A Statistical Model of Developmental Changes in Early Word Learning

Chen Yu (chenyu@indiana.edu)
Department of Psychology and Program in Cognitive Science
Indiana University
Bloomington, IN 47405 USA

Weixia Huang (whuang@xeroxlabs.com)
Xerox Corp.
800 Philips Road
Webster, NY 14580 USA

Abstract

Young language learners are able to map a word onto its referent from an infinite number of possible word-to-world mappings at early stage and further this learning ability turns to be much more effective at the later stage. What mechanisms underlie behavioral changes during word learning? This paper presents a developmental model of statistical associations, suggesting that as far as human minds are equipped with general associative learning mechanisms, ambiguity in the word-learning situation can be significantly reduced by recruiting cumulative lexical knowledge in statistical computations. Consequently, this leads to increasingly fast subsequent learning and abrupt behavioral changes, such as the vocabulary spurt and fast mapping. We provide a formal account of this argument by developing a computational model fed with the data collected from a series of picture-book reading sessions to simulate word acquisition in natural contexts. The results show that previously learned lexical knowledge can not only narrow the search space but also bootstrap the subsequent learning process, which greatly improves learning results without the change of the underlying learning mechanism. Hence, this work suggests that using lexical knowledge accumulated in subsequent statistical learning is a scaffold for vocabulary growth in early language acquisition.

1. Introduction

From scratch, young children need to solve many complex learning problems in early word acquisition, such as speech segmentation to discover word units, conceptual learning to identify the meanings (prelinguisitic concepts) of words, and then word-to-world mapping to associate words with meanings. Gleitman, Cassidy, Nappa, Papafragou, and Trueswell (2005) argued that among these tasks, a considerable part of the bottleneck for young language learners resides in the tools for solving the word-to-world mapping problem – how to map a phonological form to a conceptual representation. A common conjecture is that children map sounds to meanings by seeing an object while hearing an auditory word-form. The most popular mechanism of this word learning process is *associationism*. Richards and Goldfarb (1986) proposed that children come to know the meaning of a word through repeatedly associating the verbal label with their experience at the time that the label is used. Smith (2000) argued that word learning is initially a process in which children's attention is captured by objects or actions that are the most salient in their environment, and then they associate it with some acoustic pattern spoken by an adult.

However, the associative approach has been criticized on the grounds that it does not provide a clear explanation about how infants map a word to a potential infinity of referents when the word is heard, which is termed *reference uncertainty* by Quine (1960). Quine presented the following puzzle to theorists of lexical learning: 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 rab-

bit in the distance. How can you determine the intended referent? Quine offered this puzzle as an example of the indeterminacy of translation. Given any word-event pairing, there are, in fact, an infinite number of possible intended meanings – ranging from the rabbit as a whole, to its color, fur, parts, or activity. One explanation termed "cross-situational learning" has been proposed by many theorists, such as Pinker (1989) and Gleitman (1990). The idea is that when a child hears a word, she can hypothesize all the potential meanings for that word from the non-linguistic context of the utterance containing that word. Upon hearing that word in several different utterances, each of which is in a different context, she can intersect the corresponding sets to find those meanings which 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. In light of this and with recent empirical evidence demonstrating that children and even infants, possess powerful statistical learning capacities (Saffran, Newport, & Aslin, 1996), Yu and Ballard (2004a) (see also Yu, Ballard, & Aslin, in press) developed a computational model of how young language learners perform statistical computations on cross-situational observations, which is reviewed in Section 3.

Based on our previous work, this paper attempts to apply the statistical learning mechanism to interpret developmental changes in word learning. Specifically, we investigate whether the associative learning mechanism can account for discontinuities in behaviors during the second year of life, such as the vocabulary spurt and fast mapping (see Section 2). These phenomena has attracted much attention in the field of cognitive development since they are difficult to explain in terms of simple and transparent causes, and pose theoretical challenges that require sophisticated and highly-constrained learning principles. Nonetheless, complementary studies in connectionist 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), nonlinearity of neural networks (Plunkett, Sinha, Miller, & Strandsby, 1992), and gradual emergence of attention to some aspects of the world (Regier et al., 2001), may contribute to nonlinear behaviors of human learning. In light of this, the present paper suggests another possible explanation of discontinuous behaviors in early word learning – the performance of the same learning mechanism can significantly improve by storing lexical knowledge previously exposed and then recruiting it in subsequent learning. This idea may seem obvious and it 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 the data obtained from everyday learning environments.

