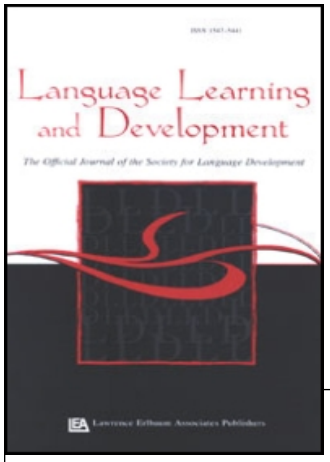

Language Learning and Development
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653671

A Statistical Associative Account of Vocabulary Growth in Early Word Learning

Chen Yu [a]
[a] Department of Psychological and Brain Sciences, and Cognitive Science Program, Indiana University,

Online Publication Date: 01 January 2008

To cite this Article: Yu, Chen (2008) 'A Statistical Associative Account of Vocabulary Growth in Early Word Learning', Language Learning and Development, 4:1, 32 — 62

To link to this article: DOI: 10.1080/15475440701739353
URL: http://dx.doi.org/10.1080/15475440701739353

LANGUAGE LEARNING AND DEVELOPMENT, *4*(1), 32–62, 2008

ISSN: 1547-5441 print / 1547-3341 online
DOI: 10.1080/15475440701739353

Psychology Press
Taylor & Francis Group

A Statistical Associative Account of Vocabulary Growth in Early Word Learning

Chen Yu
Department of Psychological and Brain Sciences, and Cognitive Science Program, Indiana University

There are an infinite number of possible word-to-world pairings. One way children could learn words at an early stage is by computing statistical regularities across different modalities—pairing spoken words with possible referents in the co-occurring extralinguistic environment, collecting a number of such pairs, and then figuring out the common elements. This paper provides computational evidence that such a statistical mechanism is possible for object name learning. Moreover, young children learn words much more effectively and efficiently at later stages. Could statistical learning account for this behavioral change? The current paper explores this question by presenting a developmental model of word learning that relies on a general associative mechanism and recruits previously learned words to guide subsequent word learning. This mechanism leads to increasingly fast learning and corresponding behavioral changes. Simulation studies are conducted using the data collected from a series of picture-book reading episodes wherein parents were asked to narrate books to their 20-month-old children. The results show that previously learned lexical knowledge can narrow the search space and therefore reduce the degree of ambiguity in word-to-world mappings. This results in the bootstrapping of lexical acquisition without changing the underlying statistical learning mechanism. Hence, this work suggests that using lexical knowledge accumulated in prior statistical learning could play an important role in vocabulary growth. To our knowledge, this is the first model that attempts to simulate the effects of cumulative knowledge on subsequent learning using realistic data collected from child-caregiver interactions.

Correspondence should be addressed to Chen Yu, Department of Psychological and Brain Sciences, and Cognitive Science Program, Indiana University, Bloomington, Indiana 47405. E-mail: chenyu@indiana.edu

INTRODUCTION

Learning the meanings of words presents a difficult problem, as illustrated in the following theoretical puzzle (Quine, 1960): Imagine that you are a stranger in a strange land with no knowledge of the language or customs. A native says "Gavagai" while pointing at a rabbit running by in the distance. How can you determine the intended referent? Quine offered this puzzle as an example of reference uncertainty in translation from language to world. Quine argued that, given the novel word "Gavagai" and the object rabbit, there would be an infinite number of possible intended meanings—ranging from the basic level kind of rabbit, to a subordinate/superordinate kind, its color, fur, parts, or activity. Quine's example points up a fundamental problem in first language lexical acquisition—the problem of word-to-world mapping. Granted that children can acquire the concepts of basic-level objects, and granted that they are able to segment continuous speech into a sequence of isolated words, how do they decide which word (among all the words in a spoken utterance) corresponds to each such object (among all the objects in the extralinguistic context)? How do they find the correct word-referent mappings from the many co-occurring but irrelevant ones?

The present article argues that general associative learning mechanisms play a key role in solving the mapping problem. This is not a new idea. A common conjecture is that children map sounds to meanings by seeing an object while hearing an auditory word-form. For example, Smith (2000) argued that word learning is initially a process in which children's attention is captured by the objects or actions that are most salient in their environment, and then those objects or actions are associated with the acoustic patterns produced by an adult[1]. To solve the problem of reference uncertainty, several theorists, including Pinker (1989) and Gleitman (1990), have proposed a solution based on associative learning, which is called "cross-situational learning." The idea is that when a child hears a word, she can hypothesize a set of potential meanings for that word from the nonlinguistic context of the utterance containing that word. After hearing that word in several different utterances, each in a different context, she can intersect the corresponding sets to find those meanings that are consistent across the different occurrences of that word. Presumably, hearing words in enough different situations would enable the child to rule out all incorrect hypotheses and uniquely determine word meanings.

This paper proposes and implements a computational mechanism to show that these kinds of statistical computations can be performed on cross-situational

[1]Waxman and Booth (2000) challenged the domain-general principles of language acquisition. But meanwhile, they agreed with others that word learning involves a learning process to map from words to individual objects.

observations. In fact, the computational power of statistical learning to overcome the word-to-world mapping problem rests on the calculation of such cross-situational statistics—not just tracking, for example, the co-occurrences of "ball" with ball or "cup" with cup, but noting the co-occurrences (and lack of co-occurrences) of "ball" with scenes containing balls and dogs, balls alone, cups, cups and dogs, and so forth. In the first part of this paper, I characterize such a learning mechanism in a computational model to simulate an early stage of word learning. Since Quine's in-determinacy problem requires the statistical learning system (e.g., the model or the child learner) to be constrained somehow in language acquisition, I also incorporate into the model a number of the constraints proposed in the literature, including the constraints that children's first guess about what words refer to is entire objects and that objects have a single label (Markman, 1994).

Bloom (2000) criticized the statistical learning account by arguing that word learning cannot be the product of an associationist process because the associative solution based on cross-situational observations cannot explain either why it is that children make few errors in word learning, or why children start with a slow pace in word learning but then gradually become more efficient word learners. Can statistical associative learning explain both children's accuracy and efficiency in word learning, and also the increasing rate of vocabulary growth in the second year (the so-called vocabulary spurt—Bates, Bretherton, & Snyder, 1988; Ganger & Brent, 2004; Gershkoff-Stowe & Smith, 1997; Goldfield & Reznick, 1990; Gopnik & Meltzoff, 1987)?

This is a theoretically very difficult but important problem. We need a mechanism that is a rapid learner of word-referent mappings but that makes few mistakes. Statistical associative learning seems problematic in this regard for two reasons. First, if language learners associate a word with a referent quickly—that is, based on just a few co-occurrences—then they should make many incorrect associations because there are many irrelevant co-occurring word-referent pairs in natural environments. The fact is, however, that they make such mistakes only rarely (see Huttenlocher & Smiley, 1987). Second, statistical learning relies on inferences over relatively large amounts of data. Thus many learning trials should be required to discover the reliable word-referent pairings, which would seem to rule out learning from a very few exposures.

I suggest that the theoretical resolution to the above criticisms is to imagine a probabilistic (but not winner-take-all) learning system that does not learn single associations between individual words and referents, but that instead learns a system of associations (also see Yoshida & Smith, 2003). In such a system, a single word-referent pairing is correlated with all the other pairings that share the same word and all the other pairings that share the same referent, which are in turn correlated with more word-referent pairs—the whole system of them. Importantly, this paper shows that the acceleration of word learning may be partially due to

the fact that lexical knowledge accumulated over time as latent knowledge of the whole lexical system can be recruited in subsequent word learning. I explore such a cumulative mechanism based on a statistical associative learning model and show that, with more lexical knowledge learned and then recruited for subsequent learning, the same associative learning mechanism can learn words in a more effective way, requiring fewer exposures before a word is learned. Thus, the ability to map words to referents via associative learning increases during learning and therefore can account for a developmental increase in the rate of vocabulary growth.

