Hypothesis Testing and Associative Learning in Cross-Situational Word Learning: Are They One and the Same?

Chen Yu, Linda B. Smith, Krystal A. Klein and Richard M. Shiffrin ({chenyu}@indiana.edu)
Department of Psychological and Brain Sciences, and Cognitive Science Program, Indiana University
Bloomington, IN 47405 USA

Abstract

Recent studies (e.g. Yu & Smith, *in press;* Smith & Yu, *submitted*) show that both adults and young children possess powerful statistical computation capabilities -- they can infer the referent of a word from highly ambiguous contexts involving many words and many referents. This paper goes beyond demonstrating empirical behavioral evidence -- we seek to systematically investigate the nature of the underlying learning mechanisms. Toward this goal, we propose and implement a set of computational models based on three mechanisms: (1) hypothesis testing; (2) dumb associative learning; and (3) advanced associative learning. By applying these models to the same materials used in learning studies with adults and children, we first conclude that all the models can fit behavioral data reasonably well. The implication is that these mechanisms – despite their seeming difference -- may be fundamentally (or formally) the same. In light of this, we propose a formal unified view of learning principles that is based on the shared ground between them. By doing so, we suggest that the traditional controversy between hypothesis testing and associative learning as two distinct learning machineries may not exist.

Keywords: language acquisition, word learning, computational modeling.

Introduction

There are an infinite number of possible word-to-world pairings in naturalistic learning environments. Quine (1960) illustrated this **indeterminancy problem** with this example: Imagine an anthropologist who goes to a foreign country and observes a native speaker saying "gavagai" while pointing in the general direction of a field with a rabbit in it. The intended referent (rabbit, grass, the field, or rabbit ears, etc.) is indeterminate from this experience. Hard as it seems to be to infer referents correctly from such data, typically developing children have no problem using data of this sort to learn their native vocabulary smoothly and effortlessly.

For 30 years, research on the indeterminacy problem has concentrated on **single trial learning** such that a language learner – despite the logical ambiguity pointed out by Quine – nonetheless correctly and rapidly maps the word to the intended referent **on that trial** and by most accounts does so on the basis of social, linguistic and/or representational constraints (e.g. Gleitman, 1990; Tomasello, 2000).

However, most previous experiments showing such fast-mapping of a word to a referent were conducted in highly constrained laboratory environments. A typical scenario is like this: an experimenter presents one or two objects to young subjects and utters a simple phrase, such as "look, this is a toma!" Everyday learning contexts are much more cluttered than this with many candidate objects for a word and many candidate words for an object, and in the discourse context many shifts in attention among the candidate words and referents. These highly ambiguous learning environments with many words and many objects may significantly limit the plausibility of fast-mapping solutions.

There is, however, an alternative way that learners might solve the indeterminacy problem, not in a single encounter with a word its referent, but across many trials and simultaneously for many words and referents. This solution is possible if learners can accumulate the statistical evidence across multiple learning situations. A learner who is unable to unambiguously decide the referent of a word on any single learning trial could nonetheless store possible word-referent pairings across trials, evaluate the statistical evidence, and ultimately map individual words to the right referents through this cross-trial evidence. For example, a young learner (an infant perhaps) might hear the words "bat" and "ball" in the context of seeing a BAT and BALL. Without other information, the learner cannot know whether the word form "ball" refers to one or the other visual object. However, if subsequently, while viewing a scene with the potential referents of a BALL and a DOG, the learner hears the words "ball" and "dog" **and if the learner can combine** the conditional probabilities of co-occurrences from two streams of data **across trials**, the learner could correctly map "ball" to BALL. This mechanism seems to be quite straightforward. However, until recently, there was no evidence as to whether human learners perform these kinds of statistical computations.

In a series of recent experiments, we showed that both adults (Yu & Smith, *in press*) and 12 month-old infants (Smith & Yu, *submitted*) do calculate cross-trial statistics and find correct word-referent mappings amidst highly ambiguous learning contexts, and they do so with impressive accuracy over relatively few trials. Cross-situational statistical learning is clearly within the repertoire of human learners. This paper intends to go beyond demonstrating what language learners can do, and focus on investigating the internal learning mechanisms that may underlie their powerful statistical learning capabilities. What is the nature of the underlying learning processes? Can we provide a formal account of cross-situational learning?