The organization of the paper is as follows: Section 2 reviews empirical evidence in developmental literatures. Section 3 briefly presents our statistical learning model which provides a basis for further discussion. Section 4 describes a cumulative learning mechanism that utilizes partial lexical knowledge in subsequent learning within the framework of statistical learning. We conclude with general discussions in Section 5.

2. Developmental Changes in Word Learning

Most global descriptions of early vocabulary growth report that for the majority of children, development proceeds from a slow and gradual increase in the number of new words produced to a faster and more noticeable increase (Bates, Bretherton, & Snyder, 1988; Gopnik & Meltzoff, 1987; Lifter & Bloom, 1989). This increased rate has been called the "naming explosion" or "vocabulary spurt". Early researchers often conceptualized this change in rate in terms of a qualitative shift in some underlying process (e.g., an insight that objects have names, MaShane, 1979). Under these conceptualizations of the "naming explosion", a rate shift was often located at a particular point in development, measured as the first "substantive" jump in vocabulary. However, recent research suggests that the conceptualization of the "vocabulary spurt" is probably wrong. Although some children show a readily identifiable shift in the rate of new word productions, others show a steady but more gradual rise (Goldfield & Reznick, 1990; Ganger & Brent, 2004).

As children know more and more words, they become faster learners of new words, becoming able to learn a new word based on only a few exposures (Carey & Bartlett, 1978; Markson & Bloom, 1997; Schafer & Plunkett, 1998; Woodward, Markman, & Fitzsimmons, 1994; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). Woodward et al. (1994) showed that 13-month-olds can learn novel words from as few as nine presentations. Schafer and Plunkett (1998) provided similar evidence about 15-month-olds under controlled conditions. Although researchers have pointed to a variety of kinds of knowledge likely relevant to this accelerating pace of new word acquisitions – from increasing knowledge about syntax (Gleitman, 1990; MacWhinney, 1998), to knowledge about categories (Smith et al., 2002), to linking rules between syntactic categories and meaning (Waxman & Markow, 1995), there is little evidence to explain how this knowledge is acquired and only vague proposals about the relevant internal mechanisms.

Can we explain these phenomena by statistical learning? This is a theoretically very difficult but important problem. We need a mechanism that is both a rapid (nearly single trial) learner of word-referent mappings but that does not make mistakes. Statistical associative learning seems problematic in this regard for two reasons. First, if language learners associate a word with a meaning based on just a very few co-occurrences (that is fast mapping), then one should predict that they make lots of wrong associations because there are many irrelevant co-occurring word-meaning pairs in natural environments. However, the fact is that they make such mistakes only rarely. Second, statistical learning relies on the inference based on relatively large amount of data, which is

contradictory with the key idea of fast mapping – a very few exposures.

The theoretical resolution to this problem is to imagine a learning system that does NOT learn single associations between individual words and referents but that learns a system of associations, and that generalizes on the basis of partial knowledge. More generally, the acceleration of word learning and fast mapping are partially due to accumulated knowledge during development. With more knowledge accumulated from more exposures to a language and then recruited in subsequent learning, children become more efficient word learners. In the present paper, we seek to characterize such a learning system.

3. Statistical Word Learning from Cross-Situational Observation

In early word learning, children need to start by pairing spoken words with the co-occurring possible referents, collecting multiple such pairs, and then figuring out the common elements. Although no one doubts this process, there has been little systematic investigation. Yu and Ballard (2004a) introduce a formal model of statistical word learning which provides a probabilistic framework for encoding multiple sources of information. Given multiple scenes paired with spoken words collected from natural interactions between caregivers and children, the model is able to compute the association probabilities of all the possible word-meaning pairs. We first apply this model to the new data collected in this work.

Data. Three native speakers of English participated in data collection. Each of them was asked to narrate 2 picture books. The books were for 1-3 year old children. They were also instructed to act as a caregiver and pretend that they were telling this story to a child so that they should keep verbal descriptions of pictures as simple and clear as possible. During the experiment, the video was recorded from a head-mounted camera to provide a dynamic first-person view. Furthermore, an eye tracker was utilized to track the time-course of the speaker's eye movements and gaze positions. These gaze positions were indicated by a cursor that was superimposed on the video of the book to indicate where the speaker was attending (as shown in Figure 1). The data used for this simulation study were our descriptions of video clips. More specifically, our description of the audio input – what we feed into the statistical simulated learner – is the entire list of spoken words. Our description of the video stream, again what we feed into the statistical learner, is the list of all (basic-level) objects in picture books that a narrator was attending to from moment to moment when spoken utterances were produced. Table 1 shows the statistics of the data.

