness*	1.55
5	*blond*	*stupid*	*hair*	1.52
		Flexible versus cue-only		
1	*recognition*	*jeans, underwear*	*shoes*	1.65
2	*jeep*	*family, mother*	*father*	1.64
3	*diary*	*cheese, milk*	*cream*	1.64
4	*repression*	*sex, men*	*women*	1.53
5	*repulsive*	*odor, teacher*	*student*	1.51

Note. The five observations in the test data for which the ratio of predicted probabilities of our flexible model ($\delta = 1$ and $\gamma = .3$) and our Markov random walk model ($\delta = \gamma = 0$) is the highest, and the five observations in the test data for which the ratio of predicted probabilities of our flexible model and our cue-only model ($\gamma = 1$) is the highest. In all cases, our models attempt to predict the target response using the cue and sequence of previous responses.

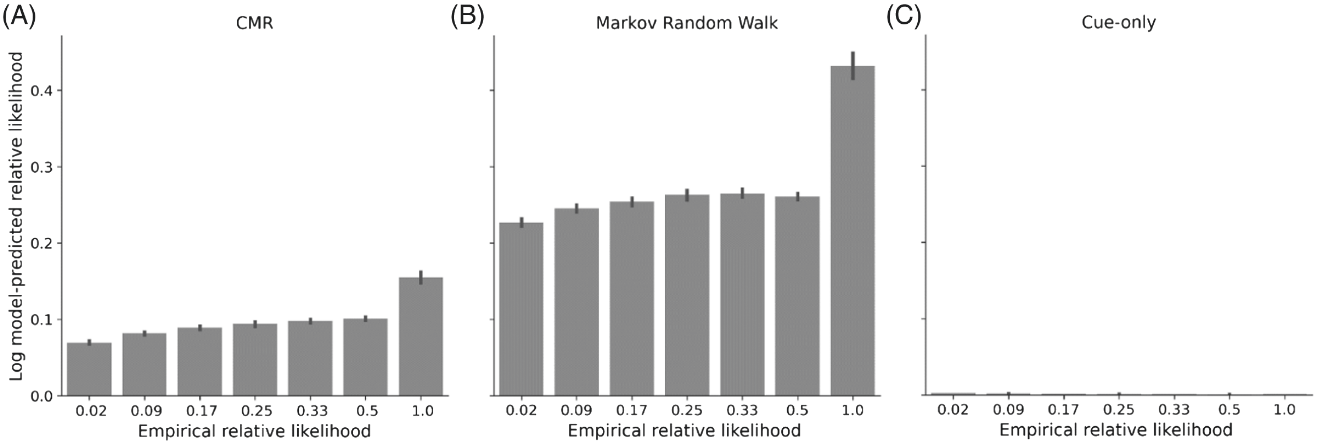
(shown in Figure 6B), $M_i^{WC} \cdot M_j^{CW}$ is the dot product of an original GloVe M^{WC} vector with a fine-tuned M^{CW} vector. Figure 6A shows that the distribution of dot products is very similar in the symmetric and asymmetric models. Both have the same mean ($\mu = 2.71$), though the symmetric model has a slightly higher standard deviation ($\sigma = 4.61$) than the asymmetric model ($\sigma = 4.29$). This indicates that fine tuning does not have a large effect on the aggregate strength of associations.

It is also the case that the symmetric weight model has symmetric dot products, that is the degree to which a word *i* activates a word *j* is identical to the degree to which a word *j* activates word *i* (though of course response probabilities could be asymmetric for the reasons discussed above). The asymmetric weight model, however, allows for asymmetries in cuing. In Figure 6C, we show the extent of these asymmetries by plotting the distribution of differences in the dot product of M^{WC} with M^{CW} vectors and M^{CW} with M^{WC} vectors, that is, the distribution of $M_i^{WC} \cdot M_j^{CW}–M_j^{WC} \cdot M_i^{CW}$. Here, we see that the majority of differences are near zero, though some differences can be very large, indicating that fine tuning leads to strong asymmetries in a minority of associations in the model.

Finally, Figure 6D plots the distribution of the cosine similarity of vectors in M^{CW} with their corresponding vectors in M^{WC}, that is COSSIM (M_i^{WC} , M_i^{CW}). Intuitively, this cosine similarity measure captures the degree to which the representation of a word is modified during fine tuning. Here, we can again see that most words have a cosine similarity that is very close to 1, indicating that most word representations have not been changed. Some words do, however, see large shifts in their representation after fine tuning.

Table 6 presents the twenty response words with the largest shifts in their representation after fine tuning, that is words that have the lowest cosine similarity between their vectors in M^{CW} and their vectors in M^{WC}. Table 6 also presents the three cues with the highest cosine similarity to these response words for both the asymmetric weight model and the symmetric weight model. These cue words track the effect of fine tuning on the representations of the response words. Table 6 suggests that there are many reasons why the asymmetric model outperforms the symmetric model. First, as with CMR's application to free recall from lists, the asymmetric model is better able to capture asymmetries in response probabilities between pairs of words. While such asymmetries usually occur due to asymmetric positive and negative temporal lag effects in free recall from lists, in the case of free association, they are likely to occur because of common phrases or idioms in language. For example, as discussed in Clark (1970), *white* strongly cues *house* but *house* is less likely to cue *white*, as *white house* is a common phrase in English (other examples provided by Clark are *cottage cheese* and *ham eggs*). In Table 6, we see such an effect for the word *bee*, likely because of the common phrase *busy bee*, and for the word *game*,

