Evidence of error-driven cross-situational word learning

Chris Grimmick[1], Todd Gureckis[1], and George Kachergis[2]
chrisgrimmick@gmail.com, todd.gureckis@nyu.edu, gkacherg@stanford.edu
[1]Department of Psychology, New York University, New York, NY, USA
[2]Department of Psychology, Stanford University, Stanford, CA, USA

Abstract

One powerful way children can learn word meanings is via cross-situational learning, the ability to discern consistent word-referent mappings from a series of ambiguous scenes and utterances. Various computational accounts of word learning have been proposed, with mechanisms ranging from storing and testing a single hypothesized referent for each word, to tracking multiple graded associations and selectively strengthening some of them. Nearly all word learning models assume storage of some feasible word-referent mappings from each situation, resulting in a degree of learning proportional to the number of co-occurrences. While these accumulative models would generally predict that incorrect co-occurrences would slow learning, recent empirical work suggests these accounts are incomplete: paradoxically, giving learners incorrect mappings early in training was found to boost performance (Fitneva & Christiansen, 2015). We test this finding's generality in a new experiment with more items, consider system- and item-level explanations, and find that a model with error-driven learning best accounts for this benefit of initially-inaccurate pairings.

Keywords: cross-situational word learning; error-driven associative learning model; word learning;

Introduction

Among the many challenging aspects of learning a language is the problem of determining which words pick out which referents in our environment. When we encounter a new word, there is rarely an explicit explanation of its meaning, and the context it appears in may present any number of possible referents. While any given situation may present a high degree of ambiguity with many possible referents, if learners are able to roughly track words and referents that often co-occur, they may learn word meanings cross-situationally (Gleitman, 1990). Both infants and adults have been found capable of cross-situationally learning names for novel objects in the laboratory (Smith & Yu, 2008; Yu & Smith, 2007; Kachergis & Yu, 2013), and such learning may be one important means of acquiring the meanings of nouns (Smith, 2000).

It is generally assumed that learners accomplish cross-situational learning by tracking the co-occurrence of each uttered word with a subset of the visible referents in a scene. A variety of biases have been proposed that could enable the learner to restrict the number of word-referent mappings they must attend to and remember. For example, the learners have been shown to exhibit a mutual exclusivity bias, preferring to map each word to one referent–and vice-versa (Markman & Wachtel, 1988; Markman, Wasow, & Hansen, 2003; Ichinco, Frank, & Saxe, 2009). Despite a variety of proposed biases and constraints, considerable debate remains about the exact mechanisms underlying this ability.

Models of Cross-situational Word Learning

A variety of computational models have been proposed, ranging from models that store and test a single hypothesized referent per word (Trueswell, Medina, Hafri, & Gleitman, 2013), to Bayesian models, (Frank, Goodman, & Tenenbaum, 2009), to associative learning models (Kachergis, 2012; Fazly, Alishahi, & Stevenson, 2010). Typically, most models can match overall human learning performance in several experiments, and can be hard to distinguish on the basis of goodness of fit. However, detailed modeling of human learning trajectories (Kachergis & Yu, 2017) and performance in systematically varied conditions (e.g., repetitions and context diversity: (Kachergis, Yu, & Shiffrin, 2016); repetitions and number of distractors: (Yurovsky & Frank, 2015) have revealed interacting memory and attentional constraints that help differentiate models.

In many accounts of cross-situational learning, it is assumed that forming an association (or hypothesis) between a word and referent makes future exposures more valuable, as the familiar trace will draw more attention if confirmed. This advantage for prior knowledge (i.e., "rich-get-richer") is present both in hypothesis-testing accounts such as the propose-but-verify model (Trueswell et al., 2013), as well as in associative accounts that allocate more attention to pre-existing associations (Kachergis, 2012).

