Cross-situational Word Learning is Better Modeled by Associations than Hypotheses

George Kachergis, Chen Yu, and Richard M. Shiffrin

Abstract—**Research has shown that people can learn many nouns (i.e., word-referent mappings) from a short series of ambiguous situations containing multiple word-referent pairs. Associative models assume that people accomplish such cross-situational learning by approximately tracking which words and referents co-occur. However, some researchers posit that learners hypothesize only a single referent for each word, and retain and test this hypothesis unless it is disconfirmed. To compare these two views, we fit two models to individual learning trajectories in a cross-situational word-learning task, in which each trial presents four objects and four spoken words—16 possible word-object pairings per trial. The model that maintains a single hypothesis for each word does not fit as well as the associative model that roughly learns the co-occurrence structure of the data using competing attentional biases for familiar pairings and uncertain stimuli. We conclude that language acquisition is likely supported by memory, not sparse hypotheses.**

Index Terms—statistical learning; language acquisition models; cross-situational learning

I. Introduction

By adulthood, humans command a vocabulary of roughly 60,000 words [1]. Considering that polyglots acquire multiple such vocabularies, it is easy to take the perspective that human memory is boundless. In principle, learning these lexicons of word-to-world meanings seems a challenging problem, given that natural scenes contain many possible referents, and it is often ambiguous what any given utterance refers to. However, assuming that speakers sometimes refer to visible objects, a listener who can remember some co-occurring words and referents from a scene can gradually learn word meanings after experiencing a variety of situations containing different subsets of words and referents. Relying only on cross-modal memory and natural statistics of the language environment, cross-situational learning may be an important way for people to acquire vocabulary.

Previous studies have demonstrated that both infants and adults can learn word-object pairs cross-situationally [2,3]. In adult studies, participants are instructed to learn which word goes with which object and then study a series of training trials. On each trial, people see an array of unfamiliar objects while hearing pseudowords (words). Although each word refers to a particular onscreen object, the correct referent for each word is not indicated, leaving meanings ambiguous on individual trials. For example, you might see objects $\{o_1, o_2\}$ on the first trial, while hearing words {*manu*, *bosa*}. You cannot know if *manu* refers to o_1, o_2, both, or neither; the same is true of *bosa*. At issue in this study is whether learners store in memory multiple candidate meanings of each word (e.g., *bosa*-$\{o_1, o_2\}$ and *manu*-$\{o_1, o_2\}$)—albeit noisily, or instead store only a single hypothesis per word (e.g., *bosa*-o_1 and *manu*-o_2). Although biases such as mutual exclusivity—a preference for having one-to-one word-object mappings—are observed in two-year-olds [4,5] as well as adults [6], these learning biases can be explained using an associative model with competing biases for prior knowledge and uncertain (e.g., novel) stimuli [7,6]. This model comes from a line of research suggesting that evidence for multiple word meanings is accumulated incrementally across several situations [3,8]. Under this view, learners acquire word-object mappings by roughly tracking the co-occurrence of multiple words and objects. That is, even if the association *manu*-o_1 is stored in memory, *manu* will become slightly associated with o_2 when they appear on the same trial.

In contrast, recent research by Medina and colleagues reports evidence consistent with the idea that word-learners retain only a single hypothesized referent for each word, and retain only this single hypothesis until it is disconfirmed [9]. Medina *et al.* suggests that if a word has a hypothesized referent, learners retain no episodic memory of other referents with that word. Seeing a claim that language acquisition does not utilize memory is quite surprising, for it seems to us that human memory is apt to support learning, and has indeed been linked since Ebbinghaus studied savings, the decreased time it takes to relearn forgotten nonsense words, in the 19^{th} century [10]. More recently, [11] found that both children and adults forget fast-mapped words in a pattern consistent with episodic memory loss. Generally, in our search to understand language learning, we aim to first use known general human abilities—such as memory—to try to explain the results before hypothesizing language-specific mechanisms. Thus, in this paper we will compare an associative memory model that noisily remembers everything it sees to a single-hypothesis model. We fit these models to learning trajectories in a cross-situational word-learning task, in which each trial presents

Manuscript received June 29, 2012. This work was supported in part by the National Science Foundation under Grant OISE-1209770.

