Continuous Measure of Word Learning Supports Associative Model

George Kachergis
Department of Psychology
Leiden University
Leiden, the Netherlands
Email: george.kachergis@gmail.com

Chen Yu
Psychological & Brain Sciences
Indiana University
Bloomington, Indiana 47408
Email: chenyu@indiana.edu

***Abstract*—Cross-situational learning, the ability to learn word meanings across multiple scenes consisting of multiple words and referents, is thought to be an important tool for language acquisition. The ability has been studied in infants, children, and adults, and yet there is much debate about the basic storage and retrieval mechanisms that operate during cross-situational word learning. It has been difficult to uncover the learning mechanics in part because the standard experimental paradigm, which presents a few words and objects on each of a series of training trials, measures learning only at the end of training after several occurrences of each word-object pair. Thus, the exact learning moment–and its current and historical context–cannot be investigated directly. This paper offers a version of the cross-situational learning task in which a response is made each time a word is heard, as well as in a final test. We compare this to the typical cross-situational learning task, and examine how well the response distributions match two recent computational models of word learning.**

I. Introduction

All of us have solved a problem as infants that researchers still struggle to explain. That problem is language acquisition, which can better be viewed (if perhaps not solved) as a constellation of problems ranging from segmenting the continuous speech streams we hear into discrete words (e.g., [1]), to learning syntax (e.g., [2]) and to learning the referential intent (i.e., meanings–concrete or abstract) of words (e.g., [3]. This paper focuses on the latter challenge of learning word-object mappings from experiencing a series of ambiguous situations containing multiple words and objects, a process referred to as cross-situational learning [4]. In the standard adult cross-situational learning task [3], a few unusual objects are presented on each trial and are then named in a random order. Thus, from a single trial participants can only guess which word refers to which object. However, since pairs occur on multiple trials spread across training, and appear with different concurrent pairs, people can learn some of the intended word-object pairings. The present study compares a standard passive cross-situational training paradigm–with 4 words and objects per trial–to a response version, in which participants must respond to each word on a training trial by clicking on one of the objects or a "Don't Know" button. Although this response condition may alter task demands, it also offers a glimpse into the ongoing learning during training that can grant greater insight into the mechanisms at play.

Consider the possible progress of learning on a couple training trials with 3 words and objects. Suppose participants hear novel words $bosa, manu, plimbi$ while viewing objects o_1, o_2, o_3 on the first trial. A few trials later, hearing $manu, stigson, bosa$ while seeing o_3, o_2, o_4, what might the participant learn? There are two basic perspectives on how people learn word-object mappings. In the hypothesis-testing view, learners store only a single hypothesized referent for each word–randomly at first, discarding the hypothesis only if it is disconfirmed [5], [6]. This perspective typically views language acquisition as a tremendous inference problem to be solved by applying logical constraints [7]. In this view, a learner may have stored $bosa - o_1$, $bosa - o_2$, or $bosa - o_3$– but not more than one of these. Moreover, if $bosa - o_1$ was stored, then o_1 could not be stored as the referent for any other word on the trial–a strict mutual exclusivity (ME) constraint. Infants and adults are known to show a bias for learning mutually exclusive word-object mappings, although adults will adaptively relax the bias when given evidence of non-ME pairings [8]. On the latter trial, a hypothesis-tester would throw out any hypotheses inconsistent with the current data (e.g., $manu - o_1$ would be thrown out). A hypothesis-tester would consider the second trial to be confirming evidence of $bosa - o_3$ and $manu - o_2$ – or vice-versa, depending on which hypotheses happened to be made on the first trial. Such a learner would *not* know that they should still be uncertain of which mapping is correct.