However, errors also play an important role in a variety of types of learning. For example, in motor control learning is thought to be based on a mismatch between predicted sensory outcomes of an action and the actual sensation (Seidler, Kwak, Fling, & Bernard, 2013). Similarly, models of animal and human conditioning experiments (Kamin, 1968; Rescorla & Wagner, 1972; Kruschke, 2011) adjust associations based on how surprising an outcome is when given particular cues. A classic example of a prediction error-based learning mechanism is the (Rescorla & Wagner, 1972) model in which the amount of learning on a trial is proportional to the amount of prediction error (i.e., surprise at an outcome). When there is a large difference between the actual outcome and predicted outcome, a large change in the predictive value of a stimulus results. Applied to cross-situational learning, surprise will be generated by the failure of a word and referent to appear together when they have been previously associated. This surprise, generated by the difference between the expectation of the word, given that object, and the actual outcome (failure of the word to appear), results in a higher learning rate for a new word to be associated with that object. Despite the widespread evidence of prediction error-based learning in the animal kingdom, empirical investigations of its role in word learning have been limited (though see Ramscar, Dye, and McCauley, 2013). Most cross-situational word learning experiments do not facilitate continuing prediction errors: in most designs, each time a word is heard its intended referent is visible, and thus as learning proceeds, the surprise that is initially generated due to the discrepancy between the words a learner predicts and what they actually hear will only decrease.

Findings from two recent empirical studies investigating erroneous mappings early in learning suggest that greater prediction error may play an important role in cross-situational word learning (Fitneva & Christiansen, 2011, 2015). In Fitneva and Christiansen (2011), eye-tracking during cross-situational learning was used to investigate the performance of learners who by chance initially looked longer or shorter at the correct referent when a word was heard. Participants were trained on 24 word-referent pairs in four blocks, seeing two referents on each trial while hearing two sequential pseudowords. A post-hoc median split based on location of longest fixation when a word was first heard was used to place participants into High and Low Initial Accuracy conditions (HIA and LIA, respectively). Thus, participants in the HIA condition happened *by chance* to look at more of the intended referents upon each word's first occurrence than the LIA participants, who happened to look more at the incorrect referents. The accuracy of each trial in the first block (i.e., initial accuracy) was determined by the fixation time on each referent after each pseudoword was displayed. A subset of 12 of the words was used for 2-alternative forced choice (2AFC) test, in which a participant heard a word and selected the better of two referents. Participants in the LIA condition outperformed the HIA group at test. Additionally, eye tracking data provided implicit evidence for increased learning among LIA participants. In instances where the correct referent was the first location of fixation, the LIA group took longer to look away than the HIA group, and when the location of first fixation was inaccurate the LIA group was quicker to move their gaze. Proportion of time spent fixated on the accurate referent increased in LIA participants, past that of HIA participants.

A follow-up study used a "familiarization" phase before a similar cross situational learning task to induce differences in initial accuracy, and tested three age groups: 4 year-olds, 10 year-olds, and adults (Fitneva & Christiansen, 2015). In the familiarization phase, 10 unambiguous word-object pairs were serially presented to participants. However, four of these pairs would be switched in the subsequent cross-situational training for participants randomly assigned to the HIA condition (60% initial accuracy), while six of the 10 pairs would be switched in the LIA condition (40% initial accuracy). This exposure was meant to seed more (LIA) or fewer (HIA) inaccurate hypotheses/associations before the subsequent cross-situational training, which presented 15 2x2 trials (i.e., two word-referent pairs per trial). Adult participants in the LIA condition again showed higher performance than those in the HIA condition, in line with the prior results. Notably, the initial accuracy of an item within a given condition seemed to have no significant effect on performance. (Fitneva & Christiansen, 2015) interpreted this lack of an item-level effect as evidence of a 'system-level' effect, meaning that "the effect emerges from the cognitive resources recruited by initially inaccurate items affecting initially accurate items as well" (p. 5). Interestingly, four year-olds showed an opposite effect of condition, with HIA participants performing better, and 10 year-olds showed only an effect of item category, performing better on initially accurate items in both conditions.

Fitneva and Christiansen (2015) suggest that the lack of item-level effects of initial inaccuracy in adults (and in 4-year-olds) may be taken as evidence of system-driven learning: rather than individual initially-inaccurate items garnering extra attention (compared to IA items), more cognitive effort is expended overall by adults in the Low IA condition, triggered by the many inaccuracies. The present study again considers system-level vs. item-level effects of IA in adults by conducting an experiment with more to-be-learned items than Fitneva and Christiansen (2015) (18 vs. 10), and with a more sensitive 19-alternative forced choice (19AFC) test. A potential concern about finding item-level effects of IA in Fitneva and Christiansen (2015) is that adults had quite high performance in the task, which tested half of the 10 studied words using a 2AFC test. In addition, the difference between the HIA and LIA conditions was one of only two words (6 out of 10 and 4 out of 10 accurate, respectively).

