Actively Learning Nouns Across Ambiguous Situations

George Kachergis, Chen Yu, and Richard M. Shiffrin
{gkacherg, chenyu, shiffrin}@indiana.edu
Department of Psychological & Brain Science / Cognitive Science Program
Bloomington, IN 47405 USA

Abstract

Previous research shows that people can use the co-occurrence of words and objects in ambiguous situations (i.e., containing multiple words and objects) to learn word meanings during a brief passive training period (Yu & Smith, 2007). However, learners in the world are not completely passive, but can affect how their environment is structured by moving their heads, eyes, and even objects. These actions can indicate attention to a language teacher, who may then be more likely to name the attended objects. Using a novel active learning paradigm in which learners choose which four objects they would like to see named on each successive trial, this study asks whether active learning is superior to passive learning in a cross-situational word learning context. Finding that learners perform better in active learning, we investigate the strategies that were most successful, discuss the implications, and model the results.

Keywords: active learning; statistical learning; cross-situational learning; temporal contiguity; language acquisition

Introduction

Human infants learn words quite quickly despite many challenges facing them, including uncertainty and ambiguity in the language environment. Recent research has studied how learners may acquire word meanings from regularities in the co-occurrence of words and referents (e.g., objects). Such cross-situational statistical word learning relies on two assumptions: 1) that spoken words are often relevant to the visible environment, and 2) that learners can to some extent remember the co-occurrence of multiple words and objects in a scene. Thus, as words and their intended referents are observed in different situations over time, learners can apprehend the correct word-object mappings. Relying only on the regularity of the linguistic environment and basic memory and attention processes, this may be an important method of learning nouns for infants, and even adult travelers.

In adult cross-situational learning studies (e.g., Yu & Smith 2007), participants are asked to learn the meaning of alien words by watching a series of training trials. On each trial learners see an array of unfamiliar objects (e.g., four sculptures) and hear pseudowords (e.g., *stigson, bosa*). The meaning of each pseudoword is ambiguous on a given trial, because although each word refers to a single onscreen object, the intended referent is not indicated. In a typical learning scenario, participants attempt to learn 18 word-object pairings from 27 trials, with four words and four objects given per trial. In this design, each word-referent pair is presented six times over the five-minute training period. Learning a correct word-object pairing requires some form of accumulation of word-object co-occurrences. When tested on each word and given four trained objects to choose from, participants chose the correct object for half of the 18 words, on average (Yu & Smith, 2007).

However, in the real world even infant learners are not passive observers, merely watching the world go by. As learners shift their attention, their eyes, head and hands move, changing the objects in their view. If caregivers notice these attention shifts, they may be more likely to name objects that are currently being attended to. Thus, learners may in essence be able to increase the likelihood an object is named by shifting their attention to include this object. This is a form of active learning, a concept studied extensively in machine learning (cf. Settles, 2009), in which a learner can query an information source for the labels of particular data points.

In this study, we introduce active cross-situational word learning, in which learners choose which four objects they would like to see named on each successive trial. Thus, learners control when to repeat pairs, when to stop experiencing pairs they feel they know, and when to attempt to learn more pairs. This gives us a glimpse of their preferred strategies. For example, participants may choose to repeat a single pair from the previous trial, and leverage working memory to quickly learn that the repeated word refers to the repeated object, while ignoring the other three word-object pairs on the trial. Equivalently, a learner may prefer to repeat three pairs from the previous trial, and quickly learn the novel pairing that was not present. Kachergis, Yu, & Shiffrin (2009a) manipulated this sort of temporal contiguity in a passive cross-situational study and found not only that repeated pairs are learned more easily, but so are unrepeated pairs in conditions with some repeats. This suggests that simple inference supported by working memory is not the only learning mechanism at work.