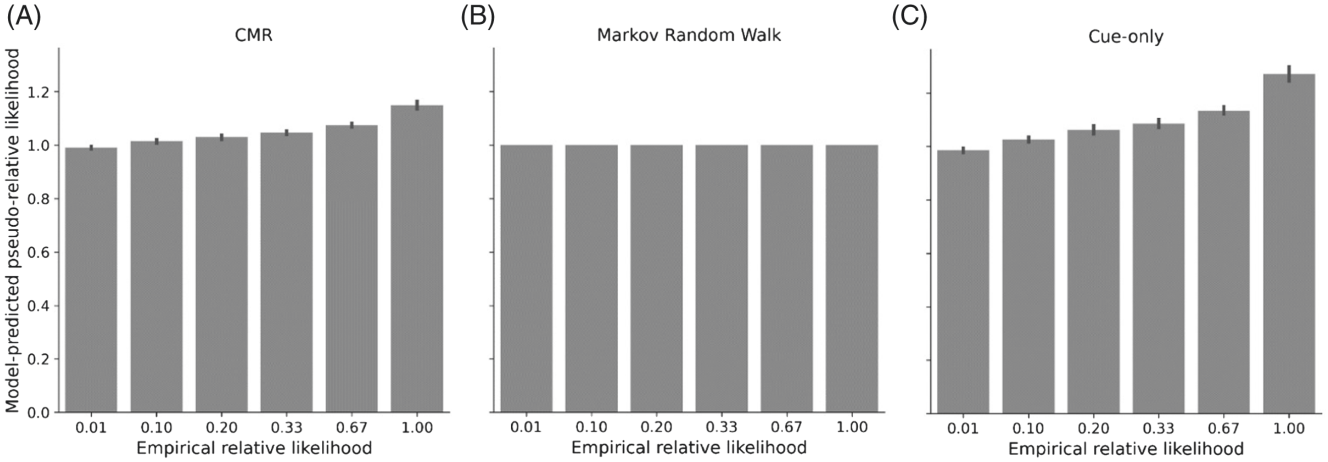
Figure 4
Model Predictions for Response Chaining



Note. Log of model-predicted relative likelihood $p(R_2|C, R_1)/p(R_2|C, \sim R_1)$ for each (C, R_1, R_2) triplet, according to CMR ($\delta = 1$ and $\gamma = .3$) (A), a Markov random walk ($\delta = \gamma = 0$) (B), and a cue-only model ($\gamma = 1$) (C), as a function of empirical relative frequency $f(R_2|C, R_1)/f(R_2|C, \sim R_1)$. *X*-ticks represent the left edge of each bin of empirical relative likelihood. Error bars represent 95% confidence intervals. CMR = context maintenance and retrieval.

Figure 5

Model Prediction of Persistent Cue Effects



Note. Log of model-predicted pseudorelative likelihood $p(R_2|C, R_1)/p(R_2|{\sim}C, R_1)$ for each (C, R_1, R_2) triplet, according to CMR (δ = 1 and γ = .3) (A), a Markov random walk (δ = γ = 0) (B), and a cue-only model (γ = 1) (C), as a function of empirical relative frequency $f(R_2|C, R_1)/f(R_2|{\sim}C, R_1)$. *X*-ticks represent the left edge of each bin of empirical relative likelihood. Error bars represent 95% confidence intervals. CMR = context maintenance and retrieval.

likely because of the common phrase *football game*. In the symmetric model, *bee* is associated with *honey*, *hive*, and *bumble*, and *game* is associated with *play*, *player*, and *match.* Although such associations persist in the asymmetric model, we see also see that this model modifies the representation of *bee* to give it a strong association with the word *busy* and modifies the representation of *game* to give it a stronger association with the word *football.* Thus, in the asymmetric model, *busy* is highly likely to cue *bee* and *football* is highly likely to cue *game* (though *bee* is not as likely to cue *busy* and *game* is not as likely to cue *football*).

Another benefit of fine tuning is disambiguation. Several words have multiple meanings, but in some cases, one of these meanings may be more privileged in free association than in distributed vector representations obtained from text. This is something that a flexible asymmetric weight model can capture, but that a constrained symmetric weights model cannot. Table 6 shows several examples of such disambiguation problems: *plant* is more associated with *grow* in the asymmetric weights model and more associated with *factory* in the symmetric weights model. Likewise, *rose* is more associated with *flower* in the asymmetric weights and more associated with *fell* in the symmetric weights model. A related effect involves perceptual or experiential qualities that are more prominent in free association than in text. Standard distributed semantics models, which proxy co-occurrence or synonymy in text have difficulty modeling these associations (see De Deyne et al., 2019, for a discussion of this issue; as well as Andrews et al., 2009 for an alternate solution), whereas flexible weight models can be fine tuned to accommodate these perceptual and experimental associations, giving them superior predictive power. For example, in Table 6, we see that the asymmetric weights model associates *rich* with adjectives like *luxurious* and *elegant*, and objects like *mansion*, but that the symmetric weights model associates *rich* with synonyms like *wealthy*. Likewise, the asymmetric weights model associates *cry* with *onion*, whereas the symmetric weights model associates *cry* with nearly synonymous verbs like *scream* and *shout*.

The flexibility of the asymmetric weights model also allows it to capture population-level idiosyncrasies. GloVe and other distributed semantics models are trained on large data sets of English language text, which consist primarily of American English. By contrast, a substantial proportion of the SWOW participants spoke British (11%) or Australian (5%) English. Words like *bum*, which are more likely to be generated by the latter group, are more associated with words like *buttocks* in the asymmetric weights model. By contrast, the primary associations of *bum* in the symmetric weights model involve *rap* and *thug*.

The explanation provided for the patterns in Table 6 is fairly speculative, as it is difficult to provide a systematic characterization of how fine tuning on free association data changes the hundreds of thousands of weights in our flexible model. In order to attempt a more quantitative approach, we regressed the cosine similarity of word vectors in M^{CW} and M^{WC} on four important psycholinguistic variables that have been shown to influence lexical access in general or free association in particular: age of acquisition (Kuperman et al., 2012),