The superior performance on initially inaccurate items in both experiments may be accounted for with a prediction error mechanism. An additional attentional account may be able to account for the overall difference in performance between the HIA and LIA conditions. Thus, the present design offers a stronger manipulation, more data per participant, and a more sensitive test, while addressing the same underlying issue of the effects of initial accuracy on learning. We then present modeling in an associative learning framework to determine if learning behavior is better accounted for by an attentional (system-level) mechanism, or by a prediction error-based (item-level) mechanism.

Experiment

To investigate the robustness of the effect of low initial accuracy observed in Fitneva and Christiansen (2015) in a setting with more to-be-learned items and a consequently longer training period, we use a similar 2x2 procedure with a "familiarization phase". However, in our design, we used studied 18 stimulus pairs (vs. 10), and a greater degree of difference between between high and low initial accuracy (12 vs. 6 of 18 pairs switched instead of 6 vs. 4 of 10 pairs switched). This presents a stronger manipulation of initial accuracy: 66.6% vs. 33.3% in the current study, compared to 60% vs. 40% in Fitneva and Christiansen (2015). In addition, at test we presented the full array of possible referents for each word (18 studied + 1 unstudied: 19AFC vs. 2AFC), and tested all 18 words (vs. 5 of 10).

Methods

Participants Participants were 45 people recruited online who completed the experiment in their web browser through Amazon Mechanical Turk. All participants completed the entire experiment, and were paid $1.50 for their participation. Participants were randomly assigned to either the High or Low Initial Accuracy condition (23 and 22 participants, respectively).

Stimuli Stimuli consisted of images of uncommon real-world objects and mono- and bisyllabic nonce words. Each participant was given a random selection of 18 images and 18 nonce words from a collection of 72 images and words.

Procedure The experiment consisted of three phases: familiarization, study, and test. The familiarization phase was one block of 18 trials. Participants were told they would be shown examples of the type of objects and words they would be learning. Each trial showed one object-pseudoword pair for 3 s with a 1 s interstimulus interval (ISI), with each of the 18 pairs being shown once, in a randomized order.

For participants in the Low Initial Accuracy (LIA) condition, 12 of the 18 pairs were switched (inaccurate) in the subsequent study phase, yielding 33.3% initial inaccuracy. In the High Initial Accuracy (HIA) condition, 6 of the 18 pairs were switched at study, yielding 66.6% initial accuracy. For each participant, the total number of objects was constant, such that when word-object pairs were switched at study, both the word and object had been seen in the familiarization phase.

On each trial during the study phase, two word-object pairs were shown simultaneously for 3 s, with a 1 s ISI. Objects were shown side by side, with words vertically arrayed in the center, below the object images. The location of objects and words was randomized to ensure participants could not reliably determine the pairing of stimuli by their location with respect to one another. Trial order, along with which word-object pairs were shown on a given trial, was randomized with the constraint that each word-object pair was presented once before being shown again. Thus, there were three (contiguous) blocks of 9 trials, for a total of three presentations per pair.

In the test phase, each trial displayed an array of all 18 studied objects along with one novel object (the same across all test trials) and a single pseudoword from the study. For each pseudoword, participants were instructed to click on the corresponding object. In addition to testing each of the 18 studied pseudowords, a trial with a novel pseudoword was added, to determine if participants were able to fast-map this novel word to the novel object in the array. The order of test trials was randomized for each participant.

A post-test questionnaire asked participants how many words they thought they mapped correctly (0-19), their rating of the engagement and difficulty of the task on scales of 1-7, and whether they used any external memory aids (Yes/No; indicating that they would still be paid, regardless).