G. Kachergis is with the Department of Psychological & Brain Sciences / Cognitive Science Program, Indiana University, Bloomington, IN 47405 USA (phone: 812-855-0626; e-mail: gkacherg@indiana.edu).

C. Yu is with the Cognitive Science Program, Indiana University, Bloomington, IN 47405 USA (e-mail: chenyu@indiana.edu).

R. M. Shiffrin is with the Department of Psychological & Brain Sciences / Cognitive Science Program, Indiana University, Bloomington, IN 47405 USA (e-mail: shiffrin@indiana.edu).

four word-object pairs without indicating the intended mappings. Learners experienced each of the 18 word-object pairs six times in each training block, and were then tested on each word. Learners did four iterations of the same training/testing block to produce acquisition trajectories. We also evaluate whether the models can reproduce the shape of individual learning trajectories, finding that the associative model can, while the single-hypothesis model cannot.

II. Experiment

Participants were asked to simultaneously learn many word-referent pairs from a series of individually ambiguous training trials using the cross-situational word-learning paradigm [3]. Each training trial was comprised of a display of four novel objects with four spoken pseudowords. With no indication of which word refers to which object, learners had a small chance of guessing the four correct word-referent pairings. However, since a word always appeared on trials with their intended referents, the correct pairings may be learned over the series of trials. In this experiment, we measure learning trajectories by having participants complete four identical training and test blocks, containing the same 18 word-object pairs. Each training block consisted of 27 training trials containing four pairs each, allowing each pair to be displayed six times.

Subjects

Participants were 54 undergraduates at Indiana University who received course credit for participating. None had participated in other cross-situational experiments.

Stimuli

Each training trial consisted of an array of four uncommon objects (e.g., abstract sculptures) shown while four pseudowords were heard. The 18 pseudowords generated by computer were phonotactically-probable in English (e.g., “stigson”), and were spoken by a monotone, synthetic female voice. These 18 arbitrary objects and 18 words were randomly paired for each learner.

Each training block consisted of the same 27 trials. Each training trial began with the appearance of four objects, which remained visible for the entire trial. After 2 seconds of initial silence, each word was heard (randomly ordered, duration of one second) followed by two additional seconds of silence, for a total duration of 14 seconds per trial. The training trials were presented in the same order for each block. No pairs appeared in consecutive trials, limiting the use of working memory to infer correct pairings.

After each training block, participants were tested for knowledge of word meanings. A single word was played on each test trial, and all 18 referents were displayed (i.e., 18AFC). Participants were instructed to click on the correct referent for the word. Each of the 18 words was presented once, and the test trials were randomly ordered in each block.

Procedure

Participants were informed that they would see a series of trials with four objects and four alien words. They were also told that their knowledge of which words belong with which objects would be tested at the end. After each training block, their knowledge was assessed using 18-alternative forced choice (18AFC) testing: on each test trial a single word was played, and the participant was instructed to choose the appropriate object from a display of all 18. They were not told that they would be seeing the same pairs—or indeed the same training—multiple times, but completed four train-test blocks with a short instruction break between each.

Results

Figure 1 shows that individual learning performance was quite variable both within and between blocks, with some individuals performing perfectly by the second test, others never acquire more than a handful, and still others steadily accumulate meanings. Nonetheless, on average participants steadily accumulate knowledge, with mean accuracies of .18, .37, .55, and .70 on successive test blocks, and respective standard deviations of .14, .24, .29, and .30. Thus, although participants only know ~3 word-object pairs, on average, after seeing the 18 pairs six times, by 24 exposures they have acquired over 12 pairs. Using models to embody the theories of single-hypothesis testing and associative learning, we ask which of these theories better accounts for the shape and range of learning trajectories observed in the data.