Traditionally, two classes of cross-situational learning mechanisms have been considered. One is associative learning. Across trials, the learner could accrue associations between words and their potential referents by strengthening and weakening associative links between experiences of

names and objects (see Plunkett, 1997, for review and discussion). Building on the "ball/bat" example above, the learner could on trial 1, equally associate "ball" with BALL and BAT. But after trial 1, and based on the experience of "ball" in the context of BALL and DOG, the association between "ball" and BALL would be much stronger than that between "ball" and BAT. Over enough trials, these association strengths would converge on the real world statistics and yield the right word-referent pairs. The success of such a learning mechanism would seem to depend heavily on constraints on the kind of associations potentially formed given the logically infinite number of referents in any scene (e.g. Regier, 2003).

An alternative way that learners could use cross-trial information is through hypothesis testing, by formulating and evaluating hypotheses about which names map to which referents (e.g., Siskind, 1996; Tennenbaum & Xu, 2000). Building on the "ball/bat" example above, the learner could wrongly hypothesize on the initial trial that "ball" refers to BAT but correct that hypothesis on trial 2 which presents disconfirming evidence. Across trials, the co-occurrence probabilities would support the "right" hypotheses for the language over others. These kinds of learning mechanisms also require constraints on the kinds of hypotheses that can be formed, given the infinite number of logically correct hypotheses true of any single datum (see also Quine, 1960).

Based on the above two learning principles, this paper first presents three computational models as simulated learners who receive the same training data that human subjects received in Yu and Smith (in press) experiments. Given highly ambiguous learning trials with many possible words to be learned and many possible referents, simulated learners need to keep track of and memorize many word-referents pairs and accrue cross-situational evidence just as the human learners did. At the end of training, the simulated learners are tested on the same tests that were used with the human learners. In this way, we can directly compare human and simulated learners. Moreover, we use simulated learners to explore a variety of potential constraints and to systematically evaluate their importance in word learning. Based on this comparative study of three computational models, we propose a unified view of hypothesis testing and associative learning, suggesting these two can be treated as variants of the very same learning mechanism.

Experimental Data for Simulation

This section presents the results in (Yu & Smith, in press) which are used as empirical data in the current simulation study. In our first experiment, we asked how easily adults could simultaneously learn 18 word-referent pairs from learning trials that are individually highly ambiguous. The experiment included three conditions that manipulated within-trial ambiguity: 2 words and 2 possible referents, or 3 words and 3 possible referents, or 4 words and 4 possible referents on each trial. The 2×2 condition yields 4 possible word-referent associations per trial. The 3×3 condition yields 9 potential associations per trial. The 4×4 condition yields the seemingly overwhelming number of 16 word-referent associations per trial. Although there is no information on any individual trial as to which label goes with which word, across trials the underlying word-referent mappings are certain in that individual labels are presented in a training trial if and only if the referent is also presented.

The stimuli were slides containing pictures of uncommon objects (e.g. canister, facial sauna, and hitch haul) paired with auditorally presented artificial words. The three conditions were presented within subjects and so in total, there were 54 unique objects and 54 unique pseudowords partitioned into the three sets of 18 words and referents for each condition. By design, the three learning conditions differed in the number of words and referents presented on each training trial (2, 3 or 4) and the number of times each word and referent pair was presented across trials was held constant at 6. Order of trials within a condition was randomly determined. Order of the three conditions (a within-subject manipulation) was counterbalanced across subjects. Training was passive. The adult participants (n=38) just watched and listened as the trials were presented; they were *not* told that there is a one-word-one-referent correspondence. After training in each condition, learning was assessed via a four-alternative forced-choice test; presented with one word, participants were asked to choose the picture to which the word referred. The three foils were all drawn from the set of 18 training pictures. Participants learned more word-referent pairs in each condition than expected by chance ($t(37)=8.785$, $p<0.001$, one-tailed, for 4 × 4). They discovered on average more than 16 of the 18 pairs in the 2×2 condition (M=16.2, SD=2.5) and more than 13 of the 18 pairs in the 3×3 condition (M=13.6; SD=3.5), all this in less than 6 minutes of training per condition. Even in the 4×4 condition with 16 potential associations per trial, subjects discovered almost 10 of the 18 word-referent pairs (M=9.5; SD=2.9). The level of performance in the three conditions is remarkable -- in a very short time, over relatively few trials, each highly ambiguous, subjects nonetheless found the underlying word-referent pairs. Human learners can and do keep track of the simultaneous co-occurrences of many labels and referents across trials such that they can find individual mappings. Moreover, this is readily accomplished in relatively few learning trials.

