Developmental Science

Developmental Science (2010), pp 1–16 DOI: 10.1111/j.1467-7687.2010.00958.x

PAPER

What you learn is what you see: using eye movements to study infant cross-situational word learning

Chen Yu and Linda B. Smith

Department of Psychological and Brain Sciences, Indiana University, USA

Abstract

Recent studies show that both adults and young children possess powerful statistical learning capabilities to solve the word-to-world mapping problem. However, the underlying mechanisms that make statistical learning possible and powerful are not yet known. With the goal of providing new insights into this issue, the research reported in this paper used an eye tracker to record the moment-by-moment eye movement data of 14-month-old babies in statistical learning tasks. Various measures are applied to such fine-grained temporal data, such as looking duration and shift rate (the number of shifts in gaze from one visual object to the other) trial by trial, showing different eye movement patterns between strong and weak statistical learners. Moreover, an information-theoretic measure is developed and applied to gaze data to quantify the degree of learning uncertainty trial by trial. Next, a simple associative statistical learning model is applied to eye movement data and these simulation results are compared with empirical results from young children, showing strong correlations between these two. This suggests that an associative learning mechanism with selective attention can provide a cognitively plausible model of cross-situational statistical learning. The work represents the first steps in using eye movement data to infer underlying real-time processes in statistical word learning.

Introduction

There is growing interest in the idea of language learning as a form of data mining. Structure that is not obvious in individual experiences or small bits of data is derivable from statistical analyses of large data sets (Landauer & Dumais, 1997; Li, Burgess & Lund, 2000; Steyvers & Tenenbaum, 2005; Chater & Manning, 2006). These techniques have been shown to be powerful in capturing syntactic categories (Mintz, Newport & Bever, 2002; Monaghan, Chater & Christiansen, 2005), syntactic structures (Elman, 1993; Solan, Horn, Ruppin & Edelman, 2005) and word boundaries (Christiansen, Allen & Seidenberg, 1998). Also growing are suggestions (as well as relevant evidence) that statistical learning characterizes *early language learning* and that infants and young children are powerful statistical learners who make what seem to be sophisticated statistical inferences from even quite limited data (Saffran, Aslin & Newport, 1996; Newport & Aslin, 2004; Xu & Tenenbaum, 2007) .

What is not so clear, however, is the nature of underlying statistical learning mechanisms. The working assumption seems to be that learners first accumulate, more or less comprehensively, the data that are available and then apply special statistical computations to that data (Siskind, 1996; Xu & Tenenbaum, 2007; Frank, Goodman & Tenenbaum, 2009). In this paper, we explore moment-by-moment attention of infants in one kind of statistical learning task and find that statistical learning is itself tightly linked to the momentary dynamics of attention and when the momentary dynamics of attention are considered, cross-situational statistical learning is explainable by simple associative mechanisms. The results suggest that momentary selective attention in the course of statistical learning is both dependent on and indicative of learning. The experiments specifically concern infants' cross-situational learning of names and referents. We use eye-tracking measures of attention during individually ambiguous training trials and data-mine such fine-grained temporal data to discover reliable patterns that are predictive for successful learning. To better understand the link between individual attentional patterns, we use an associative model that links individual differences in looking patterns to individual differences in learning.

The findings are relevant to one of the most fundamental problems in word learning. Mapping meanings onto their corresponding lexical forms in naturalistic environments is hard in that often there are many possible referents and many possible words simultaneously present at any single learning moment. Moreover, there are different kinds of words with different kinds of meanings: some words refer to concrete meanings, such as object names; some refer to more abstract noun meanings such as *idea* and *thought*; some refer to verbs, adjectives and spatial terms, and others may be function

Address for correspondence: Chen Yu, 1101 East 10th Street, Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405, USA; e-mail: chenyu@indiana.edu

words and not referential at all. How do children acquire the meaning of various kinds of words? Gleitman and colleagues (Snedeker & Gleitman, 2004; Gleitman, Cassidy, Nappa, Papafragou & Trueswell, 2005) suggested that the learning input for word acquisition is much broader and more varied than previously acknowledged to the degree that the major problem for word learning is not a 'poverty of the stimulus' (Chomsky, 1959) but a 'richness of the stimulus'.

It is well accepted that this rich input requires constraints to reduce the degree of ambiguity. In the case of object name learning, those constraints include attentional biases to attend to the shape of the whole object (Landau, Smith & Jones, 1988), conceptual biases that make some kinds of word-meaning mappings more likely than others (Gelman & Taylor, 1984; Markman, 1990; Golinkoff, 1994; Golinkoff, Jacquet, Hirsh-Pasek & Nandakumar, 1996; Klibanoff & Waxman, 2000; Booth & Waxman, 2002), and by all sorts of linguistic bootstraps whereby children use the words and linguistic structures they already know to help figure out new meanings (Gleitman, 1990; Naigles, 1990; Newport, 1990; Fisher, Hall, Rakowitz & Gleitman, 1994; Imai & Gentner, 1997; Monaghan *et al*., 2005). At the macro-level, we know a great deal about what those constraints are and how they work individually. However, at the micro process level, we are only beginning to understand how they work in real time and how a general learning process integrates and coordinates multiple constraints and cues (Merriman & Stevenson, 1997; Hollich, Hirsh-Pasek, Golinkoff, Brand, Brown, Chung, Hennon & Rocroi, 2000; Halberda, 2003; Snedeker & Gleitman, 2004; Halberda, 2006; Swingley & Aslin, 2007). The purpose of the present study is to contribute to an understanding of the micro-level processes that support real-time learning using object name learning from multiple ambiguous cross-situational observations as a case study. We selected object name learning because compared with learning other kinds of words with more abstract meanings, the basic-level word–object mappings are relatively concrete and learning itself, at least by very young children, is less likely to be influenced by high-level conceptual components. Although one would like a micro-level real-time account of conceptual influences as well, it seems prudent to start at a more concrete level. Thus we focus on one fundamental component of early noun learning: mapping words to candidate referents.

The general learning paradigm used here is cross-situational word learning. Accruing information about name–object pairings across individual learning experiences has been proposed as a solution to the uncertainty inherent in trying to learn nouns from their co-occurrences with scenes (Siskind, 1996; Yu & Smith, 2007; Smith & Yu, 2008). Scenes typically contain many possible referents, with speakers talking about and shifting their attention rapidly among the potential referents. This uncertainty is still considerable even if one assumes that a learner is biased to link names to whole objects (Markman, 1990). For example, a young learner may hear the words 'bat' and 'ball' in the ambiguous context of seeing both a BAT and BALL without any information as to which word refers to which scene element. However, although the learner may have no way of knowing from any such *single* learning situation which word goes with which referent, the learner could nonetheless determine the right mappings *if* the learner kept track of co-occurrences and non-occurrences *across situations*, and evaluated the cross-situational evidence for word–referent pairings in the proper way. Using the above example, if the learner viewed a second scene while hearing the words 'ball' and 'dog' and if the learner could remember and combine the conditional probabilities of co-occurrences from the two situations, the learner could correctly infer that 'ball' maps to BALL.

In a recent study, Smith and Yu (2008) showed that 12- and 14-month-old babies do this. They presented the infants with learning trials on which there were always two seen objects and two heard names but no information as to which name went with which object. From such individually ambiguous learning trials, the infants learned the mappings of six names to six objects and did so in a learning experience that lasted in total less than 4 minutes. The cross-trial word–referent statistics were the only information available to disambiguate those word–referent pairings. Thus the infants must have combined the information across trials. The present question is the nature of the processes that underlie this learning.

One way to attempt to understand this process is to start with the *simplest* mechanisms that are *known* to exist in the human and infant learning repertoire and see how well these simple and known mechanisms can do. One such possible learning process is Hebbian-like associative learning, a form of learning known to be fundamental to many perceptual and cognitive capabilities (Smith, 2000). In the present case, the learner could simply store all associations between words and references. With respect to the above example, if the learning system stored only associations between words and whole objects, there would be four associations formed on trial one (bat to BAT, bat to BALL, ball to BAT, ball to BALL). On the second experience containing the words 'ball' and 'dog' and the objects BALL and DOG (with four possible associations between them), one of the associations (ball to BALL, etc.) would be strengthened more than the others. Across trials, the relative strengths of associations between words and their potential referents would come to reflect the correct word–referent mappings. Simple associative models such as this have been criticized on the grounds (Keil, 1992) that there are just too many possible associations across situations to store and to keep track of.