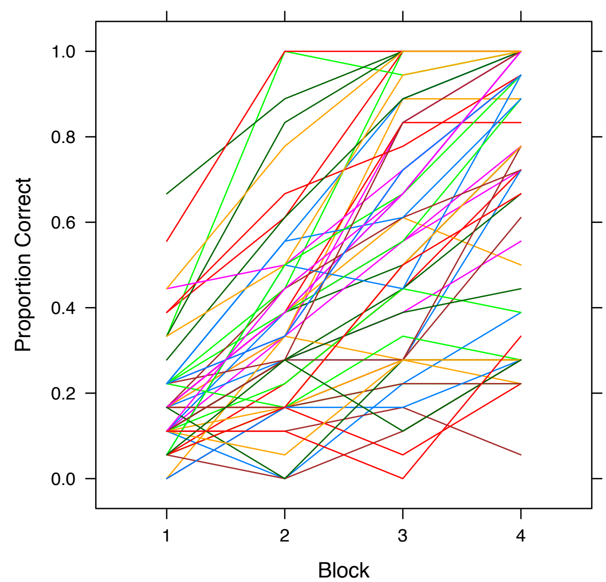

Figure 1. Individual human learning trajectories over the four blocks (chance = .056). Subjects showed great variability in trajectory shape and range. A good model should be able to reproduce the aggregate behavior and the shape of individual trajectories.

III. Models

To determine whether the single-hypothesis view can account for these variable learning trajectories, we constructed a model based on the assumptions stated in [9], specified below. This model does not use episodic memory of prior contexts nor multiple hypotheses per word to learn meanings. For comparison, we use a recent associative model of word learning that has competing biases for familiar pairings and for uncertain stimuli, but that approximately stores all co-occurrences on each trial with noisy retrieval [6]. In the following, we describe the models, fit them to the experiment, and determine whether they can both reproduce the human learning trajectories.

Associative Model

This model assumes that learners do not attend equally to all possible word-object pairings but do roughly store all co-occurrences. Storage is biased, and is guided by several factors: attention is given to pairings on the current trial, and particularly those made likely by previous co-occurrence. However, this familiarity bias competes with selective attention to stimuli that are not already known. This uncertainty bias is based on the learner's current state of knowledge, and is captured by the entropy of a word or object's associations.

Formally, given *n* words and *n* objects to be learned over a series of trials, let M be an n word × n object association matrix that is incrementally built during training. Cell $M_{w,o}$ will be the strength of association between word w and object o. Strengths are subject to general decay or forgetting but are augmented by viewing of particular pairings. Before the first trial, M is empty. On each training trial t, a subset S of m word-object pairings appears. If there are any new words and objects are seen, new rows and columns are first added. The initial values for these new rows and columns are k, a small constant (here, 0.01).

All association strengths decay, and on each new trial a fixed amount of associative weight, χ, is distributed among the associations between words and objects, and added to the (decayed) strengths. The rule used to distribute χ (i.e., attention) balances a preference for attending to unknown stimuli with a preference for strengthening already-strong associations. Consider the first time a word and referent are repeated, extra attention (i.e., χ) might be given to this pair—a bias for prior knowledge. However, as learning proceeds, novel pairings might start to stand out on trials, whereas pairings between novel objects and known words, or vice-versa, are not considered. To capture these ideas, we allocate strength using entropy (H), a measure of uncertainty that is 0 when the outcome of a variable is certain (e.g., $p(w_x|o_y) = 1$, and for all other o_z, $p(w_x|o_z) = 0$), and maximal ($log_2 n$) when every possible outcome is equally likely.

$$H(w) = -\sum_{o \in M} p(o|w) \cdot \log(p(o|w))$$

The update rule for adjusting and allocating strengths for the stimuli presented on a trial is

$$M_{w,o} = \alpha M_{w,o} + \frac{\chi \cdot e^{\lambda \cdot (H(w)+H(o))} \cdot M_{w,o}}{\sum_{w \in S} \sum_{o \in S} e^{\lambda \cdot (H(w)+H(o))} \cdot M_{w,o}}$$

where λ is a scaling parameter governing differential weighting of uncertainty and prior knowledge, α is a decay parameter, and χ is the weight being distributed. For stimuli not presented on a trial, only decay operates. After training, a participant tested with a word and asked to choose its associated referent from m alternatives does so in proportion to the strengths of each available referent to that word (i.e., probabilistic/Luce choice).