Experiment 2 of Yu & Smith (in press) explored adult learning in the condition of high within-trial ambiguity -- the 4 ×4 condition. We were particularly interested in learning under such high within-trial uncertainty as a function of the number of word-referent pairs to be learned. Accordingly, in this experiment, each condition is a version of the original 4 x 4 condition above. We manipulated: (1) the total number of word-referent pairs to be learned (9 or 18) and (2) the number of repetitions of each word-referent pair (6, 8 or 12). In the 9 words/8 repetitions condition, subjects attempt to discover a total of 9 word-referent pairs each repeated 8 times over the course of training. In the 9 words/12 repetitions condition, subjects attempt to discover 9 word-referent pairs but are given 4 additional repetitions

of each word-referent pair. Finally, the third condition is a replication of the original 4×4 condition, 18 word-referent pairs to be learned and 6 repetitions of each. Intuitively, the 9 words/12 repetitions condition should improve the learning performance because, compared with the 18 words/6 repetitions condition, the number of words needed to be learned are reduced while their occurrence frequencies are doubled. All aspects of the experiment are identical to the first experiment except for the composition of the three training conditions.

In terms of the *proportion* of word-referent pairs to be discovered, participants performed comparably in the three conditions, ($F(2,54) = 0.52$; $p > 0.5$), discovering more pairs than expected by chance ($t(27) > 6.4$ in all three conditions, $p < 0.001$). In terms of the *total number of pairs learned*, subjects actually learned more pairs in the 18 word-referent condition (M=9.461, SD=2.907) than in the two 9 word-referent conditions (8 repetitions: M=5.111, SD=1.706; 12 repetitions: M=5.481, SD=2.089). The 18 word condition presents the same within-trial ambiguity, more word-referent pairs to be learned, and fewer repetitions of the individual word-referent pairs than the other two learning conditions. If numbers of co-occurrences were all that mattered, this condition should lead to the poorest overall performance. However, for statistical learners, smaller data sets are not as good as large ones because spurious correlations are more likely to occur (and thus also in the lower foil probabilities at test).

Overall, our experimental results show the power of cross-situational statistical learning: Even when the referent of a word cannot be unambiguously determined on any single learning trial, across multiple trials involving many different words and many different potential referents, the word and referent will occur most systematically than any other. The more words and referents there are to learn and that may co-occur together on any learning trial, the more discernible the systematicity – across trials – of the underlying correct mappings.

Hypothesis Testing Model (HTM)

There is no information at the beginning of learning to guide learners; thus it is quite plausible that on the first trial learners randomly select word-referent pairs as their initial hypotheses, gradually justifying or replacing these hypotheses as more trials ensue. Following this general principle, the specific questions for such a hypothesis testing mechanism are (1) how hypothesized pairs are selected and stored from a trial? (2) how subjects justify whether a word-object pair is correct? (3) whether they use the mutual exclusivity constraint if two working hypothesized pairs are not compatible? and (4) whether they use previously learned pairs to help the learning of new pairs in subsequent trials?

Our first simulation study attempts to answer these questions and also to generate a dynamic picture of the real-time learning when the simulated learner is fed with the same set of ordered trials as the adult learners. Here we use the 4×4 condition as an example to show how the model works because this condition has been tested in two completed experiments. The simulations on other conditions are achieved by applying the corresponding stimuli to the same model.