This raises the key question for the present study, whether learners do not actually store *all* co-occurrences, but only *some* of them. Further, we ask whether infants' attention to and thus selective storage of word–referent pairs might be guided by their previous experience. And if this is so, could eye movement patterns in training be straightforwardly predictive of successful learning at test? If all that matters for what is learned in this task is the simple co-occurrence of looking at an object while hearing its name, then the eye movement patterns should straightforwardly predict learning. To provide further evidence relevant to this idea, we also ask whether a simple associative model based on gaze data could explain not only infants' overall success in learning in this task but also individual differences in that learning. However, if looking does predict learning, then we need to know more about the looking patterns themselves. Accordingly, a major component of the present study is a deeper understanding of the dynamics of those looking patterns, how they change over the course of the learning trials, and how they relate to individual differences in learning outcomes. The issue of individual differences is particularly critical if infants are not simply passive accumulators of data but instead actively learn by selecting among the available data. If infants select some pairings over others to notice and store – and if these pairings guide later selections – then individual learners may distort the regularities in the input both in ways that enhance the learning of the right word–referent pairs and in ways that hinder them. In brief, although we use a simple (and minimal) associative model to link looking to learning, the main goal of this work is an understanding of real-time attentional processes that lead to learning.

Our empirical approach, then, is to continuously track eye-gaze direction throughout learning. The assumption here is that when a learner associates a word with a referent among other simultaneously presented referents, the learner is likely to be looking at that referent and this looking behavior indicates that the learner registers this word–object association. In this way, different learners may attend to different referents in a visual scene when hearing the same word. Further, by the assumption that learners link the word to the object they are attending to at that moment, these differences in attention will lead directly to different learning results.

Recent psycholinguistic studies already suggest that speech and eye movements are closely linked in both language comprehension and production (Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995; Griffin & Bock, 1998; Meyer, Sleiderink & Levelt, 1998; Griffin, 2004; Trueswell & Gleitman, 2004; Huettig & Altmann, 2005; Knoeferle & Crocker, 2006). For example, Griffin and Bock (1998) demonstrated that speakers have a strong tendency to look toward objects referred to by speech and that words begin roughly a second after speakers gaze at their referents. Meyer *et al.* (1998) found that when speakers were asked to describe a set of objects from a picture, they usually looked at each new object before mentioning it, and their gaze remained on the object until they were about to say the last word about it. Several recent developmental studies, though not addressed to the specific questions in this paper, have shown the utility of using these finer-grained real-time measures in studies of early development and learning (von Hofsten, Vishton, Spelke, Feng & Rosander, 1998; Johnson, Amso & Slemmer, 2003; Aslin & McMurray, 2004; Trueswell & Gleitman, 2004; Halberda, 2006; Plunkett, Hu & Cohen, 2008). Motivated by those studies, we also apply this eye-tracking paradigm to early word–referent learning and use eye movements, and the synchrony of those movements with respect to the heard object names, as a measure of moment-by-moment learning and as a clue to the momentary internal states of the learner.

Method

The stimuli used are exactly the same as those in Smith and Yu (2008) and the overall procedure is also very similar. Therefore, we expect to replicate these previous results, that from individually ambiguous learning trials, infants nonetheless learn the underlying word–referent pairings. More importantly, the present study records moment-by-moment eye movement data while infants are engaged in statistical learning. Such fine-grained data are used to generate new results and new insights into the underlying mechanisms, and more specifically to test the idea that an associative learning model with selective attention to some pairings can explain this learning. Further, our data analysis and computational modeling provide a strong test of this proposal by providing evidence on both the group and individual learner levels.

Participants

The final sample consisted of 18 14-month-olds (10 boys, eight girls), with a mean age of 14.2 (SD = 0.5) months. An additional 14 infants were tested but not included in the sample due to fussiness (n = 4), persistent inattention to the display (n = 2), and mostly occasional but large movements that prohibited the complete collection of continuous eye movement data with the eye tracker (n = 8).

Stimuli

The six 'words' *bosa*, *gasser*, *manu*, *colat*, *kaki* and *regli* were designed to follow the phonotactic probabilities of American English and were recorded as single-word utterances by a female speaker in isolation. They were presented to infants over loudspeakers. The six 'objects' were drawings of novel shapes; each was a unique bright color. On each trial, two objects were simultaneously presented on a 17-inch computer monitor. The projected size of each object is 7.6 × 6.4 cm (7.2 × 6.1° visual angle at the infant's 60-cm viewing distance) separated by 11.4 cm.

There were 30 training slides. Each slide simultaneously presented two objects on the screen for 4 sec; the onset of the slide was followed 500 ms later by the two words – each said once with a 500-ms pause between. Across trials, the temporal order of the words and the spatial order of the objects were varied such that there was no relation between the temporal order of the words and the spatial position of the referents. Each correct word–object pair occurred 10 times. The two words and two objects appearing together on a trial were randomly determined such that each object and each word co-occurred with every other word and every other object at least once across the 30 training trials (thereby creating the within-trial ambiguities and possible spurious correlations). The first four training trials each began with the centered presentation of a Sesame Street character (3 sec) to orient attention to the screen. After these first four trials, this attention-grabbing slide was interspersed every 2–4 trials to maintain attention. The entire training – an effort to teach six word–referent pairs – lasted less than 4 min (30 training slides and 19 interspersed Sesame Street character slides).

There were 12 test trials each lasting 8 seconds. Each test trial presented one word, repeated four times with two objects – the target and a distracter – in view. The distracter was drawn from the training set. Each of the six words was tested twice. The distracter for each trial was randomly determined such that each object occurred twice as a distracter over the 12 test trials. This duration and structure of training and test trials was the same as in Smith and Yu (2008). However, the manner of presentation differs. Whereas Smith and Yu (2008) used a large (47 by 60 in) screen and measured total looking time, here we attempted to replicate these results using a computer screen presentation and measure moment-to-moment eye gaze using an eye tracker.

Apparatus

The learners' eye gaze was measured by a Tobii 1750 eye tracker with an infant add-on (http://www.tobii.se). The principle of this corneal reflection tracking technique is that an infrared light source is directed at the eye and the reflection of the light on the cornea relative to the center of the pupil is measured and used to estimate where the gaze is fixated. The eye-tracking system recorded gaze data at 50 Hz (accuracy = 0.5°, and spatial resolution = 0.25°) as a learner watched an integrated 17-inch monitor with a resolution of 1280 × 1024 pixels.

Procedure

Infants sat on a parent's lap 60 cm from the 17-inch computer monitor used to present the stimuli. Before the experiment, a calibration procedure was carried out. In preparation for the calibration, the experimenter adjusted the eye tracker to make sure that the reflections of both eyes were centered in the eye-tracking camera's field of view. We used a procedure including nine calibration points. The total duration of the calibration procedure was about 3 minutes before the training trials started. Parents were instructed to close their eyes during the whole procedure and not to interact with the child during the experiment.

Data

The eye tracker outputs (x,y) coordinates on the computer display of the visual presentation at the sampling rate of 50 Hz. There are in total 120 sec (4 sec/per trial × 30 trials) during training and 96 sec (8 sec/per trial × 12 trials) during testing. Therefore, there are 6000 data points in training and 4800 data points in testing, if the eye tracker works perfectly. In practice, the tracking system occasionally failed to detect the subject's eye gaze for two potential reasons: either because participants looked away and their head and gaze moved outside of the tracking plane, or the eye tracker could not correctly compute the subject's eye movements for some other reasons. For the 18 infants with good tracking results, the average tracking success is 79% in training and 69% in testing.[1] Thus, on average, we collected 4740 data points in training and 3321 data points in testing per subject, which were used in the following data analysis and modeling.

Behavioral results in testing

Infants were presented with 30 training trials (two words and two objects per trial) and then 12 test trials in which one target word was played and two objects (the correct referent and the distracter) were displayed. Infants' preferential looking on such test trials is commonly used as a measure of language comprehension (Hirsh-Pasek & Golinkoff, 1996; Schafer & Plunkett, 1998; Golinkoff, Hirsh-Pasek & Hollich, 1999; Halberda, 2003; Waxman & Lidz, 2006) in that infants systematically look at the portion of the display that corresponds to what they are hearing, and this was the behavioral measure of learning used by Smith and Yu (2008). Accordingly, the first question we addressed was whether this study replicated the previous result: did infants during the test trials look longer at the correct referent for the heard word than the distracter? In order to directly compare our eye-tracking data with human coding data reported in the prior study (Smith & Yu, 2008), we processed eye movement data by simply splitting the screen into left and right sides, and

[1] In our experiment, we used a digital camera pointing to the face of the infant in the whole session and did manual coding of video clips from three participants as a reliability analysis. The results show that the lost data were mostly caused by infants looking away from the screen (85%) but also occasionally by the eye tracker's error in which case the infant was attending to the screen but the eye tracker failed to generate gaze data.

then converting (x,y) coordinates from the Tobii eye tracker into left/right coding. We found that infants looked at each of the 8-second test trials for an average of 5.92 seconds. More importantly, they showed a longer looking time to the targets ($M = 3.25$, $SD = 0.49$) than distracters ($M = 2.67$, $SD = 0.38$; $t(17) = 5.26$, $p < .01$). Further, we compared their looking time trial by trial for each of six words. A word is counted as learned if infants looked more on the target than the distracter. Based on this measure, on average, four of the six words are learned by infants, indicted by longer looking time to targets. Among 18 infants, the four best learners acquired five of the six words, six infants acquired four words, five of them acquired three words, one acquired two words, and two of them one word.[2] Thus, this study replicates the earlier finding that very young word learners can learn word–referent pairings from individually ambiguous learning experiences by combining the information across those experiences. However, they also show that there are individual differences; whereas most infants appear to learn more than half of the words, some appear to learn very few if any.