Single-hypothesis Model

The basis for this model is the set of assumptions put forth by [9]. In Medina *et al.*'s view, learning a word is accomplished on one trial, without regard to memory of prior experience. Specifically, Medina *et al.* states that:

> "(i) learners hypothesize a single meaning based on their first encounter with a word; (ii) learners neither weight nor even store back-up alternative meanings; and (iii) on later encounters, learners attempt to retrieve this hypothesis from memory and test it against a new context, updating it only if it is disconfirmed. Thus, they do not accrue a "best" final hypothesis by comparing multiple episodic memories of prior contexts or multiple semantic hypotheses." (p. 3)

We reify these verbal hypotheses by building a computational model embodying them. The single-hypothesis model assumes that learners store a list of word-object pairs, with only up to one object stored for a given word. At the beginning of training, this list is empty. On each training trial, for each presented word w the learner retrieves the hypothesized object o_h with probability 1-f. With probability f, o_h fails to be retrieved, and the hypothesis w-o_h is forgotten. If o_h is retrieved, but is not present on the trial, the hypothesis w-o_h is erased. For any words on a trial now without a hypothesis (w_N), new hypothesized objects are chosen[1] from those objects that are not part of a hypothesized pairing. Thus, the model can bootstrap: if three of four objects on a trial are successfully retrieved, the final object will be assigned to the word that has no hypothesized meaning. A new hypothesis for w_N is only successfully stored with probability s. Although probabilistic storage was not mentioned in Medina *et al.*, it seemed a reasonable assumption to allow the model a bit more flexibility. Thus, there are two free parameters in the single-hypothesis model: the retrieval failure probability f and the storage probability s. Testing is straightforward: the model simply chooses the hypothesized object for each word, and chooses randomly from objects that have no name if there is no hypothesis stored for the current object. This model is similar to the minimal "guess-and-test" cross-situational learning model proposed by [12], although for ease of analysis Blythe *et al.* assumes that learners suffer neither failures at storage or retrieval.

Procedure

We fit the two models to the entire data set by minimizing sum of squared error (SSE) across the subjects' trajectories. Each of the 54 subjects contributed four means for a total of 216 data points. We sought a single parameter set for each model to minimize the overall SSE. The single-hypothesis model has two free parameters: retrieval failure probability f and storage probability s. The associative model has three free

[1] Uniformly at random, without replacement, implementing a local mutual exclusivity constraint. However, doing the selection with replacement—allowing 1-many word-object mappings to be formed on a single trial—had no discernible effect on the best achieved model fits.

parameters: learning rate χ, uncertainty vs. familiarity scalar λ, and decay parameter α. To equate the models' number of free parameters, we fixed the value of one parameter to the best-fitting value found in [6]. We fixed α = 0.97 for one experiment, and λ = 5.0 for another. As a learning rate, χ is a function of both the amount of time and the number of possible pairings on a trial, and this experiment differs on both dimensions from that in [6], χ was allowed to vary freely.

Results

The associative model achieved an SSE of 13.33 with α fixed at .97 (χ = 0.05, λ = 1.74), and an SSE of 13.34 with λ fixed at 5.0 (χ = 0.05, α = 0.98). The single-hypothesis model does not fit quite as well, with an SSE of 13.41 (f = 0.01, s = 0.12). As shown in Fig. 2, both models fit the aggregate shape of the learning trajectory well. Given the close fit of both models and the large amount of variability seen in participants, we next examine whether the models are each flexible enough to produce the variety of observed trajectories. Although we do not have enough data per participant to warrant fitting individuals, by simply varying one parameter at a time we can see if each model can produce human-like learning behavior. In doing so, we will also discuss the role of each parameter, and the implications of the best-fitting values.