In the associative learning view, learners approximately store word-object associations between all co-occurring stimuli [8]–[11]. Such associative models assume that although every stimulus makes an impression, these associative memories compete with each other at test, causing retrieval failures. Moreover, some associative models apply attentional biases at learning so that not all co-occurrences are stored with equal strength. For example, Kachergis *et al.* [8] offers a model that has competing biases to attend to familiar word-object associations (i.e., strong from prior exposure), but also devotes storage more to stimuli with uncertain associates (e.g., novel stimuli). On the first example trial, this model would spread attention to all of the word-object associations equally, since all are novel and have no prior association. If prompted with *bosa* after this trial, the model would select any of the three

referents with equal probability–and is also aware of its own uncertainty via the entropy of the word's associations, which is used to drive future attention. On the second example trial, this model's familiarity bias would draw attention to strengthening all associations between *manu*, *bosa* and o_2, o_3–all of which are familiar, but all of which will remain equally probable. However, the greater novelty of *stigson* and o_4 also draw some attention to the conjunction of those: *stigson* $- o_4$–rough form of mutual exclusivity. Little attention is given to associations between the familiar and novel stimuli (e.g., *stigson* $- o_2$).

Note that another class of models (e.g., the Bayesian model of [12]) would not only learn about the stimuli on the current trial, but also leverage information from several trials ago in the light of new evidence. Such batch learning models do not show order effects that are common in word learning and associative learning studies [11], [13].

Since the standard cross-situational learning task only measures knowledge in a final test after training is complete, it is difficult to infer the dynamics of learning. By asking participants to indicate which object they believe each word refers to every time it occurs, we can map out the development of knowledge over time. Of course, it is possible that this constant probing will affect task performance, but it is not a priori clear whether it will benefit or hinder learning. On the one hand, recognition memory research shows that being tested benefits memory more than a second study opportunity [14]. On the other hand, asking learners to make a guess even on the first trial–when they cannot yet be certain of anything–may be tedious, or worse, misleading. Thus, we also compare learning in the continuous responding task to performance in the passive cross-situational learning task.

II. Experiment

In this experiment, we compare the standard cross-situational word learning paradigm, in which participants are *passively* trained by observing object displays co-occurring with words, to a *response* version in which participants are asked to choose one of the objects on display–or a "Don't Know" button–each time a word is heard during training. Although we use the same training statistics, it may be that performance on the two tasks will differ: it seems equally plausible that it is an advantage to be tested often, or that it may be a nuisance that distracts learners from remembering the co-occurrences. However, if performance on the two tasks is equal, it may be that the learning trajectories in the response condition can grant insight into the factors and mechanisms underlying cross-situational learning.

A. Participants

Participants in this experiment were 62 Indiana University undergraduate students who received course credit for their participation. None had participated in other cross-situational experiments.

B. Stimuli and Procedure

Verbal stimuli were 36 computer-generated pseudowords that are phonotactically-probable in English (e.g., "bosa"), and were spoken by a monotone, synthetic female voice. Objects were 36 photos of uncommon, difficult-to-name objects (e.g., unusual tools or objets d'art). These 36 words and objects were randomly assigned to two sets of 18 word-object pairings; one set for each study condition. The entire set of stimuli from which the words and objects were randomly drawn is available online: `http://kachergis.com/downloads/stimuli.zip`

Each training trial consisted of a display of four objects (see Figure 1) shown while four pseudowords were played in succession, and 27 such trials were in each block. Although the words and objects in the two conditions were different, their co-occurrence structure was the same: e.g., w_1 and o_1 appeared at the same trial indices and with the equivalent other stimulus pairs in both training conditions. In total, each of the 18 word-object pairs occurred 6 times during training.

Fig. 1: Example trial display, during which participants would hear four words (e.g., "bosa..regli..manu..stigson").

Training trials began with the appearance of four objects, which remained visible for the entire trial, and words were heard (1 s duration, randomly ordered) after 2 seconds of initial silence. In the passive training condition, words were separated by 2 s of silence, for a total duration of 14 s per trial. In the response condition, after each word was heard, the cursor appeared along with a "Don't Know" button in the center of the screen, and participants were given unlimited time to click on one of the objects or the button. When a selection was made, the next word was presented.

Participants were informed that they would see a series of trials with four objects and four alien words. Furthermore, they were informed that their knowledge of which words belong with which objects would be tested at the end. Participants in the response condition were further instructed that for each word during training, they were to choose with the mouse the best object, or click on the "Don't Know" button. After each training block, learners' 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 from a display of all 18 objects the most appropriate one. A within-subjects design was used in order to see whether participants improved from one condition to the other. Condition order was counterbalanced.