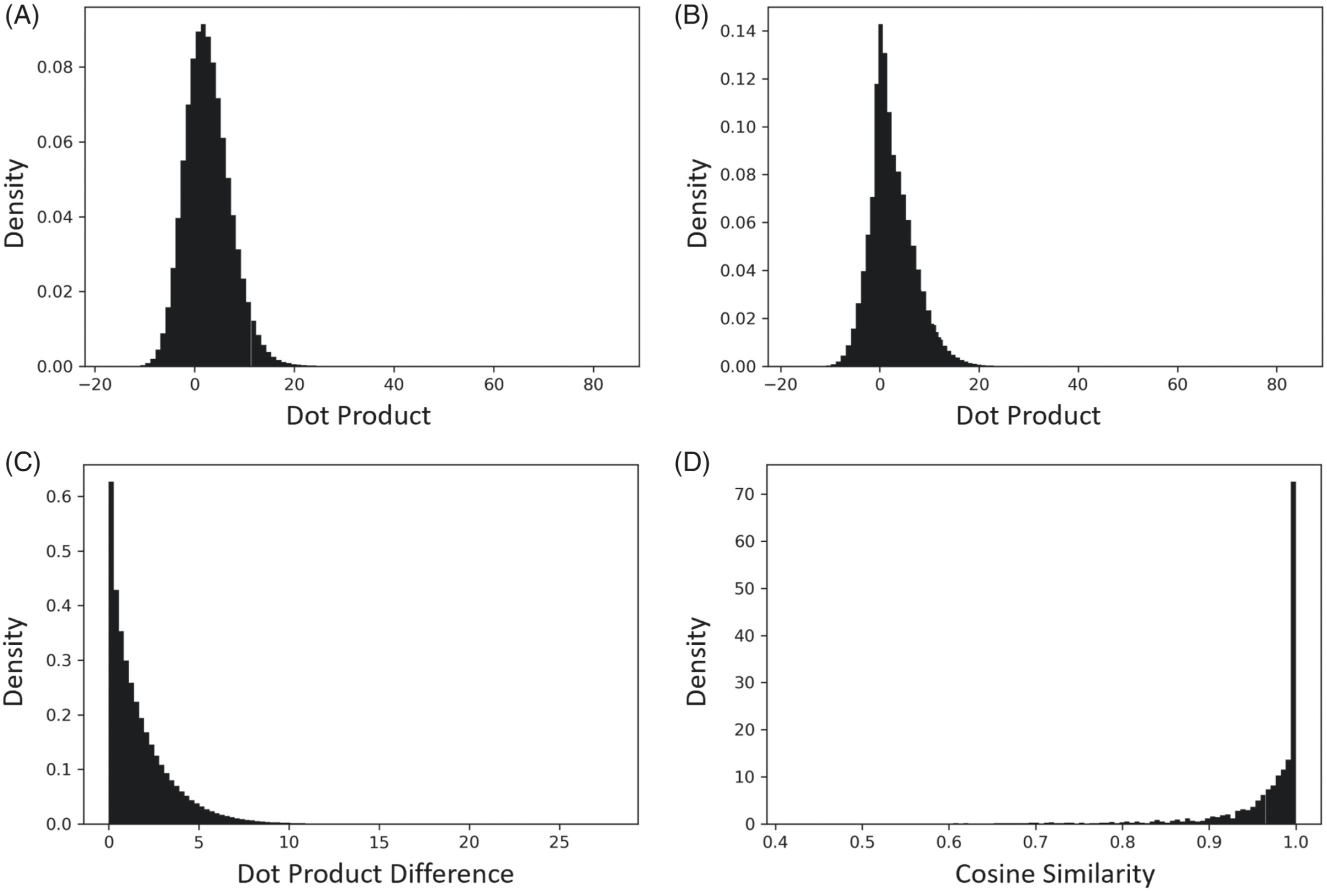
Table 5

Predicting Second and Third Responses

Training set				Test set			
Response slot	δ	γ	LLR	Response slot	δ	γ	LLR
First	0.9	0.6	1,200	First	0.8	0.4	210
Second	0.9	0.3	6,700	Second	1	0	1,000
Third	1	0.3	8,400	Third	1	0.1	1,200

Note. Likelihood ratios (LLR's) between (a) asymmetric weight models with γ and δ that optimized performance on only the first, second, or third response in the training or test set, and (b) an asymmetric weight model with γ = 1, when only evaluating on the first, second, or third response of each trial in the training set or test set. A LLR over 1 indicates that observed data are more likely under the flexible model (a) than under a comparable constrained model (b). All LLR values are statistically significant at $p < .0001$, according to chi-squared tests with two degrees of freedom.

Figure 6
Effects of Fine-Tuning



Note. Histogram of the dot products of the M^{CW} and M^{WC} vectors for our symmetric (A) and asymmetric (B) first response models. Histogram of the distribution of differences in the dot products of M^{WC} with M^{CW} vectors and M^{CW} with M^{WC} vectors for the asymmetric model (C). Distribution of the cosine similarity of vectors in M^{CW} with their corresponding vectors in M^{WC} for the asymmetric model (D).

concreteness, and (log) frequency (Brysbaert et al., 2014), and number of word senses according to the Wordsmyth online dictionary. As our dependent variable, cosine similarity, was not distributed normally, we used the ranked cosine similarity (with a high rank for words like *bum*, which differed substantially between M^{CW} and M^{WC}).

The results of our regression analysis are provided in Table 7. Here, we can see that words with greater frequency of use in language and words with more word senses were more likely to have different representations in M^{CW} and M^{WC}. The first of these effects is likely due to the fact that high frequency words are more likely to be generated as responses in free association and thus are more likely to have training data for our fine-tuning exercise. All else equal, more data lead to more changes in representation. The second of these effects can be explained by the disambiguation interpretation outlined above. Additionally, we found that words with an older age of acquisition were less likely to have similar representations in M^{CW} and M^{WC}. This is again likely caused by the fact that words that are acquired later in life are less likely to be generated as responses in free association (Matusevych & Stevenson, 2018) and thus have fewer observations in our training data. We did not find an effect of word concreteness in our data. Overall, the simple pairwise correlations of ranked cosine similarity with age of acquisition, concreteness, (log) frequency, and number of word senses are −.49, 0, .55, and .29 respectively (with all correlations except those for concreteness being significant at $p < .05$).

Discussion

Free association among words is fundamental to cognition and behavior. Yet, free association has largely eluded satisfying and scalable cognitive modeling, as it involves complex memory processes operating over thousands of semantically rich words. To address this, we combined distributed semantics (DS) models of word meaning (Howard et al., 2011; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mikolov et al., 2013; Pennington et al., 2014) with recurrent neural networks based on the CMR model for free recall (Howard & Kahana, 2002a; Polyn et al., 2009). We trained our models on large-scale free association norms (SWOW-EN; De Deyne et al., 2019) and found that borrowing CMR's assumption of asymmetric representations for cues and responses (i.e., asymmetric word-context weights M^{WC} and M^{CW}) improved upon previous DS-based models of free association (Jones et al., 2018;