A number of models have previously been developed to account for different aspects of word learning (Bailey, 1997; Li, Farkas, & MacWhinney, 2004; MacWhinney, 1998; Plunkett, Sinha, Miller, & Strandsby, 1992; Regier, 1996; Roy & Pentland, 2002; Siskind, 1996; Yu, Ballard, & Aslin, 2005). One problem with most simulation work is that the inputs have been artificially constrained: in particular, artificial or synthesized data are used to assess the performance of learning algorithms. For instance, Siskind's (1996) model was tested using utterance-meaning pairs that were generated by a computer program with a set of controllable parameters, such as vocabulary size and conceptual-symbol size. The data were good enough to demonstrate the computational power of the algorithm. Nevertheless, there are several inherent problems with artificial data. First, the results might not be directly comparable to those that would be obtained with data collected from young children. Second, simulation studies based on artificial data always simplify the word-learning problem. Most models (e.g., Li et al., 2004; Plunkett et al., 1992; Tenenbaum & Xu, 2000) have not actually addressed the word-to-referent mapping problem because their input data were already paired in the form of one-to-one correspondences between words and objects. Therefore, the underlying learning mechanisms in those simulation studies may not be directly applicable to real data collected from natural learning environments. In contrast to those previous efforts, one important contribution of the present work is that the model is tested using data collected from child-parent interactions. Consequently, the results reported here can shed light on the kinds of information that are embedded in everyday learning environments in which young children are situated, and on how a computational mechanism can acquire lexical knowledge from those naturalistic environments.

The organization of the paper is as follows: Section on "A Statistical Associative Model" presents a statistical associative model to build word-referent pairings. Experimental results from this model are also reported. Section on "A Cumulative Model" describes the mechanism for utilizing previously acquired lexical knowledge in subsequent word learning. Section on "General Discussion" reports the results of simulation studies. Section "Conclusions" concludes with a general discussion.

A STATISTICAL ASSOCIATIVE MODEL

Rationale and Method

I argue that distributional statistics across words, across objects, and across the co-occurrences of the items in these two streams can jointly determine whether a word-object pair is relevant. In this way, the ability to disambiguate many-to-many co-occurrences and to build one-to-one mappings lies in not only the statistics of a single word, but rather in how this word is distributed across different contexts and what other words surround it in those contexts. As shown in Figure 1, I conceptualize word-to-world mapping as a translation problem—how to cognitively translate words in a language into concepts in the brain. Thus, the learning mechanism I propose rests on advances in machine translation. Briefly, machines "learn" word correspondences (which word in one language corresponds to which word in another language) by extracting statistical regularities across large parallel corpora in two languages. Here, I use a similar computational approach with English itself as one language and the extralinguistic context as the other. This conceptualization provides a unique way to understand statistical learning of word-to-world mappings. Associating referents (objects and actions, etc.) with words (object names and action verbs, etc.) is viewed as the problem of identifying word correspondences between English and the meaning language. With this perspective[2], I develop an associative model based on the translation algorithm

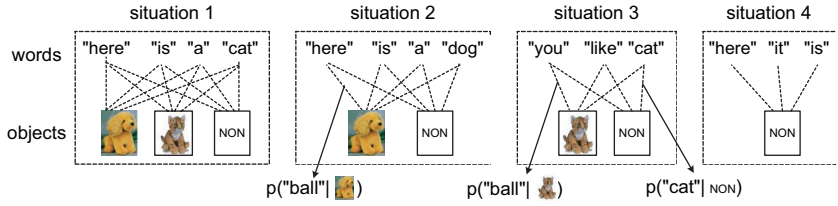

FIGURE 1 Statistical cross-situational learning. On each learning trial, multiple words and objects co-occur, which form several possible word-object pairs. Each link represents the association probability between a word and an object. Also note that we add a referent "NON" as a special item in the extralinguistic contexts because young language learners may fail to attach referents to some words and therefore realize that some words may not have object referents. In this way, high association probabilities to "NON" indicate that the corresponding words (e.g., function words) don't refer to objects because those words are present in so many non-overlapping contexts. In another view, the only shared component across those contexts is NON—not referring to any concrete referent.

[2]Fodor (1975) (Boroditsky, 2001; see also recent work by Bowerman, 1996) suggested a more interactive view of language and conceptual development wherein the structure of the conceptual space does not develop independently, but is shaped by language.

proposed in Brown, Pietra, Pietra, and Mercer (1994). The central idea is that word-object pairs are latent variables underneath the observations that consist of spoken utterances and extralinguistic contexts. Thus, the association probabilities of pairs that represents an object and represents a word, are not directly observable, but they somehow determine the observations because spoken language is produced based on the mother's lexical knowledge. Therefore, the objective of young language learners or computational models is to decode the observation data (e.g., multiple learning trials, each of which contains multiple words and co-occurring objects), and discover the values of these underlying association probabilities, so that language learners or models can better interpret the observations. Correct word-referent pairs are those that maximize the likelihood of the observations.

The learning process can be formalized as an expectation-maximization algorithm (EM) (Dempster, Laird, & Rubin, 1977). The idea of EM is that there is a way of representing the data as the sum of component probability distributions. More specifically, the probability that a word is uttered is expressed as a weighted mixture of the conditional probabilities of the word; given its possible referents in the current non-linguistic context. A simulated learner then tries to find those reliable associations of object names and their referents that maximize the likelihood function of observing the whole data set. The model computes association probabilities of all the pairs simultaneously by considering them as a system of words that interact with and influence each other, and attempts to find the whole set of reliable word-referent pairings across words, across objects, and across multiple situations. More specifically, the model intends to estimate the association probability of every co-occurring word-referent pair in a learning trial. In this way, a word-referent association matrix is built in which the rows represent all the words in the training data, and the columns represent all the referents in the training data. As shown in Figure 2, each cell indicates the association probability of a specific word-referent pair. If a word-referent pairing never co-occurs in any learning trial, the association probability is set to zero. Otherwise, each pairing is considered to be a possible lexical item and its association probability is calculated. Thus, the model searches for an overall optimal solution of those individual association probabilities as a whole—the pattern over the whole association matrix (not just individual pairings) that achieves a better interpretation of the training data across all of the learning situations. This theoretical simulation is fundamentally different from approaches that simply count the co-occurrences of individual pairs and then normalize those co-occurrence frequencies in individual learning situations.

Table 1 demonstrates how the present learning mechanism works compared with simply counting the occurrence frequencies of words and objects. The input to the simulated learner in this toy example is rather simple, consisting of four spoken utterances ("here is a cat," "here is a dog," "you like cat," and "here it is")

FIGURE 2 The row is a sorted list of most frequent words and the column is a list of (a subset of) objects. Each cell is the association probability of a specific word-referent pair. Dark color means low probability while white means high probability. The associative model is able to discover some correct word-object pairings.

and two objects. However, even in this simple example, there are 19 co-occurring word-referent pairs. Most of these are irrelevant—for example, the word "here" co-occurring with the object *cat*, the article "a" with *dog*, and "like" with *cat*.

TABLE 1
Word-Referent Pairs From the Example Shown in Figure 1

	cat	*dog*	*NON*
"cat"	0.511	.115	.375
" dog"	.458	.542	0
"here"	.128	.333	.539
"is"	.128	.333	.539
"a"	.178	.462	.360
"you"	0	.5	.5
"like"	0	.5	.5
"it"	0	0	1.0
"cat"	2	1	2
"dog"	1	1	0
" here"	1	2	3
"is"	1	2	3
"a"	1	2	2
"you"	0	1	1
"like"	0	1	1
"it"	0	0	1

Note. The rows are words and the columns are referents. **Left:** From co-occurrence statistics of words and referents, it is hard to spot relevant word-referent pairings. **Right:** Our model estimates association probabilities of word-object pairs, which tend to give more weights to relevant ones.

Only two pairs yield meaningful lexical items ("cat" with *cat* and "dog" with *dog*). This kind of input poses a special challenge to statistical learning not only because there are few regularities that the model can utilize, but also because the limited regularities in the co-occurrence of words and referents may not be reliable. As shown in the left columns of Table 1, counting co-occurrence frequencies between words and objects clearly cannot separate correct pairs from irrelevant ones. But, even with such sparse data, the statistical associative method I proposed can extract meaningful statistical regularities, as shown in the right columns of Table 1. The association probabilities of correct pairs are relatively high, for example, the association probability of "dog" with *dog*, "cat" with *cat*, and "here" with NON. Conversely, the association probabilities of irrelevant pairs are relatively low.

Such results, from such a small amount of data, demonstrate that this learning mechanism can discover reliable word-referent pairings that are not visible in the matrix to a naive eye. For instance, the word "dog" occurs only once, in the second learning trial (situation), where it has two possible referents (*dog* and NON). There is no way to tell from looking at the occurrence of "dog" alone which

referent "dog" goes with because the co-occurrence statistics are the same for the two possible pairings. Herein lies the power of statistical associative learning. The associative model favors the pair "dog"-*dog* and assigns it with a higher association probability because it considers a system of word-referent associations—what other words are in this trial, what words co-occur with these two referents in other trials, what other referents are in those learning trials, and so on. In this way, the correct associations can be reliably discovered even if there is ambiguity in individual learning trials, and even if some spurious associations exist in the training data. More detailed descriptions can be found in Appendix A.