In fact, investigating active learning can reveal what information and mechanisms a learner has at their disposal, and characterizing the observed strategies—and their performance—will motivate learning models. For example, our recent associative model of cross-situational learning assumes that learners have access to both their familiarity and their uncertainty about the word-object pairings present on a given trial, and that attention competes for uncertain stimuli and for already-strong pairings (Kachergis, Yu, & Shiffrin, 2012). This model matches adult behavior in passive cross-situational experiments investigating mutual exclusivity, a bias to find 1-to-1 word-object mappings that is present even in 2.5-year-olds (Markman & Wachtel, 1988). If active learners have access to their knowledge of pairing strength and stimulus uncertainty, these cues can be combined to produce a few active learning strategies. One strategy is to choose one object you have never seen before (i.e., one with maximal uncertainty), and fill the remaining three slots on the trial with familiar objects. Alternatively, learners may choose novel combinations of familiar objects in order to disambiguate mappings; we have previously found such contextual diversity to aid learning (Kachergis, Yu, & Shiffrin, 2009b). Detailed analysis of active learning strategies can reveal what knowledge is available to learners and how they attempt to employ it to learn the correct mappings. It may even be that people are worse at actively structuring the learning environment than the randomly-constructed passive trial sequences they normally experience in word-learning experiments.

In the Experiment, participants do two blocks of passive cross-situational learning, as well as of two blocks of active cross-situational learning in which they choose the objects that they see named on each successive trial. Although there are many other possible formulations of active cross-situational learning, we choose this instantiation because it most closely matches the passive task, and it somewhat matches the real world, where learners can attend to objects and likely increase the chance of a teacher labeling those objects.

Experiment

Participants were asked to learn 18 word-referent pairs from a series of individually ambiguous training trials using the cross-situational word learning paradigm (Yu & Smith, 2007). Each training trial was comprised of a display of four novel objects and four spoken pseudowords. With no indication of which word refers to which object, learners have little chance of guessing the four correct word-referent mappings from the 16 possible pairings. However, since words always appear on trials with their proper referents, the correct pairings may be learned over the series of trials.

The key manipulation of this study is to allow learners in active conditions to choose which four objects they want to see named on the next trial. In both conditions, 18 word-referent pairs were experienced over a series of 27 training trials. Importantly, the same pair was never allowed to appear in neighboring trials in passive conditions. In both conditions, each pair could only appear six times during the training session. Thus, both the number of exposures per pair and the ambiguity on each trial (i.e., number of pairs) were matched in active and passive learning conditions. In order to compare passive and active learning performance, each participant underwent two training and test blocks of each.

Subjects

Participants were 41 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., sculptures) and four spoken pseudowords. The 72 pseudowords generated by computer are phonotactically-probable in English (e.g., “bosa”), and were spoken by a monotone, synthetic female voice. These 72 objects and 72 words were randomly assigned to four sets of 18 word-object pairings, one set for each training condition.

Training for each condition consisted of 27 trials. Each training trial began with the appearance of four objects, which remained visible for the entire trial. After 2s of initial silence, the four words were heard in a random order (1s per word, with 2s of silence after each) for a total duration of 14s per trial.

Procedure

Participants were told they would see a series of trials with four objects and four alien words, but that the order of presentation of the words was random. They were also told that their knowledge of which words belong with which objects would be tested at the end. In the active learning conditions, participants were instructed that they would be able to choose four objects they wanted to see named next. In active learning training blocks, after each trial a display of all 18 objects in the to-be-learned set was shown, and participants chose four to be named on the next trial. Objects that had already been chosen six times were not

After each training block, participants’ knowledge of word-object mappings 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 trained objects. Each of the 18 words was tested once in a random order.

Every participant did four blocks of training and testing: half did two active learning blocks followed by two passive learning blocks, and the other half did the reverse.

Results & Discussion

A repeated measures ANOVA on accuracy[1] by training type (active or passive) and training type repetition (1st or 2nd), nested by condition order (active-first or passive-first) revealed a significant main effect of training type (F(1,39) = 15.17; p < .001). Test performance after active learning is far better than after passive learning (active M = .59; passive M = .35), confirming that adults can use knowledge of their internal state to structure their environment for better learning. Moreover, participants did not improve much on their second block of either training type: there was no significant effect of repetition (F(1,39) = 2.08; p = .15). There was no significant interaction of condition order and repetition (F(2,38) = 1.62; p = .20), nor of training type and repetition (F<1), but there was a significant interaction of training type and condition order (F(2,38) = 4.53; p < .05). As shown in Figure 3, doing active learning first improves performance in the passive conditions (passive M = .30 if passive-first, M = .39 if active-first). This is somewhat surprising, as it is easy to imagine that doing passive first

