Title
A Computational Model of Lexical False Memory based on Semantic Distance from Word Embeddings

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors
Ham, Huang
Jenkins, Adrianna C.

Publication Date
2023

Peer reviewed

A Computational Model of Lexical False Memory based on Semantic Distance from Word Embeddings

Huang Ham
hamhuang@sas.upenn.edu
University of Pennsylvania

Adrianna C. Jenkins
acjenk@sas.upenn.edu
University of Pennsylvania

Abstract

Human memory does not simply function like an information storage disk; instead, it flexibly reorganizes information. This flexibility can sometimes produce false memories of items related to those actually encountered–a possible byproduct of an adaptive memory system that enables generalization across related items or experiences. In the Deese/Roediger-McDermott (DRM) task, participants often falsely remember seeing words that are semantically related to presented words. Here, we pro- pose and test a model of lexical (false) memory that predicts these errors, made possible by integrating (i) theories of memory that posit encoding of verbatim and gist-level information with (ii) a computational framework adapted from the perceptual false memory literature, and (iii) semantic relatedness measures from word embeddings analysis of large-scale text corpora. This Lexical Target Confusability Competition (Lexical TCC) model successfully predicts human participants' false recognition in the DRM task, with implications for understanding how and when the mind produces false semantic memories.

Keywords: lexical false memory; Deese/Roediger-McDermott Task; signal detection; computational modeling; word embeddings

Introduction

People sometimes recall items or events they never actually encountered – that is, they form false memories. The formation of false memories has been observed across multiple forms of memory, including episodic memory (Loftus & Palmer, 1974), visual working memory (Schurgin, Wixted, & Brady, 2020), and semantic memory (Jou & Flores, 2013). One explanation for why people form false semantic memories is that the mind encodes not only verbatim information about encountered items but also their semantic gist (Diana, Reder, Arndt, & Park, 2006; Zeng, Tompary, Schapiro, & Thompson-Schill, 2021). For example, in the classic Deese/Roediger-McDermott (DRM) paradigm (Deese, 1959; Roediger & McDermott, 1995; Brainerd, Gomes, & Moran, 2014), participants encounter sets of words (e.g., "doze", "pillow", "snore", "awake"…) and sometimes falsely remember encountering semantically-related *lure* words (e.g., "sleep") that were never presented.

False memory effects are of scientific interest not only because of their implications in settings ranging from the classroom to the courtroom but also because they may be the byproduct of – and, accordingly, help us understand – an adaptive cognitive system that enables the mind to generalize knowledge across related, but not identical, experiences (Schacter, 1999). Evidence further suggests that this system may be disrupted in certain neurodevelopmental disorders, including autism, such that some individuals are less prone to false memory errors (Beversdorf et al., 2000, 1998). At the same time, although semantic false memory effects are well documented, it has been challenging to distinguish empirically among different theoretical accounts of the mechanisms that produce them.

Here, we propose a new computational model of lexical false memory that integrates existing models of perceptual false memory with measures of word relationships generated by deep-learning language models trained on large-scale text corpora. Because this model is an adaptation of the *Target Confusability Competition* (TCC) model developed in the perceptual false memory literature, we call it a *Lexical Target Confusability Competition* (Lexical TCC) model. First, we motivate and describe the Lexical TCC model in detail. Next, we show that this model successfully reproduces well-documented false memory effects in human performance on the RDM paradigm. Lastly, we validate the model against new DRM data from human participants and show that the model successfully predicts individual differences in false memory. The code associated with this paper is accessible on the open science framework (https://osf.io/5jubw/).