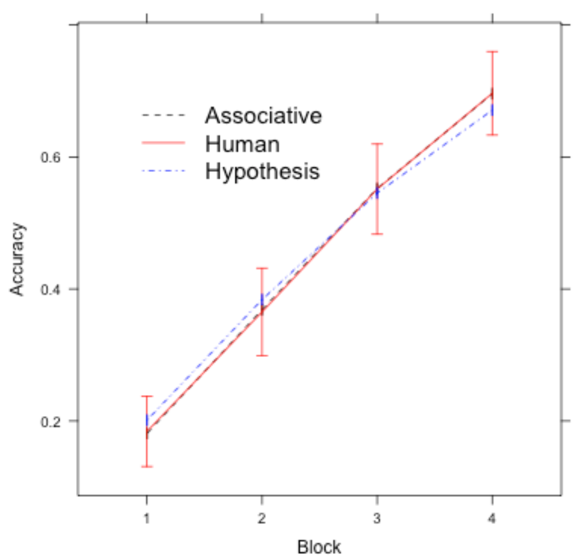

Figure 2. Mean accuracy across train-test blocks for the best fits of the single-hypothesis and associative models, and for human participants. In aggregate, both models fit well, although the associative model fits better.

Single-hypothesis Model

The two parameters are f, the probability of failing to retrieve a given word's hypothesized object during training, and s, the probability of successfully storing an object for a word that had no stored hypothesis. The best-fitting value of f is 0.01, implying that learners rarely forget a stored hypothesis. However, the best-fitting value of s is 0.12, meaning that learners have trouble actually forming a hypothesis. The story told by this model is difficult storage and easy retrieval. As shown in Fig. 3, when the retrieval failure rate is fixed at 0.01, increasing the storage probability results in overall increased learning performance. However, the trajectories either increase slowly or decelerate; none accelerate, unlike some human learners.

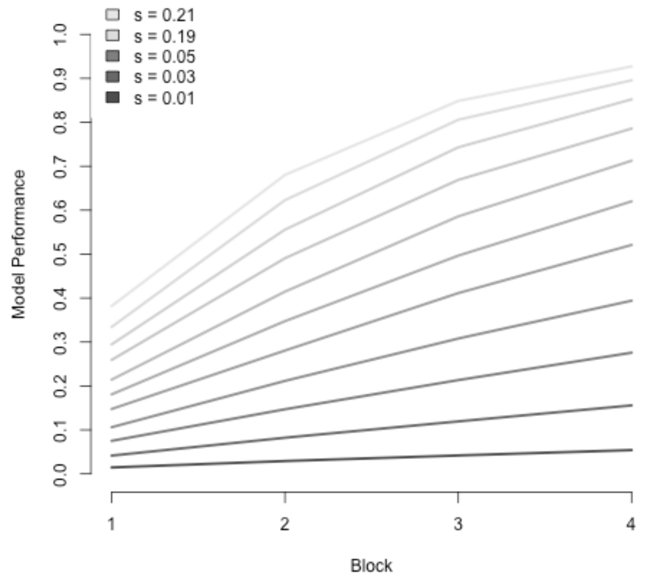

Figure 3. Mean learning trajectories of the single-hypothesis model (1000 simulated subjects) for a fixed retrieval failure probability f = 0.01 and varying storage probability s = 0.01, .03, .05, .., .19, .21. Increasing s results in higher overall learning, but only flat or decelerating curves are observed.

Shown in Fig. 4, for a fixed storage probability of 0.12, increasing the retrieval failure probability above 0.1 quickly reduces learning performance to just above chance across all blocks. Only for small values of f (e.g., 0.01) is the trajectory slope very positive. As discussed before, some human learners in Fig. 1 show accelerating trajectories—not only flat or decelerating trajectories. No parameter values allow the single-hypothesis model to show accelerating learning.

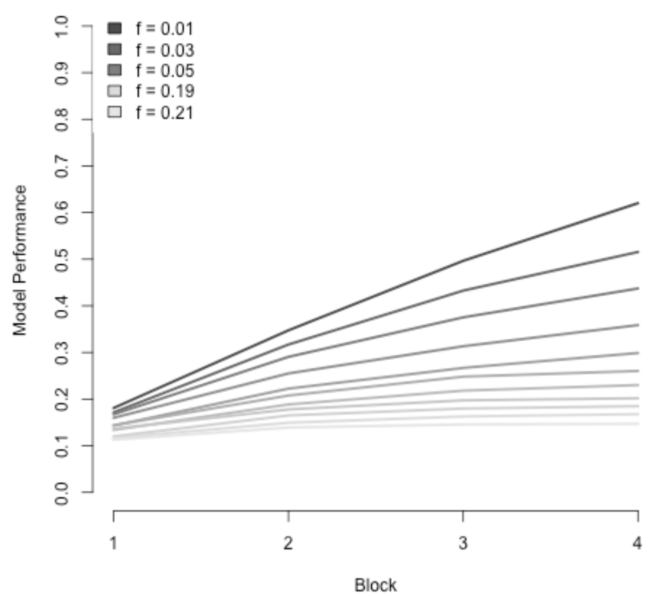

Figure 4. Mean learning trajectories of the single-hypothesis model (1000 simulated subjects) for a fixed storage