Table 1: Statistics of the training data

picture book	1	2	3	4	5	6
vocabulary	467	215	403	490	1034	468
objects	29	20	30	23	38	32

Method. In this kind of natural interaction, the vocabulary is rich and varied and the central items (object names) are far from the most frequent words. This complex but perfectly natural situation can be easily quantified by plotting a histogram of word frequency which shows that few key words
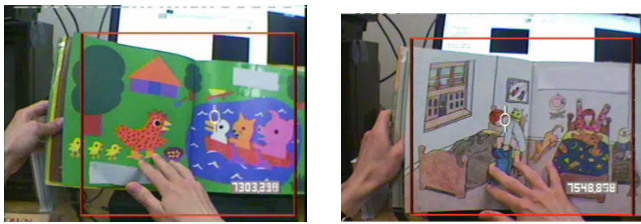
Figure 1: The snapshots from a first-person view when a speaker is narrating picture books. The white circles indicate current gaze positions.

– object names – make themselves into the top items of the list as shown in Figure 2. Similarly, even with gaze information that indicates a speaker's focus of attention, there are still multiple objects temporally co-occurring with spoken narrations. However, by considering linguistic and contextual information together, we find what *is* helpful is to partition the object sequences (attended contextual information when the speech was produced) into intervals where within each interval a single object or small number of co-occurring objects is the central subject or meaning, and then categorize spoken word sequences using the contextual bins labeled by different objects. The hypothesis is that mothers use temporal synchrony to highlight novel word-referent relations for young infants. That is, presenting information across multiple modalities simultaneously serves to highlight the relations between the two patterns of stimulation.

Formally, associating meanings (objects in picture books, etc.) with words (object names, etc.) can be viewed as the problem of identifying word correspondences between English and a "meaning language", given the data of these two languages in parallel. With this perspective, we apply a word-correspondence model from machine translation to address the word-to-world mapping problem and use an Expectation-Maximization (EM) based learning algorithm. Briefly speaking, the algorithm assumes that word-meaning pairs are latent variables underneath the observations which consist of spoken words and extralinguistic contexts. Thus, association probabilities of these pairs are not directly observable, but they somehow determine the observations because spoken words are produced based on speakers' lexical knowledge. Therefore, the objective of language learners or computational models is then to figure out the values of association probabilities so that they can increase the chance of obtaining the observations. In this way, correct word-meaning pairs are those which can maximize the likelihood of the observations. We argue that this strategy is an effective one that young language learners may apply during early word learning. They tend to guess most reasonable and most co-occurring word-meaning pairs based on the observations from different contexts. The technical details of our learning method can be found in Yu and Ballard (2004a) and Yu et al. (in press).

Results. Figure 2 shows the results of statistical associative learning on the data of one picture-book reading session. Many words at the top of the frequency list are associated with the NON meaning because they are function words. The word *girl* is correctly mapped to the meaning "girl", *flowers* to "flowers" and *bear* to "bear". Meanwhile, the simulated learner also makes some errors, such as *of* and *to* to "bear". That is because the picture book is about a brown bear. There-

Figure 2: The row is a sorted list of most frequent words and the column is a list of (a subset of) meanings. Each cell is the association probability of a specific word-meaning pair. Dark color means low probability while white means high probability.

fore, the narrator spent significantly more time on describing the "bear" and his activities. Consequently many words co-occur more frequently with the meaning "bear" compared with other meanings. Note that *mom* is likely to be associated with "mom" but not significantly, so is *bed* with "bed", which we term partial lexical knowledge – the knowledge (represented by gray areas in the figure) that has not been learned yet. The role of partial knowledge in subsequent learning will be discussed in next section.

Siskind (1996) developed a cross-situational learning model based on inference rules and logic learning. In contrast, our model is based on probabilistic learning and is able to explicitly represent and estimate the association probabilities of all the co-occurring word-meaning pairs in the training data. The results demonstrate the potential value of this mechanism – how multimodal correlations may be sufficient for learning words and their meanings. We can also go beyond demonstrating the mechanism to making new and unexpected predictions that derive from knowing more about the correlations. For example, the model makes predictions about the naming and comprehension errors that should be most likely. These predictions are based not merely on phonological nor visual similarity nor temporal proximity in the stream of events but 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 the figure), the simulated learner also potentially accumulates lots of partial knowledge (gray and dark cells) of all the word-meaning pairs previously exposed. And further, the model provides a probabilistic framework to explore the role of the partial knowledge in subsequent word learning.

