Modeling Cross-Situational Word–Referent Learning: Prior Questions

Chen Yu and Linda B. Smith
Indiana University

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 by aggregating cross-situational statistical information across contexts. This ability has been explained by models of hypothesis testing and by models of associative learning. This article describes a series of simulation studies and analyses designed to understand the different learning mechanisms posited by the 2 classes of models and their relation to each other. Variants of a hypothesis-testing model and a simple or dumb associative mechanism were examined under different specifications of information selection, computation, and decision. Critically, these 3 components of the models interact in complex ways. The models illustrate a fundamental tradeoff between amount of data input and powerful computations: With the selection of more information, dumb associative models can mimic the powerful learning that is accomplished by hypothesis-testing models with fewer data. However, because of the interactions among the component parts of the models, the associative model can mimic various hypothesis-testing models, producing the same learning patterns but through different internal components. The simulations argue for the importance of a compositional approach to human statistical learning: the experimental decomposition of the processes that contribute to statistical learning in human learners and models with the internal components that can be evaluated independently and together.

Keywords: statistical learning, computational modeling, word learning

Supplemental materials: http://dx.doi.org/10.1037/a0026182.supp

Human learners are adept at picking up regularities in data (Perruchet & Pacton, 2006; Regier, 2003; Siskind, 1996; Sobel & Kirkham, 2007; Steyvers & Tenenbaum, 2005; Tenenbaum & Griffiths, 2001; Yu, 2008). Both adults and infants track sequential probabilities to segment continuous speech into individual words (Newport & Aslin, 2004; Saffran, Aslin, & Newport, 1996); they extract rudimentary grammars from the latent structure of sequences of words (Gomez & Gerken, 1999; Gómez & Maye, 2005; Saffran & Wilson, 2003); they generalize sequential patterns from visual sequences (Kirkham, Slemmer, & Johnson, 2002), and they find word–referent pairings in noisy and ambiguous data by tracking co-occurrences and non-co-occurrences across many individual word–referent pairings (Frank, Goodman, & Tenenbaum, 2009; L. Smith & Yu, 2008; Vouloumanos, 2008; Yu, Ballard, & Aslin, 2005; Yu & Smith, 2007).

This article is concerned with the last form of statistical learning—mapping words to referents—under conditions of uncertainty. Everyday word learning occurs in noisy contexts with many words and many potential referents for those words, as well as much ambiguity about which word goes with which referent. One way to resolve this ambiguity is for learners to accumulate evidence across individually ambiguous contexts (Gleitman, 1990; Pinker, 1984). Figure 1 illustrates a simple example. A learner hears the words *ball* and *bat* in the context of seeing the object ball and the object bat. Without other information, the learner cannot know whether the word form *ball* refers to one or the other visual object. If subsequently, while viewing another scene with the potential referents of ball and dog, the learner hears the words *ball* and *dog,* and if the learner can combine the conditional probabilities of co-occurrences across trials, the learner could correctly map *ball* to the object ball (and perhaps also infer the connection between the word *bat* and the object bat). This solution seems straightforward. However, until recently, there was no evidence as to whether human learners do this kind of learning with any facility. Now, a growing set of data shows that adults are quite good at this, even when faced with many novel words and many novel referents under conditions of high ambiguity (Kachergis, Yu, & Shiffrin, 2009; Yu & Smith, 2007; Yurovsky & Yu, 2008); experimental studies also indicate that infants and young children do this kind of learning as well (Akhtar & Montague, 1999; Fisher, Hall, Rakowitz, & Gleitman, 1994; L. Smith & Yu, 2008; Vouloumanos & Werker, in press; Yu & Smith, in press). The open question is what the responsible learning mechanism is.

Discussions of word learning in general (Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005; Kaminski, Call, & Fischer, 2004; Markman, 1992; L. B. Smith, 2000) and cross-situational word learning in particular (Frank et al., 2009; Yu, 2008; Yu & Smith, 2007) are often couched in very different terms
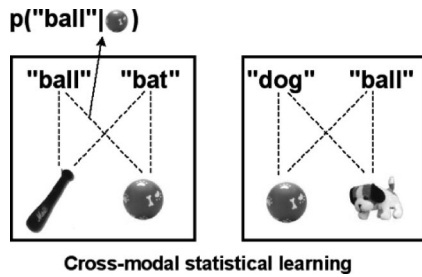
Figure 1. Cross-situational statistics between words and referents can be calculated across trials to unambiguously determine correct word–referent pairings.

that divide into two general classes of mechanisms: hypothesis testing and associative learning. Proponents of a hypothesis-testing framework often characterize associative learning as the simple and uninformed registration of co-occurrences and/or the calculation of conditional probabilities and thus as "dumb" (Keil, 1992). Hypothesis testing, in contrast, is generally characterized as a form of learning in which coherent hypotheses are formed often in conceptually constrained ways and then either confirmed or disconfirmed—not just by counting co-occurrences—but through more statistically sophisticated evaluations of the evidence. In the context of word learning, associative learning and hypothesis testing are thus typically seen as fundamentally different kinds of learning mechanisms and with fundamentally different implications about the nature of the learner.

This article presents a series of simulations that serve as a first step to targeted experimentation and modeling in the service of understanding human cross-situational word learning. The goal—and the end product of the simulations—is not a decision as to whether some specific hypothesis-testing or associative-learning model better fits some specific set of experimental data. Instead, the main thesis is that the contrast between hypothesis testing and associative learning in the context of statistical word learning is not well formed. There is a muddle as to where, when, and how the psychological work of statistical computation is being done—in the selection of information, in the learning machinery, or at decision (knowledge retrieval). We argue for a decomposition of models—and a decomposition of empirical research on statistical learning—into comparable and coherent psychological components. This decomposition is necessary because selection and decisional processes interact with the core learning mechanisms of the two different classes of theories. The upshot of these interactions is that one kind of learning mechanism (e.g., associative learning) can mimic (through multiple routes) the properties of another (e.g., hypothesis testing); a similar point, though in different contexts, has been made by other researchers (J. A. Anderson, 1970; J. R. Anderson, 1978; J. R. Anderson & Bower, 1973; Baum, 1989; Quartz & Sejnowski, 1997), but it is a lesson that needs to be relearned to make progress in understanding statistical word–referent learning. The main conclusion of this article is that these two classes of models will be discriminated by direct assessments of the different components and by an understanding of how those components interact with each other.

Hypothesis Testing

The word learner's task has often been viewed as akin to that of a linguist testing hypotheses about how words map to possible referents (Bloom, 2000; Carey, 1978; Markman, 1992). In these accounts (Vouloumanos & Werker, in press; Xu & Tenenbaum, 2007), the learner is assumed to represent a set of hypotheses about word-meaning mappings (Waxman & Gelman, 2009; Waxman & Hall, 1993) and then to accept or reject specific hypotheses based on the experienced input. Recent probabilistic models liberalize this by increasing and decreasing the likelihood (rather than outright acceptance or rejection) of hypotheses in light of new evidence (Xu & Tenenbaum, 2007). In brief, the central idea within the many variants of hypothesis-testing accounts is that learners represent specific hypotheses about word–referent pairings and then, in the face of experienced evidence, select among those hypotheses based on some principled inference procedure. This general class of models has both experimental (Halberda, 2006; Vouloumanos & Werker, in press; Xu & Tenenbaum, 2007) and computational support (Frank et al., 2009; Siskind, 1996).

However, because there is an unlimited set of possible hypotheses that are objectively true for any set of data (Goodman, 1965), research in this tradition has also been interested in constraints on the possible space of hypotheses. Some of these constraints concern possible kinds of meanings such as category hierarchies (Xu & Tenenbaum, 2007) or syntactic category–meaning links (Waxman & Booth, 2001). Within this tradition and particularly relevant to the focus of this article, there have also been proposed constraints on how the evidence for hypotheses is evaluated (Frank et al., 2009; Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992; Markman, 1990; K. Smith, Smith, Blythe, & Vogt, 2006). Most notable of these proposed constraints is the mutual exclusivity (ME) assumption (Golinkoff, Mervis, & Hirsh-Pacek, 1994; Halberda, 2006; Markman, 1992; Merriman, 1999), which posits that if an object has one name, it should not have another. There is considerable experimental evidence that children do adhere to something like an ME assumption in mapping new words to referents, showing a strong novel-word-to-novel-referent bias in many (but not all) experimental contexts (e.g., Golinkoff et al., 1992; Golinkoff, Jacquet, Hirsh-Pasek, & Nandakumar, 1996). The ME constraint has also been shown to be powerful in hypothesis-testing models (HTMs) of cross-situational learning, enabling the model to rapidly resolve competing hypotheses about word–referent pairings (K. Smith et al., 2006; Siskind, 1996).

Thus, within the hypothesis-testing framework, a learner faced with the cross-situational word learning task of Figure 1 might wrongly hypothesize on the initial trial that *ball* refers to the object bat but correct that hypothesis on Trial 2 given the disconfirming evidence. Given enough data across individually ambiguous trials and perhaps also constraints such as ME, the co-occurrence regularities would, in the end, support the right hypotheses over others. Knowledge, the outcome of this learning, is a set of propositions, a list of confirmed hypotheses, each specifying a word and its referent. Although there are different versions of HTMs that vary from all-or-none decisions to more powerful and probabilistic forms of inference, in this article, we concentrate on the core assumption—formation and evaluation of hypotheses—in an effort to more clearly understand the differences between HTMs as a class and associative models as a class.

Associative Learning

Proponents of hypothesis-testing accounts usually characterize associative learning as consisting of the unconstrained and simple counting of co-occurrences of real-world statistics (Keil, 1992; Waxman & Gelman, 2009; Xu & Tenenbaum, 2007). If one takes this simple idea and applies it to the cross-situational learning task in Figure 1, it would work like this: On Trial 1, the learner could equally associate *ball* with both the object ball and the object bat, but after Trial 2 and with the experience of *ball* in the context of the object ball and the object dog, the association between *ball* and the object ball would be stronger than that between *ball* and the object bat. Over enough trials, these association strengths will converge on the real-world statistics. Within this account, statistical learners may not have learned that an individual word refers to an individual kind of object but may only have built stronger associations of words to the targets than to the foils. Thus, instead of a list of words and their referents, the outcome of associative learning would be a matrix of stronger and weaker associations with the columns representing words, the rows representing referents, and the cells representing the associative strengths of word–referent pairs.

Critically, a contemporary understanding of associative learning suggests that it involves much more than merely counting up contingencies but instead also involves highly interactive effects of learned associations on each other, on attention, and on subsequent learning (Billman & Knutson, 1996; Kruschke, Kappenman, & Hetrick, 2005; Rogers & McClelland, 2004; Yoshida & Smith, 2005). For example, there are now a number of associative models that generate ME-like effects (MacWhinney, 1987; Merriman, Bowman, & MacWhinney, 1989; Regier, 2003). Contemporary associative models also include interactive effects among associations—both competitive (Colunga & Smith, 2005; Colunga, Smith, & Gasser, 2009; Gupta & MacWhinney, 1997; Merriman, 1999) and positive-feedback or "rich get richer" effects (Pauli & O'Reilly, 2008; Rehder & Murphy, 2003; Yoshida & Smith, 2005) that can lead to almost rule-like (all-or-none) learning outcomes (Colunga & Smith, 2005; Mareschal & Shultz, 1996) as well as nonlinear accelerations in the rate of learning (Colunga & Smith, 2005; Plunkett, Sinha, Moller, & Strandsby, 1992; Regier, 2003). The interactive effects among sets of associations in these models are founded on experimental work on associative and attentional mechanisms in animals (Kamin, 1967; Mackintosh, 1975; Rescorla & Wagner, 1965) and human adults (Chun & Jiang, 1999; Heit & Bott, 2000; Kahana, 2002; Kruschke, 2001; Kruschke & Blair, 2000; L. B. Smith, Colunga, & Yoshida, 2010). In brief, there are many variants of associative-learning models that range from quite simple (counting co-occurrences) to much more sophisticated. Here, however, with the main goal being to clarify the differences between the core ideas of associative learning and hypothesis testing, we concentrate on the simplest (dumbest) associative learner, one that counts co-occurrences.