Table 6
Effects of Fine-Tuning

Word	Similarity	Top 3 associates—Asymmetric			Top 3 associates—Symmetric		
		1	2	3	1	2	3
bum	0.42	*buttocks*	*homeless*	*breasts*	*ok*	*rap*	*thug*
money	0.44	*cash*	*pay*	*payment*	*funds*	*cash*	*fund*
plant	0.52	*grow*	*fruit*	*seeds*	*factory*	*facility*	*produce*
bee	0.52	*busy*	*buzz*	*hive*	*honey*	*hive*	*bumble*
bend	0.53	*flexible*	*flex*	*squat*	*curves*	*bow*	*horseshoe*
game	0.53	*football*	*play*	*player*	*play*	*player*	*match*
rich	0.54	*luxurious*	*elegant*	*mansion*	*wealthy*	*wealth*	*vast*
computer	0.57	*software*	*user*	*keyboard*	*software*	*pc*	*technology*
weight	0.57	*heavy*	*amount*	*tons*	*load*	*pounds*	*height*
words	0.58	*language*	*writing*	*written*	*phrase*	*meaning*	*language*
police	0.58	*arrest*	*fbi*	*criminal*	*arrested*	*policeman*	*arrest*
pen	0.58	*writer*	*written*	*sharpie*	*pencil*	*ink*	*quill*
rose	0.59	*flower*	*petals*	*pink*	*fell*	*percent*	*slipped*
tire	0.59	*wheel*	*spare*	*pump*	*wheel*	*automobile*	*brake*
cry	0.60	*onion*	*crying*	*stir*	*crying*	*scream*	*shout*
cut	0.60	*scissors*	*snip*	*chop*	*slash*	*reduce*	*would*
tooth	0.60	*filling*	*cavity*	*teeth*	*teeth*	*jaw*	*enamel*
spider	0.61	*web*	*online*	*internet*	*monkey*	*snake*	*frog*
dead	0.61	*probably*	*death*	*killed*	*killed*	*wounded*	*killing*
end	0.61	*beginning*	*start*	*finish*	*ending*	*ended*	*beginning*

Note. Twenty words with the lowest cosine similarity between their vectors in M^{CW} and M^{WC}, along with the three most associated cues in the asymmetric and symmetric models.

Nematzadeh et al., 2017; Pereira et al., 2016). We also found that responses in continued free association were best modeled by a variant of CMR with context feedback as well as a persistent input from the cue. Thus, while classic Markov models (Abbott et al., 2015; De Deyne & Storms, 2008; De Deyne et al., 2013, 2016, 2019; Dubossarsky et al., 2017; Kenett et al., 2017; Kumar et al., 2019; Steyvers & Tenenbaum, 2005) can account for some response chaining effects in free association, our full model was quantitatively superior in predicting violations of the Markov property in continued free association.

Our work shows the power of combining semantically rich representations from DS models, with process theories of how such representations are used for a particular task. While some past work has questioned the applicability of DS models (Griffiths et al., 2007) or spatial representations more generally (Tversky, 1977) to the study of semantic memory, our work emphasizes Jones et al.'s (2018) point that many apparent flaws of DS models arise from confusing a (dis)similarity metric like cosine similarity for a model of what people *do* with a semantic vector representation. When DS models are combined with an appropriate cognitive theory of choice, judgment, categorization, or in our case, memory retrieval, they can provide accurate and scalable models of behavior (see Bhatia & Stewart, 2018; Hills et al., 2012; Lu et al., 2019; Richie & Bhatia, 2021; Zou & Bhatia, 2021, for similar demonstrations; also see Bhatia & Aka, 2022, for a review).

Table 7
Predictors of Word Representation Shifts

Variable	Coef.	*t*	95% CI	Correlation
Age of acquisition	−116.13	−12.80	[−133.92, −98.34]	−0.49
Concreteness	−14.39	−0.89	[−46.16, 17.37]	0
Log word frequency	449.22	16.98	[397.35, 501.08]	0.55
Number of word senses	7.73	2.01	[0.17, 15.28]	0.29

Note. Output of regression of ranked cosine similarity between response words' vectors in M^{CW} and M^{WC} on age of acquisition, concreteness, (log) word frequency, and number of word senses. Higher ranked cosine similarity corresponds to words whose representations were changed greatly through fine-tuning M^{CW} on free association data. The fifth column presents the simple pairwise correlation between the ranked cosine similarity and each of the variables. CI = confidence interval.

We believe several directions for future research will be fruitful. First, the choice model we used might be modified or extended in various ways. For example, following previous suggestions (e.g., Jones et al., 2018), we initially tried including response-specific biases w_b at the word layer, such that $w_t = w_b + M^{\mathrm{CW}} \cdot c_t$. Intuitively, these biases capture a response word's baseline activation (with bias $w_{i,b}$ for word i). We attempted to fit these biases alongside the other parameters of our model, but found that the resulting models made poor out-of-sample predictions (likely due to overfitting of the bias weights to response frequencies in the training data). Instead of adjusting our model implementation to downscale these biases, which would arguably constitute leakage from our test set, we opted to simply drop biases from our models altogether. Future work might therefore investigate better ways of incorporating biases, possibly based on word frequency statistics, into modeling of free association.

Second, future work might also investigate how individual heterogeneity is reflected in model parameters. For example, age-related changes in memory capacity may be reflected in both γ and δ in our continued-response model: those with poorer memory might have both lower δ, reflecting less memory for previously recalled items, and lower γ, reflecting less memory that the current trial involves responding to the cue, and not the most recent response(s). This can be tested by separately fitting our models on different populations, like older and younger adults. Prior research on free recall from lists and semantic fluency has analyzed age-related differences in a similar manner, and has found that aging affects both the structure of associations between words (e.g., Dubossarsky et al., 2017; Zemla & Austerweil, 2019; also see Ramscar et al., 2017 for a related result) and the retrieval processes that operate on these representations (e.g., Healey & Kahana, 2016; Hills et al., 2013). The methods in this article show how this type of modeling can be scaled up to unconstrained free-association tasks. Fine-tuning M^{CW} may also be able to shed light on differences in underlying response representations across groups, and by doing so, better predict differences in free associations across groups. We believe that this could be a valuable tool for studying cultural, developmental, and ideological differences in mental representations (e.g., Bhatia et al., 2018; DeFranza et al., 2020; Holtzman et al., 2011).

Third, in the present work, we have only focused on modeling responses in free association, largely for the sake of simplicity. Of course, behavior in this task unfolds over time and so a more comprehensive model of free association ought to account for the time each response takes. The SWOW-EN group has also collected response times but is currently (at the time of writing) preparing these for publication (De Deyne, personal communication). When these become available, extending our framework to predict response times will be a natural, if challenging, direction. Models that account for response times as well as choices, like the Drift Diffusion Model (Ratcliff & McKoon, 2008) or the latching attractor network for semantic priming (Lerner et al., 2012), may be useful here, as might CMR, which models choice and response times in free recall as a competitive leaky accumulation process (Usher & McClelland, 2001). Other models include the linear ballistic accumulator (Brown & Heathcote, 2008) which can provide a more tractable account of free recall dynamics in the presence of multiple possible response words, as the case with free association (Osth & Farrell, 2019).