probability s = .12 and varying storage probability f = .01, .03, .05, …, .19, .21. Again, no accelerating curves are observed.

Associative Model

The three parameters of the associative model are: χ, the trial-by-trial learning rate (i.e., weight distributed per trial); λ, which scales the relative importance of the uncertainty bias vs. the familiarity bias; and α, which decays (i.e. flattens) the entire memory matrix on each trial. When we fixed α at 0.97, the best-fitting value found in [6], a different λ value (1.74) achieved as good a fit as when λ was fixed at 5.0, the value found in [6]. What is the effect of this uncertainty/familiarity bias in the model? As shown in Fig. 5, with χ = 0.2 and α = .98, varying λ from 1 to 11 has little effect, pushing the positive, decelerating learning trajectory down only slightly. Thus, it is not of concern that somewhat different values of λ were recovered.

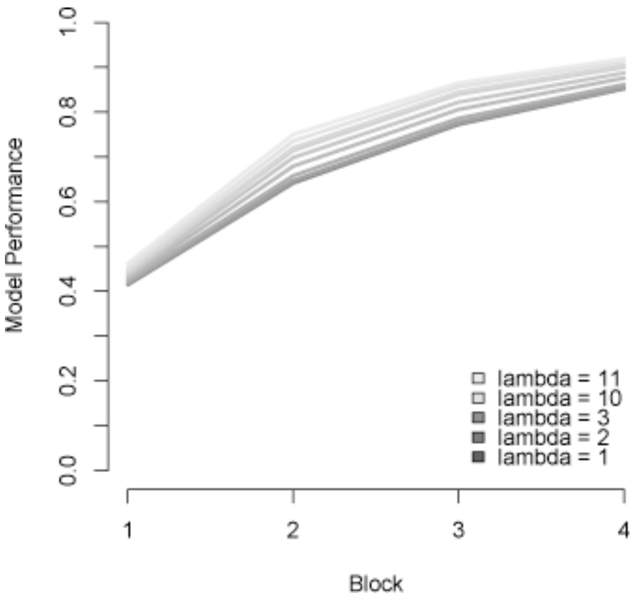

Figure 5. Learning trajectories of the associative model with fixed decay α = .98 and χ = .2 for varying λ = 1..11. At these values, lambda has little overall effect on performance across blocks, only slightly changing the shape of the trajectory.

When we fixed α or λ at values found in [6], the best-fitting value of χ was 0.05 in both cases, and achieved nearly the same fit—both better than the single-hypothesis model. The learning rate should be a function of how much time is on a trial and how many possible word-object pairings there are to consider, but may also vary across individuals. Shown in Fig. 6, with α = 0.98 and λ = 7, varying χ can produce both accelerating positively-sloped trajectories (χ < 0.05) and decelerating learning curves (χ > 0.05), both of which we observe in human data (see Fig. 1).

Finally, Fig. 7 shows how varying α from 0.9 to 1.0 (perfect fidelity—no decay) for fixed χ = 0.01 and λ = 7 can also produce large changes in learning trajectories, from decelerating (α < 0.94) to accelerating (α > 0.95). Thus, two of the associative model's parameters allow it to show both the accelerating and decelerating learning curves we observe.