Participants and Procedure

Unlike previous simulation studies, I presented the proposed model with realistic data—the kind of exposure to words and objects that young children actually receive. The goal was to demonstrate the kinds of statistical regularities that are embedded in everyday naturalistic learning environments, and how a computational mechanism can utilize those regularities to extract meaningful word-referent pairings.

To obtain realistic input data, I recorded mothers narrating picture books to their children. These recordings served as the corpus for our simulations. Nine parents and their young children between the ages of 18 and 23 months ($M = 20.6$; $SD = 3.2$) were recruited to participate. The parents were asked to narrate a set of six picture books for their children. Picture book narration is representative of everyday parent-child interactions, from which children learn the names of objects shown in the picture books. Parents were asked to narrate three picture books during each of two laboratory visits. The total time of interaction was about 10–15 min[3]. The parents held the children on their laps and sat in front of a computer screen. They were instructed to narrate the picture books naturally without any constraints on what they should or should not say. Parents wore a microphone to record their speech. They were offered an optional break before the beginning of each new picture book. Some parents whose children were not strongly engaged in the interaction opted to take the break.

Data

Six picture books for 1–3 year old children were randomly selected with the only criterion being that each book should contain multiple everyday objects. I

[3]In a pilot study, I found that most young children at 18–25 month old can be highly engaged for a period of up to 15 min—not enough time for six book readings. Therefore, I asked parents to visit twice.

scanned each page of each book and saved it as a digital image. Next, I used Adobe Photo-shop to remove all text from the images and to fill the text area with a color that matched the background. Each picture book was shown page by page on the computer screen: parents could move to the next page by pressing the space bar. Spacebar presses precisely recorded the timestamps of page turnings, which were then automatically synchronized with speech by a home-developed computer program. The data used in the simulation studies consisted of two parallel streams—speech and visual context. I first segmented the continuous streams into several learning trials (situations) based on speech silence. Each trial consisted of a spoken utterance and the objects in view at that moment. Next, the entire transcripts of those spoken utterances (not just spoken object names) were coded by experimenters. Meanwhile, each picture book page was coded as a set of basic level objects. Thus, the data used for this simulation study were our descriptions of recorded video-audio clips. The audio input fed into the statistical simulated learner was the entire list of spoken words, not only object names but all the words. The video input fed into the statistical learner was the list of all basic level objects in the picture books. Table 2 shows some examples of the data used in the experiments. For a learner without any prior knowledge, any word in the speech stream could potentially be associated with any object in the visual context. This naturalistic but highly ambiguous learning situation makes word learning a difficult task.

TABLE 2
Examples of Picture Book Narrations Consists of Two Data Streams

Speech …..…	*Visual Context …..…*
what is that	boy, flowers, bird
is that a little baby	boy, flowers, bird
and what is the little baby holding	boy, flowers, bird
that is right flowers	boy, flowers, bird
……	……
that is a pumpkin and look	boy, pumpkin, leave
what is this back there	boy, pumpkin, leave
is that a tree	boy, pumpkin, leave
……	……
look what he is doing now	boy, hat, bird, wall
he is feeding the birdie	boy, hat, bird, wall
and what is this on the ground	boy, hat, bird, wall
……	……

Note. Columns: Speech transcripts and visual contexts. Each row represents one learning situation calculated based on the speaker's speech silence.

TABLE 3
Frequency Statistics of Experimental Data in Six Picture Books

Picture Book		*1*	*2*	*3*	*4*	*5*	*6*
Words	vocabulary (types)	230	226	119	167	290	173
	# of words (tokens)	1039	904	366	460	1047	536
Objects	unique objects (types)	29	30	20	23	38	32
	# of objects (tokens)	515	968	629	435	962	766
Within-Trial	# of trials (situations)	150	186	75	109	192	98
Co-Occurrence	# of words (tokens)	7	5	5	4	6	6
	# of objects (tokens)	3	5	8	4	5	8
Overall Word-Object	# of pairs (types)	2065	1928	1664	2056	3533	2611
Co-Occurrence	% of relevant pairs (types)	1.9%	1.5%	2.5%	1.9%	1.5%	1.7%
	# of pairs (tokens)	4419	3487	2869	4360	6726	5921
	% of relevant pairs (tokens)	5.8%	5.2%	3.7%	4.9%	3.3%	3.1%

Note. I calculated four kinds of statistics: (1) Words, (2) Objects, (3) Within-Trial Co-Occurrences showing the average statistics in a learning trial, and (4) Overall Co-Occurrences showing the statistics in the whole data set. For each category, the statistics are computed based on both types and tokens.

The statistics of the training data shown in Table 3 indicate that, of the entire vocabulary of words uttered by the language teachers, only a small proportion consists of object names. In addition, those object names occur infrequently and irregularly with the corresponding objects. A further analysis of the co-occurrence statistics of word-referent pairs quantifies this observation by showing that co-occurrence frequencies between words and referents cannot be utilized to solve the word-to-world mapping problem. In Table 3, only a very small set of co-occurring pairs are relevant. Thus, if young language learners did compute and keep track of co-occurrence frequencies to build word-to-referent mappings, they would make many wrong associations. The analyses here illustrate the complex learning environments that young children experience. They need to discover correct word-referent pairs in an environment in which most co-occurring events are irrelevant. Considering the smoothness and efficiency of word learning, they must possess a very effective strategy to deal with this reference uncertainty problem. The model in the present study provides an existence proof for just such a computational mechanism.

Results and Discussion

I fed the speech and visual context data obtained in six learning sessions (corresponding to six picture books) to an associative model, which then selected a set

of word-object pairs based on high association probabilities. Two measures were used to evaluate learning performance: (1) Precision (accuracy) measures the percentage of the words spotted by the model that actually are object names, and (2) Recall (completeness) measures the percentage of all of the correct object names available that the model actually learned. Generally, I can tune up the parameters in the model to trade off precision and recall. For instance, I can easily increase recall by discovering more word-object pairs, some of which are relevant and many of which are not. Thus, if I tune for high recall, I may have to accept poor precision. Similarly, high precision can be obtained by just selecting the pairs about which the model is mostly confident, which in turn will reduce the rate of recall. Figure 2 shows the results of statistical associative learning of the data from one picture-book reading session. Many words at the top of the word frequency list are associated with the NON referent because they are function words. However, the model is quite successful with object names: the word "girl" is correctly mapped to the referent *girl*, "flowers" to *flowers* and "bear" to *bear*. Meanwhile, the simulated learner also makes some errors, such as associating "of" and "to" with *bear*, due to spurious correlations. These spurious correlations are real. Because the picture book is about a brown bear, parents said a lot about the appearance of the bear and about its activities. Consequently many words co-occurred more frequently with the referent *bear* than with other referents. With a high precision .9, the completeness of the six learning episodes is 35, 28, 37, 33, 25, and 36%, respectively. In this way, the simulated learner, like young children, does not make large numbers of wrong associations and meanwhile acquire a quite amount of lexical knowledge. Considering that fewer than 3% of co-occurring word-object pairs represents correct namings of objects, the results of this simulation clearly show that the statistical associative mechanism is able to obtain at least some lexical items from highly ambiguous learning contexts. However, the performance of the simulated learner is far from perfect. Specifically, it identifies some statistical regularities that are correct but are not strong enough to distinguish them from other irrelevant pairs. For instance, *mom* is likely to be associated with "mom" and *bed* with "bed" but not significantly so. I term these instances partial lexical knowledge—the knowledge (represented by gray areas in Figure 2) that has not yet been completely learned. The role of learned and partial knowledge in subsequent learning will be discussed below.

The simulation results suggest that we can go beyond demonstrating the mechanism to make new and unexpected predictions, because we know more about the correlations that are obtained in the dataset. For example, the model makes predictions about the naming and comprehension errors that are most likely to occur. As shown in Figure 2, statistical regularities in the learning environment may lead the simulated learner to incorrectly associate the word "things" with the object *boy*. Such predictions are not based merely on phonological or visual similarity or on temporal proximity in the stream of events, but rather on

the correlational blend across all of these. Moreover, this formal model of statistical word learning suggests that in addition to learned words (white cells in Figure 2), the simulated learner also accumulates lots of partial knowledge (gray and dark cells) about all of the word-referent pairs to which it is exposed. Overall, the first simulation study shows that a probabilistic associative learning mechanism can acquire a substantial amount of lexical knowledge (around 30%) from naturalistic and therefore highly ambiguous learning contexts in which fewer than 3% of co-occurring word-object pairs is relevant. Although the results are far from perfect, this basic model provides a probabilistic framework to explore the next important question for this study: can a statistical learning mechanism be more effective if it considers already accumulated knowledge in subsequent learning?