The main purpose of this study is to use eye movement data to reveal new insights about the underlying learning processes and, more specifically, to examine the relation between the selectivity of looking on the training trials and learning on the test trials. To this end, we first report a set of data analyses on eye movement data, and then introduce a simple associative model that makes minimal processing assumptions, takes the micro-structure of the eye-tracking data during training as input and predicts individual performance on the test trials.

Eye movement data processing and analysis

Under the assumption that learning requires attending to the object as a potential referent, the goal was to measure object fixations and not merely the side of the screen to which the infant was looking. Accordingly, we developed a simple region-based fixation finding method. While looking at an object, infants may switch their gaze from one part of the object to another part, yet still be attending to the same object and linking that object to the heard word. Thus, we defined two rectangle regions-of-interest (ROIs) that cover the areas of two visual objects displayed on screen. Each ROI covers the area occupied by one of two visual objects with a 10-pixel margin along four directions. We then grouped raw eye position data (x and y coordinates) that fell within the same ROI as a single fixation. This process converts continuous gaze data into three categories, namely, left object, right object, or somewhere else. One potential problem with this thresholding-based approach is that it cannot handle data points close to the boundaries of ROIs. For example, a single data point within a segment of a continual gaze data stream that is just out of the predefined ROI would split the whole segment into two fixations instead of maintaining one big fixation. In order to generate more reasonable results and remove artificial effects from the thresholding method, two additional data processing steps were applied to smooth fixation data. First, we merged two consecutive fixations sharing the same target object into one big fixation if the gap between these two was small enough (< 60 ms or 3 data points). This smoothing step was based on the assumption that a short period of time out of an ROI is likely to be caused by artificial effects of the thresholding-based method because a participant is less likely to switch their visual attention to the background (e.g. the middle of the screen with nothing displayed, etc.) and immediately switch back to the target object in such a short period of time. The second step was to remove those short fixations that lasted less than 60 ms (3 data points). Again, we suspected that those fixations were likely caused either by accidental eye-tracking errors or by the thresholding-based fixation finding method. The final result of this thresholding and smoothing is an event stream with each fixation entry consisting of three elements (t1, t2, target) representing the onset of a fixation, the offset of the fixation, and the target object fixated upon, respectively. Figure 1 shows the results of eye fixation data in which each color represents one of six visual objects that participants attended trial by trial.

The next goal in the analyses was to discover the nature of looking patterns during training, and particularly those that might lead to more successful learning. As noted in the previous section, looking times to targets and distracters at test clearly suggested that some infants learned relatively many word–referent correspondences and others learned very few if any. Accordingly, we grouped participants into strong or weak learners based on their (overall) preferential looking results in testing. We then extracted and compared the eye movement patterns characteristic of these groups during the learning phase. The grouping rule was straightforward and meant an approximate division by learning. To group the subjects, we used the overall accumulated looking time on targets versus on distracters to group participants. Specifically, participants who spent absolutely more time on target objects than distracters during testing were counted as a stronger learners: 12 out of 18 were categorized in the strong learner group and the other six were in the weak learner group.

Note that this grouping is one of several possible ways to distinguish strong and weak learning in this task. An alternative way to identify strong learners is to count the number of learned words based on individual learning trials. We decided to use the overall looking metric since the following analyses focus on general dynamic gaze patterns over the course of training but not those patterns at the individual word level. However, the grouping results from individual words are nearly identical to those based on accumulated looking patterns, suggesting

[2] This result is also graphically shown in Figure 6 below with modeling results.

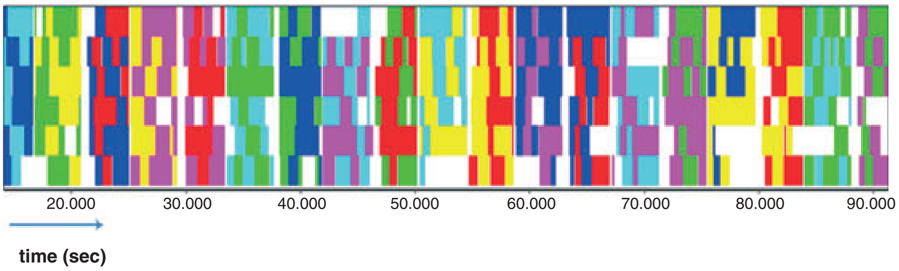

Figure 1 *Eye movement data trial by trial over the course of training. Each row shows an eye fixation stream collected from one participant. There are in total six different colors used in those streams, each of which corresponds to one of six visual objects. From this example, we can see that eye movements are very dynamic and spontaneous with different numbers of fixations, different durations for those fixations and different timings when they start and end. There are clearly individual differences across young infants.*

that slightly different grouping schemes would not change the results.

The empirical question is this: What looking patterns signal better learning? The average looking time during training for the strong and weak learners does not differ significantly (M = 2.96 s for the strong and M = 3.07 s for the weak for these 4-sec training trials). Thus, both stronger and weaker learners look at the training slides. We next calculated the average number of attention switches (between left and right objects) and the average length of the longest fixations within every training trial. The results show that weak learners generate more fixations per trial (M = 3.82) compared with strong learners (M = 2.75, $t(58) = 3.65$, $p < .001$) who generated longer eye fixations on attended objects (M = 1.69) than weak learners (M = 1.21; $t(58) = 2.52$, $p < .01$). Thus, the finer-grained nature of looking patterns between the two groups differs in that strong learners have more stable eye movement patterns of sustained attention characterized by fewer but longer fixations. In this regard, these results differ from the findings from infant visual recognition memory (Rose, Feldman & Jankowski, 2004) showing that shorter looks and more shifts (interpreted as faster processing) lead to better recognition memory. Remembering the objects one has seen and binding words to objects (and remembering that binding) are profoundly different tasks, and thus the discrepancy may indicate that infants require more sustained attention to cognitively register word–object associations.

To this point in the analyses, the measures lump looking over the whole set of training trials together. Thus, they cannot tell us whether the differences between infants who show stronger learning at test and those who show weaker learning at test are there from the start (as perhaps intrinsic individual differences). Alternatively, the differences may emerge over the course of the training events. Figure 2 shows that these differences indeed emerge over training. Figure 2(a) shows the average number of eye fixations over 30 individual training trials. Both infants who show strong learning at test and those who show weak learning at test have similar eye movement patterns at the beginning, but roughly around 3–5 trials into the experiment, their looking patterns begin to diverge. Weak learners generated more and shorter fixations while strong learners maintained more stable attention switches through the whole training. At the end of training, both strong and weak learners had similar overall attention switches again.

The number of attention switches is just one aspect of the dynamics of eye movements that might be relevant to learning. With the same number of attention switches (e.g. at the end of training), strong and weak learners can generate different looking durations. For example, one group might have more or less uniform looking durations. The other group might have a more uneven dis-