Model Framework

While false memory for lure words surely arises due to some properties of the words in the corresponding lists, existing frameworks differ in their proposals for what the key properties are, including the similarity among words, the association strength among words, and the meaning of the list as a whole (Brainerd, Chang, & Bialer, 2020; Cann, McRae, & Katz, 2011). One account of 'semantic gist' that regards it as a special instance of conceptual generalization (Destefano, Brady, & Vul, 2021), in which people generalize beyond the specific words they encountered to words that are semantically related (including the lure words). Modeling lexical false memory as conceptual generalization has the theoretic advantage of uniting it with other phenomena under the Bayesian framework of cognition (Tenenbaum & Griffiths, 2001), which has already been applied successfully to model various aspects of semantic memory (Griffiths, Steyvers, & Tenenbaum, 2007; Steyvers & Griffiths, 2008) and memory biases (Wilson, Arora, Zhang, & Griffiths, 2021).

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

Conceptual Motivation

Our lexical TCC model borrows insights directly from the original *Target Confusability Competition* (TCC) model developed to capture the phenomenon that false recognition of a given color increases with the perceptual similarity of that color to what people actually saw (Schurgin et al., 2020). For example, if you saw a shade of green, then you are more likely to falsely remember seeing another shade of green than seeing red. The TCC model accounts for this pattern by positing that, during the encoding of a target color, memory activation spreads to nearby colors in perceptual space and thus causes a boost in familiarity with other colors as well. The intensity of such a boost is proportional to the similarity with the target color as quantified by the color wheel and degrades exponentially as similarity decreases, in line with the universal law of generalization (Shepard, 1987). Such memory activation is further corrupted by noise, according to signal detection theory, be- fore guiding people's decision as to which color is one they saw before.

The classic lexical false memory effects identified in the DRM task display some similar features. People are more likely to falsely remember words that have a similar meaning as the words they actually saw than to remember less related words. However, with few exceptions (Johns, Jones, & Mewhort, 2012), past approaches have typically relied on some combination of researcher intuition and participant ratings to test ideas about the role of semantic relatedness in these effects. Thanks to recent developments in natural language processing, it is now possible to use word embeddings as a tool to quantify the semantic similarity between words, much like the color wheel. Word embedding represents each English word using a vector in a high-dimensional space, making it possible to quantify relationships between words. Here, we used the google-300 un-normalized Word2vec embeddings, which were trained using word co-occurrence data from a massive Google news corpus and which mapped each word onto a vector in a 300-dimensional vector space (Mikolov, Chen, Corrado, & Dean, 2013). The similarity between two words can thus be quantified as the Euclidean distance between their vector embeddings. Formally, let $\vec{x}_1$ and $\vec{x}_2$ represent the vector embedding of two words, their semantic distance is defined as $d = \|\vec{x}_1 - \vec{x}_2\|_2$. The smaller the d, the greater the similarity between two words.

It is worth noting that there is a crucial difference between the DRM task and the color memory task for which the original TCC model was designed. In the color memory task, participants see only one color and then have to choose what they recognize as the color they just saw among several color options in an alternative forced choice paradigm. In the DRM task, people typically view (or hear) a series of words to remember, presented individually and temporally organized into lists. Later, they view another series of words presented individually and report if the word was or was not presented earlier. Therefore, while in the color memory task, only one item triggers the relevant memory activation, in the DRM task, multiple words together trigger it. The other key difference is between the alternative forced choice paradigm in the perceptual task and the binary decision in the DRM task.

Details of the lexical TCC model

To model joint memory activation from many items, we conceptualize memory activation as a scalar field in the semantic embedding space. It is analogous to an electric field in physics whether each word is like an electric charge. An electric charge incurs an electric field in the 3-dimensional space surrounding the charge. Similarly, the encoding of each word produces a memory activation field in the 300-dimensional embedding space centered at the vector embedding for the word. If more than one electric charges are present in the space, their electric fields simply add up to one another. Similarly, when many words are encoded in the memory, the memory activation fields incurred by encoding each word also add up in the embedding space.