4. Cumulative Subsequent Learning

Section 2 reviews experimental evidence of developmental changes in early word learning. We suggest that one reason for these changes is that young children learn how to use previously acquired knowledge to learn new words. To support this idea, we propose a cumulative mechanism based on statistical associative learning and show that with more lexical knowledge learned and recruited in subsequent learning, the model is able to learn words in a more effective way and requires less exposures before a word is learned, which is similar to the behaviors of human language learners. In this study, we use the data collected from a series of picture-book reading and treat one picture-book reading session as an episode in developmental learning. The computational model processes the data in each individual episode sequentially. Two conditions in this experiment are termed cumulative learning and one-session learning. In one-session learning, we apply the statistical associative model (described in Section 3) on each individual session of picture-book reading and then merge the results at the end of each session. The merging process involves adding lexical items obtained from a current session to a list of learned words. In contrast, the cumulative learning method recruits previously learned word-meaning associations in subsequent learning as shown in Figure 3.

Method. We propose that acquired lexical knowledge can facilitate subsequent learning in several ways. The central principle is that with more words learned, the subsequent learning only needs to deal with a simpler learning problem in a smaller hypothesis space. More specifically, we propose and implement three mechanisms to utilize previously exposed word-meaning pairs in cumulative subsequent learning. First, previous learning episodes can identify many words that are function words and irrelevant to concrete meanings, such as "is", "the" and "it". To do so, we add a "NON" item in each meaning stream, compute association probabilities of word-"NON" pairs, and then select those words with high association probabilities. At the end of each learning episode, the cumulative method updates a list termed non-grounded list by adding new items, many of which are function words (e.g. pronouns). At the beginning of next episode, we then first filter the input data stream and remove those items that are in the non-grounded list. Consequently, the number of words is significantly reduced in new episode because as an experienced learner, the model already knows that the words in non-grounded list are irrelevant to any concrete meaning. In this way, previously acquired lexical knowledge is directly used to narrow the search space.

Second, the simulated learner has already acquired a set of correct word-referent pairs from previous episodes. Therefore, we maintain a list of learned word-meaning pairs and then recruits those pairs in subsequent learning. For example, if three words and three objects are temporally co-occurring in a current episode, there are 9 possible word-meaning associations. However, if the model knows that one of the words and one of the objects are reliably associated, they should be removed from the input and as a result, we need to estimate only 4 possible association probabilities. To summarize so far, the effect of the above two steps, as the introduction of knowledge learned before, is to reduce the number of items in both the word stream and the meaning stream. In this
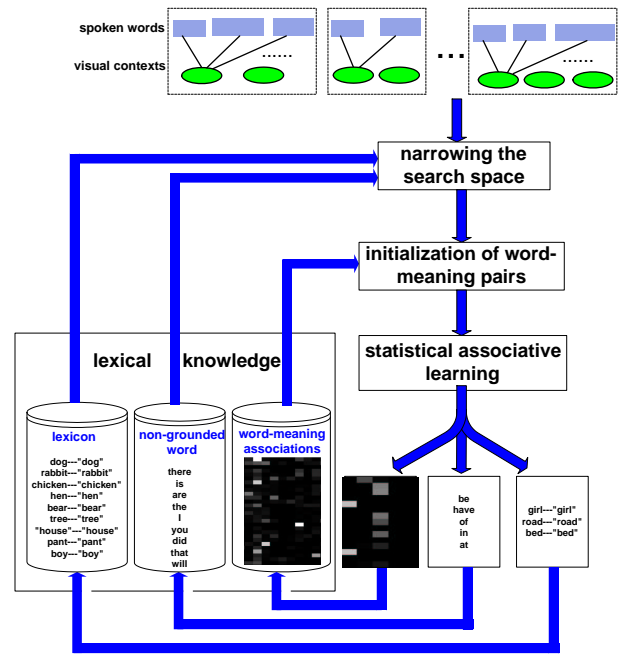

Figure 3: **Cumulative Learning.** The previously learned knowledge is stored and then utilized in subsequent learning.

way, word-to-world mappings are constrained by previously learned lexical knowledge, which leads the model to focus on some possible word-meaning mappings and rule out others. This process not only significantly reduces the computational load of the algorithm but also makes statistical associative computations in subsequent episodes more accurate and effective by removing irrelevant word-meaning associations.