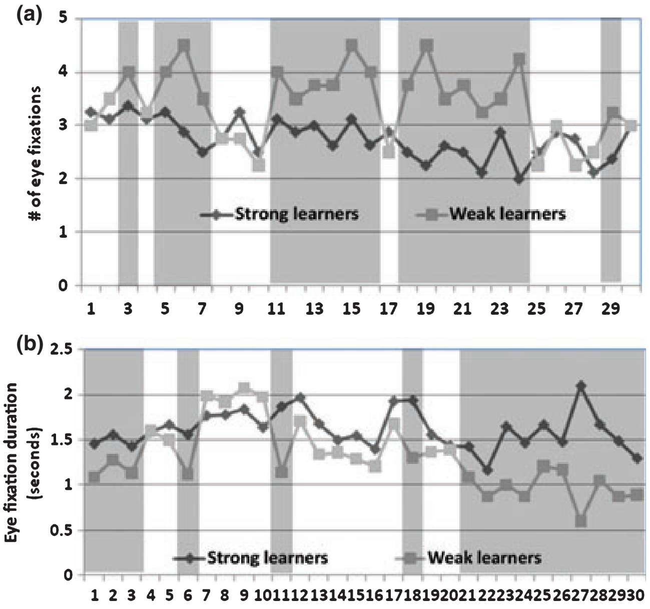

Figure 2 *The dynamics of eye fixations trial by trial: (a) we measured the average number of eye fixations per trial and showed the dynamic changes of this measure across trials; (b) we measured the average length of the longest fixation over training trials. From both measures, there are significant differences between strong and weak learners. Shaded areas (trials) are statistically significant* (p < *.005) based on pairwise* t-*test.*

tribution of looking durations containing one or two longer fixations with several shorter attention switches. To capture the dynamics of looking durations, Figure 2(b) shows the average lengths of the longest fixations over training trials. A longer fixation may indicate more stable attention and therefore better learning. Infants who showed stronger learning at test consistently generated longer fixations trial by trial while the infants who showed weaker learning at test always produced shorter fixations especially at the end of training. The overall results from Figure 2(a) and (b) suggest that strong and weak learners show quite different eye patterns from the beginning to the end of learning. On the early learning trials, weak learners tend to generate more and shorter fixations. However, the most significant differences are in the middle of learning; weak learners generate more attention switches than strong learners. Moreover, given the approximate same number of attention switches (given that there are no major differences in the longest looking duration), the other fixations generated by strong learners are more stable than those in weak learners. Finally, the end of training is clearly characterized by longer fixations by the stronger learners. Importantly, the measures in Figure 2 do not take into account where the infants are looking, but only the overall timing and dynamics of eye movements. Thus, even though some aspects of overall eye movement patterns are similar at the end of training, there may be differences in whether learners are looking at the correct or incorrect target at the moment it is named.

To further quantify the dynamics of eye movement patterns during training, we introduce a new metric based on information entropy (Cover & Thomas, 2006). Assume that a participant generated L fixations in the tth learning trial ($1 \leq t \leq 30$) and each fixation f_m lasted a certain period of time $T(f_m)$, a sequence of eye fixations generated by a participant within the learning trial can be viewed as a probabilistic distribution after we normalized each fixation time by the overall looking time within the trial. The entropy of eye movements within the tth trial can then be calculated as follows:

$$E(t) = -\sum_{m=1}^{L} \frac{T(f_m)}{\sum T(f_m)} \log \frac{T(f_m)}{\sum T(f_m)}$$

According to this measure, more eye fixations within a trial increase the span of the distribution and hence cause the increase of the entropy. For example, the entropy of a learning trial with two eye fixations is very likely to be lower than that of a learning trial with four fixations. Moreover, the same number of fixations more evenly distributed creates high entropy while an uneven distribution with some longer and some shorter fixations decreases the entropy of eye movements. If more rapid attention switches within a learning trial and more even looking times indicate a participant's uncertainty about word–referent pairings, then the entropy of eye movements is a potential measure of the participant's internal learning state over the course of learning. Figure 3 shows the results of the entropy measures from both strong and weak learners. Again, we found overall similar entropy results between the two groups at both the beginning and the end of training *but* significant differences in the middle of training.

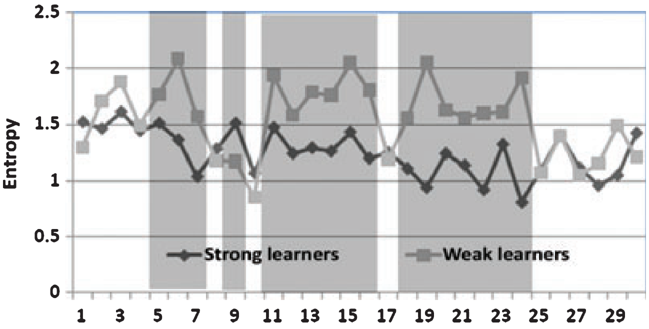

Figure 3 *The dynamics of eye movement entropy in training trials. The differences between strong and weak learners are significant during the middle of the training. Shaded areas (trials) are statistically significant (*p < *.005) based on pairwise* t*-test.*

Two implications follow from this result. First, weak learners seem to be more uncertain about the pairings in the middle of training. Second, even though the degrees of uncertainty for the two groups are similar at the end of training, we do not know, with respect to the present measure, what might have been learned or not learned. That is, the present metric does not contain information as to whether the learner is looking at the correct or incorrect target. Thus, both learners may converge to some word–object pairings at the end; however, weak learners may have lower degrees of uncertainty (just like stronger learners) but on wrong pairings. To check this, we next measured the entropy of eye movements based on whether the participants looked at the *correct* referent after hearing a word. To do so, we further decomposed eye movements in each trial into two groups based on the timings of two spoken words. More specifically, eye movements that were generated between the 300 ms after the onset of a spoken word and the onset of the other word (or the end of the current trial) were grouped together and labeled as eye movements driven by the concurrent spoken word. We chose 300 ms based on the approximation that it took participants at least 200 ms to generate an eye fixation and they might be able to recognize to-be-learned words after they heard the first part of a word (100 ms).[3] Since each word occurs 10 times in the 30 training trials, we averaged the looking time entropy of each appearance of each word by aggregating the results from individual words. Figure 4 shows significantly different patterns between strong

[3] Two other timing offsets (200 ms and 400 ms) were selected but this parameter didn't make any difference in our results.

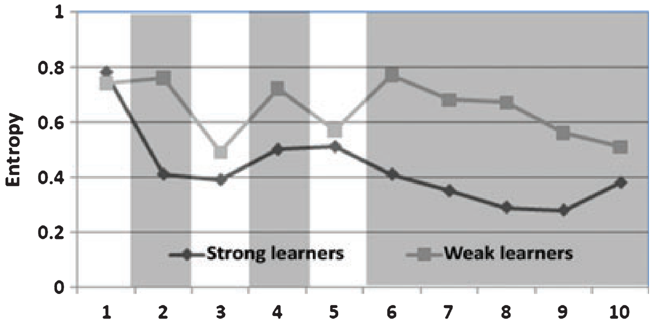

Figure 4 *The dynamics of eye movement entropy based on the temporal occurrences of to-be-learned words. Shaded areas (trials) are statistically significant (*p < .005*) based on pairwise* t*-test.*

learners and weak learners for each of the 10 presentations of the word.

As can be seen in Figures 3 and 4, by all these measures weaker learners show greater entropy in the middle of the training trials. What might this dynamic pattern mean? One possibility is that strong and weak learners are alike at the beginning because they enter the task with the same knowledge, not knowing any of the word–referent pairs. All learners on the initial trials must randomly sample some word–referent pairs by selectively looking at one of the two objects when hearing one of the words. Young children in both groups may start this sampling process similarly and thus there are no differences in their eye movement patterns. Following this line of reasoning, the diverging patterns of learning that then follow this beginning could be the consequences of different learning in this initial similar phase that then sets up the different looking patterns in the middle, leading to more successful or less successful resolution of the statistical ambiguities inherent in the learning trials. With this conjecture in mind, we built a simple associative learning model and fit real-time looking data during training to predict the outcome measures at test.

The model

The primary goal of this modeling effort is to link the fine-grained analyses of looking behavior observed in the experiment to learning as measured at test. Although the underlying processes may be more complex, for this linking goal we assume only that co-occurrences of attended objects and heard words matter. Thus, for example, the model does not include the selective attention component of the learning process, nor the processes through which associative learning might train attention (see Smith, 2000, for review of relevant evidence). Instead, selective attention is implicit in momentary gaze data fed into the model. In brief, the model is purposely conceptually simple so as to minimize assumptions and thus to allow us to see the structure in the looking patterns generated by the infants and, most critically, the link between looking and learning outcome at test. By assuming very little about the learning mechanism beyond co-occurrence, we may be able to discern just how tightly looking patterns are linked to learning outcome.

An associative learning mechanism strengthens the link between a word and a referent if these two co-occur regularly across multiple trials and weakens the link if the word and the referent do not co-occur. In the experiment, infants were exposed to six words and six pictures in total. Therefore, a 6 by 6 association matrix as shown in Figure 5(b) is used in modeling as a representation of all the possible associations that a learner *might* track. In such an association matrix, each cell corresponds to a particular word–referent association. The diagonal cells are the six correct pairings while non-diagonal cells represent spurious correlations due to the ambiguity inherent in the training trials. The model tracks association probabilities, updating them trial by trial. The model assumes that the learner makes a decision during testing in the sense of looking at the object that has been more strongly associated with the word in the learner's internal association matrix. Thus, the internal association matrix of a successful learner should be one in which the diagonal items were assigned with higher probabilities than were non-diagonal cells. In brief, by this simple model, the critical issue for learning is the specific associations that are accumulated over trials. What, then, do these internal association matrices look like?