A CUMULATIVE MODEL

The simulations I have described demonstrate how early word-to-world mappings might be accomplished by a relatively simple associative learning mechanism. The success of this model raises an interesting question: given that associative learning is possible from birth or even before (e.g., DeCasper & Fifer, 1980), why do infants not evidence word-learning until late in their first year? The model is predicated on one possible answer to this question, in that it assumes that children must first learn the separate components of word-referent associations before linking them together. That is, children learning their first language must solve three problems: they must (1) segment the speech signal into lexical units, (2) identify potential meanings of words from varied perceptual input, and then (3) associate these potential meanings with lexical units. Infants show little evidence of the ability to segment words from even simple contexts until the second half of the first year (Jusczyk & Aslin, 1995). Work on infant object recognition and categorization suggests that the development of the infant's concept space also takes time (Bowerman & Choi, 2001). Thus, although associative learning is possible from birth, it seems likely that word learning is not possible until the infant has acquired a store of individual words and individual referents to associate with one another.

A test of this proposal is beyond the scope of the present simulations, which focus on the third problem—word-referent mapping. This model uses transcripts of spoken words and coded visual objects as input on the assumption that the learner has solved the first two problems at earlier stages of development. What the simulations clearly demonstrate is how, given speech containing isolable words, and given extralinguistic context streams, the ability to make multimodal correlations across multiple learning situations may be sufficient for acquiring a significant amount of lexical knowledge.

A second important question is how an associative learning mechanism would explain a change in the rate of learning—specifically, the change in the rate of acquisition of new object names that has been described in children in the second half of their second year. One potential explanation is that young children learn how to use previously acquired knowledge to learn new words. With more lexical knowledge available to be recruited in subsequent learning, a cumulative mechanism based on statistical associative learning may be able to learn words in a more effective way, and may require fewer exposures before a word is learned.

Although this idea is intuitively appealing, no computational study has demonstrated how such a learning mechanism would work, and in particular, how it would work with real data collected from everyday parent-child interactions. The next model is an attempt to fill that gap.

Rationale and Method

The following simulations use the same data collected from six picture book narrations as described above. Each book is treated as a single learning episode. The computational model processes the data sequentially, episode by episode, to investigate the effects of the accumulation of lexical knowledge and the utilities of that knowledge in subsequent learning.

As indicated in the first simulation study, a critical issue in statistical word-to-world mappings is the ambiguity caused by temporal co-occurrences of words on the language side and referents on the extralinguistic context side. Intuitively, one way to reduce the degree of ambiguity is to exclude the known items in both the word and the referent streams. I propose that previously acquired lexical knowledge can be used to reduce the search space so that subsequent learning can deal with a simpler learning problem in a smaller search space. More specifically, I propose and implement three mechanisms to utilize previously identified word-object pairs—both those completely learned and those partially learned—in cumulative subsequent learning. My hypothesis is that young children can somehow store lexical knowledge acquired from previous learning episodes and then recruit it in three forms:

(1) a list of **function words**, (2) a list of **learned word-referent pairs**, and (3) **partial knowledge**—a matrix of relatively weak associations between words and objects (e.g., Figure 2). Compared with the statistical learning method introduced in the previous section, the cumulative process here carries out two important additional steps: (1) At the end of each learning episode, the model updates its lexical knowledge by adding more items to both the function-word list and the lexicon list. Meanwhile, the new association matrix is merged with the previously stored matrix to form an updated one. (2) As shown in Figure 3, at the beginning of a new episode, the cumulative model is fed not only the

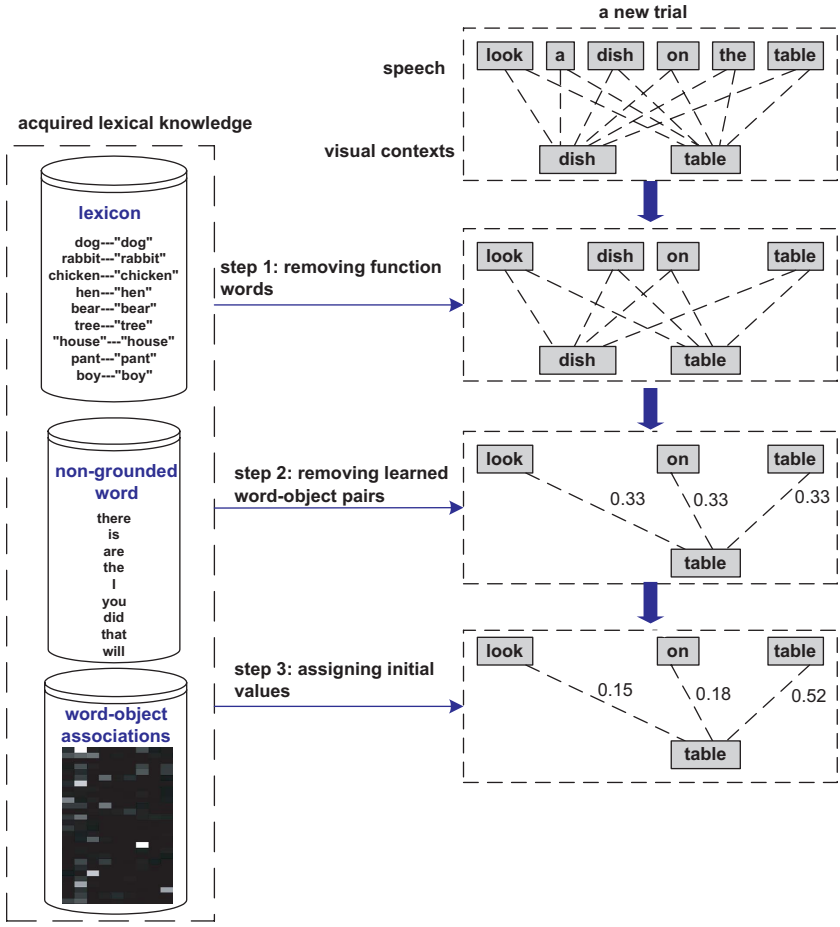

FIGURE 3 **Cumulative Learning. Left:** Based on previous learning episodes, previously exposed knowledge is acquired and stored into three forms: (1) a list of learned word-object pairs, (2) a list of words not referring to concrete objects, (3) an associative matrix with all association probabilities between co-occurring words and objects. **Right:** The current learning trials initially consists of 6 words and 2 referents. Three mechanisms of utilizing previously exposed lexical knowledge are applied to this learning situation: (1) Function words are excluded in the next learning episode; (2) Learned word-object pairs are excluded as well; (3) Initial association probabilities are assigned based on partial lexical knowledge.

utterance-scene pairs in the current episode, but also the lexical knowledge acquired in previous learning episodes. The details about how the model discovers, memorizes, and utilizes these three kinds of lexical knowledge are explained next.

Function Words

The most frequent words in maternal speech are function words that don't refer to things in the world. If the young language learner could develop a filtering mechanism to exclude those words in lexical acquisition, then the statistical mapping problem would be significantly simplified, because there would be no need to consider those words in the following computations. In the present study, I implement a computational mechanism that can identify function words like "is," "the," and "it," because such words are likely to possess one of the two characteristics. First, such function words are likely to have high association probabilities with the "NON" category (as shown in Figure 2), indicating that those words are likely to refer to non-object meanings. Secondly, function words are likely to show frequent occurrences in multiple non-overlapping contexts and thus association probabilities spread out across several referents, indicating that these words are probably associated with none of those referents. The first characteristic is embodied by the association probabilities with the "NON" category. To capture the second characteristic, I compute an entropy-based metric that measures the distribution of a word with its possible co-occurring referents:

$$\mathrm{E}(w_n) = \sum_m \max_{j:j\neq n} p(m_m \mid w_j) \times p(m_m \mid w_n) \log p(m_m \mid w_n)$$

where the first item measures whether those objects that co-occur with the word are already paired with other linguistic labels with higher association probabilities, and the second item measures the degree to which a word is distributed across different referents. The first item instantiates the proposal that, if an object, temporally co-occurring with the word, is already associated with at least one other word (but not), then it is relatively less likely that serves as an additional linguistic label for that object (this consideration is based on the mutual exclusivity constraint proposed by Markman, 1994). The second item instantiates the proposal that the more a word occurs in different non-overlapping contexts, the less likely it is to be associated with any specific object.