Formally, if the participant is asked to remember $n \in \mathbb{N}$ words in a DRM task where the vector presentation for each word is denoted as $\vec{x}_1 ... \vec{x}_n \in \mathbb{R}^{300}$, then the memory activation at each point in the semantic space $\vec{x}$ is given by the function $f : \mathbb{R}^{300} \to \mathbb{R}$

$$f(\vec{x}) = \sum_{i=1}^{n} f_i(\vec{x})$$

where f_i represents the memory activation field incurred by encoding the i^{th} word. Following the visual TCC model, we define f_i as an exponential function

$$f_i(\vec{x}) = \exp(-\|\vec{x} - \vec{x}_i\|_2^2 \sigma^{-1})$$

where σ^{-1} is a free parameter that captures the extent to which the memory activation caused by each word is generalized, which is assumed to be the same for all words. The smaller the σ^{-1}, the greater the generalization; the greater the σ^{-1}, the more locally precise is each memory activation. Therefore σ^{-1} is called the *precision* parameter.

It is worth noting that f_i is not an exponential function of the Euclidean distance $\|\vec{x} - \vec{x}_i\|_2$, but rather of the square of the distance $\|\vec{x} - \vec{x}_i\|_2^2$. This design is motivated by the sparsity of the word embedding space, where very few pairs of words are adjacent to each other in the embedding space. Consequently, if the memory activation decays according to an exponential function of the Euclidean distance, all other words are far away enough that they would get minimal memory activations, leaving little possibility for systematic lexical false memory. Squaring the Euclidean distance can make the intensity of memory activation fall slower in the semantic vicinity of the encoded words, and fall faster at a reasonable distance.

Returning to the example where the participants were asked to remember 4 words:'bed', 'yawn', 'rest', and 'doze', figure 1b illustrates the memory activation field generated by the 4 words with a relatively high σ^{-1}. The canvas represents the semantic space and the location of the words represents their vector representation in the semantic space. This is

only a 2-dimensional illustration whereas the actual semantic space is 300-dimensional. The brightness of color at each point represents the intensity of memory activation at that point. With high σ^{-1}, we see that memory activation is the highest near the 4 studied words. Even though the lure word 'sleep' is not one of the studied words, due to its semantic similarity, it still received some boost in memory activation. However, with sufficiently low σ^{-1}, the memory activation incurred by the 4 studied words can generalize far enough that once accumulated, cause the lure word 'sleep' to have the highest memory activation, even higher than the actual studied words (figure 1a). This scenario corresponds to the situation where participants are more likely to falsely recognize the lure words than the actual studied words.

The second phase of the model is to translate the memory activation of a word into the probability of recognizing it. Following signal detection theory, we assume the activation is corrupted by Gaussian noise with standard deviation ε (figure 1c). Let $\vec{x}$ be the vector representation of the word that the participant is asked to recall, then the likelihood of them remembering it is given by

$$p(\text{remember}) = 1 - \Phi(\frac{\tau - f(\vec{x})}{\varepsilon})$$

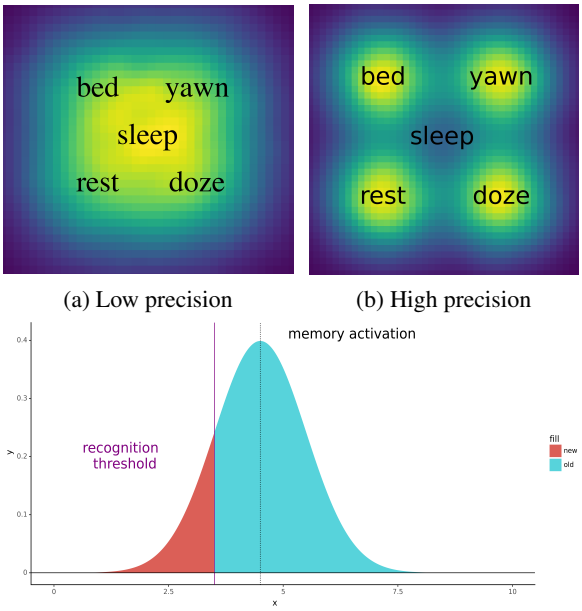
where Φ denotes the cumulative density function of the standard normal distribution and τ is a free parameter that captures parameter memory activation threshold. The larger the τ, the greater memory activation is required to remember a word. For numeric stability, we standardize the memory activations within each subject before converting it to the probability of remembering.