One major limitation of our approach is that the fine-tuning exercise can only adapt M^{CW} weights to response words with sufficient training data. This is why we have had to exclude highly infrequent response words from our analysis. We believe that it is vital for researchers to develop methods that address such data sparsity problems if they are to extend cognitive models to unconstrained tasks such as free association. One approach would be to simply assume that response vectors for infrequent words are the same in M^{CW} as in M^{WC}. This would allow us to make predictions over words in the full GloVe vocabulary. A more complex approach would be to try to model how fine-tuning transforms M^{WC} into M^{CW}, and by doing so, modify vector representations for infrequent words in M^{WC} into analogous representations in M^{CW}. Identical methods (based on the observation that similar words in different languages have similar relative positions in their corresponding semantic spaces) have already been proposed for machine translation (Lample et al., 2018; Zou et al., 2013) and could potentially be extended to the problem of free association.

Our work also shows how the transfer learning approach that has driven success in machine learning and artificial intelligence applications (e.g., Devlin et al., 2019; Pan & Yang, 2009) holds great promise for cognitive modeling. By initializing our model's response vectors with pretrained word embeddings, we were able to train effective models using far less data than would have been necessary if we had trained these parameters from scratch. We have previously proposed related techniques for modeling judgments of words on various dimensions (Bhatia, 2019; Bhatia et al., in press; Gandhi et al., 2022; Richie et al., 2019; Zou & Bhatia, 2021; also see Hollis et al., 2017; Van Rensbergen et al., 2016; Sedoc et al., 2017). Other work, including ours, has also shown the applicability of such methods for the study of concept knowledge (Bhatia & Richie, in press; Derby et al., 2019; Lu et al., 2019; Richie & Bhatia, 2021). In fact, our training exercise can be seen as a type of multinomial regression that learns word-specific weights for mapping cue and context vectors into response probabilities, making it almost identical to the approach proposed in this prior work.

Given the flexibility inherent in this approach, readers may be concerned that our model may have overfit the data, or may have been able to simply mimic the data without uncovering any fundamental truths about free association. We believe that this is not the case. First, we have evaluated our models through cross validation, and our tests have shown that our trained model is able to generalize to new cues with a high degree of accuracy. Second, despite flexibly fine tuning a large set of weights, the CMR model used in our tests places strict constraints on the types of dynamics that we may observe in the free association task. In particular, CMR describes the effect of previously retrieved words using a linear function with a single decay parameter. This forces response chaining effects to take on a very narrow functional form. Likewise, our implementation of cue effects as additive limits the ways in which cues and previous responses can combine to determine subsequent retrieval. One implication of this is that our model cannot capture the complexities of language. Thus, for example, our model would predict that the sequence of retrievals (*my*, *house*, *is*, *made*, *of*) would be followed by a word that is weighted average (in semantic space) of this set, with a higher weight on more recent words. It would not be able to predict the retrieval of words like *wood* or *bricks*. In this way, our CMR model is much more constrained than leading deep neural network models of language like BERT and GPT3 (Brown et al., 2020; Devlin et al., 2019), which are equipped with multiple layers of representation and are able to process basic types of linguistic structure, and (unlike our model) compose the vectors of words in complex and nonlinear ways. Of course, our model's additive constraint is what allows us to efficiently learn weights from free association data, and subsequently predict free association with accuracy. One could, in this sense, interpret CMR's constraints as inductive priors for modeling free association in neural language models. These priors (which propose that response dynamics are linear) are realistic and lead to good performance. More flexible models like BERT and GPT3 could eventually replicate our performance levels but it is likely that these models would need much more training data than we have (and would thus perform poorly in the tests that we have conducted).

Of course, the CMR model is not the most constrained model that we have considered. The symmetric case (which sets both sets of weights to be identical) is far more restrictive. We have shown that this model performs poorly. In this sense, the asymmetry in the CMR model is the right level of complexity for describing behavior in our task. In a similar vein, we have shown that a model with full decay (the Markov model) or a model without a persistent cue effect perform poorly. Thus, our assumptions of decay and cue context are the right assumptions for capturing the dynamics of our task. However, despite these successes, we believe that further progress depends on the development of psychologically and developmentally plausible algorithms for learning the weights at play in free association. Our approach currently does not describe how people actually learn these weights. Rather, it assumes that these weights are free parameters that can be fit on free association data, similar to how cognitive modelers fit relevant model parameters to participant data. Perhaps the most promising approach to modeling realistic learning processes is the pTCM proposed by Howard et al. (2011; see also Shankar et al., 2009), and discussed in previous sections. Although pTCM has been shown to predict cue–response probabilities in free association data, it is also able to extract word-to-context and context-to-word associations from natural language, using learning mechanisms adapted from the original CMR model. In this way, associations obtained from pTCM can replace the starting GloVe weights in our modeling framework. This would provide a much more cohesive account of the learning and retrieval processes at play in free association. Note that there have, until recently, been practical constraints to applying pTCM to large vocabularies, as the algorithm involves multiplying several *N*x*N* matrices (where *N* is the vocabulary size). However, newly developed deep learning libraries, such as those used in the current article, may solve such constraints, as they come equipped with specialized tools for large matrix multiplication.

Another direction for theoretical development involves the combination of CMR-based retrieval dynamics with additional cognitive processes to enable more sophisticated types of verbal cognition. CMR provides a theory of how sequences of words come to mind in unconstrained memory tasks, but in its current form, it lacks a mechanism for using these association-based sequences for sentence processing or relational reasoning. The syntagmatic paradigmatic (SP) model (Dennis, 2005) provides one account of how associative recall is modulated and controlled in verbal cognition, and future work could consider applying its assumptions to the associations possessed by CMR.

Finally, one major advance of our work is that it presents a framework that generalizes theories for related tasks previously studied and modeled separately. Our approach suggests that, faced with a particular task (e.g., free recall from lists or free association), participants may modulate their cognition and behavior in task-dependent ways (which we can capture with adjustments in parameters like γ and δ), while keeping the underlying representations (M^{CW} and M^{WC}) fixed. In particular, in the free association setting we studied, we found both feedback from previously recalled items ($\delta > 0$), and a persistent effect of the cue ($\gamma > 0$). However, the same system can be deployed in free recall from lists, and can mimic (a simplified version of) CMR when $\gamma = 0$ (which eliminates the persistent cue effect), as is appropriate for that task. Additionally, our tests showed that a CMR model fit to free association data provided superior predictions for semantic clustering effects in free recall from lists.