In the 4×4 condition, the 18 novel word-picture pairs can be represented as $\{(p_1, w_1),(p_2, w_2),.....(p_{18}, w_{18})\}$. In the *i*th trial, the stimuli are $T_i = \{p_{i_1}, p_{i_2}, p_{i_3}, p_{i_4}, w_{i_1}, w_{i_2}, w_{i_3}, w_{i_4}\}$ while i_1, i_2, i_3 and i_4 can be selected from 1 to 18. And there is no information as to which picture goes with which name. We also assume that the simulated learner maintains a list of hypothesized pairings as learned results from previous trials. Thus, its lexical knowledge at the *i*th trial can be then represented as a list of pairs $M = \{(p_{n_1}, w_{m_1}),$ $(p_{n_2}, w_{m_2}),.....,(p_{n_k}, w_{m_k})\}$ while n_j and m_j can be selected separately from 1 to 18, and the equivalence of these two indicates a correct pairing. At the beginning, the simulated learner randomly picks one word and one picture from a trial and builds a hypothesized pairing. With more trials, more pairings are built and stored in the memory. Two additional mechanisms are utilized to make this learning process more effective. First, one important constraint in adding new pairs is to maintain the consistency of hypothesized pairings so that one word can be associated with only one picture. This constraint explicitly encodes such proposals as mutual exclusivity (Markman, 1990) and contrast (Clark, 1987) into the learning machinery and by doing so makes learning more efficient because without such a constraint, the simulated learner would randomly select many conflicting (and therefore incorrect) word-picture pairs across multiple trials. Second, the model keeps track of the frequency of each hypothesized pair. When the number of occurrences of a pair is above a certain threshold, this pair will be treated as a learned lexeme and then used to filter out the input in subsequent trials; this significantly simplifies the learning task. For instance, if a learned pair occurs in a new trial, it will be removed from the stimuli to reduce a 4×4 condition into a 3×3 condition. More importantly, subjects in empirical studies informed experimenters that they used a similar filtering strategy in the later part of the training phase when they were confident that some word-picture pairs were correct.

We applied the same training and testing data in the two adult experiments to the model. For each condition, the simulation was run for 5000 times. Thus, we had 5000 simulated subjects (with the same set of parameters) for each condition. Note that the fundamental mechanism encoded in our model is to randomly select and store hypothesized pairs. Therefore, quite different results could be obtained on each run depending on what pairs were selected from trial to trial. We used 5000 simulated subjects to ensure the statistical power of this simulation study and the results are shown in Figure 1. We observed that in general the results in simulation are quite in line with those of human subjects, suggesting that if subjects apply a simple statistical learning machinery like the one in our model, then this could explain their superior performance in learning

from individually ambiguous learning trials. The similarities of the results between human subjects and simulated subjects are consistent not only in one condition but among five conditions of two experiments, indicating that the learning principles encoded in our model are plausibly similar to those that guide the learning of human subjects.

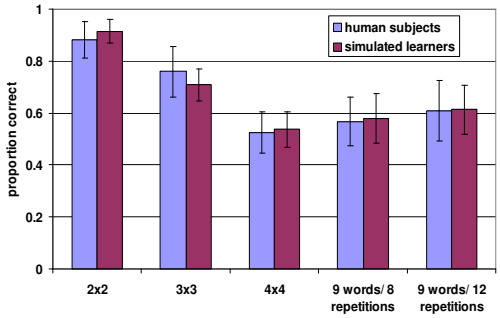
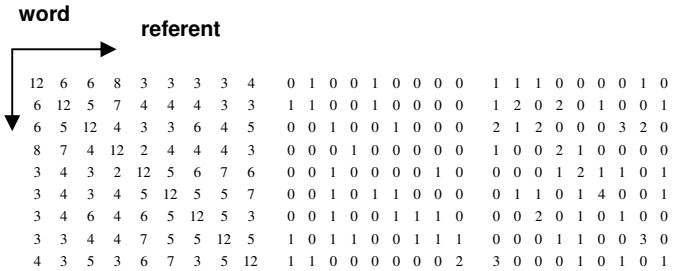


Figure 1: Simulated learners and human learners achieve similar results in all conditions.

Dumb Associative Model (DAM)

The hypothesis testing model fits behavioral data qualitatively well. One major assumption in the HTM is that it applies the constraints, such as Mutual Exclusivity, in real-time learning, and ignores other information. In contrast to this type of explicit learning, an alternative mechanism would be to associate one word with one referent at a time and to accumulate this evidence across multiple trials. During testing, this associative model picks out the object most strongly associated with the test word. Along this line, an associative learner who kept track and stored *all* co-occurrences on all trials and at test chose the most strongly associated referent would be an ideal learner, internally representing the matrix of input, as in the 9 words/12 repetitions condition shown in Figure 2 (a). Human subjects may well be able to approximate this, storing many, if not all, of the associations on a trial and accruing them across trials. Such an associative learning mechanism would, in fact, do quite well in our experimental tasks. However, it is also possible that human learners are more selective associative learners, that due to processes of competition, inhibition, and attention shifting (e.g., Kruschke, 2001), they may build a few one-word-one-object associations exhibiting a form of mutual exclusivity.