To examine these possibilities, we build the association matrix trial by trial as follows:

$$p_{ij}(t) = \frac{t-1}{t} p_{ij}(t-1) + \frac{1}{t} \frac{\lambda(t)\eta_{ij}(t)}{\sum_j \lambda(t)\eta_{ij}(t)}$$

where t is the trial number, and $p_{ij}(t)$ refers to the association probability of the object i and the word j at the tth trial. Thus, $p_{ij}(t)$ corresponds to one cell in the association matrix which is composed of two weighted parts. The first part $p_{ij}(t-1)$ reflects the accumulated association probability so far until the $(t-1)$th trial that is carried over to the current trial. The second part (with two variables $\eta_{ij}(t)$ and $\lambda(t)$) updates the previous association probability based on a learner's eye movements in the current trial. First, rapid shifts of visual attention between possible objects after hearing a word are taken as reflecting uncertainty (that is, the lack of one stronger and several other weaker associations). In brief, we expect that the learner is more likely to consistently fixate on the corresponding referent to the degree that it is strongly associated with the target word; this is, again, the very basis of using preferential looking to measure word knowledge. This principle is encoded by $\lambda(t)$ that measures the overall degree of uncertainty in the tth learning trial from individual learners' perspectives. The more uncertain the learner is, the less likely s/he reliably registers word–referent associations. Second, the multimodal synchrony between eye movements and spoken

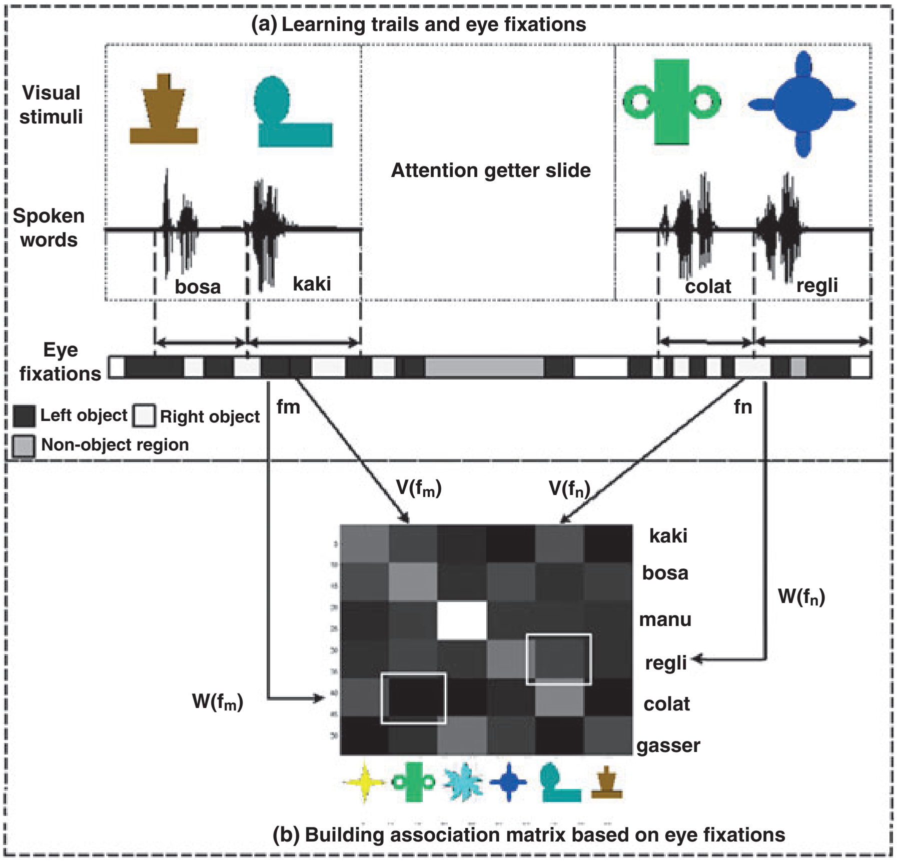

Figure 5 *(a) We measure where the learner is fixating on after hearing a spoken word. For example, after hearing the word 'bosa', there are four eye fixations on both left and right objects. Those fixations (and corresponding fixed objects) are associated with the word 'bosa'. The strength of the association between an object (left or right) and the word 'bosa' is determined by the overall duration of fixations on that particular object. (b) A 6 × 6 association matrix built based on the synchrony between a subject's eye movements and spoken words during training. Each cell represents the association probability of a word–object pair. The diagonal items are correct associations and the non-diagonal items are spurious correlations. Dark means low probabilities and white means high probabilities.*

words may indicate the strength of the registration of a certain word–referent pairing, and the duration of such synchronized behaviors may indicate how strong a word–referent association is in the learner's association matrix. This observation is captured by $\eta_{ij}(t)$ that measures the possible association between a word i and an object j at the current trial based on eye movements. In the following, we explain exactly how to estimate $\lambda(t)$ and $\eta_{ij}(t)$.

We first computed eye fixations from raw eye movement data and converted the continuous gaze data stream into a set of eye fixations marked by the onset and ending timestamps of each fixation using the same data processing method described earlier. Next, we segmented the whole set of eye fixations into individual trials by aligning eye fixations with the timing of training trials. Within each learning trial, there are multiple eye fixations on the two objects on the computer screen that occur as the two words are sequentially presented. Assume that there are L fixations $\{f_1, f_2, f_3, \ldots, f_L\}$ in the tth learning trial. For a fixation f_m, $v(f_m)$ is the object that was fixated on, $w(f_m)$ is the spoken word that the participant heard right before or during the fixation, and $T(f_m)$ is the fixation time. As shown in Figure 5, all of the eye fixations generated between the 300 ms after the onset of a spoken word and the onset of the next spoken word (or the end of the current trial) are assigned to be associated with that spoken word.

Next, $\lambda(t)$, as an indicator of the learner's uncertainty in the current trial, can be encoded as how frequently the learner moves his eyes between those objects after hearing a word in the tth trial. Therefore, we defined $\lambda(t) = \frac{1}{E(t)}$ where $E(t)$ is the same entropy measure of a sequence of eye fixations within the trial as a metric to characterize this factor. Thus, the more fixations learners generate, the more uncertain their learning state is, and therefore the less likely they successfully register word–referent associations and gain more knowledge in associative learning in the current trial.

Moreover, the second variable, $\eta_{ij}(t)$, measures the possible association between a word and an object, which is composed of two parts. The first part estimates the probability of associating an object to a particular word based on the amount of looking time at that object (compared with other objects) after hearing that word: Given multiple candidate objects, how likely is a heard word associated with each object – the competition between candidate objects for a given word. The second part estimates the probability based on comparing the looking time to the same object across several spoken words: Given multiple candidate words, how likely is an object associated with each word – the competition between candidate words for a given object. Formally, $\eta_{ij}(t)$ can be represented as follows:

$$\eta_{ij}(t) = \frac{\sum_{m=1}^{L} \delta(i, \upsilon(f_m)) T(f_m) \delta(j, w(f_m))}{\sum_{m=1}^{L} \delta(i, \upsilon(f_m)) T(f_m)} + \frac{\sum_{m=1}^{L} \delta(i, \upsilon(f_m)) T(f_m) \delta(j, w(f_m))}{\sum_{m=1}^{L} \delta(j, w(f_m)) T(f_m)}$$

where δ is the Kronecker delta function, equal to one when both of its arguments are the same, and equal to zero otherwise. Thus, the numerator of the two parts is the same, measuring the accumulated number of fixations within a trial ($T(fm)$, etc.) based on the synchrony between a certain attended object i ($v(fm) == i$) and a certain heard word j ($w(fm) == j$). The denominator in each part just normalizes the above numerator either across all the words or across all the objects, respectively. Thus, a learner's visual attention in statistical learning is directly encoded in the association matrix that the model built. For example, a longer and more stable fixation on a visual object after hearing a word increases $\eta_{ij}(t)$ and therefore strengthens the association probability $p_{ij}(t)$. To follow this example, after hearing the second word, the less amount of time that a learner attends to the same object, the less likely the learner considers that object as a candidate referent of the second word, and therefore the more likely the learner treats the object as the candidate referent of the first word instead. This observation is implemented in the calculation of $\eta_{ij}(t)$ by decreasing the denominator $\sum_{m=1}^{L} \delta(i, \upsilon(f_m)) T(f_m)$ which has the effect of increasing $\eta_{ij}(t)$ and $p_{ij}(t)$. Since individual infants generated different eye fixation sequences, the model builds different association matrices for different individuals.

In summary, our associative model attempts to build a word–object association matrix based on two variables $\eta_{ij}(t)$ and $\lambda(t)$ wherein $\eta_{ij}(t)$ captures the synchrony between visually attended objects and concurrently heard words, and $\lambda(t)$ assigns registration weights to $\eta_{ij}(t)$ based on the dynamic characteristics of eye movements (for instance, the degree of uncertainty) in the present trial. Indeed, given the entropy differences over trials between strong and weak learners shown in Figures 3 and 4, we can infer that even if the values of $\eta_{ij}(t)$ are the same with two learners, a better set of registration weights captured by $\lambda(t)$ from the learner with more stable eye fixations will make this learner build a better association matrix.