That said, the assumptions of our model are insufficient for describing the full range of memory tasks studied in the literature. For example, by itself, our framework would be unable to describe the learning of paired associates, or, more generally, episodic associations that exist alongside the semantic associations possessed by our model. There have been several attempts to build general memory models that describe the interplay between (and in many cases, the interference between) episodic and semantic memory (e.g., Humphreys et al., 1989; Murdock, 1993; Nelson, Thomas, et al., 1998; see also Cox et al., 2018). Many of these models posit episodic representations that can be combined with the semantic representations of our model to generate realistic responses in tasks with episodic learning. Further developing the modeling framework we have presented here, and deploying it on a variety of related tasks involving the recall of words, is likely to be a fruitful area of research for years to come.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review*, *122*(3), 558–569. https://doi.org/10.1037/a0038693

Aka, A., Phan, T. D., & Kahana, M. J. (2021). Predicting recall of words and lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(5), 765–784. https://doi.org/10.1037/xlm0000964

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*(3), 463–498. https://doi.org/10.1037/a0016261

Arora, S., Li, Y., Liang, Y., Ma, T., & Risteski, A. (2016). A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, *4*, 385–399. https://doi.org/10.1162/tacl_a_00106

Asr, F. T., Zinkov, R., & Jones, M. (2018). *Querying word embeddings for similarity and relatedness* [Conference session]. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, United States.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation*, *2*(4), 89–195. https://doi.org/10.1016/S0079-7421(08)60422-3

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*(2), 283–316. https://doi.org/10.1037/0096-3445.133.2.283

Bhatia, S. (2017a). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20. https://doi.org/10.1037/rev0000047

Bhatia, S. (2017b). The semantic representation of prejudice and stereotypes. *Cognition*, *164*, 46–60. https://doi.org/10.1016/j.cognition.2017.03.016

Bhatia, S. (2019). Predicting risk perception: New insights from data science. *Management Science*, *65*(8), 3800–3823. https://doi.org/10.1287/mnsc.2018.3121

Bhatia, S., & Aka, A. (2022). Cognitive modeling with representations from large-scale digital data. *Current Directions in Psychological Science*, *31*(3), 207–214. https://doi.org/10.1177/09637214211068113

Bhatia, S., Goodwin, G. P., & Walasek, L. (2018). Trait associations for Hillary Clinton and Donald Trump in news media: A computational analysis. *Social Psychological & Personality Science*, *9*(2), 123–130. https://doi.org/10.1177/1948550617751584

Bhatia, S., Olivola, C., Bhatia, N., & Ameen, A. (in press). Predicting leadership perception with large-scale natural language data. *The Leadership Quarterly*.

Bhatia, S., & Richie, R. (in press). Transformer networks of human concept knowledge. *Psychological Review*.

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, *29*, 31–36. https://doi.org/10.1016/j.cobeha.2019.01.020

Bhatia, S., & Stewart, N. (2018). Naturalistic multiattribute choice. *Cognition*, *179*, 71–88. https://doi.org/10.1016/j.cognition.2018.05.025

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *The Journal of General Psychology*, *49*(2), 229–240. https://doi.org/10.1080/00221309.1953.9710088

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., . . . Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Cattell, J. M. (1887). Experiments on the association of ideas. *Mind*, *12*(45), 68–74. https://doi.org/10.1093/mind/os-12.45.68

Chaudhari, D. L., Damani, O. P., & Laxman, S. (2011). *Lexical co-occurrence, statistical significance, and word association* [Conference session]. Proceedings EMNLP–2011, Edinburgh, Scotland, United Kingdom.

Chollet, F. (2015). *Keras*. GitHub. https://github.com/fchollet/keras

Clark, H. H. (1970). Word associations and linguistic theory. *New Horizons in Linguistics*, *1*, 271–286.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247. https://doi.org/10.1016/S0022-5371(69)80069-1

Cox, G. E., Hemmer, P., Aue, W. R., & Criss, A. H. (2018). Information and processes underlying semantic and episodic memory across tasks, items, and individuals. *Journal of Experimental Psychology: General*, *147*(4), 545–590. https://doi.org/10.1037/xge0000407

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987–1006. https://doi.org/10.3758/s13428-018-1115-7

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, *45*(2), 480–498. https://doi.org/10.3758/s13428-012-0260-7

De Deyne, S., Perfors, A., & Navarro, D. J. (2016). *Predicting human similarity judgments with distributional models: The value of word associations* [Conference session]. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan.

De Deyne, S., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior Research Methods*, *40*(1), 213–231. https://doi.org/10.3758/BRM.40.1.213

Deese, J. (1959). Influence of inter-item associative strength upon immediate free recall. *Psychological Reports*, *5*(3), 305–312. https://doi.org/10.2466/pr0.1959.5.3.305

Deese, J. (1962). On the structure of associative meaning. *Psychological Review*, *69*(3), 161–175. https://doi.org/10.1037/h0045842

DeFranza, D., Mishra, H., & Mishra, A. (2020). How language shapes prejudice against women: An examination across 45 world languages. *Journal of Personality and Social Psychology*, *119*(1), 7–22. https://doi.org/10.1037/pspa0000188

Dennis, S. (2005). A memory-based theory of verbal cognition. *Cognitive Science*, *29*(2), 145–193. https://doi.org/10.1207/s15516709cog0000_9

Derby, S., Miller, P., & Devereux, B. (2019, November). *Feature2Vec: Distributional semantic modelling of human property knowledge* [Conference session]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, January). *BERT: Pre-training of deep bidirectional transformers for language understanding* [Conference session]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, United States.

Dubossarsky, H., De Deyne, S., & Hills, T. T. (2017). Quantifying the structure of free association networks across the life span. *Developmental Psychology*, *53*(8), 1560–1570. https://doi.org/10.1037/dev0000347

Freud, S. (1913/1958). On beginning the treatment: Further recommendations on the technique of psychoanalysis. In J. Strachey (Ed.), *The standard edition of the complete works of sigmund freud* (Vol. 12, pp. 121–144). Hogarth Press. (Original work published 1913).