Thus far, three simple associative methods have been developed. These only pursue models based on the general principle that the system randomly selects and accumulates word-referent pairs without applying any constraints in real-time learning. Hence, dumb associative model can be viewed as based on Hebbian learning principles – the connection between a word and an object is increased if the pair co-occurs in a trial. One associative model simply accumulates one single word-referent pair from a trial. Thus, the total number of pairs selected is equal to the number of trials. Figure 2 (b) shows one instance of this model. The second and the third models select two and three pairs from a trial. Similar to the hypothesis-testing model, these three methods are also based on random selection of pairs. Therefore, the simulation was run for 5000 times to obtain the results from 5000 simulated associative learners for each method. Figure 2 (b) and (c) shows examples of association matrices built by one-pair and two-pairs learners respectively. Clearly, these two matrices are quite different and far from perfect compared with the matrix shown in Figure 2 (a). Nonetheless, when the simulated learners were asked to do the same forced-choice tests, they still demonstrated learning based on partial and incomplete matrices, as shown in Figure 3.



(a) an ideal learner (b) a simulated learner (one pair) (c) a simulated learner (two pairs)

Figure 2: the learning results in 9 words/18 repetitions condition. The row is a list of words and the column is a list of referents. Each cell represents the co-occurrence frequency of a word-referent pair. The diagonal items count relevant co-occurrences.

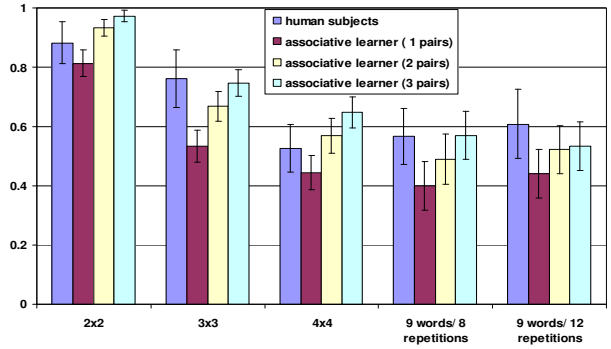


Figure 3: A comparison between human learners and three different associative learners.

Associative Translation Model (ATM)

In contrast to a dumb association model that accrues co-occurrence frequencies trial by trial and updates the strength of the connections between a word and an object, we also developed a more advanced associative model based on machine translation techniques. Briefly, machines "learn" to automatically translate one language into another through statistical regularities across large parallel corpora (e.g., statistical regularities across *Anna Karenina* in English and Russian). The basic assumption behind this approach is that there are latent meanings that both languages point to, and the machine learning techniques attempt to discover these latent structures through the statistical regularities. Here, we use this same computational approach but conceptualize the object stream as one language and the audio stream as the other, attempting to find the latent word-referent pairings between these two streams. More specifically, we used the translation model in (Brown, Pietra, Pietra, & Mercer, 1994) and applied an Expectation-Maximization (EM) based learning algorithm. Our algorithm assumes that word-referent pairs are hidden factors underneath the

observations, which consist of spoken words and extralinguistic contexts. Thus, association probabilities are not directly observable, but they somehow determine the observations because spoken language is produced based on the caregiver's lexical knowledge. Therefore, the objective of language learners or computational models is to figure out the values of these underlying association probabilities so that they can interpret the observations better. Correct word-meaning pairs are those which can maximize the likelihood of the observations. Technical detailed can be found in Yu, Ballard & Aslin (2005).

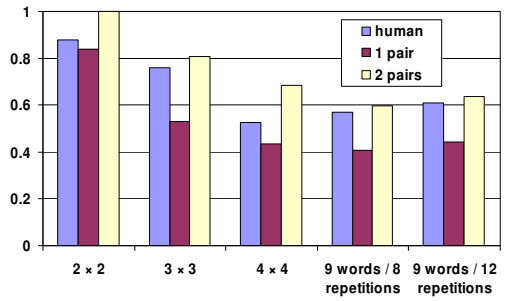


Figure 4: A comparision between human learners and assciative translation model in the 4 conditions.