With this metric, several function words are identified at the end of each learning episode and added to a function-word list. Then, at the beginning of the next episode, the model as an experienced learner filters the input language stream to exclude those items that are in the function-word list. In this way, the number of words in the speech stream encountered in the new episode can be significantly reduced, and the search space narrowed accordingly.

Learned Word-Object Pairs

After a learning episode, the model has also acquired a set of word-referent pairs and stored them in its long-term memory. If a learned word-referent pair occurs again in a new learning trial, the model does not need to re-estimate association

probabilities of those pairs: it simply excludes them from the input (as it does identify function words). The effect of this mechanism is to use learned lexical knowledge to disambiguate new learning situations. For instance if, as illustrated in Figure 3, the model has already learned that "dish" is associated with *dish*, it removes both "dish" from the word stream and *dish* from the referent stream. All of the possible links related to the word "dish" and the referent *dish* are removed as well. Thus, the number of possible associations remaining in this example is reduced to three (three words and one referent) from eight (four words and two referents).

Note that the mechanism will not influence the learning of homonyms because we assume that in a spoken utterance (defined by speech silence); there are one-to-one mappings between words and referents. Thus, it is not likely that the speaker will use one linguistic label to refer to two things in the world in a single utterance. It is equally unlikely that the speaker will use two linguistic labels within an utterance to refer to the same referent. Moreover, a learned word-referent pair will be removed from consideration if and only if the word and the referent co-occur. For example, if the speaker uses a different name to refer to a referent with another already known linguistic label, then it is the new name (but not the learned label) that co-occurs with the referent. Because the already known label does not in this instance co-occur with the referent, the learned pair will not prevent the model from learning the second linguistic label of that referent.

To summarize so far, the effect of the above two mechanisms is to reduce the number of items in both the word and the referent stream. In this way, word-to-world mappings are constrained by previously learned lexical knowledge, which leads the model to attend to some possible word-object mappings and rule out others. This process not only significantly reduces the computational load but also makes statistical associative computations in subsequent episodes more accurate and effective.

Partial Knowledge

In the simulations above, the model was exposed to a large number of co-occurrences of words and possible referents but only a relatively limited number of correct word-object pairs were learned (white cells with high association probabilities in Figure 2). Meanwhile, the model accumulated partial lexical knowledge—previously exposed (but not learned) word-object pairs (gray cells with relatively low association probabilities in Figure 2, e.g. *mom*-"mom"). This finding suggests that at the end of the process, the model may have learned that a particular word has been encountered but be uncertain about which object goes with this word. Similarly, human learners may also accumulate partial lexical knowledge that cannot be directly detected from standard familiar testing methods. Nonetheless, this knowledge could play a role in subsequent learning.

How could this kind of partial knowledge be utilized? Many learning algorithms, including most connectionist models and the EM-based algorithm used in the current associative model, could be formalized in terms of the optimization problem with constraints. One mathematical challenge in the optimization problem is to find the parameters (association probabilities in our case) that correspond to global maxima or minima of an objective function. EM-like or hill-climbing techniques are widely applied to search for global maxima (or minima) iteratively. In those methods, initial values that determine where the algorithm starts influence the final results significantly by deciding whether the learning method will fall into a local or global extreme value.

In the statistical associative model in the previous section, initial association probabilities were set as a flat distribution (the system lacked any prior knowledge). In the present model, in which an association probability matrix is obtained from prior learning episodes, one way to utilize this partial lexical knowledge is to include it in the initialization of association probabilities. For example, in the third step shown in Figure 3, assume that the association probability of "table"-*table* is .33 because three words co-occur with *table* in the current learning trial. Also assume that the model has already acquired the association probabilities of the object table to several words (including the word "table") from previous learning sessions. Now the two sets of association probabilities can be merged and then normalized to determine new initial values as follows:

$$p(m_m \mid w_n) = \alpha\, p^p(m_m \mid w_n) + (1-\alpha) p^c(m_m \mid w_n)$$

where p^p represents the association probability from previous episodes and represents the association probability based on an initial estimate from the current episode. is a weight of previous lexical knowledge and its value increases with more learning episodes the learner experiences. The new association probability of "table"-*table* following this process is .52. This increased association probability, based on prior knowledge, increases the likelihood that when the algorithm converges, the model will ultimately discover this pair. In general, then, previously acquired knowledge can lead the model to converge into a better local maximum and by doing so favor some interpretations of the data over others. With prior partial learning incorporated in this way, the same statistical associative learning mechanism can potentially get much better results.

In summary, to the extent that a word-referent pair has co-occurred in previous episodes, the model is able to encode and then utilize knowledge of this co-occurrence in subsequent statistical learning based on the three mechanisms described above. It follows that the cumulative learning method should obtain better and better learning results with more learning episodes.

Procedure and Data

Both procedure and data are the same as those in Experiment 1.

Results and Discussions

I have conducted a comparative study on two learning conditions: one-session learning and cumulative learning. In one-session learning, the statistical associative mechanism is applied to each individual session of picture-book reading and the results are merged at the end of each session. The merging process involves adding lexical items obtained from the current session to a master list of learned word-object pairs. The merged results reflect progress in learning across the completed sessions, but they are not used to enhance learning in subsequent sessions. In contrast, the cumulative learning method recruits previously established word-referent associations as input to subsequent learning as described in Subsection Rationale and Method.

I first compare the vocabulary growth that results from one-session learning versus from cumulative learning. Just as I did for the results in the first simulation, I fix the value of precision at .9 and measure recall in each of the two mechanisms so that the results illustrated in Figure 4 are based on a criterion of very high accuracy in selecting the correct word-object pairs. Both approaches learn more lexical items with increasing numbers of learning sessions because both can extract more statistical co-occurrence regularities over more instances of matched speech-visual streams. However, the cumulative-learning model works more efficiently than the one-session learning model because the statistical associative machinery of the former combines the statistical properties of the input with acquired lexical knowledge from previous exposures. Table 4 shows several examples of association probabilities in these two simulation conditions in the sixth learning session.

Our results are quite in line with evidence from other studies (e.g., Bloom, 2000; Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett, 1996; Ganger & Brent, 2004), suggesting that the pace of vocabulary development exhibits a gradual linear increase, and that there is no qualitative shift. Moreover, our work shows that cumulative lexical knowledge contributes to the gradual increase of the learning rate. From this perspective, the computational model provides a plausible alternative mechanistic explanation of why the rate of vocabulary learning increases. However, since the model does not consider other possibly relevant cognitive changes in development, it does not rule out the possibility that the vocabulary spurt does exist due to other reasons.

As shown in Figure 5, with the decrease of the times of co-occurrence of word-referent pairs, the performance of the simulated learner gets worse accordingly. However, cumulative learning maintains relatively good performance, demonstrating that infrequent words can also eventually be learned. Thus, the model is able to learn correct word-object associations based on only a few

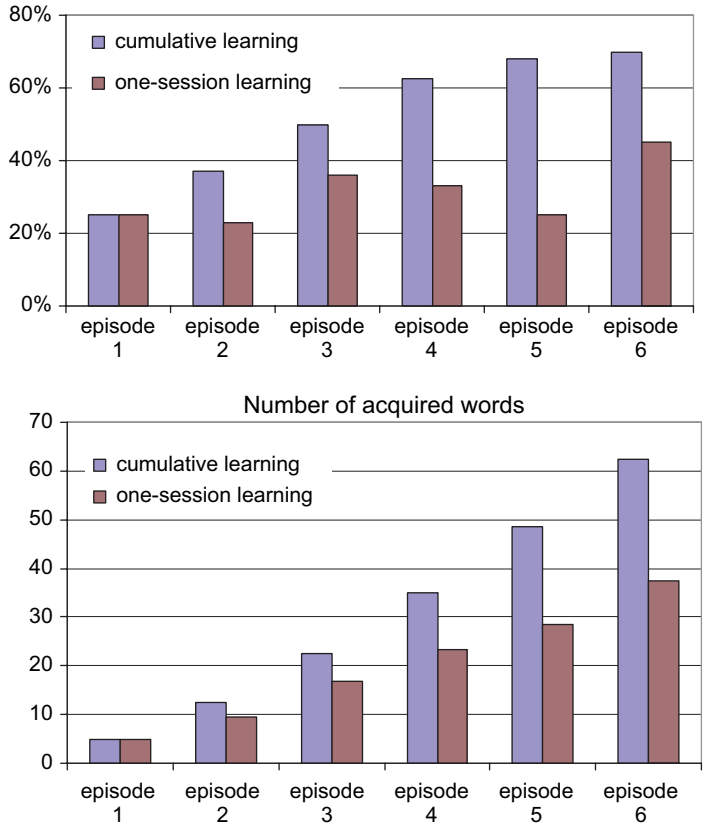

FIGURE 4 **Vocabulary growth.** Figure 4 shows single session and cumulative data when precision is set high criterion (.9). **Top:** the percentage of word-referent pairs selected by the simulated learner on each individual session. One-session learning purely depends on the co-occurrences of words and meanings in the training data while the performance of cumulative learning improves with more knowledge acquired and used. **Bottom:** the accumulated results on the number of learned word-object pairings. The increase in cumulative learning is significantly greater than that of one-session learning with additional input from previously exposed knowledge. Note that these two mechanisms share the same underlying statistical associative principle.

