Time Course of Visual Attention in Statistical Learning of Words and Categories

Chi-hsin Chen[1], Chen Yu[2] ({chen75, chenyu}@indiana.edu)
Damian Fricker[2], Thomas G. Smith[2], Lisa Gershkoff-Stowe[1]
Department of Speech and Hearing Sciences[1], Department of Psychological and Brain Sciences[2]
Indiana University, IN 47405 USA

Abstract

Previous research indicates that adult learners are able to use co-occurrence information to learn word-to-object mappings and form object categories simultaneously. The current eye-tracking study investigated the dynamics of attention allocation during concurrent statistical learning of words and categories. The results showed that the participants' learning performance was associated with the numbers of short and mid-length fixations generated during training. Moreover, the learners' patterns of attention allocation indicated online interaction and bi-directional bootstrapping between word and category learning processes.

Keywords: Eye-tracking; statistical learning; word learning; category learning.

Introduction

Over the past few decades, researchers have found that humans are sensitive to statistical regularities in the environment. People are able to use statistical information in non-linguistic tasks, such as making inferences (e.g., Xu & Denison, 2009) or finding predictive features of complex visual scenes (e.g., Fiser & Aslin, 2001). They can use statistical information in linguistic tasks as well, such as learning phonetic distributions (e.g., Maye *et al*., 2002), word boundaries (e.g., Saffran *et al*. 1996b), word and meaning mappings (e.g., Smith & Yu, 2008), and rudimentary syntax (e.g., Gomez & Gerken, 1999). These studies suggest that statistical learning is a domain-general ability in human cognition.

An earlier cross-linguistic study conducted in our laboratory (Chen *et al*., 2009) also showed that adult English and Mandarin speakers were able to use co-occurrence information to learn word-to-object mappings and to form object categories at the same time. However, even though these two groups of learners had comparable performance in learning word-to-object mappings, they showed different levels of sensitivity to the cues associated with category learning. Participants were better at learning the types of regularities that were present in their native language than the ones that were incongruent with their linguistic input. In Experiment 1 of the study, objects from the same category had similar attached object parts and their labels ended with the same final syllable. This syllable-to-category association simulated a prevalent linguistic feature in Mandarin in that the final syllables of object names often indicated category membership. The results showed that Mandarin speakers were able to learn individual word-to-object mappings and to form syllable-to-category associations under cross-situational learning contexts. On the other hand, English speakers tended not to use the final syllables of labels as cues in category learning. In Experiment 2 of that study, the category markers were moved to the beginning of labels to simulate a more frequent feature in English (e.g., the adjectives in noun phrases). As the structures of the training stimuli were more congruent with the input in the naturalistic environment, the English speakers' category learning performance became significantly better. More importantly, they also had better performance in the word learning task. One possible explanation of the improvement of word learning performance is that category learning bootstraps word learning. That is, learning which objects belong to the same category helps the learners to focus on relevant features of the stimuli and to rule out certain distractors as possible referents of a word. However, from the design of that study, we were not able to draw a conclusive link between the English speakers' success in forming categories and their improvement in word learning.

The present study was designed to address this issue by using eye-tracking techniques. Category learning studies using eye-tracking techniques have shown that learners generally attend to all possible dimensions early in learning. But during the process of learning, they gradually shift their attention to relevant dimensions (e.g., Rehder & Hoffman, 2005; Blair et al., 2009). Based on previous studies, similar patterns might be observed in statistical word learning and category learning. Our prediction is that at the beginning of training, learners will pay attention to all objects on the screen when hearing a word. Across learning, they will gradually tune their attention to the most probable referent of a word. Moreover, after successfully forming a few word-to-object mappings, the learners should notice that the objects (and their labels) can be grouped into different categories, each having its own distinctive feature. After establishing primitive category structures, the learners should then use this information to rule out certain distractors as possible referents of a word. The goals of the current study are to examine the dynamics of attention allocation in statistical learning of words and categories and to investigate the real-time interaction between word learning and category formation.

Method

Participants

Participants were 23 undergraduates (14 females, mean age: 19.1 years) who received course credit for volunteering.

None had previously participated in any cross-situational learning experiments.