Results

Figure 5(b) shows an example of an association matrix built based on a learner's eye movements. In this example, some strong associations are correct (e.g. the word *manu* with the object manu) and others are not (e.g. the word *colat* with the object regli). Two measurements evaluated whether the associative model based on eye movements predicts individual differences in learning. First, there is a strong correlation ($r = 0.71$; $p < .005$) between the number of learned words according to the model for each infant and the number of words actually learned by that infant (as measured by absolutely greater looking times to the referent versus the distracters for that word). Second, we also found that the proportion of diagonal cells (the strength of correct word–referent associations) in an association matrix was strongly correlated ($r = 0.65$; $p < .005$) with the proportion of looking time (the degree of the preference to look) to the correct referents at test. The scatter plot for this correlation is shown in Figure 6 along with the model's accumulated matrices at the end of training for four individual participants. These four association matrices are quite different from each other and are ordered top to bottom by the strength of correct associations on diagonal cells. More specifically, most diagonal items (correct word–referent associations) in the top association matrix are highlighted while the association probabilities between words and referents are more distributed in the bottom matrix. Critically, these very different matrices are built based on the same associative learning mechanism and the same start, namely, random selection of pairs, but they accumulate data as a function of the specific eye movement patterns of the individuals. The fact that these patterns predict learning tells us that the fine-grained looking dynamics during training are relevant to the outcome. The fact that weaker and stronger learners can be modeled with the very same simple model that takes only looking patterns as input suggests that the individual differences observed in learning may reflect – not differences in learning strategy or fundamental learning mechanisms – but differences in the moment-to-moment selection of information in the learning task. We will discuss this second idea more fully in the General Discussion.

The model aggregates statistical information across all learning trials and from this information can predict performance at test which follows the completion of training. If differences between the two groups of learners emerge – and are not there at the start – then the model should not be able to predict learning even using the same mechanism if it were given incomplete training data. To examine this issue, the 30 training trials were divided in half, and we used each half to separately build the association matrices for each participant. Then we again correlated the number of learned words predicted by the model with the number of learned words (as measured by absolutely greater looking times at the

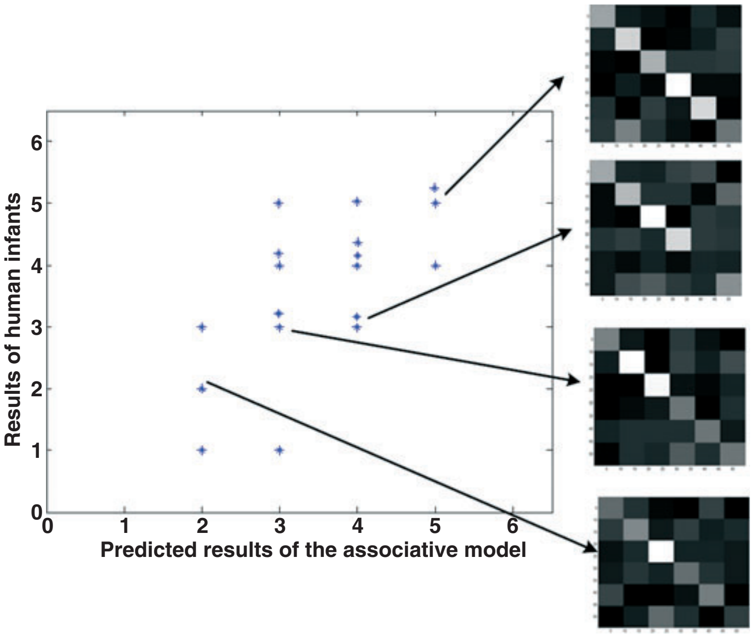

Figure 6 *The comparison of predicted results from the associative model and the actual results of human learners indicates a strong correlation between these two. Each data point in the figure corresponds to one infant participant in our study. Four examples of association matrices from eye movement data from four participants are also shown.*

target than distracters at test) for each individual learner. A weak and unreliable correlation was found for the first half (r = 0.31, p = .22) and a very strong correlation was found for the second half (r = 0.78, p < .005).

General discussion

Statistical learning appears to be a pervasive component of human learning about language and to be powerfully evident even in simplified and very brief experimental contexts. The still open question is the nature of the mechanisms that underlie this learning. The present results reveal two ingredients relevant to a mechanistic account of learning word–referent mappings across individually ambiguous trials: attention and associative learning. A simple model assuming the most elemental form of associative learning can explain both the group data and also individual differences in infant learning from simply knowing the potential referents to which infants – moment by moment – attend. This tells us that the key to understanding statistical learning in this task will be understanding attention and the trial-by-trial selection of information.

The eye-tracking data and the model indicate that learners do not passively store *all* the word–object co-occurrences but only *some* of them; learning thus depends on looking at the right referent at the right moment in time (when the referring word is also presented). Infants who showed strong learning at the end of the experiment exhibited a pattern of stable looking across learning trials; and over trials, these infants began to look more often to the right referents for the heard word, building robust associations for correct pairs and fewer spurious correlations. Weak learners start learning with looking patterns like those of the strong learners, but their looking becomes more variable within a trial, short looks and many switches back and forth, and thus, by the simple associative model, they build generally weaker associations and these weaker associations are distributed over many more incorrect pairs. In the model, individual differences emerge solely through the looking patterns; the same associative learning mechanism produces quite different results, being highly dependent on the dynamics of attention over the course of training. The simple model and the data themselves are just a first step toward understanding real-time mechanisms that lead to statistical cross-situational learning; nevertheless, they raise new insights and intriguing questions for future work; these are discussed next.

Statistical associative learning

The present results support the argument that associative learning mechanisms will be part of the explanation of statistical learning; they are sufficient (when coupled with selective attention) to predict individual differences. The power of this success quite frankly lies in the eye movement data which can be directly mapped to the hypothesized strength of association in a probabilistic associative framework. This simple model is in contrast to hypothesis-testing accounts which explicitly formulate and evaluate hypothesized word–referent pairs, eliminating incorrect hypotheses and maintaining hypotheses that are consistent with the training data observed so far.

The simulations here did not include an explicit hypothesis-testing model coupled to the eye movement data. Given the dynamic and spontaneous eye movement generated by the infants as shown in Figure 1, it is hard to conceptualize them in terms of a hypothesis-testing model in which young learners strengthen or weaken hypotheses. To take advantage of the more detailed eye movement behavioral data, a hypothesis-testing mechanism, at minimum, needs to incorporate two kinds of basic information that are in the eye movement data and were shown here to be relevant to learning. The first is the pattern of looks within a trial and the shifting of looks back and forth between two candidate objects after hearing a word, a behavior that characterized both strong and weak learners throughout the training session. While it is easy to link this dynamic information selection with the dynamic strengthening of associations, it is not at all clear how to mechanistically link these behaviors to changes in hypothesis strength. This is not to say that this could not be done. However, showing that an associative versus hypothesis-testing model is correct for explaining this kind of learning is not the present goal nor the main conclusion. The goal was to understand learning at the dynamic micro-level of moment-to-moment attention. And the main conclusion from the present data is that any model which seeks to explain learning at the level of real-time dynamics will have to take moment-by-moment attentional processes seriously. A second aspect of the eye movement data that was found to be relevant to learning concerns the durations of looks. Each eye movement has its own length and these durations vary; these differences matter to successful learning in the associative learning model. Again, it is unclear how such information might be linked to current hypothesis-testing models of cross-situational learning (Siskind, 1996). For example, is a fixation a sampling of a new hypothesis or should it be interpreted as evaluating an existing one? If the latter, how to distinguish between accepting and rejecting an existing hypothesis based on eye fixation patterns, and how do long versus short fixations play out in sampling and evaluating? There is no argument here that this is an in-principle limitation; but looking when listening – both numbers, timings and durations of eye fixations – matters straightforwardly in the infant's learning in this task and also in building the associations in an associationist model.