Building a Model

These two learning mechanisms are founded on fundamentally different assumptions and would seem to have profoundly different implications for how we understand the learning task, what is learned, and the learner. Yet the ideas of hypothesis testing and association are grand ideas that provide at best bare-bones sketches of possible models. To examine the applicability of these ideas to some learning domain such as cross-situational learning, one needs to fill in these sketches. The problem is that to build such a working model to perform some learning task, one will have to specify processes that seem outside of the main debate about the two kinds of learning. More specifically, to build a working model that can simulate performance in some task, one needs to specify information selection, the learning machinery (which is viewed as the core of the debate but which also needs specification to be actually implemented in a model), and the decision processes at test.

Information Selection

The cross-situational learning task simultaneously presents multiple words and multiple referents. The starting point for any model then is how much and what kind of information is selected and stored from each learning trial. Even if one assumes that the units for learning are whole words (not their parts or phrases) and whole objects (not their parts, properties, or sets of objects), there are still different options in information selection. One could, as an ideal learner, register all the word–referent pairs—all the associations or all the possible hypotheses consistent with that input. Alternatively, one might attend to only a subset of words and referents, registering just partial information—some words, some referents—from all that are available at a single moment. Selection could be very narrow (e.g., just one word and one referent per learning moment), or it could be broader. Furthermore, selection could change with learning, beginning broadly and then becoming narrower if selection depends on what is already known. In any case, a first step for a model that learns from a series of individually ambiguous learning trials is to specify the information that is input into the learning mechanism.

Learning Machinery

In most discussions of statistical word learning, the learning machinery—associative or hypothesis testing—is seen as the central theoretical question. However, within these two classes, there are choices to be made about the learning mechanism itself. For example, a hypothesis-testing learner could keep track of and aggregate evidence for just some (and not all possible) word–referent hypotheses. If so, the model needs to specify how those initial hypotheses are formed or selected as well as how many hypothesized pairs the learning system is capable of tracking. The model also needs to specify how strong the evidence needs to be for the learning system to accept or reject a hypothesized pair. There are roughly the same choices, though couched in different terms, for associative models. An associative learner could memorize all co-occurrences (the whole matrix) or (because of possibly competitive processes) only register and aggregate some occurrences. Furthermore, an associative learner could simply count registered occurrences, or such a learner could apply more advanced probabilistic computations based on those counts, such as conditional probabilities (Aslin, Saffran, & Newport, 1998) or latent variables in probabilistic graphical models (Jordan, 1998). In brief, both hypothesis-testing and associative models have choices about the kind of information aggregated, the kinds of computa-

tions applied to that information, and the form of the learning outcome.

Decisions at Test

Finally, for both kinds of learning systems, models need to specify how the accrued information is retrieved and used by learners to make decisions during testing. Commonly, experiments on word learning present the learner at test with a single word and a choice among some number of alternative referents (Yu & Smith, 2007; Yurovsky & Yu, 2008). Given the evidence accumulated during training—a list of hypotheses, the probabilities associated with hypotheses, or weaker and stronger associations—participants must make a momentary decision, selecting the best alternative choice given the queried word. Learners could apply a winner-take-all strategy such that the strongest hypothesis or association for the tested word governs choices in an all-or-none manner. Alternatively, responses could be graded and based on the probability that a hypothesis is correct or on the relative strengths of all the associations to the candidate word. Furthermore, the decision could be based on the single word (or word–referent pairing) being queried, or it could be based on a "best overall solution" for all the word–object pairings acquired during training. Whatever the core learning mechanism, models need to specify how acquired information is retrieved and how decisions are made at the time of test because the learning performance that one is modeling is based on those decision processes.

Rationale and Organization

The main goal of the present article and the simulations is to understand how these components—and their different specifications—interact to produce different learning outcomes. This is a necessary prior step to any attempt to distinguish the two classes of models. Moreover, we suggest that systematic empirical studies targeted at understanding these internal components will be necessary before attempts at model discrimination are useful and generalizable. In the present simulations, we stick to simple core learning mechanisms designed to capture the core ideas of hypothesis testing and associative learning and specifically vary selection and decision processes to understand how they interact with the core difference between the two classes of theories. The organization of the article is as follows: We first present an overview of the two basic models that we compare and how they work. We then present the experimental tasks (from Yu & Smith, 2007) that we use merely as a starting point in our examination of how different components—selection, learning machinery, decision—might interact. Our goal is not to model performance in these tasks but rather to understand how the three psychological components inherent to statistical word–referent learning influence learning performance. We then report a series of simulations that vary selection in the same way for each model. Next we consider different decision processes, then the flexibility of the models to simulate each other's performance. The results provide new insights on the models and on cross-situational learning and raise a new agenda for research. Overall, the results suggest that it will not be easy to distinguish one mechanism from the other without knowing much more than we do about processes of information selection and decision and how these processes may change across tasks, across trials in the task, and across development.

Two Models and the Tasks

Hypothesis-Testing Model

This model attempts to capture the fundamental principle of hypothesis testing and investigates how this basic learning mechanism operates on cross-situational statistics. Because the goal is to compare this model in various versions to associative-learning models, the model is intentionally simple, with a clean structure and several basic components. More specifically, this first version of an HTM is built on the following assumptions: There is no information at the beginning of learning to guide learners; thus on the first trial, learners are assumed to randomly select word–referent pairs as their initial hypotheses; as more trials ensue, these initial hypotheses, are gradually justified or replaced. Following these general principles, we need to specify (a) how many hypothesized pairs are selected and stored from a trial, (b) how learners justify whether a word–object pair is correct, (c) whether learners use the ME constraint to eliminate one hypothesis if two working hypothesized pairs are not compatible, and (d) whether learners treat previously supported (confirmed) hypotheses as learned knowledge and use that knowledge to help the learning of new pairs in subsequent trials. To illustrate how HTM works, we use one of Yu and Smith's (2007) conditions in which learners were presented with four words and four referents on every trial, with no information about which word went with which referent, and in which there were a total of 18 words and referents to be learned from these highly ambiguous training trials.

In this 4×4 training condition, the 18 novel word–picture pairs can be represented as $\{(p_1, w_1), (p_2, w_2), \ldots (p_{12}, w_{18})\}$. On the ith trial, the stimuli consist of four visual objects and four spoken words $T_i = \{p_{i_1}, p_{i_2}, p_{i_3}, p_{i_4}, w_{i_1}, w_{i_2}, w_{i_3}, w_{i_4}\}$ while i_1, i_2, i_3, and i_4 can be selected from 1 to 18. Assume that the learner maintains a list of hypothesized pairings, the learned results from previous trials. Thus, the learner's current knowledge can be represented as a list of pairs $M = \{(p_{n_1}, w_{m_1}), (p_{n_2}, w_{m_2}), \ldots, (p_{n_k}, w_{m_k})\}$, while n_j and m_j can be selected separately from 1 to 18, and the equivalence of these two indicates a correct pairing. Consider first the case in which HTM forms and tests one hypothesis per trial, randomly picking one word and one picture from a trial to build a hypothesized pairing. Each trial potentially adds hypotheses. Two additional mechanisms are utilized in this version of HTM to make this learning process more effective. First, we implement ME in adding new pairs so as to maintain the consistency of hypothesized word–referent pairings, that is, one word can be associated with only one picture. Without this, the simulated hypothesis tester could randomly select many conflicting (and therefore incorrect) word–picture pairs across multiple trials. Second, the model evaluates a word–object hypothesis in its list if the word or the object appears in the present trial and is selected by the model. More specifically, if the same word–object pairing occurs again, this information if selected will serve as supporting evidence for the hypothesis under consideration, and the confirmed hypothesis will be treated as learned knowledge. Moreover, this knowledge will be used to filter out

the input in subsequent trials; this also 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 the within-trial ambiguity from a 4 (words) $\times$ 4 (referents) trial with 16 candidate word–referent pairs—in terms of the unlearned information—to a 3 (words) $\times$ 3 (referents) trial and thus only nine word–referent pairs to be selected.

In the first version of HTM, the model on each trial randomly selects one word and one object to form a hypothesis about a word–referent pair, leading to three possible outcomes on that trial: (a) If neither the word nor the object occurs in the list of hypothesized pairs, then the model simply adds this new pair into the list because this new pair is compatible with others in the current list. (b) If the new pair selected from the current trial conflicts with an existing pair, the model will check whether the word and the object in the existing pair co-occur in the current trial. If it does, then the model will keep and confirm the existing pair (as learned knowledge) because the pair has co-occurred twice (once before and once in the current trial). Meanwhile, to avoid conflicting hypotheses, the model will discard the newly selected pair. If the existing pair is not in the current trial, then that pair definitely is not a correct pair. As a result, the model would disconfirm and remove the existing pair and add the new pair into the list as a new hypothesis to be evaluated later. (c) At the later part of the training phase, the model may detect a set of word–object pairs that are likely to be correct. Then, it will treat those pairs as learned and use them to first filter out the input in a new trial before randomly selecting a new pair from the trial. With more and more accumulatively acquired knowledge trial by trial, the learning task is simplified.

At test, the model, like the adults in the Yu and Smith (2007) experiments, is presented with one word and four choice objects, with the task of selecting the object that is the referent of the word. Given the hypothesis list built during training, the model will use one of two possible decisions for each tested word: (a) First, if the tested word is in the current hypothesis list, then the model selects the corresponding object in the hypothesis pair that may be correct or incorrect, or (b) if the tested word is not in the list, the model will first exclude those objects that are hypothesized to be linked to other words (a form of ME during testing) and then randomly select an object from the

remaining options. In this way, the model maintains a coherent set of hypothesized pairings and generates one answer for each testing trial.

To clarify these mechanisms, Figure 2a presents a trial-by-trial working version of the HTM considering just two presented words and two presented referents per trial. In Trial 1, the model randomly selects a pair A-b, and another pair D-d is selected in Trial 2. Note that the model does not know yet whether those two pairs in the current hypothesized list are correct. In Trial 3, the model selects A-a and registers that it is not compatible with A-b in the current list, and thus, it has to select one and discard the other to follow the ME constraint. To determine which one should be included, the model checks the current trial and notices that A-b is not in the current trial and therefore is less likely to be the correct one; therefore, the model decides to replace A-b with A-a. Meanwhile, the model may also select and store C-c (the remaining word and referent in this trial) if the model has the capacity to process more information. In Trial 4, the model gets a chance to confirm that D-d is correct as it appears a second time in training, and furthermore, it may also use this information to bind B with b (the leftover items). At this point, the model has a confirmed pair (D-d) that can be treated as prior knowledge; a hypothesis pair (A-a) requires further confirmation, as do two more pairs (B-b, C-c, etc.) if the model has the capability to keep track of this many. As more trials are presented, this model will gradually converge to confirm correct pairs and exclude incorrect ones. Given the knowledge acquired at this point, the model should be able to correctly identify the confirmed pairs (e.g., D-d) with confidence and maybe also perform above chance on the yet-to-be-confirmed pairs (e.g., A-a, C-c). Therefore, even with minimal exposure to statistical information (four trials), HTM should demonstrate learning.

Dumb Associative Model (DAM)

One major assumption in the HTM used above is that it applies the ME constraint on each trial. This limits within-trial ambiguity. In contrast to this type of explicit one-word-one-referent learning, an alternative mechanism would be to accumulate evidence across multiple conflicting associations across trials. During testing, such
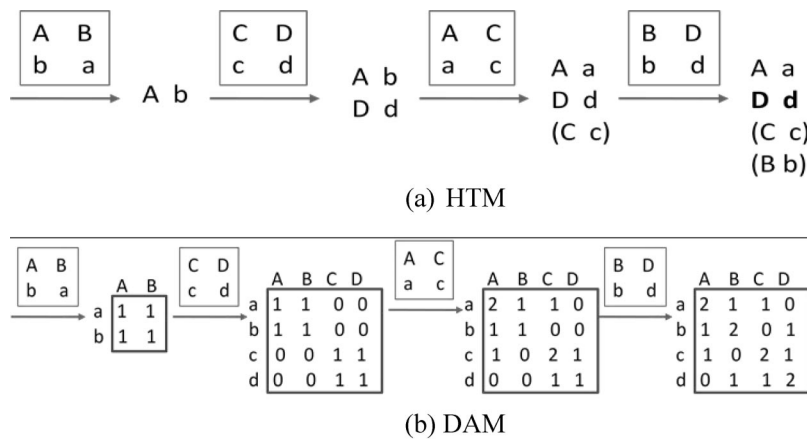

(a) HTM

(b) DAM

Figure 2. A toy example of cross-situational learning to illustrate how hypothesis-testing models (HTMs) and dumb associative models (DAMs) work with multiple learning trials. Panel a: HTM maintains and evaluates a hypothesis list. Panel b: DAM counts co-occurring statistics and stores the information in a matrix.

an associative model might simply pick out the object most strongly associated with the test word. 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 memorizing and representing the word–object co-occurrence matrix of input, as shown in Figure 2b. Thus, the amount of information that an associative learner selects and registers per trial becomes a highly relevant question. Although selecting and testing a single word–referent hypothesis seems the natural approach for hypothesis testing, the registering of multiple and potentially conflicting co-occurrences from each trial seems plausible for a human associative learner. Moreover, human learners may well be able to approximate the ideal learner, tacitly storing many, if not all, of the possible associations on a trial and accruing information about all possible pairs across trials. Such an associative-learning mechanism would, in fact, do quite well in cross-situational experimental tasks such as those used in Yu and Smith (2007). However, it is also possible that human learners are more selective associative learners, picking up only some of the available information each trial.