[1] Data from one subject were excluded after it was found that their average performance in all four blocks was below chance (chance in an 18AFC test is .056).

might give learners an idea for better active learning strategies, whereas it is difficult to see how practice at active learning can improve one's performance in conditions with no command. However, it may be that active learning also allows learners to practice different rehearsal strategies, and helps them choose better ones even when they cannot control the structure of the trials. In any case, individual performance after the different types of training was significantly correlated (Pearson's *r*=.62, *t*(38)=4.81, *p*<.001). Figure 2 shows that almost every participant performed at least as well after active training as passive training.

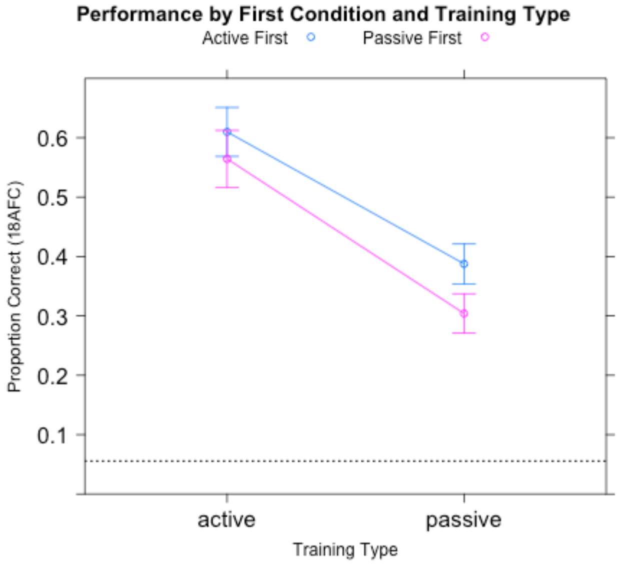

Figure 1: Accuracy by type of first condition and training type in the Experiment. Active learning resulted in far higher test performance than passive learning. Moreover, learners who did active learning first performed better in the passive learning conditions. Error bars show +/-SE.

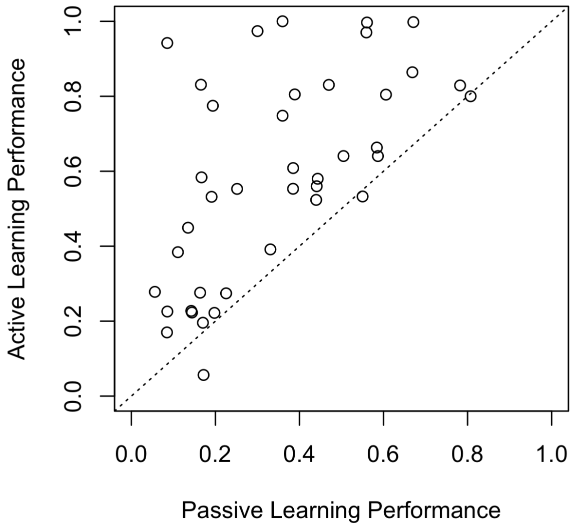

Figure 2: Comparison of performance after passive vs. active learning for each participant. Performance after the two types of training is correlated (*r*=.62), but learners are almost universally better after active training.

Given that adults can actively structure their environment in order to effectively learn the word meanings, we next investigate the strategies effective learners use to disambiguate mappings. However, we must first consider that there are many strategies, and that not all of them result in swift learning. Performance in cross-situational word-learning is typically highly variable, both within- and between-subjects. This is likely because what is learned on a given trial depends on what has been learned on all previous trials, and both the ambiguity on each trial and the fallibility of human memory means that people often learn different things. Giving learners an opportunity to structure training may yield a more diverse set of learning states, and thus may increase variability in performance. Figure 3 shows a histogram of learning performance after each block of active and passive learning. While accuracy after passive training is unimodal and positively skewed, accuracy after passive learning looks roughly bimodal, with peaks at .25 and at .95, which may reflect strategies of differing utility. In the following analysis, we will examine strategy differences both by doing a median split on the performance of active learners and analyzing the strategies used by each group, and by clustering the active training trials and looking at the mean performance of each cluster.