C. Results

1) Passive vs. Response Conditions: Seven of the 62 participants were excluded for failing to perform above chance at the final test of either condition (18AFC chance performance $= .056$). Mean accuracy at the final test in the passive training

condition for the remaining 55 participants was .31 (95% Confidence Interval [.25, .37], which was not significantly different than mean accuracy in the response condition: .35 (95% CI [.30, .41]; $t(54) = 1.03$, $p = .31$). Because these two conditions result in nearly equal performance and have the same statistical structure despite the major difference of responding throughout training, it may be that we can predict performance in one condition from performance in the other. As a first look at this, we examined the correlation between individual subjects' performance in the two conditions, but it was not significantly correlated ($r = .04$, $t(53) = .30$, $p = .77$). However, it turned out that there was a condition order effect: subjects showed worse performance in the first training condition, regardless of which condition it was (response mean: .27 vs. passive mean: .24), than in their latter training condition (response: .40 vs. passive: .43). This general improvement from one condition to the next makes it unsurprising that there is little correlation between subjects' performance in the two conditions. It also suggests that the tasks are similar enough that practice on the earlier helps the latter–whatever the order. There was also no significant correlation between the performance on statistically equivalent test items in the two conditions ($r = -.21$, $t(16) = -.84$, $p = .41$). The maximum accuracy for an item (.51) was in the response condition, and the minimum (.24) was achieved by a unique item in each condition. This lack of consistency between the passive and response conditions could result from different strategies/mechanisms being used in each condition, or simply because the random learning trajectory taken by each learner varies too much.

The remainder of our analyses focus on the response condition data, which allow us to investigate several additional interesting questions, such as: Of the pairs that were known on the final test, how many repetitions were required for learning? Was there evidence that some learned pairs were forgotten at the final test? How many pairs were typically learned on a given training trial–in general, and over time?

2) Training Responses: The median time to make a response after word onset on a training trial was 1869 ms (mean: 2416 ms), similar in duration to the 2000 ms between words on a training trial in the passive condition. 42% of the responses during training were incorrect, 38% were correct, and 20% were "Don't Know" responses. On average, learners' median response times were fastest on correct responses (1745 ms), faster than incorrect responses (2117 ms; paired $t(54) = 5.49$, $p < .001$), which were faster than "Don't Know" responses (2830 ms; paired $t(54) = 2.84$, $p < .01$).

On the first occurrence of each word, participants were more likely to use the "Don't Know" button (proportion on first occurrence: .37 vs. all later occurrences: .17, $t(54) = 5.83$, $p < .001$), showing some awareness that they had no grounds to hypothesize a meaning for that word. The mean proportion of correct and incorrect responses on the first occurrence was .20 and .44, respectively–showing that many participants are willing to guess, even when they cannot yet know the correct meaning. On the second occurrence of each word, we investigated the conditional probability of correct, incorrect, and uncertainty responses as a function of their response on the first occurrence of the word. We had two hypotheses in mind: 1) that they would be more likely to be correct on the second response if they were previously correct, and 2) that even for incorrect or uncertain responses, they may be more likely to select the correct referent–since they have acquired some knowledge. Learners who were correct on the first appearance were correct on the second appearance for 49% of the items, greater than the 19% that were correct on the second after guessing on the first (Welch's $t(78.3) = 5.30$, $p < .001$) or than the 30% that were correct after being incorrect on the first (Welch's $t(85.05 = 3.29$, $p = .001$). This matches the finding in a somewhat differently-structured paradigm in [6], which found that learners were at chance when selecting a referent for a word they had been wrong about on the previous occurrence. Among other differences, that paradigm did not allow learners to choose a "Don't Know" option. Nonetheless, it is good to replicate this result in our response paradigm.

We now examine the training responses for pairs that were known at the final 18AFC test in comparison to those that were not finally known.