Accordingly, a dumb associative-learning model (DAM) was developed that instantiates the general principle that the system randomly selects pairs on each trial and accumulates word–referent pairs without applying any constraints or inferences in real-time learning. Hence, this dumb associative learner can be viewed as straightforwardly based on Hebbian learning: The connection between a word and an object is increased if the pair co-occurs in a trial. More specifically, this general principle is implemented in the model by updating the strength of a word–object association over the course of learning in the following ways:

$$A(i, j) = \sum_{t=1}^{K} \eta(t) \sum_{n=1}^{N} \delta(w_t^n, i)\delta(o_t^n, j),$$

where $A(i, j)$ represents the strength of the association between word $w(i)$ and object $o(j)$. This strength is updated from Trial 1 to Trial k. Within each trial t, w_t^n and o_t^n indicate the nth word–object association that the model attends to and attempts to learn. δ is the Kronecker delta function, equal to one when both of its arguments are the same and equal to zero otherwise. Thus, the model updates $A(i, j)$ in the tth trial if that pair is one of N associations selected and attended by the model; $\eta(t)$ controls the gain of the strength of associations over the course of learning, which can be used to capture various cognitive factors related to the registration of an association, for example, memory decay. Note that one can develop and incorporate more complicated components consistent

with these general principles in various ways, for instance, by adding competitions between pairs when updating individual associations, by adding an attentional control gate that allows the model to switch to to-be-learned words after some words are well learned, and by specifying different ways of updating associations based on the present learning state of individual pairs, such as highlighting and blocking (Kruschke, 2003; Lotz, Vervliet, & Lachnit, 2009). Here, for clarity of comparison to the simple HTM, we keep the associative model simple, with only bare bones.

To illustrate how the simple DAM works, we use the toy (two words and two referents) example again, as shown Figure 2b. If we assume that the model can keep track of all of the possible word–referent pairs, then the model registers four pairs when it is exposed to Trial 1 {A, B, a, b} by increasing the counts of the four cells {A-a, A-b, B-a, B-b} in the association matrix. When proceeding to Trial 2, four more cells (C-c, C-d, D-c, D-d) receive co-occurring counts. In Trial 3, four cells are updated (A-c, A-a, C-a, C-c) and increased by 1. As a result, both the correct pairs A-a and C-c gain more counts than other incorrect ones. In Trial 4, the other two correct pairs B-b and D-d gain more counts. After training, the model simply counts co-occurrences. By so doing, the model successfully identifies four correct pairs as the association counts of those four are larger than other incorrect ones. This toy example admittedly presents a straightforward association task; still, one can readily imagine how, even with more to-be-learned words, more training trials, and a higher degree of uncertainty within a trial, the DAM might still succeed.

The Tasks

For the simulations that follow, we use the experimental conditions of Yu and Smith (2007) to define a set of different learning tasks. These conditions vary within-trial ambiguity and the number of pairs to be learned. The conditions are labeled by the number of words presented per trial, the number of words to be learned, and the number of repetitions of each correct word–referent pair. So, "4 × 4 18 words/6 repetitions" is a condition in which learners were presented with 18 word pairs to learn, every trial presented four words and four referents with no information about which referent went with which word, and there were, across trials, six repetitions for each word–referent pair (and thus a total of 27 four-word-by-four-referent learning trials). The full set of experimental data used in the following simulations derived from five conditions of the Yu and Smith word–referent learning task and are listed in Table 1. These conditions differed in within-trial ambiguity, two words and two referents, three words and three referents, or three versions of four words and four possible refer-

Table 1
Five Cross-Situational Learning Conditions Used in the Present Simulation Studies

Condition	Number of total words	Number of occurrences per word	Number of trials	Time per trial (s)	Total time (s)	Number of subjects
2 × 2 18 words/6 repetitions	18	6	54	12	324	38
3 × 3 18 words/6 repetitions	18	6	36	9	324	38
4 × 4 9 words/8 repetitions	9	8	18	12	216	38
4 × 4 9 words/12 repetitions	9	12	27	12	324	28
4 × 4 18 words/6 repetitions	18	6	27	12	324	28

ents on each trial. The 2×2 condition yields four possible associations per trial, the 3×3 condition yields nine, and the 4×4 conditions yield the seemingly overwhelming number of 16 word–referent associations per trial. Across conditions, the experiments also manipulated the total number of word–referent pairs to be learned (nine or 18) and the number of repetitions of each word–referent pair (six, eight, or 12). In summary, the 2×2 condition presented two words and two pictures on each trial, and there were 18 word–picture pairs each repeated six times. The 3×3 condition presented three words and three pictures on each trial, and again, there were 18 word–picture pairs in total, each of which co-occurred six times during training. There were three different 4×4 conditions each presenting four words and four pictures on each trial: 4×4 18 words/6 repetitions condition; 4×4 9 words/8 repetitions condition, and 4×4 9 words/12 repetitions condition. In these experiments, after training, subjects were tested in a four-alternative forced-choice test—one testing trial for each word and 18 testing trials in total. For each testing question, subjects heard one word and were asked to select the corresponding picture from four options on the computer screen. Figure 3 (right bars) shows adult learning performance; across conditions that manipulated the ambiguity of the individual trials and the number of word pairs, participants were successful to different degrees: Performance (in terms of percentage of pairs learned) is very good in conditions of moderate uncertainty (e.g., 2×2 and 3×3), but learners still learn more than half of the total number of the pairs even under conditions of considerable uncertainty (4×4), and the number of pairs to be learned and the number of repetitions matter little in the three conditions of high uncertainty.

Information Selection—Amount of Information

The first set of simulations provides information on the models' abilities to do the basic task, that is, to (at least qualitatively) fit the adult data. These simulations also provide evidence on the interaction of learning performance with the amount of the information selected (the number of word–referent pairs, etc.) on each learning trial. We first consider each model with respect to pair selection and human learning and then compare the two classes of models.

HTM

The selection of a single pair per trial (one explicitly tested hypothesis) is consistent with usual intuitions about hypothesis testing; however, this single selection per trial would not seem a core component of the claim about the learning machinery. Indeed, modern HTMs can entertain more than one hypothesis (e.g., Tenenbaum, Griffiths, & Kemp, 2006). Accordingly, to be consistent with the associative models, the first set of HTM simulations randomly selected one word–referent pair per trial, or two, or three. One-pair selection worked as demonstrated in the toy version in Figure 2a. The mechanism when more than one pair is selected per trial functions in a similar way as the version with one pair per trial. The only difference is that on each trial, the simulated learner selects two or three compatible hypotheses (e.g., if A-a is already selected, then B-b or D-c may be selected as a second hypothesis, but not A-c or B-a—the ones containing either A or a). In this way, more hypothesized pairs are added and then evaluated over the course of learning in the same way as the
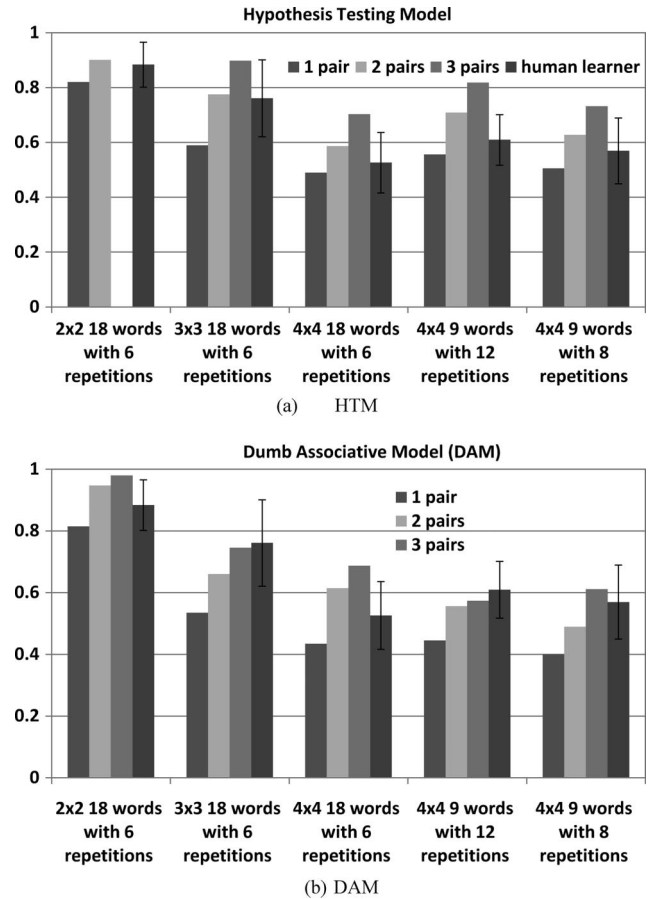

Figure 3. Simulation results from two models compared with human data (the rightmost bars). Panel a: Simulation results of the hypothesis-testing model (HTM) with one, two, and three pairs selected from each trial. Overall, HTM and human learners share the same trend in learning performance. Among three variants of HTM, both the one-pair selection and two-pair selection models can best fit the behavioral data from some of the five learning conditions. Panel b: A comparison between human learners and three different DAM learners. Although the learning performance of one-pair DAM does not fit well with human data compared with two-pair and three-pair versions, the results are still far above chance, suggesting that a simple associative mechanism with limited input data can still acquire some knowledge from cross-situational learning.

version in Figure 2a. More specifically, both the ME constraint and the mechanism for identifying and utilizing confirmed pairs are in operation. In this way, three versions of HTM are used to simulate the five Yu and Smith (2007) conditions (only one- and two-pair selection is possible in the 2×2 training condition). Because the model randomly selects and stores pairs, very different outcomes are possible given different histories of pair selections. However, different from the task of gathering empirical data from human learners, we can easily run enough simulations to obtain solid estimates from simulated learners. Pilot simulations indicated that the variation of the results from HTM (and the associative model) converges on a stable pattern by 1,000 runs. Accordingly, each simulation was run 1,000 times (as 1,000 simulated learners).

Figure 3a shows the simulation results. Relative to human performance, HTM with just one pair selected and processed did

quite well, showing a roughly similar pattern to humans across the five conditions. When selecting two or more pairs, however, HTM overlearned relative to humans in the three 4×4 conditions that present the most within-trial ambiguity. In terms of the absolute level of performance (rather than human-like performance), all variations of HTM did very well in conditions of low ambiguity (2×2 and 3×3) and performed above chance in all conditions. Moreover, HTM showed improvement when more pairs were selected and evaluated per trial. This observation is in line with the general principle of statistical learning: More statistical computations with more statistical information lead to better learning. The one-pair- and two-pair-selection HTMs were relatively unaffected (in terms of proportion of pairs learned) by the number of pairs to be learned (i.e., there are no significant differences between 4×4 18 words/6 repetitions and 4×4 9 words/8 repetitions) and (within the limits of the experiment) by the number of repetitions of each pair.