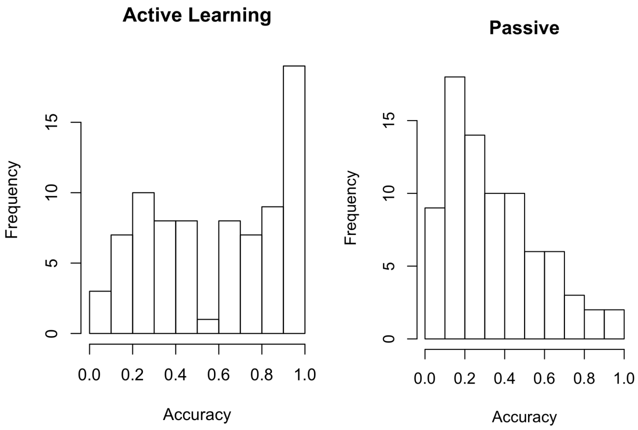

Figure 3: Histograms of performance after active learning (left) and passive learning (right) training blocks (2/subject/condition). Accuracy after active learning is bimodal, indicating that some strategies are quite successful while others are mediocre.

As mentioned earlier, one obvious active learning strategy is to choose to repeat some pairs from this trial on the next trial. If constructed randomly, a given trial would contain only .22 pairs repeated from the previous trial. In our passive training conditions, no pairs were allowed to repeat. The overall mean number of repeated pairs selected by active learners was 1.5—learners are using repetition to disambiguate pairs. To distinguish individual strategies (e.g., repeating one vs. repeating three), we clustered the trial-by-trial number of repetitions chosen in each active learning training block. Using partitioning around medoids we found two clusters, estimated by the optimum average silhouette width (Kaufman & Rousseeuw, 1990). Cluster 1 contained 33 of the active training structures, and Cluster 2 contained the other 47. Figure 4 shows the trial-by-trial average number of repeated pairs for each cluster. Although people in both clusters initially repeat around one pair per trial, learners in Cluster 1 soon began to repeat two or more pairs on average, while those in Cluster 2 stayed closer to one repeat, until the last few trials[2]. Overall, Cluster 1 repeated 1.9 pairs per trial, significantly more than Cluster 2's mean of 1.1 repetitions (Welch $t(60.7) = 9.09$, $p<.001$). From Figure 5, which shows how many pairs were repeated trial-by-trial in each cluster, it is clear that learners in Cluster 2 often chose to repeat single pairs until the very end. Cluster 1 shows a much more varied approach, repeating anywhere from one to three pairs. It turns out that these strategy clusters—constructed solely from the active training data—result in different overall levels of performance: Cluster 1's mean of .71 is significantly higher than Cluster 2's mean of .50 (Welch $t(69.8) = 3.00$, $p<.01$). Repeating more than one pair seems to be a good strategy—indeed, the mean number of pairs repeated per trial in active training is correlated with learning (Pearson's r=.30, $t(78)$=2.82, $p<.01$). Corroborating this clustering result, a median (*Mdn*=.61) split on active learning performance identifies a similar grouping: Cluster 1 contained 22 of the 33 better blocks, whereas Cluster 2 contained 30 of the 47 worse blocks (χ^2=6.05, p=.01). A graph of the active learning blocks identified by median split looks much like Figure 5, showing that better learners repeat more pairs.