Results

Participant's item-level accuracy data for each studied item were subjected to a logistic mixed-effects regression with condition (High Initial Accuracy (HIA) or Low Initial Accuracy (LIA)) as a between-subjects factor and item category (Initially Accurate or Initially Inaccurate) as a within-subject factor. Mixed-effects regression is more appropriate for forced-choice data than ANOVAs, especially for experiment designs with imbalanced cells such as this one (Jaeger, 2008). The analysis was conducted using the `afex` `R` package (Singmann, Bolker, Westfall, & Aust, 2018). This analysis indicated a significant main effect of item category (F(1,43.7)=42.19, $p < .001$), and no significant main effect of condition (F(1,43.7)=0.86, $p = .36$). Learners had higher performance for items that were initially accurate (M=.59, SD=.31) than for items that were initially inaccurate (M=.35, SD=.30). There was a marginal interaction of condition and item category (F(1,43.7) = 3.50, $p = 0.07$). Shown in Figure 1, accuracy on initially inaccurate items was higher in the LIA condition (M=.42, 95% CI=$[.30, .55]$) than in the HIA condition (M=.28, CI=$[.15, .40]$), but lower than initially accurate items, which were similarly high in both conditions (IA: LIA M=.60, CI=$[.46, .73]$; HIA M_{HIA}=.59, CI=$[.47, .72]$). Overall, in both conditions participants learned on average the same proportion of the 18 items (M_{HIA}=.49; M_{LIA}=.48). Finally, for the novel word presented at test, 47% of participants chose the unstudied test object.

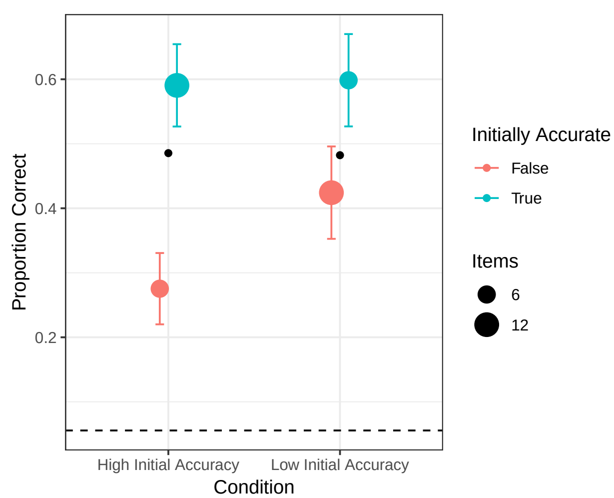

Figure 1: Participants' mean accuracy at test in each condition by item category, with dot size denoting the number of items that could be learned in that category. Black dots show mean performance per condition. Dotted line shows chance (1/18). Error bars represent ±1 SE.

Post-test Questionnaire To investigate metacognitive awareness, individuals' performance was correlated with post-test questionnaire results. Participants in both conditions were aware of their performance level, with significant correlations between their actual and estimated number of learned items they learned (HIA $r(20) = 0.67$, $p < .001$) and LIA ($r(18) = 0.60$, $p < .01$)). Rank-order tests were used to investigate the relationship of engagement and difficulty ratings with performance at test. Strong negative relationships between performance at test and difficulty ratings were found in both HIA ($r_s = -0.73$, $p < .001$) and LIA ($r_s = -0.77$, $p < .001$) conditions. A system-level account of benefits of initial inaccuracy might predict that difficulty would be higher in the LIA condition, but that engagement would be higher. However, there was no difference in participant's perceived level of difficulty in the two conditions ($t(42.7) = 0.56$, $p = .57$), and participants in the HIA condition trended toward being more engaged than LIA participants ($t(38.6) = 1.87$, $p = .07$)–the opposite of what might be predicted by a system-level account.

Discussion

Our results differ somewhat from those of Fitneva and Christiansen (2015) in that we do not find an overall advantage for the Low Initial Accuracy condition. Moreover, we find in both conditions that initially accurate items are learned more often than initially inaccurate items, in agreement with conventional assumptions. However, we do find that initially inaccurate items are learned at a greater rate when they are a greater proportion of the study items (i.e., in the LIA condition). Given that this experiment has a stronger manipulation of initial accuracy (66% vs. 33% instead of 60% vs. 40%), includes more studied items (18 instead of 10), and tests all 18 of them (instead of half) with a more sensitive test (19AFC instead of 2AFC), we contend that it makes a stronger case for the influence of varying initial inaccuracy on cross-situational learning. In the following, we explore what mechanisms account for these effects.