exposures, a result which simulates rapid word learning. Again, previously learned lexical knowledge plays a key role by reducing the hypothesis space. With more knowledge, the model becomes a more "confident" associative learner while applying the same statistical learning machinery. One way to explain rapid word learning is to use the concepts of recall and precision described in the first simulation study—how young language learners improve

TABLE 4
Several Examples of Word-Object Association Probabilities in Two Learning Mechanisms at the Sixth Learning Session. Those Relevant Pairs Are Assigned With Higher Association Probabilities in the Cumulative Learning Approach

Word-Referent Pairs	*# Of Co-Occurrences*	*Cumulative Learning Association Probabilities*	*One-Session Learning Association Probabilities*
"bear"-bear	12	**.429**	.256
"man"-bear	6	.231	.368
"boy"-bear	8	.179	.236
"cat"-cat	3	**.539**	.256
"frog"-cat	3	.186	.273
"dog"-cat	2	.174	.182
"girl"-girl	4	**.386**	.212
"man"-girl	3	.186	.255
"bear"-girl	2	.136	.222

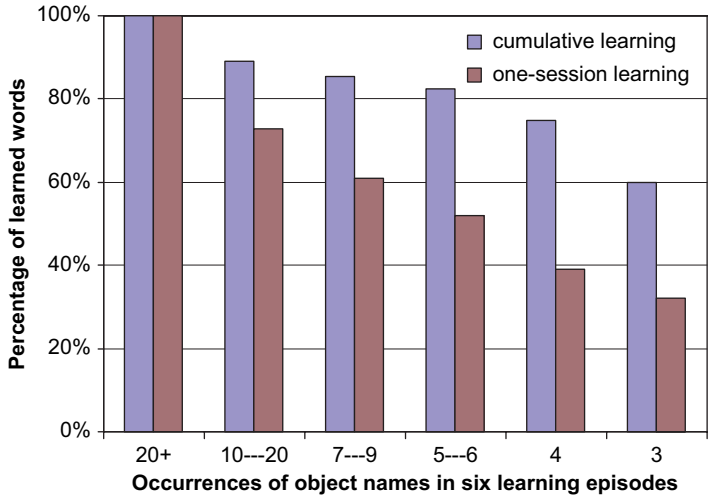

FIGURE 5 **Rapid word learning.** Both one-session learning and cumulative learning can acquire the correct word-referent pairs if the words and the meanings co-occur more than 20 times during the six episodes. With the decrease of the times of co-occurrence, the performance of both approaches gets worse accordingly. However, cumulative learning maintains relatively good performance that is much better than one-session learning, demonstrating that infrequent words can also eventually be learned and that the statistical associative mechanism can build word-to-world mappings much more effectively in highly ambiguous learning contexts.

recall significantly based on only one or a very few instances while making few mistakes (maintaining a high precision). From the results in the first simulation study, we know that recall and precision trade off, suggesting that the previous model cannot be both rapid and accurate word learners—when one is tuned up, the other falls. However, this is not what we observe in children, who manage to be precise language learners and also show an increased rate of acquisition while maintaining precision. Compared to the previous model, the present model offers a more satisfactory explanation in terms of the role of previous learning.

Overall, the results in Figures 4 and 5 suggest that the model is able to yield similar high performance by utilizing cumulative knowledge. The findings do not perfectly reproduce the learning abilities of young children, but they are strong enough to suggest a promising direction. Thus, the statistical associative learning mechanism has the potential to provide a plausible mechanistic explanation of real, observed behavioral changes in young language learners like the increase in the rate of word learning as word learning progresses.

GENERAL DISCUSSION

Statistical Language Learning

One of the most important findings in research on language acquisition is that humans are sensitive to statistical regularities in language and are able to acquire linguistic knowledge based on statistical learning. Saffran, Aslin, and Newport (1996) demonstrated that 8-month-old infants are able to find word boundaries in an artificial language based only on statistical regularities. Gomez and Gerken (1999) showed that after less than 2 min exposure to one of two grammars in an artificial language, 12-month-olds could discriminate new strings from the two grammars, suggesting that statistical learning might play a role also in acquiring rudimentary syntax (the ordering of words, etc.). In a recent study by Gomez and Maye (2005), 15-month-old children are able to learn nonadjacent dependency regularities in an artificial language, such as aXb or bXd, indicating that they might be able to acquire similar structure in natural language. Recent analyses on child-directed corpora (e.g., Mintz, Newport, & Bever, 2002; Redington, Chater, & Finch, 1998; also see empirical work on Gerken, Wilson, & Lewis, 2005; Gomez & Lakusta, 2004) have also demonstrated that simple computational mechanisms using distributional information as a powerful cue can obtain a considerable amount of knowledge on grammatical category membership.

While most studies on statistical language learning focus on the statistical regularities in the linguistic input, the present article asks whether young language

learners can apply the same learning machinery to extract the co-occurring statistical regularities embedded in words and extralinguistic contexts. In a previous study (Yu & Smith, in press), we showed that human learners were able to learn word-to-world mappings based purely on statistical co-occurrence regularities between words and objects. This complementary study here quantifies how this kind of learning mechanism works.

Constraints in Word Learning

A statistical cross-situational account of word learning requires that the learning system be somehow constrained. Indeed, there have been active debates about how to characterize constraints in children's word learning and how the constraints emerge (are they innate or learned). The present simulation encoded two crucial constraints proposed by Markman (1994): (1) the whole object constraint—words refer to whole objects and (2) the mutual exclusivity constraint—every object has just one name and therefore a new word will not go to the objects with known labels. Similarly, the "contrast constraint" proposed by (Clark, 1993) has it that children assume that new words are different from known words so that they should be mapped to different meanings, and the "Novel-Name Nameless-Category" constraint embodies the principle that children map novel labels to novel objects (Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992). These constraints are different in the motivations they posit for children's mapping of novel words to novel objects, but they share the same underlying principle-word learners apply existing lexical knowledge to learn new words. That is, they use something they know to infer something novel. This suggests that although there may be theoretical differences distinguishing these constraints, all three could be realized by the same underlying computational mechanism. In fact, the statistical associative learning mechanism in the present study implicitly encoded all of those constraints in a general learning mechanism. When presented with a toy dog and a toy cat, and at the same time given the words "dog" and "cat," our simulated learner mapped "dog" to the object *dog* in part because it had already acquired the word-object pair "cat"-cat from previous learning experiences. This result from the model is compatible with previous results from behavioral studies of the mutual exclusivity, contrast, and novel-name-novel-category constraints.

In light of the present results, I propose that there is one additional mechanism to facilitate probabilistic word-referent pairings—the constraint that comes from partial lexical knowledge accumulated over previous learning experiences. The present results show that, in addition to learned lexical items which can then be directly excluded from consideration during new learning trials, the kind of associative mechanism tested here can amass a huge amount of partial lexical knowledge—those word-referent pairs with low association probabilities. The

simulation results show that this kind of partial knowledge can bias the associative learning mechanism's search for reliable word-referent pairings, so that the overall learning results favor those word-referent pairings consistent with previous learning. In this way, partial knowledge implicitly acts as a *probabilistic* mutual exclusivity constraint. The success of the model in simulating developmental findings suggests that human learners may also apply a probabilistic mutual exclusivity constraint based on partial knowledge. In my ongoing studies, I am attempting to systematically investigate whether (and if so, how) young language learners make use of this potentially powerful mechanism.