3) Learned vs. Unlearned Pairs: In the response condition, what patterns of responses during training separate pairs that were known at the final 18AFC test from those that were not known? We measured a few statistics for each pair, and measured their correlation with accuracy for that pair on the final test across all subjects. The statistics we included for each pair are: the occurrence when its object was first correctly chosen (1-6, 7 if never; First Learned), how many times the correct object was selected for that word (0-6; Correct), the number of times an incorrect object was selected for that word (0-6; Incorrect), and the number of times "Don't Know" was selected for the word (0-6; Don't Know). These item-level statistics concerning responses during training were averaged for each subject and correlated with accuracy on the final 18AFC test for those items in the response condition. Figure 2 shows that mean accuracy on the final 18AFC test increased the more often a word's referent was correctly selected during training (Correct; $r = .66$, $t(310) = 15.37$, $p < .001$), and that all other measures were negatively correlated. Selecting the incorrect object more often resulted in lower accuracy on the final test ($r = -.58$, $t(290) = 12.13$, $p < .001$). Similarly, choosing the "Don't Know" button more often was correlated with lower test accuracy ($r = -.29$, $t(203) = 4.33$, $p < .001$). Correctly selecting the object earlier (First Learned; occurrence 1-6, or 7 if never) resulted in higher test accuracy ($r = -.35$, $t(324) = 6.77$, $p < .001$).

These response statistics are correlated to varying degree, so it is also somewhat instructive to look at the average rate of correct, incorrect, and uncertain responses on each occurrence, split by accuracy on the final test. Shown in Table I, the proportion of uncertain (i.e., "Don't Know") responses for both finally correct and incorrect items are nearly the same for the first two appearances, but then decline faster for the correct items as correct training responses pick up. For ultimately

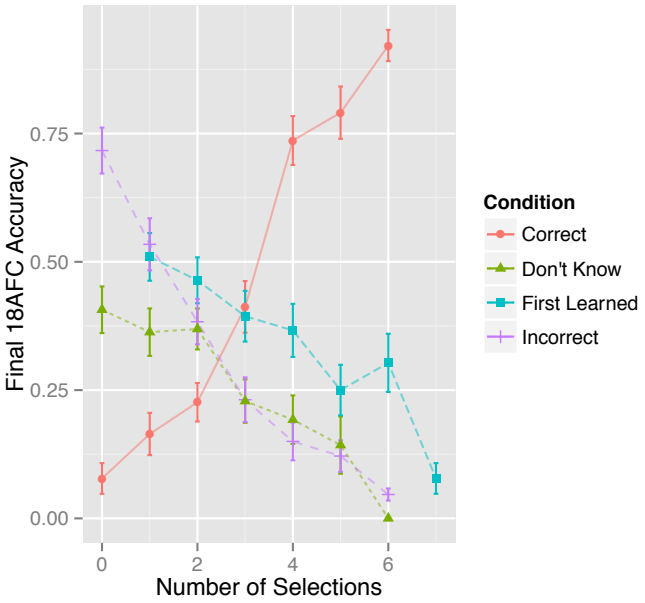

Fig. 2: Accuracy on final 18AFC test as a function of different statistics about that item's responses during training.

incorrect items, the proportion of incorrect responding remains nearly constant, starting and finishing at .5. Correct training responses for ultimately incorrect items do not increase roughly two-fold, whereas for finally correct items they increase nearly three-fold. With statistical analysis alone, it is difficult to translate these results to the underlying mechanisms. In the next section, we fit two models–representing the hypothesis and associative views of word learning–to the distribution of responses and final test accuracy, to see which mechanisms better account for the results.

TABLE I: Training response by appearance and accuracy on final test.

Final Test	Training Resp.	Appearance 1	2	3	4	5	6
Correct	Correct	.28	.44	.57	.67	.75	.86
	Uncertain	.37	.20	.13	.11	.06	.03
	Incorrect	.35	.36	.30	.22	.19	.11
Incorrect	Correct	.15	.22	.25	.26	.27	.35
	Uncertain	.36	.22	.20	.20	.20	.15
	Incorrect	.49	.56	.55	.54	.53	.5

III. Models

We compare two recent models that implement competing intuitions about word learning. As discussed in the introduction, the hypothesis view holds that learners only store a single hypothesized meaning for each word [5], [6]. Hypotheses are chosen from available objects on a trial, but only if that referent is not already linked to another word. In the associative view, multiple possible meanings for a word accumulate in memory on each trial (e.g., [9], [15]): each word is associated with all of the objects, although perhaps not equally [8]. Although opponents of this view find associative systems too powerful, since they are essentially storing a large, weighted co-occurrence matrix. However, at test a word's competing associations with multiple objects serve as noise, making recall probabilistic. After describing the two models, we test how well each model is able to fit the overall proportion of correct responses humans made at each occurrence of a word, as well as the final proportion correct on the 18AFC test.