The third mechanism is to utilize previously exposed (but not learned) word-meaning pairs (e.g. *mom*-"mom" in Figure 1), which we term partial lexical knowledge. This kind of knowledge corresponds to a larger proportion in Figure 1 (gray or dark areas) compared with learned word-meaning pairs (white areas). Based on previous exposures, language learners somehow know that a word is taught before but are uncertain about which meaning goes for this word. Thus, language learners may accumulate partial lexical knowledge that cannot be directly detected from standard familiar testing methods. Nonetheless, this knowledge could also play a role in subsequent learning. For instance, assume that a word-meaning pair (e.g. *mom*-"mom") is not spotted from previous learning episodes because the corresponding association probability is not significant enough. In subsequent episode, we can initialize the association probability of this pair using the previous result rather than based on a flat distribution, then it is more likely that when the EM-based learning algorithm converges, the new association probability will be increased based on the initial value and the model will be more likely to acquire this pair. More generally, many learning algorithms in computational modeling could be formalized in terms of the optimization problem with constraints – finding a set of the parameters (association probabilities in our case) that correspond to global maxima or minima of an objective function. In this context, initial values of those parameters determine where the algorithm starts from and significantly influence where it will finally converge to. In light of this, using partial lexical knowledge to initialize association probabili-

ties of word-meaning pairs can lead the model to favor some interpretations of latent lexical items underlying the training data over others. In this way, the same statistical associative learning mechanism can potentially get much better results.

Results. The data collected from six picture-book reading sessions is the input to the simulated learner and we compare the vocabulary growth of one-session learning and cumulative learning as illustrated in Figure 4. Our results are quite in line with evidence from other studies (e.g. Elman et al., 1996; Bloom, 2000; Ganger & Brent, 2004), suggesting that the pace of vocabulary development exhibits a gradual linear increase and there is no qualitative shift. Moreover, our work shows that cumulative knowledge contributes to the increase of the learning rate. From this perspective, the computational model provides a plausible mechanistic explanation of why the rate of vocabulary learning increases. However, since the model does not encode other cognitive changes, it does not rule out the possibility that vocabulary spurt does exist due to other reasons. Hence, we suggest that the increase of the learning rate is partially due to accumulated results of previous exposures.
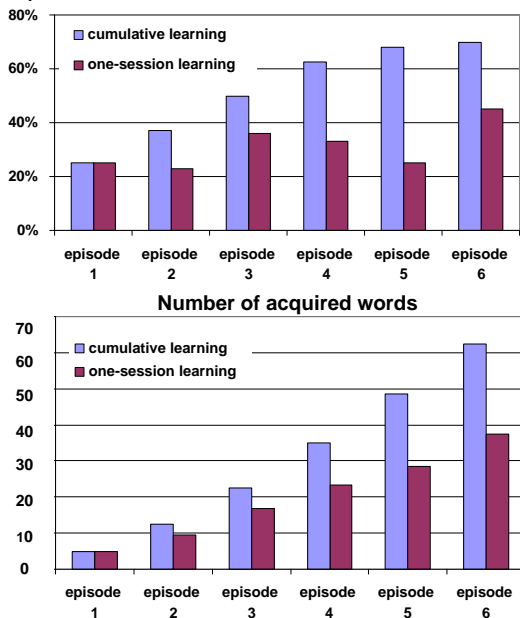

Figure 4: **Vocabulary growth.** Top: the results on each individual session. One-session learning purely depends on the co-occurrences of words and meanings in the training data while the performance of cumulative learning improves with more knowledge acquired and used. Bottom: the accumulated results. The increase in cumulative learning is significantly greater than that of one-session learning.

As shown in Figure 5, the model is also able to learn correct word-meaning associations based on a few exposures, which is intended to simulate fast mapping. Again, previously learned lexical knowledge plays a key role by reducing the hypothesis space. With more knowledge, the model becomes more "confident" associative learner while applying the same statistical learning machinery. One way to explain fast mapping is to use the concepts of recall and precision. In the context of modeling lexical acquisition, we define precision as the proportion of selected word-meaning pairs that the model correctly acquires and recall is defined as the proportion of correct word-meaning pairs (among all the pairs) that are learned by the model. Generally, we can tune up the
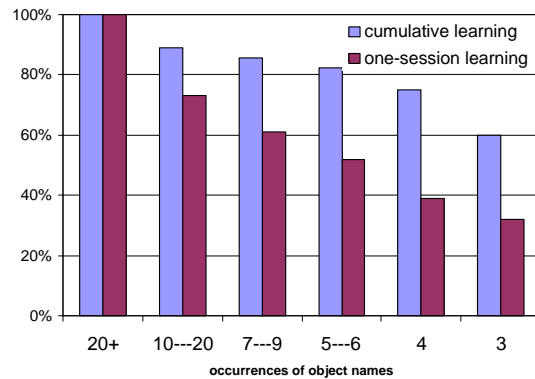