Similar to the dumb associative model, we compare three variants of ATM with empirical results. As shown in Figure 4, this sort of associative mechanism can easily extract the proper mappings of words to referents – the most effective learning mechanism compared with HTM and DAM.

A Unified View

So far, we have considered three computational models, Hypothesis Testing Model (HTM), Dumb Associative Model (DAM) and Associative Translation Model (ATM). As radically different as associations versus hypotheses may seem to be, all these models can be viewed as variants of the same kinds of processes. Critically, from both an associative and a hypothesis testing perspectives, the relevant mechanisms for the cross-situational learning of words and referents are almost the same: (1) what information is stored on any individual learning trial, (2) how past knowledge constrains future learning, and (3) how accrued information is evaluated. In the following, we will compare associative versus hypothesis testing mechanisms in terms of these three aspects.

Representations and Learning Results

The representations and learning results based on associative learning and hypothesis testing appear to be radically different. As shown in Figure 5, one builds a big two-dimensional matrix to count all possible co-occurrences between words and objects while the other just keeps track of a short list of word-referent pairings. There are two ways to quantify these differences in representations: (1) the number of word-object pairs and (2) as probabilistic versus all-or-none representations.

First, the hypothesis testing model maintains a clean and short list while the associative models store many co-occurring pairs. If the number of pairs really matters, one would expect that there should, quantitatively, be a clear boundary (threshold) that can be drawn to differentiate these two approaches. For example, if the threshold is 19, then a set of 19 pairs should be treated as hypothesis testing and a set of 20 should be based on associative learning. Nonetheless, it is not clear that this kind of threshold exists at all.

Second, the hypothesis testing model stores the accrued information in a winner-take-all way – the pairs in the list are equally treated as correct while the pairs not in the list are excluded from consideration. In contrast, the associative learning method accumulates and stores the information in a probabilistic and graded way. Every co-occurring word-referent pair is assigned to an association probability based on co-occurrence frequency while some pairs have high probabilities and others are assigned with low probabilities. The dumb associative model purely relies on co-occurrence frequencies while associative translation model computes association probabilities by considering various correlations between words and objects to find an overall optimal solution.

Conceptualizing the differences in this way makes clear that associative models can be converted into hypothesis testing models and that hypothesis testing models can be converted into associative models. More specifically, a set of hypotheses can be considered as a special type of associative representation with the probabilities equal to either 1 or 0 but nothing in between. As such, the hypothesis set can be treated as a special case of associative representations and converted into a sparse and binary association matrix as shown in Figure 5. Similarly, even if lexical knowledge is stored in a probabilistic mode in an association matrix, the associative learner will need to make decisions at testing, which may force the learner to retrieve the strongest relevant associations. For example, in our word-learning task (or any other task of the same sort, for example, naming the object), the learner needs to pick out an object after hearing a word. To do that, the learner finds the most relevant referent and ignores others. Thus, one can also extract a hypothesis set from an association matrix by picking out strongest associations (converting probabilistic associations into explicit hypotheses). In doing so, different thresholds used in the conversion may determine the number of pairs in the hypothesis set. But again there is no clear threshold of the number of pairs that can be applied to separate two mechanisms.

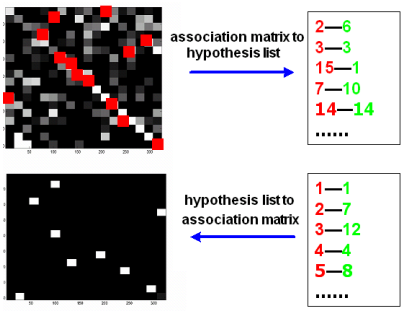


Figure 5: we can select the strong associations in an association matrix to form a hypothesis set. Similarly, we can represent a hypothesis set as a sparse association matrix. The numbers in the lists are indexes of words and referents.