DAM

Here, we use the simplest version of DAM by keeping $\eta(t)$ constant and then, similar to the simulations with HTM, manipulate the number of pairs—one, two, or three—that are selected in each trial. During testing, the model accesses the association matrix built during training and selects the object that has the strongest association with the tested word i ($max_iA(i, j)$, etc.). Again, each simulation was run 1,000 times as 1,000 simulated statistical learners. Overall, as shown in Figure 3b, the simulated learners did quite well in the forced-choice test. DAM is a really simple model, yet it managed to learn even given highly ambiguous data. The two- and three-pair versions, but not the one-pair version, compare favorably to HTM and to the human data. However, to match what HTM accomplishes with one randomly selected pair per trial, DAM requires broader (less focused) attention and the registration of more information (pairs, etc.) in each trial.

The success of DAM might be considered surprising. Examining examples of association matrices built by one-pair and two-pair learners, respectively, shows those matrices consist of many zero cells and just a few nonzero entries. Nonetheless, in the forced-choice tests, DAM demonstrated learning based on partial and incomplete matrices. This makes the simple but important point that within a matrix of knowledge, choosing the best alternative is often good enough, even given a very sparse matrix in which most cells are zeros and thus may seem to contain no useful information. In brief, DAM—a very simple just-co-occurrence-counting model—compares reasonably well with HTM and is able to approximate HTM's more sophisticated data evaluation by picking up and storing more information per trial. This is a fundamental difference between the two, and thus, the question of how much information learners pick up per trial would seem to be a key (but not as yet well investigated; see also Medina, Snedeker, Trueswell, & Gleitman, 2011; Sumarga & Namy, 2010) empirical question to distinguishing hypothesis-testing and associative models of cross-situational learning.

The perhaps key point is that although co-occurrence matrices seem a dumb form of knowledge (Keil, 1992), they contain much latent and usable structure that is realized in the decision process. As we demonstrate in the decision component simulations, choice at test can be related to multiple associations within the matrix

because the association matrix itself is a system of associations between words and referents (also see Yoshida & Smith, 2005; Yu, 2008). In the present case, the decision—given a word and four choices—depends on selecting the referent with the highest association probability with the word. Yet, even here, there are relations within the matrix that may be captured in that seemingly simple decision. More generally, a single word–referent pairing is correlated with all the other pairings that share the same word (in the same row with that word) and all the other pairings that share the same referent (in the same column with that referent), which are in turn correlated with more word–referent pairs—the whole system of them. Imagine, in an association matrix, when a word is used to query the corresponding referent, many cells in the row sharing that word might be activated or retrieved, and those activations would further light up other cells as well. In this way, this initial query can be propagated to many cells in the association matrix, and each decision captures to a degree the joint information from all associations in the matrix. In brief, these sparse matrices can be viewed as latent knowledge about the lexical system as a whole, latent knowledge that plays out as a system at time of test.

Conclusion

Also evident in Figure 3, if we deliberately choose the number of pairs selected in each model to best fit individual learning conditions, then by adjusting this one single parameter, both of the models can be made to fit the learning conditions almost perfectly. This emphasizes, again, the importance of the empirical question of just how much information a learner picks up on each trial. This observation lies in the fact that even with the same learning machinery, simply manipulating the amount of the information fed to the learning device can dramatically change the learning performance. Information selection gives the models flexibility to simulate a range of learning results.

HTMs typically are thought of as focusing on a single hypothesis at a time and thus just being highly selective but principled; in contrast, associative models are often thought of as being unselective and unprincipled in their selection (although filtering effects are common in sophisticated associative models; e.g., Kruschke, 2001; Mackintosh, 1975). However, the question of how much information is picked up or how much past learning filters new learning is, strictly speaking, not a core distinction between hypothesis testing and associative learning since both kinds of models can be made both with and without those add-on components. We show this in the next section, as we examine the role of different principles of selection on the two types of models.

Information Selection—Principled Selection Across Trials

To investigate how information selection across trials may interact with associative learning versus hypothesis testing, we examined three broad ways in which learning from previous pairs could affect pair selection on subsequent trials: (a) no influence—that is, random selection as in the previous simulations; (b) a bias for familiarity (narrow selection)—the learner pays more attention to words and referents to which it has paid attention before, building a small set of correct word–referent pairings before moving on to other pairs; and (c) a bias for novelty (broad selection)—

the learner accumulates nonoverlapping word–referent pairs from one trial to the next by being biased to select novel words and novel referents in a new trial. We examined these cross-trial dependencies in DAM and HTM with one-pair selection per trial because one-pair selection leads to the greatest differences between the models.

In HTM, a list of nonconflicting hypotheses is maintained such that each word or referent occurs at most once in the list. Under the familiarity bias, the system will try to select new pairs that maximize the overlap in the current list; thus, a learner will tend to select a pair with one item (either the word or the referent) that is already in the current list. By doing so, this pair selection provides immediate confirming or disconfirming evidence to the existing hypotheses that might enable HTM to quickly learn those pairs. For the novelty bias, the learner is biased to select pairs that are not in the current list, more broadly gaining knowledge about a variety of pairs but weaker knowledge about any particular one. In DAM, the implementation of the familiarity bias is to pay more attention to a pair with a stronger association up to the threshold at which it is considered learned and then to select other pairs (which, by the way, is also the pattern that characterizes infant preferential looking—from a bias to familiar objects early in learning to a bias for novel ones later; Hunter & Ames, 1988; Roder, Bushnell, & Sasseville, 2000). Under the novelty bias, the learner pays more attention to novel words and novel referents from the beginning.

Figure 4 shows the results: HTM and DAM show qualitatively different effects of the different selection strategies in the 3×3 and 4×4 conditions. Overall, HTM, again, does much better than DAM, particularly given fewer word–referent pairs to learn (the

two 4×4 9 words conditions). Moreover, under conditions of higher uncertainty, the familiarity bias results in the best learning performance for HTM, and the novelty bias results in the poorest (e.g., in 4×4 18 words/6 repetitions, $M_{\text{HTM\_familiarity}} = 0.62$, $M_{\text{HTM\_novelty}} = 0.38$). A novelty bias is not well suited to hypothesis testing as it encourages the selection of pairs not in the list and, consequently, most pairs in the list do not get opportunities to be well evaluated by the end of training. In this sense, under a novelty bias, HTM moves toward functioning like dumb associative learning, basically just counting word–referent co-occurrences without any additional evidence that can be used to evaluate those hypotheses. Notice, therefore, that a human learner with hypothesis-testing machinery but a novelty bias might look like an associative learner. Indeed, HTM with a novelty bias and DAM with a novelty bias do not differ (e.g., in 4×4 18 words/6 repetitions, $M_{\text{HTM\_novelty}} = 0.38$, $M_{\text{DAM\_novelty}} = 0.39$), supporting the idea that under a novelty bias, HTM and DAM are essentially the same. DAM, in marked contrast to HTM, is entirely unaffected by these selection strategies, doing just as well (or not as well) under a familiarity or novelty bias in all conditions (e.g., in 4×4 18 words/6 repetitions, $M_{\text{DAM\_familiarity}} = 0.40$, $M_{\text{DAM\_novelty}} = 0.39$). It is not, however, that DAM builds the same knowledge under the two biases. Under the familiarity bias, DAM builds a sparse matrix with many zeros, gaining more certainty about a small subset of pairs. At test, DAM with a familiarity bias can easily pick the correct referents of this small set of words but meanwhile has to guess an answer for many other words for which it does not have any knowledge. Under the novelty bias, DAM accumulates more statistics but all with less certainty. However, a more knowledgeable guess at test (with partial statistical knowledge) makes the model perform as well as the one with a small set of more certain pairs. Therefore, the overall performance in the two cases is pretty much the same. Associative learning builds the latent knowledge in a system of co-occurrence data, which in turn yields reasonable and robust (albeit not perfect) performance under different forms of data selection.

In summary, variations in pair selection again show associative learning and hypothesis testing to be different kinds of learning mechanisms. The amount of information—but not so much which information—matters to associative learners as they accumulate over trials the co-occurrence statistics. Performance from this learning machinery is also robust across different learning conditions and is quite good from quite sparse matrices, as the matrix as a whole contains considerable latent information about the underlying structure. The amount of information matters less to the HTM, but the kind of information matters more, although effects of the kind of information are muted by the internal machinery that filters noncompatible information.

To summarize the results so far, the two models do reasonably well and roughly simulate adult learning in the Yu and Smith (2007) tasks. Furthermore, with a specific set of selection parameters, each model can be made to fit human learning in the individual learning conditions. However, no single combination of model and selection strategy provides a particularly good fit across all the learning tasks as a whole. This could accurately reflect the play in the human system, with learners sometimes selecting more or less information per trial (and perhaps in accord with different strategies). Moreover, there are combinations of selection strategies and models (DAM with three-pair selection and HTM with
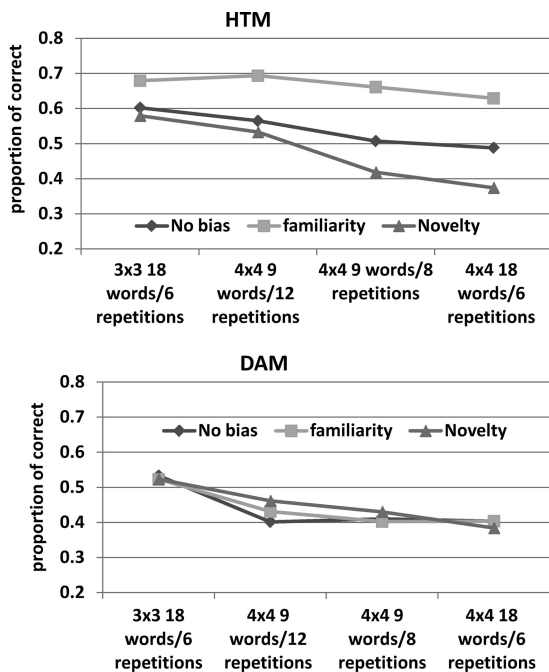

Figure 4. A comparison of pair selection across trials with DAM and HTM. DAM is not sensitive to pair selection, while HTM learns much better with the familiarity principle that allows HTM to concentrate on evaluating the current hypotheses in the list. DAM = dumb associative model; HTM = hypothesis-testing model.

one-pair selection, or DAM and HTM both with a novelty bias) when the two different classes of models produce very similar patterns of learning across various task conditions. Finally, HTM with a novelty bias generates similar results as DAM. All this yields the following preliminary conclusions: (a) HTM and DAM are different models that are more or less successful in different ways; (b) however, the same performance can be achieved by the two models by varying aspects of selection that would seem to be outside of the core claim about hypothesis testing or associative learning; and (c) HTM can yield very different patterns of performance depending on the selection strategy. The consideration of decision-making strategies discussed next yields similar conclusions for DAM.

Decision Making

At test, a statistical learning mechanism must retrieve the knowledge acquired through statistical learning and make decisions about individual words and pairs. The decision-making component in HTM, like most hypothesis-testing theories (e.g., Snedeker, 2009), is straightforward as word–referent pairings are decided during learning. As a result, HTM has explicit knowledge about words and referents. At test, the model just needs to check whether the tested word is correctly linked with the corresponding referent in the final hypothesis list. In contrast, associative models vary more widely as a class since their basic claim, the core of any statistical learner, is that people learn co-occurrences. This basic principle can be combined with different computational processes at learning as well as different decision processes at test. Thus, there are many ways within associative models to, in essence, clean up association matrices—from learning correlations and coherent co-covariation (McClelland & Siegler, 2001; Rogers & McClelland, 2004) to calculating joint probabilities and competitive processes (Fiser & Aslin, 2001). Then, given a stored matrix of co-occurrence data, there are many ways to retrieve and make decisions from those data. This is in direct contrast to HTM and its straightforward decision-making process with almost no ambiguity in the list of hypotheses: The pairs in the hypothesis list are treated as correct, and the ones not in the list are treated as incorrect. In contrast, the learning outcome in associative learning—an associative matrix—retains considerably more information about the statistics in the learning environment, statistics that could be utilized in potentially different ways by different learners or, perhaps, in different ways by the same learner in different tasks.