Models

To determine whether the effect of initial accuracy implies novel system-level or item-level learning mechanisms, we first test whether the biased associative model (Kachergis, Yu, & Shiffrin, 2012), with competing attentional biases for existing associations and for attending to stimuli with uncertain associates, is able to account for the effect of varying initial accuracy. This model, explained in detail below, has successfully captured human behavior in a variety of cross-situational learning experiments (Kachergis & Yu, 2017; Kachergis et al., 2016; Kachergis, 2012; Kachergis & Yu, 2013). We also test two modified versions of this model, representing the two theories of why forming initial inaccurate associations may improve overall learning. In the system-level variant, the learning rate on each trial was scaled by the model's relative uncertainty about the words for each presented referent, representing the theory proposed in (Fitneva & Christiansen, 2015) that learners may be more alert in the Low Initial Accuracy condition. In the item-level variant, we add a simple prediction error-based learning mechanism borrowed from Rescorla and Wagner (1972).

Biased Associative Model

The biased associative model (Kachergis et al., 2012) assumes that learners do not attend equally to all possible word-object pairings. Thus, although all co-occurrences are registered to some extent in associative memory (a word × 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). In addition to familiar associations being reinforced, attention is also pulled individually to novel stimuli because of the high uncertainty of their associations (i.e., they have diffuse associations with several stimuli). Uncertainty is tracked by the entropy of a stimulus' association strengths, and attention is allocated to a stimulus in proportion to this entropy.

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 built incrementally during training. Cell $M_{w,o}$ will be the strength of association between word *w* and object *o*. Strengths 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:

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

The update rule for adjusting the association between a given word w and object o on a given trial is:

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

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. As each word w is tested, learners choose referent o from the m alternatives in proportion to associative strength $M_{w,o}$.

Biased Associative Model with Attention

To capture the theory proposed in (Fitneva & Christiansen, 2015) that learners may be more alert in the Low Initial Accuracy condition, we scale the learning rate used on each trial by the mean entropy of the objects on a given trial, relative to the overall entropy of all associations. Thus, trials with more uncertain items–as in the Low IA condition, and particularly for initially inaccurate items–will have a higher learning rate.

Predictive Biased Associative Model

This model differs from the original Biased Associative Model in two ways. First, for the cues (objects) on the trial, let the prediction of each outcome (word w) be

$$V_w = \sum_{o\in O} M_{w,o} \quad (3)$$

V_w was added to the update equation for on-trial word-object associations as a prediction error term

$$M_{w,o} = \alpha M_{w,o} + \chi \cdot e^{\lambda\cdot(H(w)+H(o))} \cdot M_{w,o} \cdot (\beta - V_w) \quad (4)$$

where β is the maximum association value (here, 1), and as before α is a memory fidelity parameter, χ is a learning rate, and λ is relative novelty/familiarity focus. The second difference is the removal of the denominator, which makes it possible for the predictive model to distribute different amounts of associative weight per trial. Thus, the amount of adjustment for a particular association $M_{w,o}$ is scaled not only by the current strength of that association and the uncertainty (entropy) of w and of o, but also proportional to the prediction error of w from the sum of all associations involving w and objects on that trial.

Model Fitting

All models were fit hierarchically: first, differential evolution optimization (Ardia, Mullen, Peterson, & Ulrich, 2015) was used to find best-fitting parameters for each individual, and then optimization was run again with a regularization term to penalize parameter values far from the medians of the group's parameter values.[1]

Model Results

The best-fitting performance achieved by each model, along with Mean Squared Error (MSE) are shown in Figure 2, as well as in Table 1. All variants of the Biased Associative Model (BAM) match performance well in the HIA condition, for both initially accurate and inaccurate items. However, in the LIA condition, both the original BAM and BAM + Attn underestimate human performance on initially inaccurate items and overestimate learning of initially accurate items, while the Predictive BAM fits well.

Condition	High IA		Low IA			
Initially Accurate	False	True	False	True	MSE	r^2
Human	.28	.59	.42	.60	–	–
Biased Assoc.	.26	.60	.30	.66	.026	.939
Biased Assoc. + Attn	.27	.60	.34	.67	.025	.941
Predictive Biased Assoc.	.27	.59	.40	.63	.008	.983

Table 1: Human performance vs. best-fitting models.

Model Discussion

All models match human performance well in the High Initial Accuracy (HIA) condition, and predict slightly higher than observed performance for initially accurate items in the Low IA condition. Both the original Biased Associative Model (Kachergis et al., 2012) and the variant with a learning rate scaled to the uncertainty (i.e., entropy) about items on the current trial are unable to match human performance for initially inaccurate items in the Low IA condition. However, the variant of the Predictive Biased Associative Model does match human performance, suggesting that learners allocate more attention to associations involving words from initially inaccurate items as a result of prediction error.