Galton, F. (1880). Psychometric experiments. *Brain: A Journal of Neurology*, *2*(2), 149–162. https://doi.org/10.1093/brain/2.2.149

Gandhi, N., Zou, W., Meyer, C., Bhatia, S., & Walasek, L. (2022). Computational methods for predicting and understanding food judgment. *Psychological Science*, *33*(4), 579–594. https://doi.org/10.1177/09567976211043426

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*(7), 975–987. https://doi.org/10.1038/s41562-022-01316-8

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Gruenenfelder, T. M., Recchia, G., Rubin, T., & Jones, M. N. (2016). Graph-theoretic properties of networks based on word association norms: Implications for models of lexical semantic memory. *Cognitive Science*, *40*(6), 1460–1495. https://doi.org/10.1111/cogs.12299

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033. https://doi.org/10.1177/1745691619861372

Healey, M. K., & Kahana, M. J. (2016). A four-component model of age-related memory change. *Psychological Review*, *123*(1), 23–69. https://doi.org/10.1037/rev0000015

Healey, M. K., & Uitvlugt, M. G. (2019). The role of control processes in temporal and semantic contiguity. *Memory & Cognition*, *47*(4), 719–737. https://doi.org/10.3758/s13421-019-00895-8

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–440. https://doi.org/10.1037/a0027373

Hills, T. T., Mata, R., Wilke, A., & Samanez-Larkin, G. R. (2013). Mechanisms of age-related decline in memory search across the adult life span. *Developmental Psychology*, *49*(12), 2396–2404. https://doi.org/10.1037/a0032272

Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *70*(8), 1603–1619. https://doi.org/10.1080/17470218.2016.1195417

Holtzman, N. S., Schott, J. P., Jones, M. N., Balota, D. A., & Yarkoni, T. (2011). Exploring media bias with semantic analysis tools: Validation of the Contrast Analysis of Semantic Similarity (CASS). *Behavior Research Methods*, *43*(1), 193–200. https://doi.org/10.3758/s13428-010-0026-z

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 923–941. https://doi.org/10.1037/0278-7393.25.4.923

Howard, M. W., & Kahana, M. J. (2002a). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299. https://doi.org/10.1006/jmps.2001.1388

Howard, M. W., & Kahana, M. J. (2002b). When does semantic similarity help episodic retrieval? *Journal of Memory and Language*, *46*(1), 85–98. https://doi.org/10.1006/jmla.2001.2798

Howard, M. W., Shankar, K. H., & Jagadisan, U. K. (2011). Constructing semantic representations from a gradually-changing representation of temporal context. *Topics in Cognitive Science*, *3*(1), 48–73. https://doi.org/10.1111/j.1756-8765.2010.01112.x

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, *96*(2), 208–233. https://doi.org/10.1037/0033-295X.96.2.208

Ji, H., Lemaire, B., Choo, H., & Ploux, S. (2008). Testing the cognitive relevance of a geometric model on a word association task: A comparison of humans, ACOM, and LSA. *Behavior Research Methods*, *40*(4), 926–934. https://doi.org/10.3758/BRM.40.4.926

Jones, M. N., Gruenenfelder, T. M., & Recchia, G. (2018). In defense of spatial models of semantic representation. *New Ideas in Psychology*, *50*, 54–60. https://doi.org/10.1016/j.newideapsych.2017.08.001

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552. https://doi.org/10.1016/j.jml.2006.07.003

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

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*(1), 103–109. https://doi.org/10.3758/BF03197276

Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychology*, *71*(1), 107–138. https://doi.org/10.1146/annurev-psych-010418-103358

Kahana, M. J., Aggarwal, E. V., & Phan, T. D. (2018). The variability puzzle in human memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1857–1863. https://doi.org/10.1037/xlm0000553

Kahana, M. J., & Caplan, J. B. (2002). Associative asymmetry in probed recall of serial lists. *Memory & Cognition*, *30*(6), 841–849. https://doi.org/10.3758/BF03195770

Kenett, Y. N., Anaki, D., & Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons. *Frontiers in Human Neuroscience*, *8*, Article 407. https://doi.org/10.3389/fnhum.2014.00407

Kenett, Y. N., Levi, E., Anaki, D., & Faust, M. (2017). The semantic distance task: Quantifying semantic distance with semantic network path length. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(9), 1470–1489. https://doi.org/10.1037/xlm0000391

Kintsch, W. (2014). Similarity as a function of semantic distance and amount of knowledge. *Psychological Review*, *121*(3), 559–561. https://doi.org/10.1037/a0037017

Kumar, A. A., Balota, D. A., & Steyvers, M. (2019). *Distant concept connectivity in network-based and spatial word representations* [Conference session]. Proceedings of the 41st Annual Conference of the Cognitive Science Society, Montreal, Canada.

Kumar, A. A., Balota, D. A., & Steyvers, M. (2020). Distant connectivity and multiple-step priming in large-scale semantic networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(12), 2261–2276. https://doi.org/10.1037/xlm0000793

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Ladd, G. T., & Woodworth, R. S. (1911). *Elements of physiological psychology: A treatise of the activities and nature of the mind from the physical and experimental point of view*. Scribner.

Lample, G., Conneau, A., Ranzato, M. A., Denoyer, L., & Jégou, H. (2018). *Word translation without parallel data* [Conference session]. International Conference on Learning Representations, Vancouver Convention Center, Vancouver, Canada.

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

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*(1), 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Lerner, I., Bentin, S., & Shriki, O. (2012). Spreading activation in an attractor network with latching dynamics: Automatic semantic priming revisited. *Cognitive Science*, *36*(8), 1339–1382. https://doi.org/10.1111/cogs.12007

Levy, O., & Goldberg, Y. (2014). *Neural word embedding as implicit matrix factorization* [Conference session]. Advances in Neural Information Processing Systems, Cambridge, Massachusetts, United States.

Lohnas, L. J., & Kahana, M. J. (2014). Compound cuing in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 12–24. https://doi.org/10.1037/a0033698

Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10), 4176–4181. https://doi.org/10.1073/pnas.1814779116

Luce, R. D. (1959). *Individual choice behavior*. Wiley.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Matusevych, Y., & Stevenson, S. (2018). Analyzing and modeling free word associations. In C. Kalish, M. Rau, T. Rogers, & J. Zhu (Eds.), *Proceedings of the 40th annual conference of the cognitive science society* (pp. 750–755). Cognitive Science Society.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 1–9). Curran Associates.