In sum, the lexical TCC model has three free parameters that can be fit to each participant: the precision parameter σ^{-1} which controls the spread of memory activation in the semantic space, the threshold parameter τ that controls the overall tendency of recalling, and the noise parameter ε that controls the extent to which the difference in memory activation translates to the difference in the likelihood of recall. The lexical TCC model can be thought of as an instance of the global matching models (Osth & Dennis, 2020). Its mathematical formality is particularly similar to the generalized context model (Nosofsky, 1991). Despite the similarity in mathematical formality, however, the generalized context model was not proposed to model lexical memory but rather perceptual memory of stimuli with simple features.

Assessing the Model Validity

Empirical data

We analyzed data from the DRM portion of an experiment for which we recruited 120 participants on the Prolific platform using their standard sample option. Participants between 18 and 60 years old who resided in the US were eligible to participate. Participants received a median payment of $15.64. Data from 107 people remained after exclusion based on attention checks. The experimental protocol was approved by the local IRB.



(a) Low precision (b) High precision

(c) Noise corruption

Figure 1: (a) Illustrates the memory activation field with a low precision parameter, causing the lure word to have higher familiarity than the studied words. (b) Illustrates the memory activation field with a relatively high precision parameter, causing the lure word to have lower familiarity than the studied words, but higher familiarity than irrelevant words. (c) Illustrates the conversion from memory activation to the likelihood of remembering. The mean of the Gaussian distribution, represented by the black dotted line, is the memory activation of the target word. The purple line represents the memory threshold. The blue region above the memory threshold represents the probability of recognizing the word.

There are two phases to our DRM task: the learning phase and the testing phase. In the learning phase, participants were presented with words one after another and asked to pay attention to them. We chose the words from the bank of 24 lists of words used by the original DRM task (Roediger & McDermott, 1995). Each list includes 20 words and has a corresponding lure word that is not in the list but is semantically related to words in the list. For each participant, we randomly chose 12 lists from those 24 lists and then randomly chose 10 words from each chosen list to use in the learning phase. The 10 words from a given list were always presented consecutively; the order of the words within each list was randomized. Each word was displayed in the center of the screen for 1 second before the next word automatically appeared. After viewing all 10 words, participants were asked to type in a word that came to their mind while viewing the last set of words. (This question was included to test hypotheses unrelated to the present report and is not discussed further; analyses indicate that the inclusion of this question had no effect on performance in the test phase.) After typing a word, the participant could proceed to view the next 10 words from an-

other list until all 12 lists were presented. The order in which the 10 words were presented and the order of the selected 12 lists was all randomized.

In the testing phase, participants were presented with 72 words, one after another, at the center of the screen and asked to indicate whether each word was an "old" word (that they saw during the learning phase) or a "new" word (that they did not see during the learning phase) within a response window of 4 seconds. They were instructed to press the key '1' on the keyboard for old words and to press the key '2' for new words. To remind participants of these two candidate responses, we displayed the text '1 old' at the lower left side and '2 new' at the lower right side of the presented word. Participants have had up to 4 seconds to respond to each word. After each response, the text corresponding to chosen response turned purple for 0.5 seconds. The next word appeared following an inter-trial interval of 1 second, during which the screen displayed a white fixation cross at the center of the screen. Among the 72 words, 36 words were randomly selected from the studied words that participants actually saw during the learning phase, 12 words were lure words corresponding to the 12 word lists presented to participants during the learning phase, 12 words were unrelated lure words corresponding to the 12 un-shown word lists, and 12 words were randomly selected from within the 12 unshown word lists.