Discussion

Similar to earlier studies of the effects of initial accuracy on cross-situational word learning (Fitneva & Christiansen, 2011, 2015), our findings show that experiencing a single initial inaccurate mapping of more word-object pairs selectively benefits the later learning of those initially mismatched pairs. However, in contrast to prior research, which found overall higher learning in the Low Initial Accuracy condition, we found the benefit was not conferred on initially accurate items. Rather, performance on initially accurate items in our experiment was similarly high in both conditions–and higher than initially inaccurate items in either condition. As mentioned earlier, the present experiment presents a stronger test of the effects of initial accuracy due to the stronger manipulation, the larger number of studied and tested words, and the

[1]This approximates Gaussian L1-regularization.

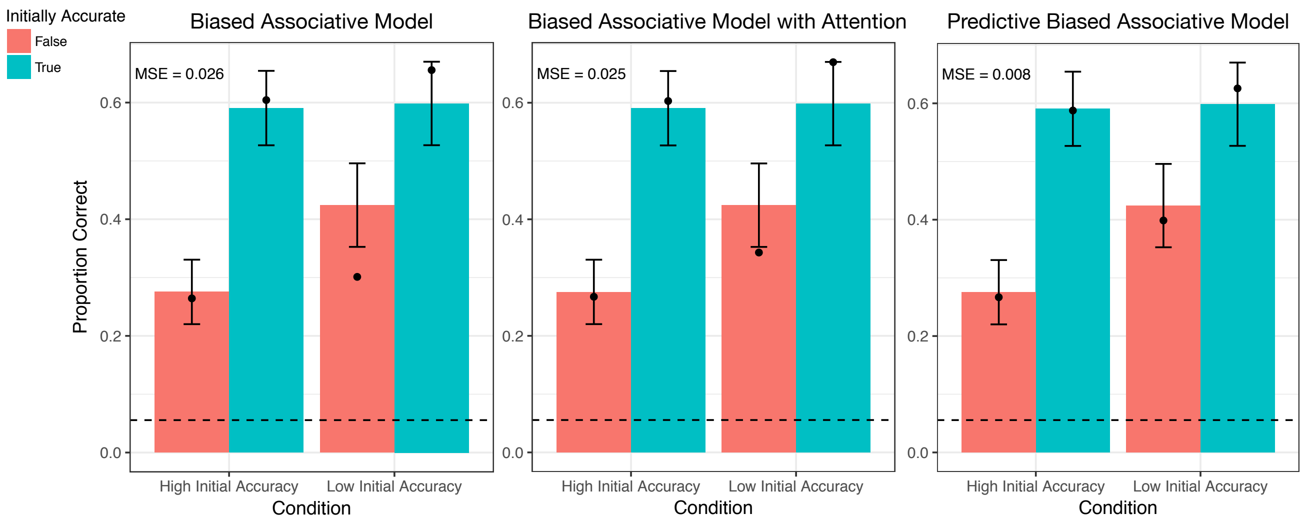

Figure 2: The Biased Associative Model (left) fails to show much benefit for initially inaccurate items in the Low IA condition, unlike people. This is largely true for the variant of the model that gives greater attention to trials with high-entropy stimuli (middle), while the associative model with prediction-error based learning (right) fits quite well (MSE=.008).

lower chance rate of the test format. However, we should note that our stronger manipulation resulted in a slightly higher proportional imbalance of item types per condition (67% vs. 33%)–although we did have more items of both types per condition. It's possible that this greater number of items influenced participants' awareness and thus their treatment of initially inaccurate items. However, we note that difficulty was similarly rated similarly in both groups, and engagement trended higher in the HIA group–opposite to what might be expected if more attention was drawn by the LIA condition.