Developmental Changes in Rate of Word Learning

For most children, vocabulary growth begins slowly with a gradual increase in the number of new words but then quickens to a noticeably fast rate of word acquisitions (Bates et al., 1988; Benedict, 1979; Dromi, 1987; Gershkoff-Stowe & Smith, 1997; Goldfield & Reznick, 1990; Gopnik & Meltzoff, 1987; Lifter & Bloom, 1989). Traditionally, this "naming explosion" or "vocabulary spurt" was understood in terms of a qualitative shift in some underlying process, such as an insight that objects have names (Lock, 1978; McShane, 1979, 1980), a conceptual change in representing objects (Lifter & Bloom, 1989), or the emergence of basic-level categories (Gopnik & Meltzoff, 1987). Indeed, researchers often sought to determine the first "substantive" jump in vocabulary (Gershkoff-Stowe & Smith, 1997; Lifter & Bloom, 1989; Mervis & Bertrand, 1994). Recently, an alternative conceptualization of productive vocabulary growth suggests a continuous and smoothly accelerating curve with no single point of acceleration (Bates & Carnevale, 1993; Bloom, 2000; vanGeert, 1991).

This developmental change in rate of word learning, which has attracted so much attention in the field of cognitive development, is difficult to explain in terms of simple and transparent causes. By the time children learn words at a fast rate, they need to possess and use a set of constraints to help them reduce the degree of reference uncertainty in the word-to-world mapping problem. Recent studies in computational modeling (e.g., Elman et al., 1996; see a good review in Regier, 2003) suggest that abrupt internal changes may not be needed to produce discontinuous external changes in behaviors. These works show that some factors, such as limited memory at the starting point of learning (Elman et al., 1996; Newport, 1990), nonlinearity of neural networks (Plunkett et al., 1992), and gradual emergence of attention to some aspects of the world (Regier, Corrigan, Cabasan, Woodward, Gasser, & Smith, 2001), may contribute to non-linear behaviors of human learning.

The present results also suggest that there may be no need to posit a qualitative mechanistic change to explain a change in the rate of vocabulary growth. More specifically, the changing dynamics of vocabulary acquisition may be partially due to accumulated knowledge during development. The performance of

the same learning mechanism can be significantly improved by storing lexical knowledge previously extracted and then recruiting it in subsequent learning. With more knowledge accumulated and then recruited, young children become more efficient word learners. This idea is certainly consistent with many formal theories of learning. However, the main contribution of this paper is to propose and implement such a learning mechanism and demonstrate how the mechanism works using data obtained from everyday learning environments.

Moreover, human development and learning is a life-long process but most modeling studies focus on snapshots of learning at different points in developmental time because of not only the difficulty in data collection but also the challenge of finding an efficient learning mechanism that can handle those data. This work rep-resents first steps toward tackling these two problems. I illustrate how the same statistical learning mechanism, operating incrementally and without any significant internal changes, is able to give rise to dramatically different behaviors during a series of learning sessions. This result may seem obvious. But that does not reduce its profound importance for how we think about development and the role of accumulating partial and incomplete knowledge in an emerging system of knowledge and in creating the developmental trajectory. A system of partially learned regularities—even if insufficient to show up in overt behavior—shapes, constrains, and potentially speeds current learning. This kind of mechanism may take the mystery out of the phenomenon known as the vocabulary spurt.

Homonym, Synonymy and Weird Words

A solution to the reference uncertainty problem must deal with how to distinguish correct one-to-one word-object pairs from many-to-many co-occurrence pairs. But that is not enough. Young children also learn homonymous and synonymous words. From the view point of the mapping problem, a homonymous word has associations with multiple referents while multiple synonymous words have a shared referent. Thus, from many-to-many co-occurrences between words and referents, young learners are able to build not only one-to-one mappings but also many-to-one and one-to-many mappings, which makes the task of word learning much more complex computationally. In our picture-book narration experiments, I have observed that speakers may use more than one linguistic label to refer to the same object but are less likely to use one linguistic label to refer to multiple meanings in a single session. In light of this, our computational model allows one object to be associated with multiple words, but does not allow one word to be associated with multiple objects within a learning session. The model is able to acquire most synonymous words within individual learning sessions and obtains

homonymous words based on the data collected across multiple sessions. Thus, the same framework of statistical associative learning can be used to tackle these two problems separately.

Another observation in early word learning is that some of children's first word meanings are incorrect, though comprehensible as the products of association learning. So, for example, a child might use the word "hot" to describe an oven (Macnamara, 1982), the word "flying" to describe birds, or the word "sock" to describe a shoe (Dromi, 1987). Interestingly, besides acquiring correct word-referent pairs, our model also acquires these sorts of pairs. Thus, the model's simulations resemble children's performance very closely: it not only discovers correct word-referent pairs, it also occasionally builds incorrect associations. Even more interestingly, the model can also "self-correct" and discard some initially incorrect associations. For example, the simulated learner may build an incorrect word-object pairing at the first learning episode simply because of the spurious correlation between a word and an object. However, a new (and correct) mapping is built in subsequent episodes when the model encounters the word again but this time with the relevant object. Eventually the new mapping outweighs the incorrect one—a solution based on cross-situational and accumulated statistics that deals with spurious and unreliable associations. Overall, these additional outputs of the simulation make statistical associative learning an even more cognitively plausible candidate for a mechanistic explanation of young children's word learning.

Assumptions and Limitations in the Model

Two major assumptions in this computational study are that: (1) young children can segment words from continuous speech; and (2) they can categorize visual objects in the picture books into corresponding types. These two assumptions are addressed in Yu et al. (2005), in which we propose and implement a computational model that is able to discover spoken words from continuous speech and associate them with their perceptually grounded meanings. Similar to infants, the model spots word-referent pairs from unprocessed multisensory signals collected in everyday contexts. The purpose of the present work, however, is to understand the mechanistic nature of developmental changes in vocabulary growth. To address this problem, I have simplified some aspects of early word learning to focus on the key issue—the word-to-world mapping problem.

Moreover, I argue that statistical learning is just one of the important driving forces in language acquisition. In addition to distributional information, there are at least two other important factors. The first is social cues. It has been shown that social cues, such as joint-attention, guide children to find the referents of words (Baldwin, 1993; Yu et al., 2005). Furthermore, children are

sensitive to how the referents of a new word are generated, and it makes a difference whether the examples are chosen by knowledgeable teachers or by the learners themselves (Tenenbaum & Xu, 2000). Secondly, Gleitman (1990) has proposed that syntactic information is also a potentially powerful cue for the acquisition of meaning. MacWhinney (1989) adds that parents make extensive use of stable syntactic patterns, and that when their children learn these patterns, they become able to quickly identify new words embedded in these frames. In future work, I will study how these two additional kinds of cues interact with statistical cues, and how all these factors can be integrated into a general learning mechanism.

CONCLUSIONS

The proposal that previously acquired knowledge may help subsequent word learning seems to be intuitively obvious. However, there has been no previous demonstration of a mechanism that instantiates this intuition. The contribution of this paper is to conceptualize word learning as a translation problem, to implement a statistical associative model based on this conceptualization and to demonstrate explicitly how a learning device gradually accumulates lexical knowledge and then uses that accumulated knowledge to increase the speed and accuracy of future lexical acquisitions. The model uses acquired word meanings to change the rate of new word learning without changing the underlying statistical associative learning mechanism, suggesting that developmental changes in children's vocabulary growth may also be at least partially due to this accumulative effect. Moreover, because the model was tested using real data collected from child-caregiver interactions, it provides quantitative evidence about both the kinds of statistical regularities that are in a child's language learning environment and how the associative mechanism can capture those regularities. Finally and most generally, this work suggests that computational models of word learning are an excellent tool with which to study the mechanisms underlying the development of complex systems.

ACKNOWLEDGMENTS

I would like to express my thanks to Dana Ballard and Linda Smith for fruitful discussions. Susan Jones kindly provided helpful and detailed comments on an earlier version of this article. I would also like to thank LouAnn Gerken, Susan Goldin-Meadow, Rebecca Gomez, and other two anonymous reviewers for insightful comments and suggestions. This research was supported by National Science Foundation Grant BCS0544995.

REFERENCES

Bailey, D. (1997). *When push comes to shove: A computational model of the role of motor control in the acquisition of action verbs.* Unpublished doctoral dissertation, Computer Science Division, University of California—Berkeley.

Baldwin, D. (1993). Early referential understanding: Infant's ability to recognize referential acts for what they are. *Developmental Psychology, 29,* 832–843.

Bates, E., Bretherton, I., & Snyder, L. (1988). *From first words to grammar.* Cambridge, England: Cambridge University Press.

Bates, E., & Carnevale, G. F. (1993). New directions in research on language development. *Developmental Review, 13,* 436–470.

Benedict, H. (1979). Early lexical development: Comprehension and production. *Journal of Child Language, 6,* 183–200.

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

Boroditsky, L. (2001). Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive Psychology, 43,* 1–22.

Bowerman, M. (1996). The origins of children's spatial semantic categories: Cognitive versus linguistic determinants. In J. Gumperz & S. Levinson (Eds.), *Re-thinking linguistic relativity.* Cambridge, England: Cambridge University Press.