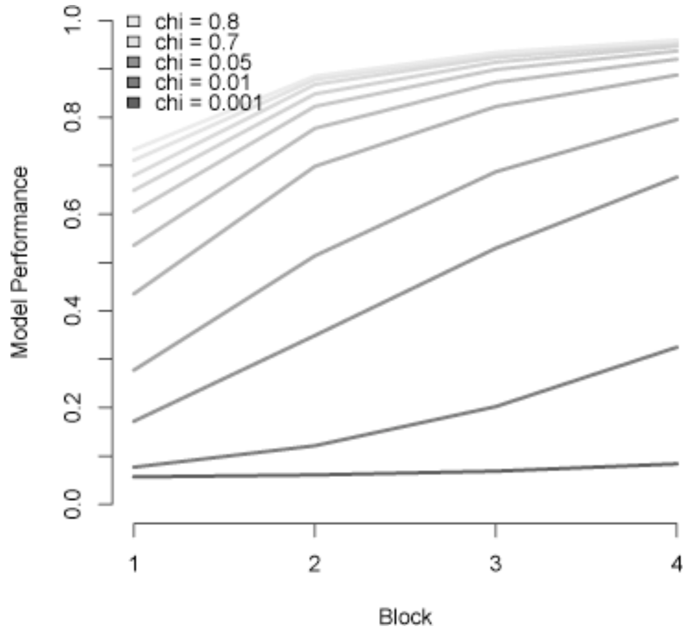

Figure 6. Learning trajectories of the associative model with fixed decay α = .98 and λ = 7 for varying χ = .001, .01, .05, .1, …, .8.

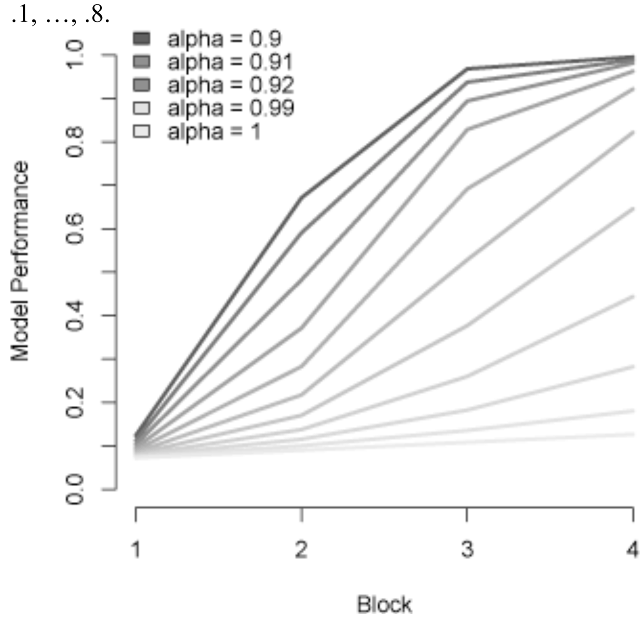

Figure 7. Learning trajectories of the associative model with fixed χ = .01 and fixed λ = 7 for varying decay α = 0.9-1.0.

Given that α is a memory decay parameter that flattens the association matrix trial by trial at values less than 1, it may be surprising to see that values close to 1 yield overall worse performance. This is explained by considering that values of α closer to 0 are weighting recent experience more than earlier experience. The associative model also uses its memory state—both in terms of familiarity (i.e., association strength) and uncertainty across a stimulus' associations—to direct attention to more likely pairings. Since nothing is known in the beginning, and words are associated with all objects they appear with, early experience is less informative than recent experience, when associations are built according to what prior knowledge and uncertainty-based inference dictate are likely. This model is thus a dynamical system with a state space that depends on the entire history of its experience. Lower values of α actually improve learning because older, less informative experience is forgotten, and the more valuable recent experience is amplified. Certain values of α even produce trajectories with two inflection points (e.g., $\alpha = 0.93$-0.94)—another pattern that humans sometimes show. Overall, the associative model was able to generate the variety of trajectories observed in human data, whereas the single-hypothesis model could only generate decelerating curves, and often had a small slope—even when not at ceiling or floor.

IV. Discussion

Some language acquisition researchers have an intuition that it is far too demanding for human learners to store more than one hypothesized object for each word, and to bring these memories to bear in future experience (e.g., [9]). Others find it difficult to imagine why the objects we encounter should not be stored in memory along with words and other stimuli we encounter—whether we are in a language learning class or eating breakfast [6,11,13]. Moreover, these researchers suppose that episodic memory, a basic human cognitive ability, likely supports language learning. We tested these two contrasting views of language learning using two models built from the assumptions of each. We used data from a cross-situational word-learning experiment with four train-test blocks, in which we measured individual learning trajectories for a vocabulary of 18 words at four points. Both the associative model and the single-hypothesis model fit the aggregate data well, but the associative model fits better.