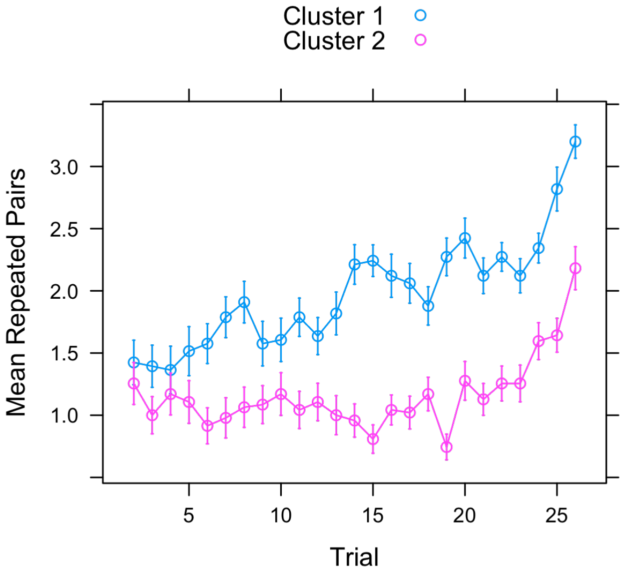

Figure 4: The mean number of word-object pairs repeated on consecutive trials by the two clusters of active learners. Learners in Cluster 1 repeated more pairs per trial than Cluster 2, except at the beginning, when both repeated ~1. Error bars show +/-SE.

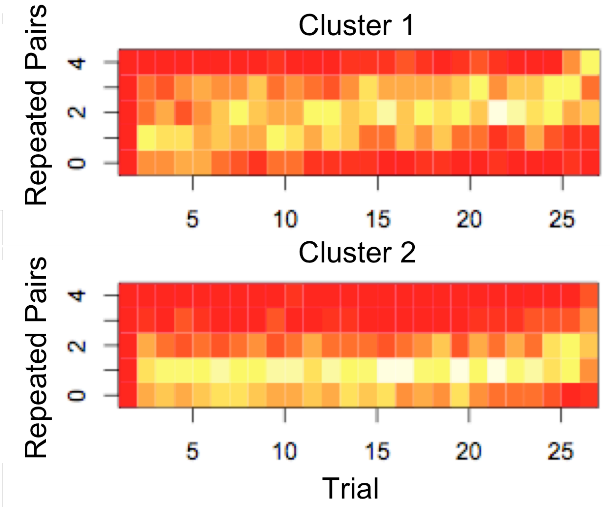

Figure 5: The number of pairs active learners chose to repeat on each consecutive trial, accumulated for each of the two clusters (red=0, white=27). Learners in Cluster 2 most often repeated one—or even zero—pairs, while Cluster 2 chose anywhere from one to three repeats per trial.

How is it that repeating more than one pair on each trial can further improve learning? Working memory can likely be used to segregate repeated and unrepeated pairs on a given trial. Thus, choosing one or three objects for repetition allows the learner to infer that the single repeated or new word goes with the repeated or new object. Repeating two pairs also yields information—only 8 associations are reasonable using repetition information, instead of 16 on a normal trial—but is most useful if a learner already knows one of the repeated pairs: then they may learn the unknown pair, and practice the known pair. To elucidate what learners are doing in active training when repeating multiple pairs, we extend a recent associative model of cross-situational word learning with a working memory mechanism.

Model

The Experiment showed that adults learn many more words from active cross-situational training than passive training. Our analysis of active learning strategies found that most people repeated one or more pairs in consecutive trials, and that repeating more pairs helped: many excellent learners repeated close to two pairs per trial. To understand how this is helpful, we will introduce and then extend an associative model of cross-situational word learning proposed by Kachergis, Yu, and Shiffrin (2012).

The model assumes that learners do not equally attend to all word-object pairings on a trial (i.e., store all co-occurrences). Rather, selective attention on a trial is drawn to strengthen associations between words and objects that have co-occurred previously. This bias for familiar pairings competes with a bias to attend to stimuli that have no strong associates (e.g., as a novel stimulus). The competing familiarity and uncertainty biases allow the model to exhibit fast mapping, since a novel word-novel object combination will demand more attention, and mutual exclusivity: a novel

[2] Due to the constraint of each pair appearing only six times—as in passive training—there are only a few objects remain to choose from, with the final trial being completely determined.

word will only become weakly associated with an already-known referent (Kachergis, Yu, & Shiffrin, 2012). For example, suppose word w_1 and object o_1 have appeared together and are thus somewhat associated, while w_7 and o_7 are novel. Given a trial with both pairs: $\{w_1,o_1,w_7,o_7\}$, w_1-o_1 demands more attention than w_7-o_1, w_1-o_7, or w_7-o_7, since w_1-o_1 is stronger than baseline. However, attention is also pulled individually to w_7 and to o_7, since both of these novel stimuli have no strong associates. Uncertainty is measured by the entropy of each stimulus' association strengths. Because of the high joint uncertainty of w_7 and o_7, more attention is given to the association w_7-o_7. Thus, attention is mostly divided between w_1-o_1 and w_7-o_7, although the other pairings will be strengthened a bit.