The next set of simulations offers a limited but (we believe) telling exploration of how different decision-making processes interact with DAM. The maximum-likelihood (ML) method decides word–referent pairs at test using the strongest associations. Given a test word, the model searches through all of the candidate referents and selects the one that has the strongest association with that word. An additional variant is based on ML but makes probabilistic decisions (ML-P) instead of selecting a single strongest association. For example, if the association probabilities of A-a, A-b, and A-c are .55, .30, and .25, respectively, ML will always select A-a, but ML-P will select probabilistically, meaning that every referent has a chance to be selected depending on its association probability. The hypothesis-testing decision method (also shown in the Simulating Other Models section, below) first converts an association matrix into a list of hypothesized pairs and

then uses those pairings to choose answers at test. Thus, whereas a standard HTM forms and evaluates a list of hypotheses trial by trial, DAM with a hypothesis-testing decision strategy at test accumulates statistical co-occurrences first during training and then extracts hypotheses. While both approaches generate a list of hypotheses to be used at test, the difference between the two is when the hypothesis-testing inference is applied. Finally, the ME method implements the ME assumption at test: The learner on each test trial considers not just which word is the most likely associate for the referent being judged but also whether that referent is the most likely associate of other words. The above strategy can go further by examining whether those other words are the most likely associates of other referents and so on. In this way, an association matrix can be treated as a system of associations—not just a list of co-occurrences but a whole lexical network in which one association is connected to many other associations with shared words and referents.[1]

Figure 5 shows the simulation results for DAM with these four approaches to retrieval and decision. The main result and key point is that different decision processes—operating on the same learned information—yield very different patterns of performance. The ME method yields the most correct decisions about word–referent pairs perhaps because this method is most in line with the core idea of associative learning and takes advantage of both accumulating a system of associations during training and utilizing the structure in the whole matrix at time of test. The hypothesis-testing method (which generates the same pattern that HTM does) does not do as well as the ML and ME methods. Unlike these two, the hypothesis-testing method does not make use of all the accumulated data in the association matrix, which is the core strength of associative learning. The ML-P method also may not make the best use of the information in the stored matrices and (given its nature) may be particularly hindered by the greater likelihood of spurious correlations in word–referent pairs in small (in terms of the number of words and referents) data sets. The differences between the different decision processes are nearly uniform across the Yu and Smith (2007) tasks except for the ML-P method, which is due to its greater dependence on the sparseness of the matrix (i.e., on how nonzero cells are distributed in the 18×18 vs. 9×9 matrices).

In summary, DAM is a flexible model (or to put it more negatively, chameleon-like and changeable). Associative learners build co-occurrence matrices; once built, there is a great deal of choice as to what might be done with the information. In this way, DAM can take advantage of information-rich representations in an association matrix and retrieve the accumulated statistical information in various ways. By so doing, associative models gain the flexibility to simulate different kinds of learning results with the same underlying learning machinery. In contrast, the straightforward HTM with an explicit hypothesis list cannot leverage this

---

[1] In implementation, for a test word j, object i is selected if and only if both object i has a higher co-occurrence count (or association probability) than other co-occurring objects $max_A A(i, j)$ and for this selected object i, $j = max_m A(i, m)$, indicating that for object i, j is the most likely word. In the case of $m \neq j$, that is, object i is more likely to be the referent of word m (but not j), object i is excluded from being considered to be the referent of word j, and the model considers the rest if and only if the selected one satisfies the above criteria.
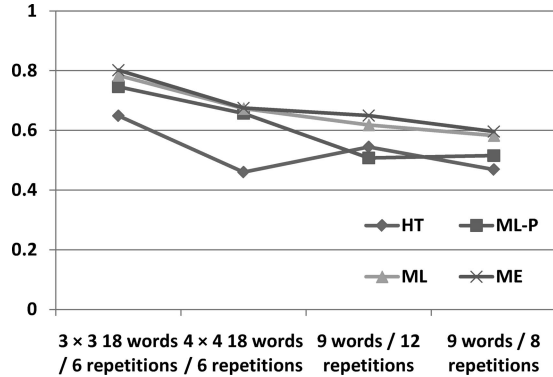
Figure 5. Different ways of retrieving lexical knowledge from accumulated associations can generate different simulated behaviors across four learning conditions, although the learning mechanisms are based on the same associative principle. HT = hypothesis testing; ME = mutual exclusivity; ML = maximum likelihood; ML-P = ML with probabilistic decisions.

flexibility in decision making. This brings us to the next set of simulations, exploring the flexibility of these models to simulate each other.

Simulating Other Models

The tradeoff between amount of data and computation has led many to suggest that hypothesis testing better fits tasks requiring rapid all-or-none learning from minimal data and associative learning better fits slower, graded, and incremental learning (Hollich et al., 2000; Siskind, 1996; Thornton, 2002; Xu & Tenenbaum, 2007; Zwaan & Ross, 2004). The present simulations generally fit this characterization. Nonetheless, we found that similar learning results can be achieved by both models by adjusting specific components in information processing. For example, HTM with a novelty bias in across-trial information selection generates similar performance as DAM. In addition, as we have already shown, there is a straightforward (and psychologically plausible) way to turn DAM into HTM, that is, to have DAM register and store co-occurrences but then to test explicit hypotheses at time of test. In the previous versions of DAM, the model selected the strongest associated referent (the strongest cell, as shown in Figure 7A) given the pairs being tested. However, when the hypothesis-testing decision rule is used from the accumulated association matrix, DAM demonstrates patterns just like HTM. Furthermore, as shown in Figure 6, if we adjust just one more parameter in DAM—the number of pairs selected within a trial (yielding better data— therefore better learning—in the stored association matrix), DAM can match HTM in terms of both the overall performance and the trend across the learning tasks.

Now this might seem sleight of hand: HTM encounters data and derives a hypothesis list, and the just-described version of DAM encounters data, acquires an internally represented sampling of that data in the form of an association matrix, and then derives a hypothesis list. Have we not just made DAM a hypothesis-testing mechanism? In a way, we have. The only difference between the hypothesis-testing version of DAM and HTM is when hypothesis-testing computations (that exclude some statistical information)

happen in a learning mechanism, during trial-by-trial learning or at test. This is an approach taken by some hypothesis-testing theories; for example, the Frank et al. (2009) model first builds a co-occurrence matrix and then generates and evaluates hypotheses from that matrix.

Yet there is a deeper importance than modeling convenience to the question of whether people store co-occurrence data. The question of what is learned and stored versus how and when decisions are made in specific judgment tasks is a critical psychological question. A DAM learner who amasses co-occurrence data but makes smart hypothesis-testing computations at test could also show more graded judgments in other contexts, in other tasks, with the same learned information. Furthermore, DAM learners who just keep aggregating co-occurrences will (eventually) converge on the real-world statistics (and potentially discover higher order and latent regularities in the data). Hypothesis testing is powerful because it filters information out. In the long run, dumb associative learning, by not throwing data out, might prove more likely to come up with right regularities in the end. This is a point perhaps more important for explanations of real-world learning, such as early vocabulary development, given the massive amounts of relevant information in the learning environment, than for explanations of experimental task performance. Moreover, some developmental theorists have suggested a general transition from more associationist-like learning early in word learning to more hypothesis-testing-like learning (Hollich et al., 2000). Knowing where in the psychological system hypothesis testing happens (in the learning mechanism when the data is stored vs. when the system makes in-task decisions from stored data) is critical to understanding these developmental differences.

A general associative model stores a whole association matrix that changes as learning progresses. Therefore, a critical influence on the learning outcome is the information that enters this matrix, what we have called selection. Pieces of information in the matrix may also mutually interact through various forms of competition or augmentation. They might even be the input to some rationalist Bayesian operations. In these ways and through other forms of retrieval and decision processes, the same knowledge and the same association matrix can yield different patterns of performance and different learning outcomes that are more hypothesis-like versus more association-like. As we noted earlier, one advanced HTM, a Bayesian model (Frank et al., 2009), proposes that the learner has acquired an association matrix and tests hypotheses on those learned co-occurrences, which begs the question of whether hypothesis testing and associative learning are indeed mutually exclusive mechanisms of statistical learning.

In principle, one can also do the mimicking the other way around, adjusting HTM to simulate the behaviors generated by associative models. The core mechanism of hypothesis testing is the rational excluding of information. In the present simple model, HTM uses explicit inferences and the ME constraint to maintain a short list of coherent hypotheses that necessarily excludes possible word–referent associations to maintain a consistent hypothesis list. This may get rid of spurious hypotheses, but it also gets rid of data. There are at least two ways to adjust HTM to make it keep more data and therefore be more association-like. First, as already demonstrated, giving HTM a novelty bias during selection so that it collects a little information about a lot of pairs without sufficient information to exclude much information reduces HTM to DAM.
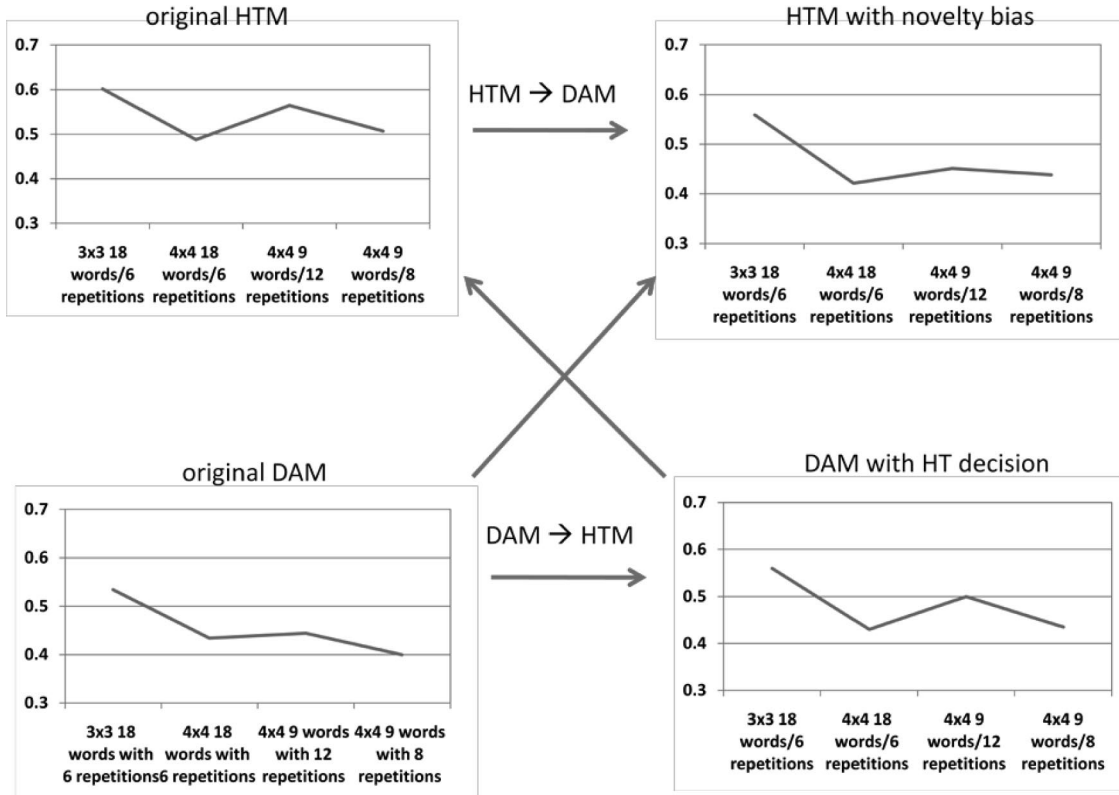
Figure 6. Top: From HTM to DAM. HTM with a novelty bias generates similar learning results as those from DAM. Bottom: From DAM to HTM. DAM accumulates co-occurring statistics from training trials and then extracts a hypothesis list by selecting strongest associations in the association matrix. The decisions at test are based on the extracted hypothesis list. HT = hypothesis testing; HTM = hypothesis-testing model; DAM = dumb associative model.

Another variant of HTM that would mimic DAM would be if HTM kept track of multiple hypothesis sets (Xu & Tanenbaum, 2007). By doing so, this type of HTM would accumulate multiple systems of word–referent associations, and while the associations within any set would be ME compliant, they would not be ME compliant across different sets of hypotheses. From this perspective, this mechanism is similar to associative learning because multiple co-occurring word–referent pairs—including competing ones—are memorized and evaluated.