A. Hypothesis Model

Medina *et al.* [5] laid out the assumptions of the hypothesis model, although a simpler version (the "guess-and-test" model) was analyzed in [16][1]. The assumptions put forth in [5] are:

> "(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)

Following similar assumptions, [6] introduced the propose-but-verify model of cross-situational learning, which begins by guessing and storing a single hypothesized object for each word on a trial. When a word appears again, the previous guess is recalled with some probability α_0. If the recalled hypothesis is present on the trial, α_0 is increased by an amount α_r. If the object fails to be recalled, or is recalled but not present, a new referent is selected–but only from objects that are not currently linked to a word.

The propose-but-verify 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 α_0. If o_h fails to be retrieved, 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[2] 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. 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 word.

B. Associative Model

The biased associative model [8] assumes that learners do not attend equally to all possible word-object pairings. Thus,

[1]For ease of analysis, [16] assumes that learners suffer neither failures at storage or retrieval.

[2]Randomly without replacement–a local mutual exclusivity constraint.

although all co-occurrences are registered to some extent in associative memory (a word $\times$ object association matrix), greater attention and storage is directed to pairings that have previously co-occurred. Moreover, this bias for familiar pairings competes with a bias to attend to stimuli that have no strong associates (e.g., novel stimuli). Familiar associations demand more attention pairings that have not been associated before. However, attention is also pulled individually to novel stimuli because of the high uncertainty (or lack) of their associations, quantified by the entropy of their association strengths, and they thereby attract attention.

Formally, given n words and n objects to be learned over a series of trials, let M be an n word $\times$ 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 forgetting (i.e., general decay) but are augmented by viewing the particular stimuli.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).

Association strengths are allowed to 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 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. When a word and referent are repeated, extra attention (i.e., χ) is given to this pair—a bias for prior knowledge. Pairs of stimuli with no or weak associates also attract attention, whereas pairings between uncertain objects and known words, or vice-versa, do not attract much attention. To capture stimulus uncertainty, strength is allocated using entropy (H), a measure of uncertainty that is 0 when the outcome of a variable is certain (e.g., a word appears with one object, and has never appeared with any other object), and maximal ($log_2 n$) when all of the n possible object (or word) associations are equally likely (e.g., when a stimulus has not been observed before, or if a stimulus were to appear with every other stimulus equally). In the model, on each trial the entropy of each word (and object) is calculated from the normalized row (column) vector of associations for that word (object), $p(M_w, \cdot)$, as follows:

$$H(w) = -\sum_{i=1}^{n} p(M_{w,i}) \cdot \log(p(M_{w,i})) \tag{1}$$

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

$$M_{w,o} = \alpha M_{w,o} + \frac{\chi \cdot e^{\lambda \cdot (H(w)+H(o))} \cdot M_{w,o}}{\sum_{w \in W} \sum_{o \in O} e^{\lambda \cdot (H(w)+H(o))} \cdot M_{w,o}} \tag{2}$$

In Equation 2, α is a parameter governing forgetting, χ is the weight being distributed, and λ is a scaling parameter governing differential weighting of uncertainty ($H(\cdot)$; roughly novelty) and prior knowledge ($M_{w,o}$; familiarity). As λ increases, the weight of uncertainty (i.e., the exponentiated entropy term, which includes both the word and object's association entropies) increases relative to familiarity. The denominator normalizes the numerator so that exactly χ associative weight is distributed among the potential associations on the trial. For stimuli not on a trial, only forgetting operates. After training and prior to test, a small amount of noise (c = .01 here) is added to M. At test learners choose the associated referent for the word from the m alternatives in proportion to their strengths to the word.

IV. Model Results

Using a differential evolution search algorithm, we sought optimal parameter values for both models in order to minimize the sum of squared error between the models' and humans' proportion of correct 4AFC responses across training and the final proportion correct at test. The best parameter values for the hypothesis model were α = .46 and α_{incr} = .03, achieving SSE = .030. The best parameter values for the familiarity- and entropy-biased associative model were χ = .03, λ = 8.48, α = .89 reaching SSE = .016. As seen in Figure 3, although both models match the 4AFC training trajectories pretty well, the hypothesis model outperforms humans on the final 18AFC test because the large number of competing distractors at test have no influence on its performance. In contrast, these many competing memories in the associative model result in lower, more human-like levels of performance with the larger test set. In summary, the associative model achieves a better quantitative and qualitative fit to the data than the propose-but-verify model offered by [6].