Formally, 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 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 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 bias for attending to unknown stimuli with a bias 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 strong associates also attract attention, whereas pairings between uncertain objects and known words, or vice-versa, draw little attention. To capture stimulus uncertainty, we allocate strength 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_2n) 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:

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

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

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

In this equation, α is a parameter governing forgetting, χ is the weight being distributed, and λ is a scaling parameter governing differential weighting of uncertainty and prior knowledge (familiarity). As λ increases, the weight of uncertainty (i.e., the exponentiated entropy term, which includes both the word's 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, a learner is tested with each word and chooses an object from n alternatives in proportion to the association strengths of each alternative to that word.

Using competing biases for familiar pairings and uncertain stimuli, this associative model learns on a trial-by-trial basis by distributing attention in a way that corresponds with both our intuitions about word-learning and a number of empirical findings. However, although this model does exhibit training order effects, it has no working memory component that would confer additional benefit for successively repeated pairs. Thus, we augment the **baseline** model with a mechanism that segregates words and objects repeated from the last trial from unrepeated stimuli, and only strengthens associations within these subsets. This working memory (**WM**) model will learn better than the baseline model whenever there are repetitions, because of the 16 possible associations on the trial, it will not attend to the spurious ones between repeated stimuli and unrepeated stimuli: 6 in the case of one or three repeated pairs, and 8 in the case of two repeated pairs. To estimate whether people are attending more to the repeated or unrepeated stimuli, we added an attention parameter β to the WM model that apportions more weight to associations between repeated stimuli as β approaches 1, and more weight to unrepeated pairs as β approaches 0. When β=.5, the attention given to repeated vs. unrepeated associations is proportional to the size of each subset.

Three parameters (χ, α, and λ) were fit to each active training order for the baseline model, and four (χ, α, λ, and β) were fit to the WM model. Fitting only to the overall mean accuracy of each active training order—two conditions per learner—does not capture detail of repetition's effect on accuracy, which may vary in different active training sessions. Instead, we fit to the accuracy for each subgroup of pairs that were repeated different numbers of times (0-5, as each pair was seen 6 times). An ANCOVA shows number of repetitions significantly affected accuracy (F(1,209) = 8.50, p<.01), discussed in more detail later.

Results & Discussion

Overall, both models achieved quite good fits to the data, with R^2=.901 for the baseline model, and R^2=.925 for the WM model. The WM model's BIC was 577.7 and the baseline model's BIC was 565.9, so the WM model is preferred, despite the additional parameter. Figure 5 shows mean accuracy for humans and both models on the subsets of pairs that were repeated on pairs of consecutive trials. Accuracy increases from 0 to 3 repetitions, while the few people who repeated pairs 4 or 5 times improved less, though with great variability.

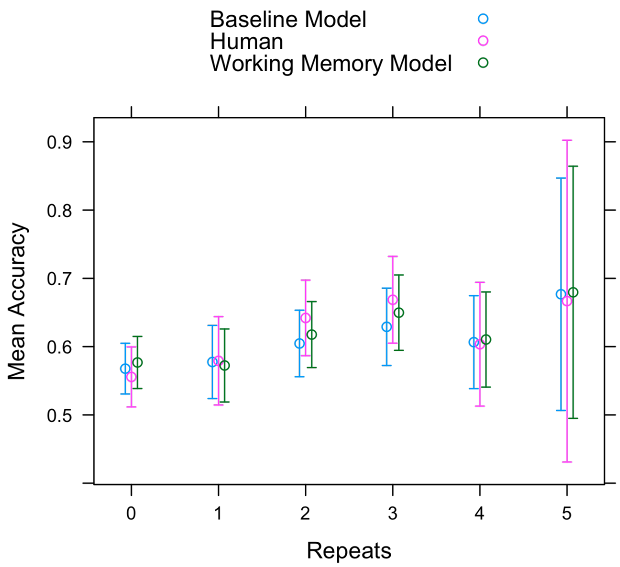

Figure 5. Human and model accuracy on actively-learned subsets of items that were repeated 0-5 pairs of trials (not necessarily consecutive for all repetitions—except in the rare case of 5 repetitions). Error bars are +/-SE.