Certainly, the very nature of knowledge, what is learned in the end, seems to be radically different in associative-learning and hypothesis-testing accounts. For example, DAM builds a big two-dimensional matrix that counts all experienced co-occurrences between words and objects; in contrast, HTMs usually learn a short list of word–referent pairings, although probabilistic Bayesian accounts mimic associative accounts in retaining many hypotheses, each associated with a probability (Thornton, 2002; Xu & Tenenbaum, 2007). There are two ways to quantify these differences in possible representations: (a) the number of word–referent pairs stored (and thus the information that is potentially retrievable, potentially influencing future learning, and relevant to decisions) and (b) as probabilistic versus all-or-none representations. As far as we can discern, there is no reason to believe that people are limited in the number of co-occurrences (or hypotheses) that

they can register (and that can be shown to influence learning performance). If there were a psychological advantage to short hypothesis lists, one might expect there would be some number of learned word–referent pairs beyond which learning would clearly suffer (e.g., some threshold beyond perhaps 7 ± 2). In the Yu and Smith (2007) experiments, adult learners did quite well with 18 pairs to be learned, learning proportionally the same proportion as when there were nine pairs to be learned (and thus actually learning more individual word–referent pairs under the same degree of within-trial uncertainty when learning 18 than when learning nine).

The second quantitative difference concerns the all-or-none versus probabilistic learning of pairs. This is not an essential difference between associative models and HTMs (see also Tenenbaum & Griffiths, 2001). Still, traditional conceptualizations of hypothesis testing (and propositional knowledge more generally) take a winner-take-all approach to what is known versus not known, that is, a word refers to an object or it does not. In the context of cross-situational learning of word–referent pairs, this means that the pairs in the list are treated as all equally correct, whereas the pairs not in the list are excluded from consideration. Simple associative-learning models, like DAM, in contrast, store the information in a probabilistic and graded way. Every co-occurring word–referent pair is assigned to an association proba-

bility based on co-occurrence frequency such that some pairs have high probabilities and others are assigned with low probabilities.

Characterizing the differences between hypothesis-testing and associative models in this way, however, reveals how associative models can be converted into HTMs and how HTMs can be converted into associative models. That is, 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, a sparse and binary association matrix as shown in Figure 7b. Similarly, even if lexical knowledge is stored in a probabilistic mode in an association matrix, the associative learner will need to make decisions at test. Processes at decision could force the retrieval of only the strongest relevant associations, effectively leading to a binary and sparse decision matrix. This might be understood as extracting a hypothesis set from an association matrix and thereby converting probabilistic associations into explicit hypotheses, as shown in Figure 7a. In doing so, different thresholds used in the conversion may vary to determine the number of pairs in the hypothesis set. Moreover, multiple compatible hypothesis lists can be extracted in parallel from the same association matrix to form more than one hypothesis set that can better represent word–referent pairing information in a more probabilistic way. Indeed, with some deliberation, a set of hypothesis lists can store the same information as in an association matrix.

Figure 7. The association matrix representation in a dumb associative model and the hypothesis list in a hypothesis-testing model seem to be different, but these two representations are exchangeable. Panel a: From association matrix to hypothesis list: By selecting a set of strongest associations, an associative learner can build a hypothesis set from the association matrix. Similarly, an association matrix can be decomposed into several hypothesis sets, each of which forms a single and coherent list. Panel b: A hypothesis set can be converted into an association matrix—a sparse matrix wherein most cells are zeros and a very few of them are ones.

Although the representational forms posited by associative and hypothesis-testing accounts seem fundamentally different and although simple models instantiating these core ideas work differently and are affected in different ways by processes of information selection and decision, it is remarkably easy—and in ways that seem not central to the different core ideas—to get one model to mimic the other through what seem like straightforward assumptions and thereafter literally turn one model into the other. One begins to wonder whether these two kinds of learning mechanisms are truly fundamentally different (see also Mitchell, De Houwer, & Lovibond, 2009; Shank, 2010; Thornton, 2002) or whether they reflect some large system of statistical learning that learns co-occurrences but can self-organize to yield different knowledge and learning outcomes via different selection and decision processes.

All statistical learners, whatever their other properties, begin with co-occurrence data. Considerable evidence about human learning (Griffiths & Mitchell, 2008; Hogarth, Dickinson, Austin, Brown, & Duka, 2008; Kruschke et al., 2005; Yu & Smith, 2007) also suggests that whatever else human learners might also do, they register co-occurrence data. All statistical learners are also more likely to link two events, say, a word and a referent, that co-occur more often relative to less frequent co-occurrence events. Thus, the two classes of models might be considered to differ in the mere details of how co-occurrence data are used, that is, differing in the operations that are performed on the data or in the interactions within those co-occurrence data. The problem for theorists as revealed in the present simulations is that those details include selection and decision as well as the learning machinery itself.

Neither performing well nor being able to emulate another kind of learning mechanism is the proper metric for judging whether any specific model or class of models provides an appropriate explanation of human learning. What the simulations make clear is that we cannot judge which class of models best describes human performance without knowing more about the three separable steps of information selection, learning machinery and storage, and the retrieval and use of that information to formulate a response in some test task. Put another way, given what we already know, we cannot make much more progress by comparing some associative model as a whole to some HTM as a whole in some demonstration that such a model can do cross-situational learning or in some comparison to the overall learning performance of humans from a limited number of experimental tasks. Instead, we need to understand what aspects of the model (from information selection to core machinery to decision) make it work, and we need to constrain those component processes by empirical evidence from humans about those very same component processes.

General Discussion

Associative models or HTMs start with fundamentally different assumptions about learning, the learner, and the knowledge that is the product of learning. The different assumptions underlie much of the research (and many of the debates) about children's early word learning in general (Gleitman, 1990) and cross-situational word learning in particular (Yu & Smith, 2007). Yet the core assumptions that learners test hypotheses or store co-occurrence data cannot stand alone. If the core assumptions are to account for performance in tasks, they must be embedded in other cognitive
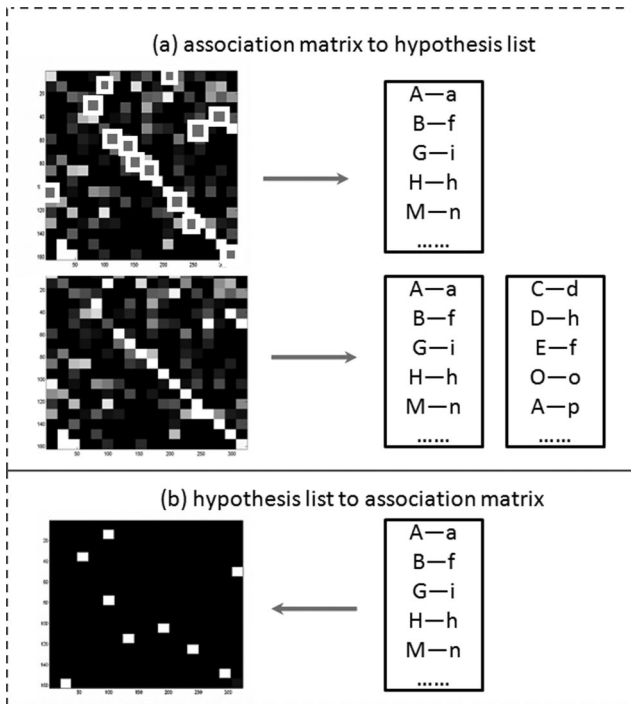
processes—including information selection, retrieval, and decision. The simulations presented in this article show that those other processes matter and that they interact with different learning mechanisms in different ways that can make one model yield very different learning outcomes depending on the kind of selection or decision strategy employed and that can also make the learning outcomes of the two different kinds of models very similar. This, then, is the main contribution and conclusion of the simulations of cross-situational statistical learning: Given what we already know and what has been demonstrated from previous experiments and modeling efforts based on this paradigm, what we can learn is limited by demonstrating that some particular HTM or some particular associative model can do statistical learning (because already many models can) and by comparing the overall performance of some statistical learning device to the overall performance of humans from a limited number of experimental conditions (since there are many ways to get the same overall performance). Instead, we need to look at the components of models to understand how they contribute; we need to measure how those component processes in humans and in models contribute to overall learning. Furthermore, if we want to compare different fundamental claims about the core machinery—hypothesis testing versus association—we need to do so by embedding those core differences in models that are the same in their noncore components, ideally with noncore components constrained by empirical evidence on their operation. Critically, we also need new forms of empirical research on cross-situational learning that move beyond demonstrating that children and adults do this kind of learning to studying the component processes. How selective is the information picked up on a trial? Do children and/or adults store a limited amount of information? Does this depend on both the task and the age of the learner? Do different measures of learning indicate different knowledge, a result that would suggest that learners store co-occurrences but then use that knowledge in different ways in different decision tasks? In sum, we cannot distinguish between these two classes of theories without additional constraints from human experimental data on how information is selected on learning trials, on how that might change over the course of the task, on how it might depend on the specific task, and on how decisions are made at test.

A Focus on Components

If we all agree that all contemporary candidate models are not even close to perfect yet but that they nonetheless can be used to explain various intriguing phenomena, then what we need is to move to the second step in the enterprise—not just model fitting but empirically examining the components of the model itself. This would seem to be a useful first step to understanding exactly how these two classes of theories differ or whether they might be better understood within a single unified theoretical account of the processes involved in selecting, storing, and making decisions about experienced co-occurrences. By a focus on components (and their interactions), we mean a deliberate move away from treating a model as a whole package and a deliberate move away from the main theoretical question (and the main point of modeling) being to show that one grand idea or principle beats another (e.g., Bayesian probabilistic inference against association). Instead, our understanding within each and across the two approaches will

benefit from decomposing the learning mechanism into individual components, then asking the basic questions about how those components work and answering those questions both empirically and computationally. An understanding of those basic building blocks and principles is critical to achieving the larger goals of a unified scientific understanding of human statistical learning.

The present simulations provide a possible first step in this direction by identifying several principles in information selection (e.g., ME or familiarity or novelty across trials), information computation/processing (counting co-occurrence, hypothesis-based logic inferences), decision making at test (winner-take-all or probabilistic), all of which merit study in their own right. In this way, the simulations open empirically answerable questions. For example, given previous empirical results (Kachergis et al., 2009; Yu & Smith, 2007; Yurovsky & Yu, 2008) and the insights from the present modeling, one might ask questions about just what people know about competing pairs. If A co-occurs with a 100% of the time, with b 60% of time (a spurious correlation in terms of reference), but with c 40% of the time, do people—even those who reliably pick A as referring to a—know that A might also refer to b? Does it matter if learners are exposed to A-a first and then A-b and A-c, or the other way around, A-b and A-c first and then A-a? If people maintain only compatible hypotheses at test, can we show that the co-occurrence statistics are still in the system through such measures as savings in learning? One might also ask how much information people are registering per trial and whether it changes with learning, with how much information is presented on a trial, with the level of ambiguity, the task, and the developmental level of the learner. Along this line, one can ask those questions in a broader context. For example, studying adult learners with cognitive deficit or children with atypical development and investigating how their learning may break down under damage and cognitive overload (Leech, Mareschal, & Cooper, 2008; Thomas & Karmiloff-Smith, 2003) can generate informative empirical evidence from complimentary perspectives to help us get a more complete picture. In summary, by empirically answering such questions, the constituent components and steps in model building can be grounded in behavioral data.

A further theoretical question is whether individual components should be treated separately or if they are so closely tied that there is no clear interface between them. As shown in Figure 8, there are different ways to group those basic building blocks together. This by no means requires whole system modeling. Regardless of how it is done, we need to understand the building blocks and how they contribute to the observed outcomes. In brief, we need to study how components function individually and how they are integrated systematically. ME provides a good example of this issue. This constraint goes by a number of names in the developmental literature—principle of contrast (Clark, 1990), the novel-name-nameless-category principle (Golinkoff et al., 1992), the pragmatic account (e.g., Diesendruck & Markson, 2001), and the logic-based explanation of disjunctive syllogism (Halberda, 2006)—which reflect the varying assumptions about the different kinds of mechanisms that might underlie the phenomenon. In Markman and Wachtel's (1988) classic experiment, a child is presented with a known object (ball) and an unknown object (gyroscope) and is asked by the experimenter to bring the "toma." Having never heard *toma* before, the child will select the novel object gyroscope as its referent, but not the object ball for which the child already has a
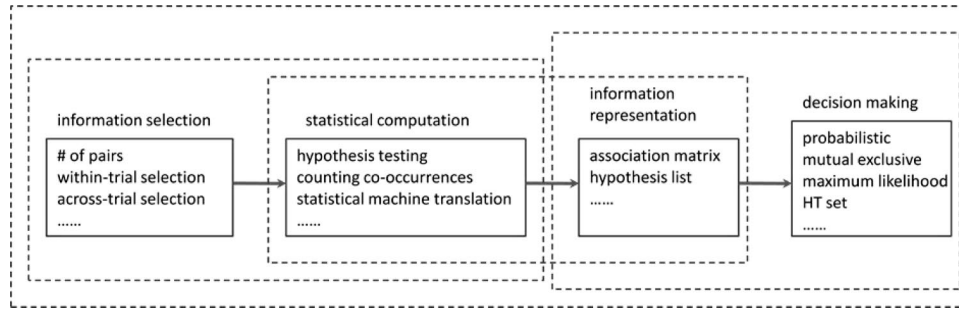
Figure 8. There are several fundamental building blocks in a model, including information selection, statistical computation, information representation, and decision making. With each component, we investigate different solutions/ways to process, store, and retrieve statistical knowledge. Moreover, some of those components may not be truly separable. Thus, two components may interact so closely so that they should be treated as one instead (as illustrated by dot boxes). Therefore, we advocate more empirical and computational studies to specify the component processes within each component and, more importantly, to better understand how those components are integrated. HT = hypothesis testing.

label. The ME constraint is thus a real empirical phenomenon that is also a potent possible learning mechanism in that it reduces uncertainty within the learning moment. Yet where and how in the statistical learning system is ME functioning?