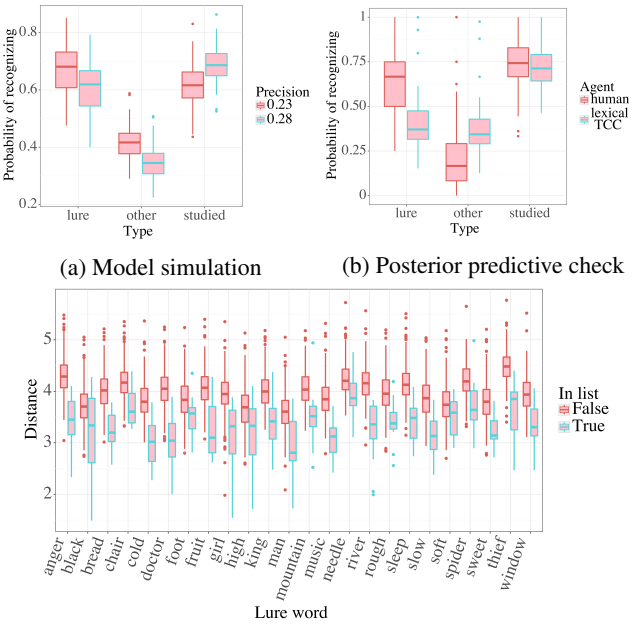
Model Simulation

First, we confirmed that the lure word of a given list was associated with a smaller Euclidean distance to the words within that list than to words from the remaining 23 lists in word embeddings space (figure 2c). Using a Mann-Whitney test, we confirmed that this comparison was significant for all lists ($ps < 0.05$). This ensured that our similarity metric based on Word2vec reproduces the property of semantic relatedness that the original DRM word lists were designed to possess.

We then simulated the model by pre-setting the parameters to see if it produces the false memory effect (figure 2a). We first simulated using $\sigma^{-1} = 0.28$, $\tau = -0.25$, and $\varepsilon = 0.5$, and discovered that the model did produce the false memory effect, where the likelihood of falsely remembering the lure words was higher than the likelihood of falsely remembering other unrelated words. By decreasing the precision parameter to $\sigma^{-1} = 0.23$ while holding the other two parameters the same, we see that the simulated data show an even higher likelihood of falsely remembering the lure words than correctly remembering the studied words. There were no particular reasons behind the selection of these specific parameter values for simulation except they can enable the model to produce the corresponding behavioral patterns. These simulation results aim to confirm that our model framework is flexible enough to predict various degrees of the DRM false memory effect, given the appropriate parameter values.

Fitting and Validation

We used hierarchical Bayesian methods to fit model parameters to participants' responses in our DRM task, which has



(a) Model simulation (b) Posterior predictive check

(c) Semantic distance

Figure 2: (a) Model simulation using pre-set parameter values: each data point represents one participant's average probability of remembering a particular type of trigger words. The red color represents the simulated data with a lower precision parameter, and the blue color represents the simulated data from the model with a higher precision parameter. (b) Model simulation using fitted posteriors of parameters compared to the participants' data: The red color represents human performance on the DRM task, and the blue color represents the simulated data from the model using the parameter estimates fitted to the human data. (c) Euclidean distances between lure words and other words: x-axis represents the lure words of all 24 word lists used in the DRM. Each data point represents the Euclidean distance (according to Word2vec) between the lure word and a word either in the lure word's corresponding list (color-coded as blue) or in a word list that corresponds to another lure word (color-coded as red)

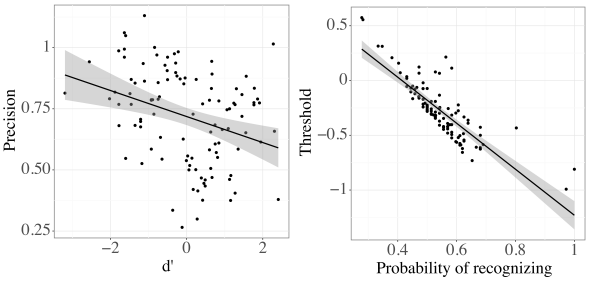
many advantages over the traditional maximum likelihood fitting (Lee, 2011). The population-level priors for all model parameters were chosen as the following:

$$\boldsymbol{\sigma}^{-1} \sim \text{Inverse-Gamma}(\alpha = 3, \beta = 1)$$

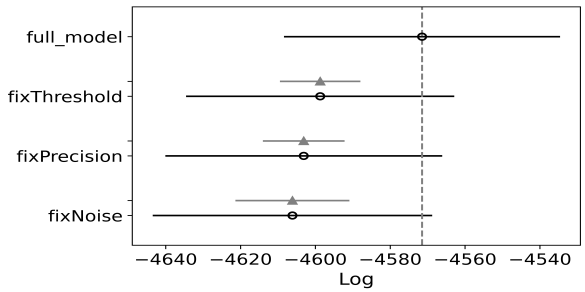
$$\boldsymbol{\tau} \sim \text{Normal}(\alpha = -0.25, \sigma = 0.5)$$

$$\boldsymbol{\epsilon} \sim \text{Inverse-Gamma}(\alpha = 1, \beta = 1)$$

We performed fitting using the python PyMC4 package (Salvatier, Wiecki, & Fonnesbeck, 2016) via the no-U-Turn sampler, which is the state-of-the-art Markov-chain Monte Carlo sampling method to estimate parameter posteriors. For each model, we ran 4 chains of 800 tuning samples (which were discarded) and 1000 kept samples used to estimate the posterior distributions. Therefore in total 4000 samples were



(a) Precision parameter σ^{-1} (b) Threshold parameter τ



(c) Model comparison

Figure 3: (a) The correlation between the fitted precision parameter σ^{-1} and d': Each dot represents a participant, the black line is the regression line, and the shade represents the standard error. (b) The correlation between the fitted threshold parameter τ and participants' overall probability of remembering a trigger word regardless of the word's type: Each dot represents a participant, the black line is the regression line, and the shade represents the standard error. (c) Model comparison plot generated by the *compare* function of the arviz package: The hollow circles represent the average log WAIC of a model. The darker lines represent the standard errors of the log WAIC. The triangles represent the average difference in log WAIC between each model and the best-fitting model (in this case the full model). The lighter lines represent the standard errors of the difference in log WAIC.

used to represent each parameter's posterior distribution. For diagnostic checks, we required $\hat{R} \leq 1.01$, $BFMI \geq 0.2$ for all chains, a sufficiently large effective sample size ($ESS \geq 400$) for all parameters, and that no divergences were observed.

Through model posterior predictive checks, we simulated the model response of the DRM task. We notice that the model successfully reproduces the qualitative pattern of the data, but quantitatively underestimates the false recall of lure words and overestimates the false recall of non-lure words figure 2b). Because the model is capable of simulating a large DRM effect given appropriate parameter values, this lack of quantitative validation is unlikely to be a feature of the model itself; instead, it may reflect possible sub-optimality of the fitting procedure in recovering the best parameters.

However, the fitted model parameters, obtained by computing the mean of the parameter posterior distributions, significantly correlate with the respective key behavioral metrics (figure 3a, 3b). The precision parameter σ^{-1} has a significant negative correlation with d' (spearman $\rho = -0.347$, $p < 0.001$). d' is a widely used measure that quantifies the degree to which the false recognition of the lure words exceeds the true recognition of the studied words (Brady, Robinson, Williams, & Wixted, 2022). Formally, d' is defined as the difference between the z score of the false recognition rate of the lure words and the z score of the true recognition rate of the studied words. The threshold parameter τ is also strongly correlated with the participant's overall rate of recognition, regardless of the type of word (spearman $\rho = -0.850$, $p < 0.001$).

Lastly, we show that all 3 parameters are important for the model performance by comparing the full lexical TCC model with 3 alternative models. One alternative model fixes the precision parameter σ^{-1} to be 1, another alternative model fixes the threshold parameter τ to be -0.5, and the third model fixes the noise parameter ε to be 1. All other model fitting procedures are the same. We compared these 4 models using the Widely Applicable Information Criterion (WAIC; Watanabe, 2013). The full model outcompeted all the alternative models, suggesting that all three parameters are essential for the model (figure 3c).