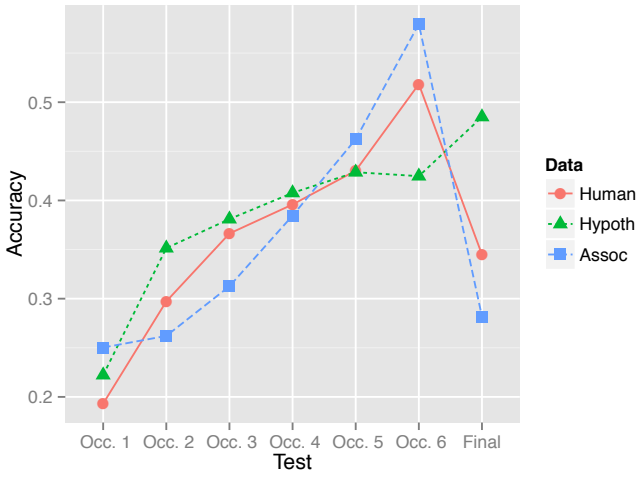

Fig. 3: The mean proportion of correct responses at each word's occurrence and at the final 18AFC test for humans and the two models. While both models fit the 4AFC response trajectory fairly well, the hypothesis model cannot help but do well on the final 18AFC test since only one hypothesis is stored, whereas human participants and the associative model suffer from having a large number of competing distractors.

V. Discussion

In this paper we presented a modification of the cross-situational word learning task that allowed us to measure learning as it proceeds. Overall, participants' showed the same final accuracy in this response task as they displayed in the passive learning task. Even the time spent during training was roughly equal: although learning was self-paced in the response condition, the median time to respond was very close to the 2 s spacing between words in the passive condition. Although there was no item-level correlation between the two conditions, the similar performance on the two conditions–and the observed improvement from one condition to the next, regardless of order–suggest that responding during training may not significantly alter the strategies used for cross-situational learning. Thus, we examined the continuous testing during the response condition in order to gain insight into the learning mechanisms. Various measures of performance during training all predicted final accuracy, further showing that these online responses can show us the moment-to-moment timecourse of learning. Thus, to test what mechanisms could produce these learning trajectories, we applied to them two recent models of word learning, representing the hypothesis-testing and associative accounts.

We showed that the propose-but-verify model [6], which stores only a single hypothesized object for each word, cannot simultaneously match human 4AFC training trajectories and the final 18AFC test performance: although it comes close to the former, it performs too well on the final test since there are no competing associations in memory to interfere with performance. In contrast, the biased associative model [8] accounts for both the human training responses and the significant drop in human performance seen on the final test. This complements earlier evidence that simple hypothesis-testing models are not able to capture human cross-situational learning behavior: a model built from the assumptions of [5] has been shown to be unable to reproduce the shape of some individuals' block-to-block learning trajectories, whereas the familiarity- and uncertainty-biased associative model can [17].

Moreover, it is not unreasonable to assume that learners have access to both stimulus familiarity and novelty in order to guide their attention. Familiarity judgments are a critical function of episodic memory, and memory has been linked to word learning in children [18]. Novelty has been shown to have an effect on activation in some regions in the brain, even when participants were unaware of the novelty [19]. In competition with each other, these biases can produce both inference-like behaviors (e.g., devoting attention to pairing a novel word with a novel object when in the presence of other familiar pairs [8]), as well as capture effects of varying word frequency and contextual diversity [20]. An important part of developing lexical knowledge is to learn the context surrounding words–something that cannot be captured by single, mutually exclusive hypotheses, but that comes naturally to an associative model. This study implies that competition at test, likely from these extra accumulated associations, contributes significant noise at test. We gained this insight by employing both a continuous online measure of learning during training and a harder final test, with seemingly little effect on strategy. We encourage other researchers to combine these different measures to further illuminate the process of word learning.

Acknowledgment

This work was in part funded by National Institute of Health Grants R01HD056029 and R01HD074601.