The present simulations show how ME can be understood as operating within different components of the learning mechanism, from information selection to information computation and decision making. The usual and straightforward way of implementing ME is accomplished by HTM: Only those new hypotheses compatible with the ME constraint are allowed to be selected and stored in the current hypothesis list. This real-time implementation of ME in HTM is probably the most explicit and strongest form of ME and also is in line with the macro-level experimental data from children (Markman & Wachtel, 1988). Within an associative model (e.g., Merriman, 1999; Yu, 2008), the ME constraint can also be encoded in a more implicit way by using conditional probabilities in statistical computation. Specially, multiple referents compete for the same word in that the sum of all association probabilities conditioned on that word are constrained to equal to 1. Thus, if one referent obtains a relatively high probability, other referents have to be forced to decrease their conditional probabilities. This competition can lead to a winner-take-all phenomenon in which one referent gains close to 1 probability or it may lead to a distribution across two or more referents. The overall statistical computation dynamically adjusts its solution with the goal to optimize the whole data set. In this way, the implementation through conditional probabilities can also be viewed as a probabilistic (and soft) form of ME compared with the winner-take-all implementation in the HTM. This characterization of ME fits contemporary understanding of competitive processes in lexical access (Levelt, 2001; L. B. Smith et al., 2010). Finally, within even a simple associative model such as DAM, the ME constraint can also be implemented in decision making, restricting the model from selecting the same word twice for different referents at test. Therefore, even if the association matrix—the knowledge the system represents—does not look ME compatible, performance might still be. The present simulations suggest that the ME constraint is needed—somewhere—to simulate human performance. HTM is built upon the ME constraint, and this may be the single

key to its success. Yet ME is easily added to other kinds of models. Accordingly, the key empirical question to be answered is not whether hypothesis-testing or associative models work better as a whole model but, rather, what the processes and mechanisms that implement ME in humans are.

Open Questions About Cross-Situational Word Learning

Experimental research on cross-situational learning is in its early stages, just beginning to move beyond demonstrations. There is a great deal not known that will be critical to both model building and evaluation, and the present simulations suggest several avenues for empirical research. One open question concerns the flexibility within individuals as to the kind of learning that they do. The simulations suggest a tradeoff between amount of information registered and the powerful computations that are done on that information. If a learner gathers lots of data, powerful computations seem less necessary. So, does an individual learner look like a co-occurrence counter with large data sets (say, learning real words in the world) but like a hypothesis tester when faced with limited data in a problem-solving task (say, learning an artificial word from a few examples in the laboratory)? How much do learners move around their learning mechanisms to fit the task at hand? The play in these models with respect to such factors as information selection and decision suggest that there might be similar play in learners' cognitive systems. There is nothing that we know of in the empirical literature on human cognition that suggests otherwise.

It is also possible that there are multiple mechanisms with different selection rules, different computations and kinds of storage, and different decision processes. That is, human statistical learning mechanisms, like many other processes, may not be one mechanism but may be broadly implemented but in somewhat different ways throughout the cognitive and neural systems. If so, research needs focused studies on how the learning system organizes itself across tasks and on how different components might be functionally connected in different ways in different tasks. Likewise, an understanding of statistical learning—and a unified ac-

count of that learning—might benefit from the study of individual differences in these tasks. Instead of treating individual differences as the error in statistical analyses and instead of ignoring the data from participants who fail to learn, one might attempt to use these data to decode the general learning process (McDaniel, Dimperio, Griego, & Busemeyer, 2009; Myung, 2000). The present simulations suggest the value of such an endeavor in that they show that relatively small changes in core processes may result in more dramatic changes in performance. A unified theory that uses interactions among components may give us deeper insights into the prowess—and limitations—of human statistical learning.

The approach to theory building and the challenges advocated here may also help us bridge the gap between macro and micro levels of explanation. At present, although there are many models that easily fit the overall performance in a statistical learning task (Fazly, Alishahi, & Stevenson, 2010; Fontanari, Tikhanoff, Cangelosi, Ilinc, & Perlovsky, 2009; Frank et al., 2009; Siskind, 1996; K. Smith et al., 2006; Yu et al., 2005), there are few models that attempt to simulate more fine-grained micro-level behavioral data. However, recent advances in sensing techniques enable researchers to collect dense, real-time, and multimodal human behavioral data, such as eye-movement data (Hayhoe & Ballard, 2005) and body movements (von Hofsten, 2004). Moreover, psycholinguistic and psychophysics studies already provide compelling evidence at the sensorimotor level of eye movements about real-time competitive processes (Allopenna, Magnuson, & Tanenhaus, 1998; Rehder & Hoffman, 2005; Rehder & Murphy, 2003; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In the context of cross-situational learning, one can collect and analyze the learner's momentary eye-movement data (L. B. Smith & Yu, in press; Yu & Smith, in press). If a subject's looks indicate the information the internal learning device is gathering, then eye-movement data could provide direct insights into the information selected and also into a learner's internal state. Along this line, in the literature of concept learning, a fairly sophisticated set of experimental designs were developed to address issues about whether concepts were being learned gradually or all-or-none and whether people tracked more than one hypothesis trial by trial (Bower, 1961; Levine, 1966; Trabasso & Bower, 1968; Watson, 1968). One way of testing these models was to look at the trial-by-trial learning results and the error profiles. In the context of statistical learning, gathering trial-by-trial decisions may alter the statistical learning strategies that people use otherwise in continuous learning; this is possible because, for participants to make trial-by-trial decisions, they have to explicitly retrieve information trial by trial, and this retrieving process may interfere with the accumulation of statistical evidence. Nowadays, we may be able to work around this problem using measures such as momentary eye tracking, which may provide an uninterrupted way to access the learner's internal state in real time without the interference of ongoing cognitive learning processes, if we can link the learner's external visual attention with the internal ongoing learning process. Recently, the research reported in Yu and Smith (in press) used an eye tracker to record the moment-by-moment eye-movement data of 14-month-old infants in cross-situational statistical learning tasks. A set of gaze patterns in the course of trial-by-trial learning was extracted and used to predict strong and weak learners—those babies demonstrating more successful learning at test and those who were less successful. The results from this fine-grained data analysis shed light on what kind of selective attention may lead to better statistical learning. With advances in the direction of micro-level data analysis, a componential and psychological approach to modeling would seem essential to unifying behavioral data at this micro level with grand claims about the fundamental nature of the learning mechanism.

Conclusion

The tension between associative learning and hypothesis testing is one with a long history in psychology and in theories of computation. Others have suggested that they are fundamentally the same process, that they are not firmly separated, and also that one can make associative learning, though not the same as hypothesis testing, approximate other forms of learning well enough (Bower & Winzenz, 1970; Mackintosh, 1975; Shank, 2010; Thornton, 2002). We cannot resolve these possibilities here, not even within the more restricted bounds of cross-situational word learning. We cannot because we do not know about the psychological processes that select information, that register and aggregate information across trials, and that retrieve and make decisions about that stored information, and because, as we have shown here, these components interact in complex ways that yield very different outcomes. Consequently, given the interactions among these components, there is more than one way to get the same learning outcome. Given what we already know through previous empirical and modeling studies, there is at present, in the domain of cross-situational word–referent learning, no point in making grand claims about the fundamental nature of the learning, in demonstrating that some particular kind of model can do this learning as lots of different models can, and in comparing one model as a representative of its class to another as a representative of its class. These kinds of debates must wait for better data and models focused on all the steps and their interactions that matter to learning mechanisms. The empirical and theoretical approach advocated here—a focus on comparable and well-specified components in both experiments and models—may provide a rich test bed for understanding just what the fundamental differences are between hypothesis testing and associative learning and whether they can be viewed as different instances of the same learning machinery.

References

Akhtar, N., & Montague, L. (1999). Early lexical acquisition: The role of cross-situational learning. *First Language, 19,* 347–358. doi:10.1177/014272379901905703

Allopenna, P., Magnuson, J., & Tanenhaus, M. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38,* 419–439. doi:10.1006/jmla.1997.2558

Anderson, J. A. (1970). Two models for memory organization using interacting traces. *Mathematical Biosciences, 8,* 137–160. doi:10.1016/0025-5564(70)90147-1

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review, 85,* 249–277. doi:10.1037/0033-295X.85.4.249

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory.* Washington, DC: Hemisphere.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of

conditional probability statistics by 8-month-old infants. *Psychological Science, 9,* 321–324. doi:10.1111/1467-9280.00063

Baum, E. B. (1989). A proposal for more powerful learning algorithms. *Neural Computation, 1,* 201–207. doi:10.1162/neco.1989.1.2.201

Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 458–475. doi:10.1037/0278-7393.22.2.458

Bloom, P. (2000). *How children learn the meaning of words.* Cambridge, MA: MIT Press.

Bower, G. H. (1961). Application of a model to paired-associate learning. *Psychometrika, 26,* 255–280. doi:10.1007/BF02289796

Bower, G. H., & Winzenz, D. (1970). Comparison of associative learning strategies. *Psychonomic Science, 20,* 119–120.

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). Cambridge, MA: MIT Press.

Chun, M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science, 10,* 360–365. doi:10.1111/1467-9280.00168

Clark, E. V. (1990). On the pragmatics of contrast. *Journal of Child Language, 17,* 417–431.

Colunga, E., & Smith, L. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review, 112,* 347–382. doi:10.1037/0033-295X.112.2.347

Colunga, E., Smith, L., & Gasser, M. (2009). Correlation versus prediction in children's word learning: Cross-linguistic evidence and simulations. *Language and Cognition, 1,* 197–217. doi:10.1515/LANGCOG.2009.010

Diesendruck, G., & Markson, L. (2001). Children's avoidance of lexical overlap: A pragmatic account. *Developmental Psychology, 37,* 630–641.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science: A Multidisciplinary Journal, 34,* 1017–1063.

Fiser, J., & Aslin, R. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12,* 499–504. doi:10.1111/1467-9280.00392

Fisher, C., Hall, D., Rakowitz, S., & Gleitman, L. (1994). When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *Lingua, 92,* 333–375.

Fontanari, J. F., Tikhanoff, V., Cangelosi, A., Ilinc, R., & Perlovsky, L. (2009). Cross-situational learning of object-word mapping using neural modeling fields. *Neural Networks, 22,* 579–585. doi:10.1016/j.neunet.2009.06.010

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science, 20,* 578–585. doi:10.1111/j.1467-9280.2009.02335.x

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition, 1,* 3–55. doi:10.1207/s15327817la0101_2

Gleitman, L., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. (2005). Hard words. *Language Learning and Development, 1,* 23–64. doi:10.1207/s15473341lld0101_4

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L., & Wenger, N. (1992). Children and adults use lexical principles to learn new nouns. *Developmental Psychology, 28,* 99–108. doi:10.1037/0012-1649.28.1.99

Golinkoff, R. M., Jacquet, R., Hirsh-Pasek, K., & Nandakumar, R. (1996). Lexical principles may underlie the learning of verbs. *Child Development, 67,* 3101–3119. doi:10.2307/1131769

Golinkoff, R. M., Mervis, C. B., & Hirsh-Pasek, K. (1994). Early object labels: The case for a developmental lexical principles framework. *Journal of Child Language, 21,* 125–155. doi:10.1017/S0305000900008692

Gómez, R., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy, 7,* 183–206. doi:10.1207/s15327078in0702_4

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70,* 109–135. doi:10.1016/S0010-0277(99)00003-7

Goodman, N. (1965). *Fact, fiction, and forecast.* Indianapolis, IN: Bobbs-Merrill.