Discussion

In this paper, we introduced a new model of lexical false memory: the Lexical Target Confusion Competition (Lexical TCC) model. This model conceptualizes lexical memory encoding as a concept generalization process where the generalization of each memory signal decreases exponentially along a semantic similarity scale, which is formalized as the Euclidean distance in word embeddings vector space. Through simulation, we showed that the model is flexible enough to produce various intensities of the DRM false memory effect simply by varying the parameter that controls the degree of generalization around studied words. We also fit this model to an actual data set of human performance on the DRM task and saw that the fitted model parameter estimates correlate significantly with behavioral indicators of false memory.

One theoretic implication of the lexical TCC model is that the DRM false memory effect may be understood as the natural consequence of concept generalization, which is one of the most fundamental and domain-general principles of human cognition, implicated in visual processing, numeric cognition, learning, and memory. The lexical TCC model is also capable of generating predictions about the likelihood of lexical false memory on any set of words, without having to preclassify words into lure words and non-lure words. According to this model, there are no fundamental differences between the lure words and the non-lure words. The lure words were more likely to be falsely recalled because they are semantically closer to the actually remembered words. This property of the model allows us to generate predictions about the intensity of the false memory effect of any words that have a vector representation according to the word2vec model. This can aid

the development of new word stimuli for the DRM task beyond the originally used ones. The lexical TCC model also implies that to explain false memory effects in the DRM task, it is not strictly necessary to posit verbatim and gist memory as two separate memory processes whose outputs are combined during memory retrieval. In theory, both verbatim and gist memory can be unified under a single construct capturing the degree of generalization around a stimulus (in this case, a word).

Several limitations of the lexical TCC model are worth noting. First, the posterior predictive check of the model did not reach an ideal level of quantitative accuracy (figure 2b). This might be an indication that the word2vec embedding is an imperfect metric word similarity. It is worth exploring how much improvement might be gained using different types of word embeddings, such as Glove, BERT, and BEAGLE, or word free-association data (De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019). Second, the model has been tested so far only on predictions about the original, standard version of the DRM task. Various modifications of the task have been devised, such as changing the response options in the memory phase to allow participants to express that the presented word is a new word but with a similar meaning to words the participant saw before (Brainerd & Reyna, 2018). More work is needed to examine to what extent the framework of the lexical TCC model can be extended to model performance on these modified DRM tasks. Third, the Lexical TCC model right now is static, that is, it does not capture the dynamic feature of memory, including memory decay over time, instead treating all words as having been encoded at the same time. Accordingly, the Lexical TCC model is not yet capable of modeling many of the time- or list-length related effects known to characterize performance on the DRM task (Osth & Dennis, 2020). Future iterations of the model could address this by building in a mechanism for memory activation to decrease gradually as a function of time.

Acknowledgments

We thank the anonymous reviewers, the Jenkins Lab at UPenn, and the Thompson-Schill Lab at UPenn for helpful feedback.

References

Beversdorf, D. Q., Anderson, J. M., Manning, S. E., Anderson, S. L., Nordgren, R. E., Felopulos, G. J., ... Bauman, M. L. (1998, November). The effect of semantic and emotional context on written recall for verbal language in high functioning adults with autism spectrum disorder. *Journal of Neurology, Neurosurgery & Psychiatry*, *65*(5), 685–692. doi: 10.1136/jnnp.65.5.685

Beversdorf, D. Q., Smith, B. W., Crucian, G. P., Anderson, J. M., Keillor, J. M., Barrett, A. M., ... Heilman, K. M. (2000, July). Increased discrimination of "false memories" in autism spectrum disorder. *Proceedings of the National Academy of Sciences*, *97*(15), 8734–8737. doi: 10.1073/pnas.97.15.8734