Bowerman, M., & Choi, S. (2001). Shaping meanings for language: Universal and language-specific in the acquisition of spatial semantic categories. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 475–511). Cambridge, England: Cambridge University Press.

Brown, P. F., Pietra, S., Pietra, V., & Mercer, R. L. (1994). The mathematics of statistical machine translation:parameter estimation. *Computational Linguistics, 19*(2), 263–311.

Clark, E. V. (1993). *The lexicon in acquisition.* New York: Cambridge University Press.

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science, 208,* 1174.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, 39*(1), 1–38.

Dromi, E. (1987). *Early lexical development.* London: Cambridge University Press.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: The MIT Press.

Fodor, J. A. (1975). *The language of thought.* Sussex, England: Harvester Press.

Ganger, J., & Brent, M. R. (2004). Reexamining the vocabulary spurt. *Developmental Psychology,* 621–632.

Gerken, L., Wilson, R., & Lewis, W. (2005). 17-month-olds can use distributional cues to form syntactic categories. *Journal of Child Language, 32,* 249–268.

Gershkoff-Stowe, L., & Smith, L. (1997). A curvilinear trend in naming errors as a function of early vocabulary growth. *Cognitive Psychology, 34,* 37–71.

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

Goldfield, B. A., & Reznick, J. (1990). Early lexical acquisition: rate, content, and the vocabulary spurt. *Journal of Child Language, 17*(1), 171–183.

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L., & Wenger, N. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology, 28,* 99–108.

Gomez, R., & Lakusta, L. (2004). A first step in form-based category abstraction by 12-month-old infants. *Developmental Science, 7*(5), 567–580.

Gomez, R., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy, 7*(2), 183–206.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70,* 109–135.

Gopnik, A., & Meltzoff, A. N. (1987). The development of categorization in the second year and its relation to other cognitive and linguistic developments. *Child Development, 58*, 1523–1531.

Huttenlocher, J., & Smiley, P. (1987). Early word meanings: The case of object names. *Cognitive Psychology*.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29*, 1–23.

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural networks. *Neural Networks, 17*, 1345–1362.

Lifter, K., & Bloom, L. (1989). Object knowledge and the emergence of language. *Infant Behavior and Development, 12*, 395–423.

Lock, A. (1978). The emergence of language. In A. Lock (Ed.), *Action, gesture and symbol: The emergence of language* (pp. 3–18). San Diego, CA: Academic Press.

Macnamara, J. (1982). *Names for things: A study of child language.* Cambridge, MA: The MIT Press.

MacWhinney, B. (1989). Competition and lexical categorization. In F. Eckman & M. Noonan (Eds.), *Linguistic categorization* (pp. 195–242). Philadelphia: Benjamins.

MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology, 49*, 199–227.

Markman, E. (1994). Constraints on word meaning in early language acquisition. In L. Gleitman & Barbara Landau (Eds.), *The acquisition of the lexicon* (pp. 199–228). Cambridge, MA: MIT Press.

McShane, J. (1979). The development of naming. *Linguistics, 17*, 79–90.

McShane, J. (1980). *Learning to talk.* Cambridge, England: Cambridge University Press.

Mervis, C., & Bertrand, J. (1994). Acquisition of the novel name-nameless category (n3c) principle. *Child Development, 65*, 1646–1662.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*, 393–424.

Newport, E. L.(1990). Maturational constraints on language learning. *Cognitive Science, 14*, 11–28.

Pinker, S. (1989). *Learnability and cognition.* Cambridge, MA: The MIT Press.

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

Quine, W. (1960). *Word and object.* Cambridge, MA: The MIT Press.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*(4), 425–469.

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism.* Cambridge, MA: The MIT Press.

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

Regier, T., Corrigan, B., Cabasan, R., Woodward, A., Gasser, M., & Smith, L. (2001). The emergence of words. In *Proceedings of the 23rd annual meeting of the Cognitive Science Society* (pp. 815–820). Mahwah, NJ: Erlbaum.

Roy, D., & Pentland, A.(2002). Learning words from sights and sounds: A computational model. *Cognitive Science, 26*(1), 113–146.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month old infants. *Science, 274*, 1926–1928.

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

Smith, L. (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.

Tenenbaum, J., & Xu, F. (2000). Word learning as Bayesian inference. In L. Gleitman & A. Joshi (Eds.), *Proceeding 22nd annual conference of the Cognitive Science Society* (pp. 517–522). Mahwah, NJ: Erlbaum.

vanGeert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological Review, 98*, 3–53.

Waxman, S. R., & Booth, A. E. (2000). Principles that are invoked in the acquisition of words, but not facts. *Cognition, 77*, B35–B50.

Yoshida, H., & Smith, L. (2003). Shifting ontological boundaries: How Japanese-and English-speaking children generalize names for animals and artifacts. *Developmental Science, 6*(1), 1–36.

Yu, C., Ballard, D. H., & Aslin, R. N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science, 29*(6), 961–1005.

Yu, C., & Smith, L. B. (in press). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*.

APPENDIX A

The general setting is as follows: suppose we have a word set and a referent set, where N is the number of words and M is the number of meanings (toys, etc.). Let S be the number of learning situations. All word data are in a set, where for each learning situation, consists of r words, and can be selected from 1 to N. Similarly, the corresponding contextual information in that m learning situation include possible meanings and the value of is from 1 to M. A simple example in Figure 1 consists of four learning situations in which every word can potentially be associated with any co-occurring meaning. The computational challenge here is to build several one-to-one mappings (e.g., *cat* to "cat") from many-to-many possible associations. We suggest that to figure out which word goes to which meaning, language learners don't consider the association of just a single word-referent pair, but they estimate all these possible associations simultaneously. Thus, they attempt to estimate the association probabilities of all of these pairs so that the best overall mapping is achieved. In doing so, the constraints across multiple learning situations and the constraints across different word-referent pairs are jointly considered in a general system which attempts to discover the best translation between words and referents based on statistical regularities in the observation.

Formally, given a data set, we use the machine translation method proposed by Brown et al. (1994) to maximize the likelihood of generating the referent strings given English descriptions:

$$
\begin{aligned}
&P(S_m^{(1)}, S_m^{(2)} \ldots, S_m^{(S)} \mid S_w^{(1)}, S_w^{(2)}, \ldots, S_w^{(S)}) \\
&= \prod_{s=1}^{S} \sum_{a} p(S_m^{(s)}, a \mid S_{\mathbf{w}}^{(s)}) \\
&= \prod_{s=1}^{S} \frac{\in}{\tau + 1} \prod_{j=1}^{l} \sum_{i=0}^{r} p(m_{v(j)} \mid w_{u(i)})
\end{aligned}
$$

where the alignment indicates which word is aligned with which meaning. is the association probability for a word-referent pair and o is a small constant.
To maximize the above likelihood function, a new variable is introduced which represents the expected number of times that any particular word in a language string generates any specific meaning in the co-occurring referent string

$$c(m_m \mid w_n, S_m^{(s)}, S_w^{(s)}) = \frac{p(m_m \mid w_n)}{p(m_m \mid w_{u(1)}) + \cdots + p(m_m \mid w_{u(r)})} \times \sum_{j=1}^{l} \delta(m_m, v(j)) \sum_{i=1}^{r} \delta(w_n, u(i))$$

where is equal to one when both of its arguments are the same and equal to zero otherwise. The second part in Equation (3) counts the number of co-occurring times of and. The first part assigns a weight to this count by considering it across all the other words in the same learning situation. By introducing this new variable, the computation of the derivative of the likelihood function with respect to the association probability results in:

$$p(m_m \mid w_n) = \frac{\sum_{s=1}^{S} c(m_m \mid w_n, S_m^{(s)}, S_w^s)}{\sum_{m=1}^{M} \sum_{s=1}^{S} c(m_m \mid w_n, S_m^{(s)}, S_w^s)}$$

As shown in Algorithm 1, the method sets an initial to be flat distribution, and then successively compute the occurrences of all word-referent pairs using Equation (3) and the association probabilities using Equation (4). In this way, our method runs multiple times and allows for re-estimating word-referent association probabilities. The detailed technical descriptions can be found in (Brown et al., 1994; Yu et al., 2005).

Algorithm 1 Estimating word-referent association probabilities
Assign initial values for based on co-occurrence statistics.
repeat
E-step: Compute the counts for all word-referent pairs using Equation (3).
M-step: Re-estimate the association probabilities using Equation (4).
until the association probabilities converge.