Given the large number of repetitions used by active learners, it is surprising that the baseline model can approach the fit of the WM model without explicit awareness of repetitions. This may indicate that individual differences (e.g., in learning rate) contribute much of the variability. However, the WM better accounts for the data, and contains a parameter, β, that should be valuable in our pursuit to understand the range of strategies. Do learners focus more on repeated (β≈1) pairs, unrepeated pairs (β≈0), or do they split attention (β≈.5)? Figure 6 shows the trimodal distribution of the estimated β values: many people focused almost exclusively on learning the repeated pairs, but several attended only to unrepeated pairs, and the majority split attention roughly equally. Once again, we see individual differences spanning the range of possibilities, although the peaks are of interest. However, β values were uncorrelated with accuracy (r=.06), and people with modal β values showed no different accuracy, on average. Thus, the WM model found three attention strategies for repeated pairs, but the strategies alone do not predict performance.

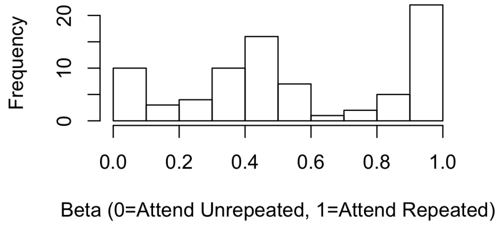

Figure 6: Histogram of best-fitting β values, showing a trimodal distribution peaked (highest to lowest) at 1—attend repeated pair, .5—split attention, and 0—attend unrepeated.

General Discussion

Active learning can speed language acquisition if the learner can implement an appropriate strategy based on the information available to them. In the context of cross-situational word learning, we have shown that many adults can generate strategies that improve their overall learning. Indeed, people who did active learning first were better at passive learning, suggesting that some strategies carried over, which is somewhat puzzling because many of the active learning strategies involved trial-to-trial repetitions of at least one word-object pair—many more, in the most successful active learning blocks. Given that active learners were using many repetitions, but with apparently diverse strategies and outcomes, we extended a word-learning model with a working memory mechanism to attempt to see how people were leveraging repetitions. Overall, the model accounted for active learning accuracy very well, but parameters told of a plurality of strategies: many people ignore unrepeated pairs while several only attend to these pairs, but the majority fall roughly in the middle, attending to both repeated and unrepeated pairs. It may be that this focus often shifts during a block, as knowledge develops. Future work should also focus on predicting which pairs people will choose next, perhaps based on their current knowledge state.

In summary, active noun learners use many repetitions, and successfully learn far more than in passive training. Infants may also benefit from such repeated labeling, and fortunately there is much autocorrelation in scenes (as you turn your head or shift your eyes, many objects remain in view) and in language (conversations drift over minutes). Moreover, we suggest that infants likely influence their learning environment in a way that is analogous to the active learning paradigm we present here. By choosing to look longer at some objects, they may increase the likelihood that a caregiver will label one of those objects. Active learning is clearly a powerful learning aid, and with better understanding it can likely be harnessed in education to speed learning in many domains.

References

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009a). Temporal contiguity in cross-situational statistical learning. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.) *Proceedings of CogSci 31*. Austin, TX: Cognitive Science Society.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009). Frequency and contextual diversity effects in cross-situational word learning. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.) *Proceedings of CogSci 31* (pp. 755-760).

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

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding groups in data: An introduction to cluster analysis*. Wiley, New York.

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

Settles, B. (2009). Active learning literature survey. *Computer Sciences Technical Report 1648*, Uni. of Wisconsin-Madison.

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