Brady, T. F., Robinson, M. M., Williams, J. R., & Wixted, J. T. (2022, October). Measuring memory is harder than you think: How to avoid problematic measurement practices in memory research. *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-022-02179-w

Brainerd, C. J., Chang, M., & Bialer, D. M. (2020, November). From association to gist. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *46*(11), 2106–2127. doi: 10.1037/xlm0000938

Brainerd, C. J., Gomes, C. F. A., & Moran, R. (2014). The two recollections. *Psychological Review*, *121*(4), 563–599. doi: 10.1037/a0037668

Brainerd, C. J., & Reyna, V. F. (2018, March). Complementarity in False Memory Illusions. *Journal of experimental psychology. General*, *147*(3), 305–327. doi: 10.1037/xge0000381

Cann, D. R., McRae, K., & Katz, A. N. (2011). False recall in the Deese–Roediger–McDermott paradigm: The roles of gist and associative strength. *The Quarterly Journal of Experimental Psychology*, *64*, 1515–1542. doi: 10.1080/17470218.2011.560272

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

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*, 17–22. doi: 10.1037/h0046671

Destefano, I., Brady, T. F., & Vul, E. (2021). Predicting Memory Errors with a Bayesian Model of Concept Generalization. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Diana, R. A., Reder, L. M., Arndt, J., & Park, H. (2006, February). Models of recognition: A review of arguments in favor of a dual-process account. *Psychonomic Bulletin & Review*, *13*(1), 1–21. doi: 10.3758/BF03193807

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

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2012). A synchronization account of false recognition. *Cognitive Psychology*, *65*, 486–518. doi: 10.1016/j.cogpsych.2012.07.002

Jou, J., & Flores, S. (2013, November). How are false memories distinguishable from true memories in the Deese-Roediger-McDermott paradigm? A review of the findings. *Psychological Research*, *77*(6), 671–686. doi: 10.1007/s00426-012-0472-6

Lee, M. D. (2011, February). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*(1), 1–7. doi: 10.1016/j.jmp.2010.08.013

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of auto-

mobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning & Verbal Behavior*, *13*, 585–589. doi: 10.1016/S0022-5371(74)80011-3

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, September). *Efficient Estimation of Word Representations in Vector Space.* arXiv. doi: 10.48550/arXiv.1301.3781

Nosofsky, R. M. (1991, February). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology. Human Perception and Performance*, *17*(1), 3–27. doi: 10.1037//0096-1523.17.1.3

Osth, A. F., & Dennis, S. (2020, July). *Global matching models of recognition memory.* PsyArXiv. doi: 10.31234/osf.io/mja6c

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 803–814. doi: 10.1037/0278-7393.21.4.803

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

Schacter, D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist*, *54*, 182–203. doi: 10.1037/0003-066X.54.3.182

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020, November). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, *4*(11), 1156–1172. doi: 10.1038/s41562-020-00938-0

Shepard, R. N. (1987, September). Toward a Universal Law of Generalization for Psychological Science. *Science*, *237*(4820), 1317–1323. Retrieved from `https://www.science.org/doi/10.1126/science.3629243` doi: 10.1126/science.3629243

Steyvers, M., & Griffiths, T. L. (2008, March). Rational analysis as a link between human memory and information retrieval. In N. Chater & M. Oaksford (Eds.), *The Probabilistic Mind:* (pp. 329–350). Oxford University Press. doi: 10.1093/acprof:oso/9780199216093.003.0015

Tenenbaum, J. B., & Griffiths, T. L. (2001, August). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640. doi: 10.1017/S0140525X01000061

Watanabe, S. (2013). A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, *14*(27), 867–897.

Wilson, S. A., Arora, S., Zhang, Q., & Griffiths, T. (2021). A Rational Account of Anchor Effects in Hindsight Bias. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Zeng, T., Tompary, A., Schapiro, A. C., & Thompson-Schill, S. L. (2021, July). Tracking the relation between gist and item memory over the course of long-term memory consolidation. *eLife*, *10*, e65588. doi: 10.7554/eLife.65588