Griffiths, O., & Mitchell, C. (2008). Selective attention in human associative learning and recognition memory. *Journal of Experimental Psychology: General, 137,* 626–648. doi:10.1037/a0013685

Gupta, P., & MacWhinney, B. (1997). Vocabulary acquisition and verbal short-term memory: Computational and neural bases. *Brain and Language, 59,* 267–333. doi:10.1006/brln.1997.1819

Halberda, J. (2006). Is this a dax which I see before me? Use of the logical argument disjunctive syllogism supports word-learning in children and adults. *Cognitive Psychology, 53,* 310–344. doi:10.1016/j.cogpsych.2006.04.003

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9,* 188–194. doi:10.1016/j.tics.2005.02.009

Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In D. L. Medin (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 39, pp. 163–199). San Diego, CA: Academic Press.

Hogarth, L., Dickinson, A., Austin, A., Brown, C., & Duka, T. (2008). Attention and expectation in human predictive learning: The role of uncertainty. *Quarterly Journal of Experimental Psychology, 61,* 1658–1668. doi:10.1080/17470210701643439

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., . . . Rocroi, C. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development, 65*(3).

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research, 5,* 69–95.

Jordan, M. I. (Ed.). (1998). *Learning in graphical models.* Dordrecht, the Netherlands: Kluwer Academic.

Kachergis, G., Yu, C., & Shiffrin, R. (2009). Temporal contiguity in cross-situational statistical learning. In N. Taatgen, L. van Rijn, J. Nerbonne, & L. Schomaker (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1704–1709). Austin, TX: Cognitive Science Society.

Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory & Cognition, 30,* 823–840. doi:10.3758/BF03195769

Kamin, L. (1967). Attention-like processes in classical conditioning. In M. R. Jones (Ed.), *Miami Symposium on the Prediction of Behavior: Aversive stimulation* (pp. 9–33). Coral Gables: FL: University of Miami Press.

Kaminski, J., Call, J., & Fischer, J. (2004, June 11). Word learning in a domestic dog: Evidence for fast mapping. *Science, 304,* 1682–1683.

Keil, F. (1992). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kirkham, N. Z., Slemmer, J., & Johnson, S. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83,* B35–B42. doi:10.1016/S0010-0277(02)00004-5

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology, 45,* 812–863. doi:10.1006/jmps.2000.1354

Kruschke, J. K. (2003). Attention in learning. *Current Directions in Psychological Science, 12,* 171–175. doi:10.1111/1467-8721.01254

Kruschke, J. K., & Blair, N. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review, 7,* 636–645. doi:10.3758/BF03213001

Kruschke, J. K., Kappenman, E., & Hetrick, W. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 830–845. doi:10.1037/0278-7393.31.5.830

Leech, R., Mareschal, D., & Cooper, R. P. (2008). Analogy as relational priming: A developmental and computational perspective on the origins of a complex cognitive skill. *Behavioral and Brain Sciences, 31,* 357–378. doi:10.1017/S0140525X08004469

Levelt, W. J. M. (2001). Spoken word production: A theory of lexical access. *PNAS: Proceedings of the National Academy of Sciences, 98,* 13464–13471. doi:10.1073/pnas.231459498

Levine, M. (1966). Hypothesis behavior by humans during discrimination learning. *Journal of Experimental Psychology, 71,* 331–338. doi:10.1037/h0023006

Lotz, A., Vervliet, B., & Lachnit, H. (2009). Blocking of conditioned inhibition in human causal learning: No learning about the absence of outcomes. *Experimental Psychology, 56,* 381–385. doi:10.1027/1618-3169.56.6.381

Mackintosh, N. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82,* 276–298. doi:10.1037/h0076778

MacWhinney, B. (1987). The competition model. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 249–308). Hillsdale, NJ: Erlbaum.

Mareschal, D., & Shultz, T. R. (1996). Generative connectionist networks and constructivist cognitive development. *Cognitive Development, 11,* 571–603. doi:10.1016/S0885-2014(96)90018-0

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science: A Multidisciplinary Journal, 14,* 57–77.

Markman, E. M. (1992). Constraints on word learning: Speculations about their nature, origins, and domain specificity. In M. R. Gunnar (Ed.), *The Minnesota Symposium on Child Psychology: Vol. 25. Modularity and constraints in language and cognition* (pp. 59–101). Hillsdale, NJ: Erlbaum.

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

McClelland, J. L., & Siegler, R. (Eds.). (2001). *Mechanisms of cognitive development: Behavioral and neural perspectives*: Mahwah, NJ: Erlbaum.

McDaniel, M. A., Dimperio, E., Griego, J., & Busemeyer, J. (2009). Predicting transfer performance: A comparison of competing function learning models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 173–195. doi:10.1037/a0013982

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

Merriman, W. E. (1999). Competition, attention, and young children's lexical processing. In B. MacWhinney (Ed.), *The emergence of language* (pp. 331–358). Mahwah, NJ: Erlbaum.

Merriman, W. E., Bowman, L., & MacWhinney, B. (1989). The mutual exclusivity bias in children's word learning. *Monographs of the Society for Research in Child Development, 54*(3–4). doi:10.2307/1166130

Mitchell, C. J., De Houwer, J., & Lovibond, P. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32,* 183–198. doi:10.1017/S0140525X09000855

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204. doi:10.1006/jmps.1999.1283

Newport, E. L., & Aslin, R. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48,* 127–162. doi:10.1016/S0010-0285(03)00128-2

Pauli, W. M., & O'Reilly, R. (2008). Attentional control of associative learning: A possible role of the central cholinergic system. *Brain Research, 1202,* 43–53. doi:10.1016/j.brainres.2007.06.097

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences, 10,* 233–238. doi:10.1016/j.tics.2006.03.006

Pinker, S. (1984). *Language learnability and language development.* Cambridge, MA: Harvard University Press.

Plunkett, K., Sinha, C., Moller, M., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. *Connection Science, 4,* 293–312. doi:10.1080/09540099208946620

Quartz, S. R., & Sejnowski, T. J. (1997). The neural basis of cognitive development: A constructivist manifesto. *Behavioral and Brain Sciences, 20,* 537–556. doi:10.1017/S0140525X97001581

Regier, T. (2003). Emergent constraints on word-learning: A computational perspective. *Trends in Cognitive Sciences, 7,* 263–268. doi:10.1016/S1364-6613(03)00108-6

Rehder, B., & Hoffman, A. (2005). Eye tracking and selective attention in category learning. *Cognitive Psychology, 51,* 1–41. doi:10.1016/j.cogpsych.2004.11.001

Rehder, B., & Murphy, G. (2003). A knowledge-resonance (KRES) model of category learning. *Psychonomic Bulletin & Review, 10,* 759–784. doi:10.3758/BF03196543

Rescorla, R., & Wagner, A. (1965). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Roder, B. J., Bushnell, E. W., & Sasseville, A. M. (2000). Infants' preferences for familiarity and novelty during the course of visual processing. *Infancy, 1,* 491–507.

Rogers, T., & McClelland, J. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Saffran, J. R., Aslin, R., & Newport, E. (1996, December 13). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928. doi:10.1126/science.274.5294.1926

Saffran, J. R., & Wilson, D. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy, 4,* 273–284. doi:10.1207/S15327078IN0402_07

Shanks, D. R. (2010). Learning: From association to cognition. *Annual Review of Psychology, 61,* 273–301. doi:10.1146/annurev.psych.093008.100519

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition, 61,* 39–91. doi:10.1016/S0010-0277(96)00728-7

Smith, K., Smith, A., Blythe, R., & Vogt, P. (2006). Cross-situational learning: A mathematical approach. *Lecture Notes in Computer Science, 4211,* 31–44. doi:10.1007/11880172_3

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106,* 1558–1568. doi:10.1016/j.cognition.2007.06.010

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, England: Oxford University Press.

Smith, L. B., Colunga, E., & Yoshida, H. (2010). Knowledge as process: Cued attention and children's novel noun generalizations. *Cognitive Science: A Multidisciplinary Journal, 34,* 1287–1314. doi:10.1111/j.1551-6709.2010.01130.x

Smith, L. B., & Yu, C. (in press). Attention is not enough: Individual differences in statistical word-referent learning in infants. *Language Learning and Development.*

Snedeker, J. (2009). Word learning. In L. R. Squire (Ed.), *Encyclopedia of neuroscience* (pp. 503–508). Amsterdam, the Netherlands: Elsevier. doi:10.1016/B978-008045046-9.01882-9

Sobel, D. M., & Kirkham, N. (2007). Bayes nets and babies: Infants'

developing statistical reasoning abilities and their representation of causal knowledge. *Developmental Science, 10,* 298–306. doi:10.1111/j.1467-7687.2007.00589.x

Steyvers, M., & Tenenbaum, J. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science: A Multidisciplinary Journal, 29,* 41–78.

Sumarga, H. S., & Namy, L. (2010, March). *Sensitivity to statistics and salient cues in cross-situational word learning.* Paper presented at the International Conference on Infant Studies, Baltimore, MD.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995, June 16). Integration of visual and linguistic information in spoken language comprehension. *Science, 268,* 1632–1634. doi:10.1126/science.7777863

Tenenbaum, J., & Griffiths, T. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629–640.

Tenenbaum, J., Griffiths, T., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10,* 309–318. doi:10.1016/j.tics.2006.05.009

Thomas, M. S. C., & Karmiloff-Smith, A. (2003). Modeling language acquisition in atypical phenotypes. *Psychological Review, 110,* 647–682. doi:10.1037/0033-295X.110.4.647

Thornton, C. (2002). *Truth from trash: How learning makes sense.* Cambridge, MA: MIT Press.

Trabasso, T. R., & Bower, G. H. (1968). *Attention in learning: Theory and research.* New York, NY: Wiley.

von Hofsten, C. (2004). An action perspective on motor development. *Trends in Cognitive Sciences, 8,* 266–272. doi:10.1016/j.tics.2004.04.002

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition, 107,* 729–742. doi:10.1016/j.cognition.2007.08.007

Vouloumanos, A., & Werker, J. (in press). Infants' learning of novel words in a stochastic environment. *Developmental Psychobiology.*

Watson, P. C. (1968). On the failure to eliminate hypotheses: A second look. In P. C. Wason & P. N. Johnson-Laird (Eds.), *Thinking and reasoning* (pp. 165–174). Harmondsworth, England: Penguin.

Waxman, S. R., & Booth, A. E. (2001). Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. *Cognitive Psychology, 43,* 217–242. doi:10.1006/cogp.2001.0764

Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences, 13,* 258–263.

Waxman, S. R., & Hall, D. G. (1993). The development of a linkage between count nouns and object categories: Evidence from fifteen- to twenty-one-month-old infants. *Child Development, 64,* 1224–1241. doi:10.2307/1131336

Xu, F., & Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review, 114,* 245–272. doi:10.1037/0033-295X.114.2.245

Yoshida, H., & Smith, L. (2005). Linguistic cues enhance the learning of perceptual cues. *Psychological Science, 16,* 90–95. doi:10.1111/j.0956-7976.2005.00787.x

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Development, 4,* 32–62. doi:10.1080/15475440701739353

Yu, C., Ballard, D., & Aslin, R. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science: A Multidisciplinary Journal, 29,* 961–1005.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18,* 414–420. doi:10.1111/j.1467-9280.2007.01915.x

Yu, C., & Smith, L. (in press). What you learn is what you see: Using eye movements to understand infant cross-situational statistical learning. *Developmental Science.*

Yurovsky, D., & Yu, C. (2008). Mutual exclusivity in cross-situational statistical learning. In B. C. Love, K. McRae & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 715–720). Austin, TX: Cognitive Science Society.

Zwaan, R., & Ross, B. (2004). *The psychology of learning and motivation.* New York, NY: Academic Press.