Figure 5: **Fast mapping.** Both one-session learning and cumulative learning can acquire the correct word-meaning pairs if the words and the meanings co-occur more than 20 times during the six episodes. With the decrease of the times of co-occurrence, the performance of both approaches get worse accordingly. However, cumulative learning maintains relatively good performance that is much better than one-session learning.

parameters in the model to trade off precision and recall. For instance, we can easily increase the performance of recall by discovering more word-meaning pairs, among which some of them are relevant and many others are not. Meanwhile, we have to accept the decrease in precision. One intriguing question is how children learn words based on only one or a very few exposures to get a good recall while making few mistakes (maintaining high precision). The results here cannot perfectly simulate the learning abilities of young children but are strong enough to suggest a promising direction. In future work, we plan to collect more data to explore the role of previously acquired knowledge in subsequent learning.

5. General Discussions and Conclusion

McClelland, McNaughton, and O'Reilly (1995) suggested that the discovery of structure in the environment is based on gradual learning in which changes in connection weights in neural networks result from whole ensembles of inputs rather than the single most recent experience. Our model estimates the association probabilities of word-meaning pairs through multiple learning episodes, suggesting a similar interpretation of developmental changes in early word learning. We 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 and it is certainly consistent with many formal theories of learning. 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.

We also want to note two major assumptions in this computational study: (1) young children can segment words from continuous speech; and (2) they can recognize visual objects in the picture books. These two assumptions are addressed in Yu et al. (in press), 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-meaning pairs from unprocessed multisensory signals collected in everyday contexts. Nonetheless, the focus of this work is to understand the mechanistic nature of developmental changes of vocabulary growth. To do so, we simplify some aspects of the learning to focus on the key issue – the word-to-world mapping problem.

Moreover, we argue that statistical learning is just one of important driving forces in language acquisition. In addition to distributional information, there are at least two other important factors. The first one is about social cues. It has been shown that social cues, such as joint-attention, guide children to find the referents of words (Baldwin, 1993; Yu & Ballard, 2004b). In addition, Gleitman (1990) proposed that syntactic information is also a potentially powerful cue for the acquisition of meaning. In future work, we will study how these cues interact with statistical cues and how all these factors could be integrated in a general learning mechanism.

Acknowledgments

References

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, United Kingdom: Cambridge University Press.

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

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. In *Papers and reports on child language development.* Stanford University.

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

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

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

Gleitman, L., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. (2005). Hard words. *Language Learning and Development*, *1*.

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

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.

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

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

Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, *385*(6619), 813-815.

MaShane, J. (1979). The development of naming. *Linguistics*, *17*, 79-90.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419-457.

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

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

Quine, W. (1960). *Word and object.* Cambridge, MA: 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 cognitive science society* (p. 815-820). Mahwah,NJ: ErlBaum.

Richards, D., & Goldfarb, J. (1986). The episodic memory model of conceptual development: An integrative viewpoint. *Cognitive Development*, *1*, 183-219.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, *35*, 606-621.

Schafer, G., & Plunkett, K. (1998). Rapid word learning by 15-month-olds under tightly controlled conditions. *Child Development*, *68*, 309-320.

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* (p. 51-80). Oxford: Oxford University Press.

Smith, L., Jones, S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*, 13-19.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13- month-old infants. *Cognitive Psychology*, *29*, 257-302.

Woodward, A., Markman, E., & Fitzsimmons, C. (1994). Rapid word learning in 13- and 18-month-olds. *Developmental Psychology*, *30*, 553-566.

Yu, C., & Ballard, D. H. (2004a). A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception*, *1*(1), 57-80.

Yu, C., & Ballard, D. H. (2004b). A unified model of early word learning: Integrating statistical and social cues. In *Proceedings of the 3rd international conference on development and learning.* San Diego, CA.

Yu, C., Ballard, D. H., & Aslin, R. N. (in press). The role of embodied intention in early lexical acquisition. *Cognitive Science*.