However, human learners showed a wide variety in both trajectory range and shape, with some quickly reaching ceiling while others remain near the floor, and many shapes in-between. The associative model, which stores every word-object pairing it observes—although in a biased way, based on pairing familiarity and stimulus uncertainty, was able to reproduce this variety of learning curves by modifying parameter values. The single-hypothesis model, which allows only one hypothesized object to be stored for each word and has no memory to generate these hypotheses, cannot produce accelerating learning curves, nor trajectories with multiple inflection points. The model simply does not have the flexibility needed to produce the variety of human learning trajectories. Moreover, the single-hypothesis model fits best with both a small probability of forgetting a single hypothesis ($f = 0.01$), and a low probability of storing a hypothesis for a word that lacks one ($s = .12$). We find it surprising that hypotheses are remembered so well, once made, and that hypotheses are so difficult to store. However, when given a storage rate above 0.4, the single-hypothesis model easily reaches ceiling performance after only one or two blocks even with a much higher forgetting rate (e.g., 0.1). Thus, perhaps counter to the reasoning of [9], having no retrieval difficulty at test makes it easy to learn all of the pairs—after all, there are no other associations in memory competing for choice. In contrast, although some call associative models powerful because they store all co-occurrences, they face difficulty at test since all of these associations are competing for choice. Our associative model can learn faster—showing accelerating learning curves—than a simple co-occurrence counting model using its familiarity and uncertainty biases, which have accounted for a variety of human word-learning data [6,7] and explains word-learning biases (e.g., mutual exclusivity [4,5]) without explicitly including any hard constraints.

Thus, the seemingly simple learning assumptions of the single-hypothesis model actually translate to a fairly-powerful learner of 1-to-1 mappings[2] with no noise at test, while the more complex learning machinery of the associative model produces more difficulty at test because of its memory of other associations. We conclude that language acquisition, like many human activities, is probably also subject to memory—which can both help learners direct their attention, but also confuse them during retrieval. While there may be other language-specific mechanisms at work in language learning, the single-hypothesis model that we formulated from assumptions made in [9] does not match the variety of human trajectories we observed, so we prefer the simpler explanation: that memory of multiple word-object co-occurrences—though it may well be weak—supports word-learning.

Acknowledgment

G. Kachergis thanks Shohei Hidaka, Daniel Yurovsky, and Gregory E. Cox for helpful discussions.

References

[1] Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

[2] Smith, L. & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106, 1558-1568.

[3] Yu, C. & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18, 414-420.

[4] Markman, E. M. & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20, 121–157.

[5] Merriman, W.E. & Bowman, L. L. (1989). The mutual exclusivity bias in children's word learning. *Monographs of the Society for Research in Child Development*, 54(3/4), 1–129.

[6] Kachergis, G., Yu, C., & Shiffrin, R.M. (2012). An associative model of adaptive inference for learning word-referent mappings. *Psychonomic Bulletin & Review*, 19(2), 317-324.

[7] Kachergis, G. (2012). Learning Nouns with Domain-General Associative Learning Mechanisms. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 533-538). Austin, TX: Cognitive Science Society.

[8] Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, 107, 729-742.

[9] Medina, T.N., Snedeker, J., Trueswell, J.C., & Gleitman, L.R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108, 9014-9019.

[10] Ebbinghaus, H. (1913) Trans. H. A. Ruger & C. E. Bussenius. *Memory: A contribution to experimental psychology*. New York, NY: Teachers College, Columbia University.

[11] Vlach, H. A., & Sandhofer, C. M. (2012). Fast mapping across time: Memory processes support children's retention of learned words. *Frontiers in Developmental Psychology*, 46(3), 1-8.

[12] Blythe, R. A., Smith, K. and Smith, A. D. M. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science*, 34(4), 620-642.

[13] Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Acquisition*, 4(1), 32-62.

[2] Although, of course, not 1-to-many or many-to-1 mappings, which the associative model can handle (see [6,7]).