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. https://doi.org/10.1016/j.jml.2015.10.002

Murdock, B. B., Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*(5), 482–488. https://doi.org/10.1037/h0045106

Murdock, B. B. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review*, *100*(2), 183–203. https://doi.org/10.1037/0033-295X.100.2.183

Nelson, C. A., Thomas, K. M., de Haan, M., & Wewerka, S. S. (1998). Delayed recognition memory in infants and adults as revealed by event-related potentials. *International Journal of Psychophysiology*, *29*(2), 145–165. https://doi.org/10.1016/S0167-8760(98)00014-2

Nelson, D. L., Bennett, D. J., & Leibert, T. W. (1997). One step is not enough: Making better use of association norms to predict cued recall. *Memory & Cognition*, *25*(6), 785–796. https://doi.org/10.3758/BF03211322

Nelson, D. L., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition*, *28*(6), 887–899. https://doi.org/10.3758/BF03209337

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. https://doi.org/10.3758/BF03195588

Nelson, D. L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review*, *105*(2), 299–324. https://doi.org/10.1037/0033-295X.105.2.299

Nematzadeh, A., Meylan, S. C., & Griffiths, T. L. (2017). *Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words* [Conference session]. Proceedings of the 39th Annual Conference of the Cognitive Science Society, London, United Kingdom.

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin*, *142*(7), 758–799. https://doi.org/10.1037/bul0000046

Osth, A. F., & Farrell, S. (2019). Using response time distributions and race models to characterize primacy and recency effects in free recall initiation. *Psychological Review*, *126*(4), 578–609. https://doi.org/10.1037/rev0000149

Osth, A. F., & Hurlstone, M. J. (2022). Do item-dependent context representations underlie serial order in cognition? Commentary on Logan (2021). *Psychological Review*. Advance online publication. https://doi.org/10.1037/rev0000352

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Peirsman, Y., & Geeraerts, D. (2009). *Predicting strong associations on the basis of corpus data* [Conference session]. Proceedings of the 12th Conference of the European Chapter of the ACL, Athens, Greece.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). *GloVe: Global vectors for word representation* [Conference session]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar.

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), 129–156. https://doi.org/10.1037/a0014420

Posnansky, C. J. (1972). Probing for the functional stimuli in serial learning. *Journal of Experimental Psychology*, *96*(1), 184–193. https://doi.org/10.1037/h0033503

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*(2), 93–134. https://doi.org/10.1037/0033-295X.88.2.93

Ramscar, M., Sun, C. C., Hendrix, P., & Baayen, H. (2017). The mismeasurement of mind: Life-span changes in paired-associate-learning scores reflect the “cost” of learning, not cognitive decline. *Psychological Science*, *28*(8), 1171–1179. https://doi.org/10.1177/0956797617706393

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Richie, R., & Bhatia, S. (2021). Similarity judgment within and across categories: A comprehensive model comparison. *Cognitive Science*, *45*(8), Article e13030. https://doi.org/10.1111/cogs.13030

Richie, R., Zou, W., & Bhatia, S. (2019). Predicting high-level human judgment across diverse behavioral domains. *Collabra. Psychology*, *5*(1), Article 50. https://doi.org/10.1525/collabra.282

Romney, A. K., Brewer, D. D., & Batchelder, W. H. (1993). Predicting clustering from semantic structure. *Psychological Science*, *4*(1), 28–34. https://doi.org/10.1111/j.1467-9280.1993.tb00552.x

Rotaru, A. S., Vigliocco, G., & Frank, S. L. (2018). Modeling the structure and dynamics of semantic processing. *Cognitive Science*, *42*(8), 2890–2917. https://doi.org/10.1111/cogs.12690

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, Article e55. https://doi.org/10.7717/peerj-cs.55

Sedoc, J., Preoţiuc-Pietro, D., & Ungar, L. (2017). *Predicting emotional word ratings using distributional representations and signed clustering* [Conference session]. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers.

Shankar, K. H., Jagadisan, U. K., & Howard, M. W. (2009). Sequential learning using temporal context. *Journal of Mathematical Psychology*, *53*(6), 474–485. https://doi.org/10.1016/j.jmp.2009.07.005

Snefjella, B., & Blank, I. (2020). Semantic norm extrapolation is a missing data problem. *PsyArXiv*. https://doi.org/10.31234/osf.io/y2gav

Socher, R., Gershman, S., Sederberg, P., Norman, K., Perotte, A., & Blei, D. (2009). A Bayesian analysis of dynamics in free recall. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22, pp. 1–9). Curran Associates.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78. https://doi.org/10.1207/s15516709cog2901_3

Tulving, E. (1962). Subjective organization in free recall of “unrelated” words. *Psychological Review*, *69*(4), 344–354. https://doi.org/10.1037/h0043150

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352. https://doi.org/10.1037/0033-295X.84.4.327

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. https://doi.org/10.1037/0033-295X.108.3.550

Utsumi, A. (2014). A semantic space approach to the computational semantics of noun compounds. *Natural Language Engineering*, *20*(2), 185–234. https://doi.org/10.1017/S135132491200037X

Utsumi, A. (2020). Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, *44*(6), Article e12844. https://doi.org/10.1111/cogs.12844

Van Rensbergen, B., De Deyne, S., & Storms, G. (2016). Estimating affective word covariates using word association data. *Behavior Research Methods*, *48*(4), 1644–1652. https://doi.org/10.3758/s13428-015-0680-2

Zemla, J. C., & Austerweil, J. L. (2018). Estimating semantic networks of groups and individuals from fluency data. *Computational Brain & Behavior*, *1*(1), 36–58. https://doi.org/10.1007/s42113-018-0003-7

Zemla, J. C., & Austerweil, J. L. (2019). Analyzing knowledge retrieval impairments associated with Alzheimer’s disease using network analyses. *Complexity*, *2019*, Article 4203158. https://doi.org/10.1155/2019/4203158

Zemla, J. C., Cao, K., Mueller, K. D., & Austerweil, J. L. (2020). SNAFU: The semantic network and fluency utility. *Behavior Research Methods*, *52*(4), 1681–1699. https://doi.org/10.3758/s13428-019-01343-w

Zou, W., & Bhatia, S. (2021). Judgment errors in naturalistic numerical estimation. *Cognition*, *211*, Article 104647. https://doi.org/10.1016/j.cognition.2021.104647

Zou, W. Y., Socher, R., Cer, D., & Manning, C. D. (2013). *Bilingual word embeddings for phrase-based machine translation* [Conference session]. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, United States.

Received October 27, 2021
Revision received August 10, 2022
Accepted August 16, 2022 ■