Our item-level results show that initial accuracy predicts a greater chance of remembering the pairing of that item, which is in accordance with conventional assumptions. However, in the LIA condition initially inaccurate pairs are ultimately more likely to be learned. One explanation may be that errors draw attention selectively to initially inaccurate items. Analysis of fixation times in the eye-tracking experiment of Fitneva and Christiansen (2011) indeed suggests that attention to targets overall increases with greater error. Additionally, the pattern of our results suggests an attention effect: the HIA condition may have included too low a proportion of inaccurate items to draw attention attention away from the majority accurate items. In the LIA group, if there was an overall increase in attention, we should expect to see an increase in performance for initially-accurate (IA) items as well, especially as there were only six IA items in this condition. Our modeling results are consistent with this idea, as without a learning rate proportional to item-level prediction error the fit is notably poor for initially inaccurate items in the LIA condition.

Together with Fitneva and Christiansen (2011, 2015), our results suggest that cross-situational word learning is subject to prediction error-based learning. Our account suggests that when learners see referents they may predict which words will be heard. Subsequently, they allocate attention based on competing biases toward known associations and referents with uncertain associations (Kachergis et al., 2012), and learn at a rate proportional to their surprisal at hearing each word with the given referents. Further research is needed to determine whether this item-level prediction error-based learning mechanism accounts for human behavior—both in typical research settings which offer few inaccurate mappings, as well as in more naturalistic scenarios—to help us further understand the domain-generality of error-based learning.

References

Ardia, D., Mullen, K. M., Peterson, B. G., & Ulrich, J. (2015). *'DEoptim': Differential Evolution in 'R'.* Retrieved from `http://CRAN.R-project.org/package=DEoptim`

Fazly, A., Alishahi, A., & Stevenson, S. (2010, May). A Probabilistic Computational Model of Cross-Situational Word Learning. *Cognitive Science*, *34*(6), 1017–1063.

Fitneva, S. A., & Christiansen, M. H. (2011). Looking in the Wrong Direction Correlates With More Accurate Word Learning. *Cognitive Science*, *35*(2), 367–380.

Fitneva, S. A., & Christiansen, M. H. (2015). Developmental changes in cross-situational word learning: The inverse effect of initial accuracy. *Cognitive Science*, 367–380.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009, May). Using Speakers' Referential Intentions to Model Early Cross-Situational Word Learning. *Psych. Science*, *20*(5), 578–585.

Gleitman, L. (1990). The structural sources of word meaning. *Language Acquisition*, *1*, 3–55.

Ichinco, D., Frank, M., & Saxe, R. (2009). Cross-situational word learning respects mutual exclusivity. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 1–6).

Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Kachergis, G. (2012). Learning nouns with domain-general associative learning mechanisms. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proc. of CogSci 34* (p. 533-538).

Kachergis, G., & Yu, C. (2013). More naturalistic cross-situational word learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proc. of CogSci 35* (pp. 527–532).

Kachergis, G., & Yu, C. (2017). Observing and modeling developing knowledge and uncertainty during cross-situational word learning. *IEEE Transactions on Cognitive and Developmental Systems*. doi: 10.1109/TCDS.2017.2735540

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

Kachergis, G., Yu, C., & Shiffrin, R. M. (2016). A bootstrapping model of frequency and contextual diversity effects in word learning. *Cognitive Science*. doi: 10.1111/cogs.12353

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior, 1967: Aversive stimulation* (pp. 9–31). Coral Gables, FL: University of Miami Press.

Kruschke, J. K. (2011). Models of attentional learning. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization.* Cambridge University Press.

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

Markman, E. M., Wasow, J. L., & Hansen, M. B. (2003, November). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, *47*(3), 241–275.

Ramscar, M., Dye, M., & McCauley, S. (2013). Error and expectation in language learning: The curious absence of 'mouses' in adult speech. *Language*, *89*(4), 760-793.

Rescorla, R. A., & Wagner, A. R. (1972). *A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement.* New York, NY: Appleton Century Crofts.

Seidler, R. D., Kwak, Y., Fling, B. W., & Bernard, J. A. (2013). Neurocognitive mechanisms of error-based motor learning. *Advances in Experimental Medicine and Biology*(782).

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). afex: Analysis of factorial experiments [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=afex` (R package version 0.22-1)

Smith, L. B. (2000). How to learn words: An associative crane. In R. Golinkoff & K. Hirsh-Pasek (Eds.), *Breaking the word learning barrier* (pp. 51–80). Oxford: Oxford University Press.

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

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psych. Science*, *18*, 414–420.

Yurovsky, D., & Frank, M. C. (2015). An integrative account of constraints on cross-situational word learning. *Cognition*, *145*, 53–62.