References

[1] J. Saffran, E. Newport, and R. Aslin, "Word Segmentation: The Role of Distributional Cues," *Journal of Memory and Language*, vol. 35, no. 4, pp. 606–621, 1996.

[2] A. S. Reber, "Implicit learning of artificial grammars," *Verbal Learning and Verbal Behavior*, vol. 5, no. 6, pp. 855–863, 1967.

[3] C. Yu and L. Smith, "Rapid word learning under uncertainty via cross-situational statistics," *Psychological Science*, vol. 18, pp. 414–420, 2007.

[4] L. Gleitman, "The structural sources of word meaning," *Language Acquisition*, vol. 1, pp. 3–55, 1990.

[5] T. Medina, J. Snedeker, J. Trueswell, and L. Gleitman, "How words can and cannot be learned by observation," *PNAS*, pp. 1–6, May 2011.

[6] J. C. Trueswell, T. N. Medina, A. Hafri, and L. R. Gleitman, "Propose but verify: Fast mapping meets cross-situational word learning," *Cognitive Psychology*, vol. 66, pp. 126–156, 2013.

[7] J. M. Siskind, "A computational study of cross-situational techniques for learning word-to-meaning mappings," *Cognition*, vol. 61, pp. 39–91, 1996.

[8] G. Kachergis, C. Yu, and R. M. Shiffrin, "An associative model of adaptive inference for learning word–referent mappings," *Psychonomic Bulletin and Review*, vol. 19, no. 2, pp. 317–324, 2012.

[9] C. Yu, "A statistical associative account of vocabulary growth in early word learning," *Language Learning and Development*, vol. 4, no. 1, pp. 32–62, 2008.

[10] A. Fazly, A. Alishahi, and S. Stevenson, "A Probabilistic Computational Model of Cross-Situational Word Learning," *Cognitive Science*, vol. 34, no. 6, pp. 1017–1063, May 2010.

[11] G. Kachergis, "Learning nouns with domain-general associative learning mechanisms," in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, N. Miyake, D. Peebles, and R. P. Cooper, Eds. Austin, TX: Cognitive Science Society, 2012, pp. 533–538.

[12] M. C. Frank, N. D. Goodman, and J. B. Tenenbaum, "Using Speakers' Referential Intentions to Model Early Cross-Situational Word Learning," *Psychological Science*, vol. 20, no. 5, pp. 578–585, May 2009.

[13] G. Kachergis, C. Yu, and R. Shiffrin, "Temporal contiguity in cross-situational statistical learning," in *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, N. Taatgen and H. van Rijn, Eds. Austin, TX: Cognitive Science Society, 2009, pp. 1704–1709.

[14] M. Carrier and H. Pashler, "The influence of retrieval on retention," *Memory and Cognition*, vol. 20, pp. 633–642, 1992.

[15] L. B. Smith, "How to learn words: An associative crane," in *Breaking the Word Learning Barrier*, R. Golinkoff and K. Hirsh-Pasek, Eds. Oxford: Oxford University Press, 2000, pp. 51–80.

[16] R. A. Blythe, K. Smith, and A. D. M. Smith, "Learning Times for Large Lexicons Through Cross-Situational Learning," *Cognitive Science*, vol. 34, no. 4, pp. 620–642, Jan. 2010.

[17] G. Kachergis, C. Yu, and R. M. Shiffrin, "Cross-situational word learning is better modeled by associations than hypotheses," in *IEEE Conference on Development and Learning-EpiRob (ICDL)*, 2012.

[18] H. A. Vlach and C. M. Sandhofer, "Fast mapping across time: Memory processes support children's retention of learned words," *Frontiers in Developmental Psychology*, vol. 3, no. 46, pp. 1–8, 2012.

[19] G. S. Berns, J. D. Cohen, and M. A. Mintun, "Brain regions responsive to novelty in the absence of awareness," *Science*, vol. 276, pp. 1272–1275, 1997.

[20] G. Kachergis, C. Yu, and R. M. Shiffrin, "Frequency and contextual diversity effects in cross-situational word learning," in *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, N. A. Taatgen and H. van Rijn, Eds. Austin, TX: Cognitive Science Society, 2009, pp. 2220–2225.