The associationist approach taken here also contrasts with the more common approach to statistical learning which assumes sophisticated and powerful learning algorithms operating on messy data and most often assumes a form of inference making (Xu & Tenenbaum, 2007; Frank *et al.*, 2009). Although hypothesis-testing inference and associative learning may be formally treated as variants of the same learning framework (Yu, Smith, Klein & Shiffrin, 2007), there are also fundamental differences. An associative learning mechanism with real-time attention treats the learning process as a dynamical system and focuses on how the learning system may actively select some potential pairs over others based on real-time feedback and the current learning states. Real-time attention thus removes a significant amount of uncertainty from the learner's internal data set (and could potentially even yield better and sparser internal matrices than the real-world data). In contrast, hypothesis-testing inferences often assume that the learners receive unprocessed ambiguous data but rely on the powerful learning machinery to infer meaningful knowledge from that data. The first approach – being tied to real-time behaviors on the part of the learner – may offer a deeper understanding of how learning progresses moment by moment from sets of experiences and could be conceptualized as the implementational real-time model of hypothesis testing. Understanding processes at this more micro-level might also give new insights into how and why the social scaffolding of adult partners is so effective in promoting word–referent learning. A more systematic simulation study and more detailed discussion of associative learning and hypothesis testing in the context of adult cross-situational learning can be found in Yu *et al.* (2007) and Yu and Smith (submitted).

Learning co-occurring cross-situational statistics

Although the results show the promise of analyzing micro-level behaviors during the course of learning, we note that the model did not reach the level of predicting which individual words were learned by the specific infants. There is one simple and one deeper reason for this limitation. First, infants generate more than 100 eye movements with various durations during the course of 4 minutes' training. Some are likely generated by randomness and some are more systematic; some are directly learning related and some may be caused by other perceptual and cognitive factors that may not be directly relevant to the learning task or, if they are, we do not yet know it. The current work represents a first effort and was motivated by psycholinguistic studies on adult spoken language processing. But there are significant challenges in applying this approach to infants. Infant eye movements (just like other behaviors that they generate at the early stage) are more spontaneous and more random compared with adult eye movements. They may also tie to knowledge and learning in different ways than do seemingly comparable behaviors in adults. Moreover, the only data we have from these infants is looking patterns; there is no independent measure of what they know trial by trial to help discover the meaning of these looking patterns. In brief, the simple answer to why we cannot predict the individual words learned by individual learners is that this is a first step, and there is much that we do not know. As we get to know more about the looking behavior and how it both drives and reflects learning, we may be able to move to that level of precision.

However, there may be a deeper reason for the limitation that has to do with the very nature of cross-situational learning. In this paradigm, a learner perceives multiple words and multiple objects trial by trial. As indicated by the association matrix shown in Figure 5(b), each word/object co-occurs with all of the other words/objects in the whole training. Therefore, as learning proceeds, knowing one word in a trial may simplify the learning task by reducing the degree of ambiguity in the current trial. Thus, knowing more words may facilitate and boost the whole learning process. Consequently, the mechanisms of learning individual words may differ depending on both the current learning context and the learner's internal learning state. Hence, even for all the successfully learned words, eye movement patterns on those words may be different depending on both whether a word was learned first or later and whether other co-occurring words were already learned or not. From this perspective, cross-situational learning (and attention in these tasks) may fundamentally not be explained as the learning of individual word–referent pairs but instead may only be understood in terms of learning a *system* of associations (Yu, 2008). 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. More specifically, given a word–referent pair in the association matrix, this association is related to all of the other cells in the matrix sharing the same row and all of the others sharing the same column. Those associations are in turn correlated with more word–referent pairs – and the whole system as the whole association matrix. Indeed, Yu (2008) proposed and developed a computational model to account for so-called latent knowledge in statistical learning in which the model shows lexical knowledge accumulated over time as latent knowledge of the whole lexical system, and that partial knowledge can be recruited in subsequent word learning. Thus, in cross-situational statistical learning, although we have to test the learning outcome as individual words, a more comprehensive view of cross-situational learning may need to be based on learning the whole system of co-occurring associations instead of learning individual ones. If this is the case, then analyzing eye movement data should start with the idea of global eye movement patterns as a product of a system of associations. In future studies, we intend to explore these ideas through both experimental studies and computational modeling.

Better looking and better learning

Does better looking lead to better learning or does better learning lead to better looking? Studies of infant learning present learning trials first and then testing trials, using looking time to measure what has been learned. But these are not two distinct kinds of trials from the infant's point of view. Thus, if what has been learned guides looking at test (and does so robustly so that researchers can use it as the dependent measure), then surely what is being learned may guide looking during the training trials. At the beginning of training, infants may randomly select visual objects after hearing a word, with no word–object associations influencing that selection. But at some point in training, after exposure to some sufficient number of co-occurrences, the statistical regularities in those occurrences seem likely to influence looking and visual selection. Eye movements, particularly those in the later part of training, may reflect what learners already know.

However, those eye movements driven by a learned association also lead to the strengthening of already (maybe partially) learned associations. Thus, a learner may register word–referent associations by selectively attending to target objects trial by trial, and the synchrony between heard words and visually attended objects will allow the infants to build and enforce corresponding word–object associations. Indeed, a recent study on the preference to attach novel labels to novel objects found that infants couldn't use Mutual Exclusivity (ME) on the first encounter of a novel label with one familiar object and one novel object but there was a carry-over effect to the next time the label was presented and thus the ME effect emerged over a more extended period containing multiple learning episodes (Mather & Plunkett, 2009). If we further consider cross-situational learning as learning a system of associations between words and referents, then learning itself may drive attention to as-yet-unlearned referents from as-yet-learned ones in the context of hearing an as-yet-unlearned name – a form of mutual exclusivity (Golinkoff, Hirsh-Pasek, Bailey & Wenger, 1992).

Based on this observation, the idea of any clear boundary between learning and test becomes even less likely. There is no separation between learning and looking: infants' looking creates learning and the looking itself is influenced by learning. The present simple model, albeit helpful in revealing how learning depends on the looking patterns, does not incorporate this real-time dynamic loop between looking and the internal learning state. For instance, longer looking durations at the end of training shown in Figure 2(b) can be the outcome of learning or the cause of learning. Similarly, more attention switches in the middle of training from weak learners shown in Figure 2(a) can be the consequence of unsuccessful attempts to register word–object associations or they can be the cause of unsuccessful learning. A more detailed data mining will need to not only discover different kinds of patterns over time but also link them together to decipher the consequential effects of those patterns. Further, a more correct future model will need to explain both information selection (as it dynamically happens in real-time learning) and the learning mechanism.

Sources of individual differences

Weak and poor learners show different looking patterns over the course of training. One plausible account is that both strong and weak learners start by randomly sampling possible word–referent pairs but the infants who will become strong learners just happened to sample the right pairs at the start and those who will become the poor learners do so because they just got off to a bad start, sampling and storing wrong pairs. Thereafter, strong learners consistently strengthen those correct associations while weak learners cannot (as those wrong associations will not occur with much frequency). The resulting uncertainty may be what results in briefer and more variable fixations that emerge in the middle portion of training.

Another possible reason for the observed individual differences is that there may be some intrinsic reasons causing weak learners to fail to consistently stabilize their attention. Perhaps they pick up just as many right and wrong early associations as the good learners but this early partial learning does not organize attention as well and thus they fail to strengthen these associations. A next step to answering these issues is to bring infants in for multiple training and testing sessions (using different sets of words and referents) to determine whether these individual differences are reliable across testing. Are the same infants good and poor learners on different days or does successful learning self-organize in the session from what must initially be randomly selected word–referent pairs? Further, individual differences on both whether they learn in training and whether they demonstrate learning outcome in test depend on how reliable their looking behaviors are related to the statistical learning task. Therefore, another next step for deeper data analyses of individual differences at the micro-level is to record, code and analyze the infant's engagement/attention level in both training and testing sessions as in this study (and many other infant studies) to better use their looking behaviors to infer their learning mechanisms.

Conclusion

As stated in Gleitman *et al.* (2005), the problem with word learning is not the poverty of the stimulus but ambiguity in the massive amounts of information available. Thus learners need to select some information – the right information – for learning. At the macro-level, we know a great deal about the forces that enable young learners to select the right information for mapping words to referents and we know that these include social, linguistic, and conceptual biases. But we know very little about how these play out in real time. The present study and the results did not consider these higher level constraints on children's word–referent learning but we believe that they are deeply relevant to them nonetheless. The present results tell us that word–referent learning (at least in this one task with highly ambiguous learning trials) is, at the micro-level of analysis, tightly tied to real-time information selection. This suggests that we might look for a deeper understanding of word learning in general – and social, linguistic, and conceptual biases – through fine-grained analyses of moment-to-moment attention (see also Halberda, 2003; Swingley & Aslin, 2007).