Design and Stimuli

The experimental design in this study was the same as the one used in Experiment 2 of Chen *et al.* (2009) with slight modification in the length of training trials. Participants were trained under a cross-situational learning paradigm, which was first proposed by Yu and Smith (2007). In each training trial, the participants viewed four novel objects on a computer screen and heard four novel words. However, the temporal order of the word presentations was not related to the spatial locations of the words' target referents. In order to find the correct word-to-object mappings, the participants had to track the co-occurrence regularities between objects and words across different trials. There was a total of 18 object-word pairs to learn. Over the training, there were 12 repetitions per object-word pairing, yielding a total of 54 trials (18 pairs *12 repetitions / 4 pairs per trial). The length of each trial was 14 seconds and the whole training lasted for 12.6 minutes.

The to-be-learned objects were divided into three different categories, with six items in each category. Members in a category had an attached part that looked similar to each other. As an example, Figure 1 shows two items from a category in which all members had an attached spiral part that spread at the end. Moreover, these objects all had labels that began with the same syllable (e.g., *la-* in this case).

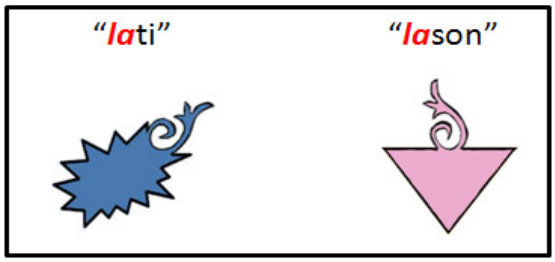

Figure 1 Sample objects and labels used in the study

Apparatus

The course of the experiment was controlled by a computer using E-prime. The visual stimuli were presented on a 17 inch monitor with a resolution of 1280*1024 pixels. The learners' eye gaze was measured by a Tobii 1750 near infrared eye-tracker (www.tobii.se). The eye-tracking system recorded gaze data at 50Hz (accuracy = 0.5°, and spatial resolution = 0.25°).

Procedure

Before the experiment, the eye-tracker system was calibrated. We used a procedure including nine calibration points. The experiment consisted of a Training session, followed by a Testing session. In the Training session, the participants were presented with 4 novel objects and 4 novel words in each trial without any information about which word referred to which object. The learners had to keep track of the co-occurrences between objects and words across trials to find the correct word-to-object mappings. Once they formed several correct word-to-object mappings, we expected they would be able to detect the associations between the first syllables of words and the attached object parts and to form object categories accordingly. The syllable-to-category associations should in turn facilitate word-to-object mappings, because the learners would be able to use the first syllable of a label to determine its possible referents. Eye movements were recorded during the Training session.

There were two tasks in the Testing session, a word-to-object Mapping task and a Generalization task. The Mapping task tested how well the participants learned the names of the training objects. The participants were instructed to select the referent of a training word from 4 alternatives. There were 18 trials in the Mapping task.

In the Generalization task, the participants were asked to select the referent of one novel word from three alternatives, each containing the object-part that corresponded to the particular feature of one category. The first syllable of the novel word was the same as the labels from one of the three categories. If the learners had formed the syllable-to-category associations, they should be able to use the first syllable of the novel word to find its referent. There were 9 trials in the Generalization task (3 for each category).

Eye-tracking dependent variables

To derive eye movement measures, we defined four rectangular region-of-interests (ROIs) that covered the objects displayed on the screen for each trial. We took the onset of a series of gaze data that fell within an ROI as the onset of a fixation and the end of the fixation was determined when the gaze fell outside of the same ROI. The minimum length of a gaze was 20ms (i.e., the length of 1 data point recorded by the eye-tracker). All gaze data outside the ROIs were viewed as saccadic eye movements and not included in the analyses.

Based on the remaining gaze data, we computed two dependent measures. The first variable was the *number of fixations* per trial. We set the thresholds at 100ms, 500ms, and 1000ms and counted the numbers of fixations exceeding these thresholds. Moreover, fixations with a length between 100ms and 500ms were defined as Short fixations; fixations between 500ms and 1000ms were viewed as Mid-length fixations; and those longer than 1000ms were taken as Long fixations. The reason for setting different thresholds was that previous category learning studies using eye-tracking techniques have found that looking more at the correct or relevant features during training was positively correlated with behavioral performance (e.g., Rehder & Hoffman, 2005; Blair et al., 2009). This indicates that more looking at the relevant features during training might lead to better learning. However, more looking could result from either having a few long fixations or having many short fixations combined together. Setting different thresholds would allow

us to examine whether longer looking also leads to better learning.