Learning Mechanisms

In general, all three mechanisms use cross-situational information accumulatively and all three mechanisms need to include some constraints in learning. One may attempt to differentiate the three mechanisms based on real-time versus batch processing -- at what moments are the constraints added to reduce the degree of ambiguity in the data? Does this happen in trial-by-trial learning or only at the retrieval of accumulated information during testing? HTM is a real-time learning process wherein the simulated learner continuously builds and justifies hypothesized pairs trial by trial. The decisions are made in real time on whether a selected pair is included in or excluded from the hypothesis list. When the training phase ends, the learner acquires a list of hypothesized pairs. In contrast, DAM functions in a batch mode because it accumulates data during training to form an association matrix including all the possible pairs that the model can select and store. The decisions are made in testing when the model needs to retrieve an object given a spoken word and several options. Different from the above two, ATM's learning mechanism is in between batch and real-time processing – similar to hypothesis testing, it estimates association probabilities in real time and similar to DAM, the association probabilities in the association matrix jointly determine the referent of a word during testing. Recent advances in machine learning suggest that many learning algorithms can be converted between batch mode and real-time mode. Thus, even if real-time and batch processings are fundamentally different, associative learning and hypothesis testing mechanisms are at least compatible (and convertible) and can be grouped under a more general learning system.

Retrieving accrued information at Testing

The hypothesis testing model uses the accrued knowledge in a straightforward way. After hearing a word at each trial, it checks the hypothesis list to match the word with those in the hypothesized pairs. For associative models, lexical knowledge is represented as latent information and as a system of associations, which can be activated in response to a specific input. In fact, there are several different ways to extract and utilize latent knowledge from an association matrix. For example, a hypothesis set can be extracted by picking out the strongest associations from the association matrix. Similar to the hypothesis testing model, the mutual exclusivity constraint can be added to ensure the word-object pairs in the list are consistent. Another method is to decompose the association matrix into several hypothesis sets, each of which forms a consistent set of word-object pairs. At test, the overall decision is based on hypothesis test averaging. Thus, this is the main conclusion -- there is *no* fundamental difference in these two learning principles.

Conclusion

This idea that hypothesis testing and associative learning may be special cases of a unified set of learning mechanisms is theoretically significant on various grounds. First, the theoretical and empirical exploration of learning mechanisms within such a unified view may reveal new insights about learning processes that fall between these two classic extremes. Second, the learning literature is replete with cases in which one or the other approach appears better. Conceptualizing the two approaches as special cases of the same principles may be a first step to understanding how certain tasks, contexts, and past history select for specific learning solutions. Third, there are reasons to suspect that young learners (in some domain) are more likely to appear to be associative learners whereas older (or more expert) learners appear to be hypothesis testers. The present conceptualization offers a framework within which to theoretically understand the developments. Finally, this conceptualization suggests that many of the heated debates about associations versus hypotheses may be fundamentally –and mechanistically –misguided.

Acknowledgments: This research was supported by National Science Foundation Grant BCS0544995.

References

Clark, E.V. (1987). The Principle of Contrast: a constraint on language acquisition. In B. MacWinney (Ed.), *Mechanisms of language acquisition* (pp. 1-33): Hillsdale, NJ: Lawrence Erlbaum Associates.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition, 1* 1-55.

Markman, E. M. (1990). Constraints Children Place on Word Learning. *Cognitive Science, 14*, 57-77.

Plunkett, K. (1997). Theories of early word learning. *Trends in Cognitive Sciences*, 1, 146-153.

Quine, W. V. O. (1960). *Word and Object.* Cambridge, MA: MIT Press.

Siskind, J.M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition, 61*, 39-61.

Smith, L.B. & Yu, C (*submitted*). Infants rapidly learn words from noisy data via cross-situational statistics.

Tenenbaum, J.B. & Xu, F. (2000) Word learning as Bayesian inference. In L. Gleitman and A. Joshi (eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Regier, T. (2003). Emergent constraints on word-learning: A computational review. *Trends in Cognitive Sciences, 7*, 263-268.

Tomasello, M. (2000). Perceiving intentions and learning words in the second year of life. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 111-128): Cambridge University.

Yu, C., Ballard, D. H., and Aslin, R. N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science,* **29**, 961-1005.

Yu, C. & Smith, L.B. (in press). Rapid Word Learning under Uncertainty via Cross-Situational Statistics. *Psychological Science.*