The more specific contribution concerns cross-situational learning of words and referents. The present work is built upon the recent work in statistical word learning (Smith & Yu, 2008). The present findings go beyond demonstrating that infants can do this form of learning by revealing deep links between the dynamics of attention and statistical learning. The work was motivated by and took advantage of three recent advances in cognitive science and psychology: (1) developmental psychology: using eye-tracking techniques to measure moment-by-moment eye movement data from infants (Aslin & McMurray, 2004); (2) psycholinguistics: measuring the synchrony between visual attention and speech (Tanenhaus *et al.*, 1995); and (3) data mining: analyzing fine-grained temporal behavioral data using computational techniques (Yu, Ballard & Aslin, 2005). The work also represents the first attempts to use momentary eye movement data as input to a computational model so as to understand word learning. The present results from this endeavor yield two promising directions for the future study of cross-situational learning. First, they show that a simple associative learning mechanism (without high-level statistical computations) can support this kind of learning if the learner selectively registers the right statistical information at every moment. Second, the results show how individual differences in learning may arise from different learners attending to different statistical information but based on the same associative learning mechanism. The results also show how eye movements can be used as a window to infer the statistical learner's internal state. This finding in particular allows us to ask in future work how selective attention works in real-time learning.

Acknowledgements

We thank Juliette McNamara, Amara Stuehling, and Char Wozniak for collection of the data. We are grateful for insightful comments from the action editor Usha Goswami and four anonymous reviewers. We would like to acknowledge discussions with Daniel Yurovsky and Krystal Klein. This research was supported by National Institutes of Health R01 HD056029 and National Science Foundation Grant BCS0544995.

References

Aslin, R., & McMurray, B. (2004). Automated corneal-reflection eye tracking in infancy: methodological developments and applications to cognition. *Infancy*, **6** (2), 155–163.

Booth, A., & Waxman, S. (2002). Word learning is 'smart': evidence that conceptual information affects preschoolers' extension of novel words. *Cognition*, **84** (1), 11–22.

Chater, N., & Manning, C. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, **10** (7), 335–344.

Chomsky, N. (1959). Verbal behavior. *Language*, **35** (1), 26–58.

Christiansen, M., Allen, J., & Seidenberg, M. (1998). Learning to segment speech using multiple cues: a connectionist model. *Language and Cognitive Processes*, **13** (2/3), 221–268.

Cover, T., & Thomas, J. (2006). *Elements of information theory.* New York: Wiley-Interscience.

Elman, J. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, **48** (1), 71–99.

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

Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, **20** (5), 578–585.

Gelman, S., & Taylor, M. (1984). How two-year-old children interpret proper and common names for unfamiliar objects. *Child Development*, **55**, 1535–1540.

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

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

Golinkoff, R. (1994). Early object labels: the case for a developmental lexical principles framework. *Journal of Child Language*, **21** (1), 125–155.

Golinkoff, R., Hirsh-Pasek, K., Bailey, L., & Wenger, N. (1992). Children and adults use lexical principles to learn new nouns. *Developmental Psychology*, **28** (1), 99–108.

Golinkoff, R., Hirsh-Pasek, K., & Hollich, G. (1999). Emerging cues for early word learning. In B. MacWhinney (Ed.), *The emergence of language* (pp. 305–330). Hillsdale, NJ: Lawrence Erlbaum Associates.

Golinkoff, R., Jacquet, R., Hirsh-Pasek, K., & Nandakumar, R. (1996). Lexical principles may underlie the learning of verbs. *Child Development*, **67** (6), 3101–3119.

Griffin, Z. (2004). Why look? Reasons for eye movements related to language production. In J.M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 213–247). New York: Psychology Press.

Griffin, Z., & Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language*, **38** (3), 313–338.

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, **87** (1), 23–34.

Halberda, J. (2006). Is this a dax which I see before me? Use of the logical argument disjunctive syllogism supports word-learning in children and adults. *Cognitive Psychology*, **53** (4), 310–344.

Hirsh-Pasek, K., & Golinkoff, R. (1996). The intermodal preferential looking paradigm: a window onto emerging language comprehension. In D. McDaniel & C. McKee (Eds.), *Methods for assessing children's syntax* (pp. 105–124). Cambridge, MA: MIT Press.

Hollich, G., Hirsh-Pasek, K., Golinkoff, R., Brand, R., Brown, E., Chung, H., Hennon, E., & Rocroi, C. (2000). Breaking the language barrier: an emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, **65** (3, Serial No. 262).

Huettig, F., & Altmann, G. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. *Cognition*, **96** (1), 23–32.

Imai, M., & Gentner, D. (1997). A cross-linguistic study of early word meaning: universal ontology and linguistic influence. *Cognition*, **62** (2), 169–200.

Johnson, S., Amso, D., & Slemmer, J. (2003). Development of object concepts in infancy: evidence for early learning in an eye-tracking paradigm. *Proceedings of the National Academy of Sciences*, **100** (18), 10568–10573.

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

Klibanoff, R., & Waxman, S. (2000). Basic level object categories support the acquisition of novel adjectives: evidence from preschool-aged children. *Child Development*, **71** (3), 649–659.

Knoeferle, P., & Crocker, M. (2006). The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive Science: A Multidisciplinary Journal*, **30** (3), 481–529.

Landau, B., Smith, L., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, **3** (3), 299–321.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104** (2), 211–240.

Li, P., Burgess, C., & Lund, K. (2000). The acquisition of word meaning through global lexical co-occurrences. In E.V. Clark (Ed.), *Proceedings of the 30th Stanford Child Language Research Forum* (pp. 167–178). Stanford, CA: Center for the Study of Language and Information.

Markman, E. (1990). Constraints children place on word meanings. *Cognitive Science: A Multidisciplinary Journal*, **14** (1), 57–77.

Mather, E., & Plunkett, K. (2009). Learning words over time: the role of stimulus repetition in mutual exclusivity. *Infancy*, **14** (1), 60–76.

Merriman, W., & Stevenson, C. (1997). Restricting a familiar name in response to learning a new one: evidence for the mutual exclusivity bias in young two-year-olds. *Child Development*, **68**, 211–228.

Meyer, A., Sleiderink, A., & Levelt, W. (1998). Viewing and naming objects: eye movements during noun phrase production. *Cognition*, **66** (2), 25–33.

Mintz, T., Newport, E., & Bever, T. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science: A Multidisciplinary Journal*, **26** (4), 393–424.

Monaghan, P., Chater, N., & Christiansen, M. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, **96** (2), 143–182.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, **17** (2), 357–374.

Newport, E. (1990). Maturational constraints on language learning. *Cognitive Science: A Multidisciplinary Journal*, **14** (1), 11–28.

Newport, E., & Aslin, R. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, **48** (2), 127–162.

Plunkett, K., Hu, J., & Cohen, L. (2008). Labels can override perceptual categories in early infancy. *Cognition*, **106** (2), 665–681.

Rose, S., Feldman, J., & Jankowski, J. (2004). Infant visual recognition memory. *Developmental Review*, **24** (1), 74–100.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, **274** (5294), 1926.

Schafer, G., & Plunkett, K. (1998). Rapid word learning by fifteen-month-olds under tightly controlled conditions. *Child Development*, **69** (2), 309–320.

Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, **61** (1–2), 39–91.

Smith, L. (2000). Avoiding associations when it's behaviorism you really hate. In R. Golinkoff & K. Hirsh-Pasek (Eds.), *Breaking the word learning barrier* (pp. 169–174). Oxford: Oxford University Press.

Smith, L., & Yu, C. (2008). Infants rapidly learn word–referent mappings via cross-situational statistics. *Cognition*, **106** (3), 1558–1568.

Snedeker, J., & Gleitman, L. (2004). Why it is hard to label our concepts. In D.G. Hall & S.R. Waxman (Eds.), *Weaving a lexicon* (pp. 257–294). Cambridge, MA: MIT Press.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, **102** (33), 11629–11634.

Steyvers, M., & Tenenbaum, J. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science: A Multidisciplinary Journal*, **29** (1), 41–78.

Swingley, D., & Aslin, R. (2007). Lexical competition in young children's word learning. *Cognitive Psychology*, **54** (2), 99–132.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, **268** (5217), 1632–1634.

Trueswell, J., & Gleitman, L. (2004). Children's eye movements during listening: evidence for a constraint-based theory of parsing and word learning. In J.M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 319–346). New York: Psychology Press.

von Hofsten, C., Vishton, P., Spelke, E., Feng, Q., & Rosander, K. (1998). Predictive action in infancy: tracking and reaching for moving objects. *Cognition*, **67** (3), 255–285.

Waxman, S., & Lidz, J. (2006). Early word learning. In D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology* (6th edn., Vol. 2, pp. 299–335). New York: J. Wiley & Sons.

Xu, F., & Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review*, **114** (2), 245–272.

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Development*, **4** (1), 32–62.

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

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, **18** (5), 414–420.

Yu, C., & Smith, L. (submitted). Hypothesis testing and associative learning in cross-situational word learning: are they one and the same?

Yu, C., Smith, L., Klein, K., & Shiffrin, R. (2007). Hypothesis testing and associative learning in cross-situational word learning: are they one and the same? In D.S. McNamara & J.G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 737–742). Austin, TX: Cognitive Science Society.

Received: 29 April 2008
Accepted: 17 December 2009