The second measure was *proportion looking time* (ranging from 0 to 1), which took the time spent fixating on one object divided by total time spent fixating on all objects. Moreover, based on the word being presented, we divided the objects into 3 categories: Correct Object, Within-Category Distractor, and Between-Category Distractor. Because there were 4 objects in each training trial while there were only 3 categories to learn, there could be more than 1 object from a specific category in a trial. Therefore, for each word, the Correct Object was the target referent while a Within-Category Distractor was an object from the same category. On the other hand, the Between-Category Distractors were the ones from a different category. Figure 2 illustrates a situation in which there are two objects from the *la-* category, one from the *jo-* and one from the *mu-* category. The label of each object can be found above it (please note that in real training, the labels were presented auditorily). For the word "*lati*", there is one Within-Category Distractor and two Between-Category Distractors in this trial. In contrast, for the word "*joler*", there are three Between-Category Distractors. However, in this case none of the objects is a Within-Category Distractor for this word. The mean numbers of Correct Object, Within-Category Distractor, and Between-Category Distractor for the training words in each trial are: 1, 0.74, and 2.26, respectively.

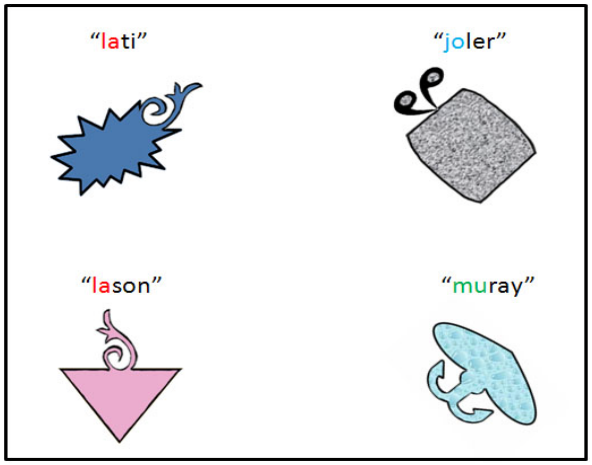

Figure 2 Sample stimuli in Training

Behavioral Results

On average, more than 50% of the participants' responses were correct in the Mapping task and in the Generalization task as well (see Figure 3). Consistent with earlier findings, participants learned more word-to-object mappings than expected by chance ($t(22) = 4.211$, $p < .001$). They also performed significantly above chance in the Generalization task ($t(22) = 3.227$, $p = .004$). That is, they could use the first syllable of a novel label to find its referent. In addition, we found a strong positive correlation between the learners' Mapping and Generalization performance ($r = .773$, $p < .001$). This suggests that the more words participants learned, the more likely they were to use the first syllable as a cue in categorizing novel objects.

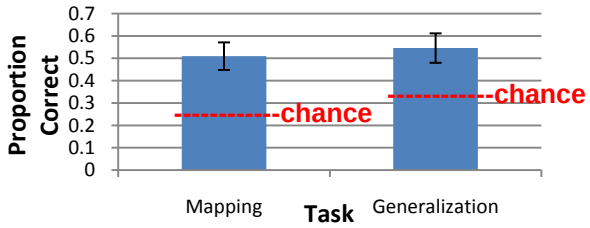

Figure 3: Proportion of accurate responses in Mapping and Generalization tasks

Eye Movement Data Analyses

According to the participants' performance in the Mapping task, we divided them into three groups. The participants that had more than 70% correct responses were viewed as High Learners. The people that made less than 35% correct responses were viewed as Low Learners. People having 35% to 70% correct responses were viewed as Mid Learners. There were 8, 6, and 9 people in the High, Mid, Low group, respectively. We compared the *number of fixations* and *proportion looking time* to different types of objects of the High, Mid, and Low Learners to see if there were differences in their eye movement patterns during the training.

Number of Fixations

As mentioned previously, we counted the numbers of fixations exceeding 100ms, 500ms, and 1000ms for each participant. The results can be found in Figure 4. The solid lines indicate the numbers of fixations exceeding 100ms. The High, Mid, and Low Learners had comparable numbers of fixations at the beginning of training. Across the Training session, the numbers of fixations of the Mid and Low Learners gradually decreased and the decreasing rate was slightly higher for the Low Learners. The dashed lines show that when the threshold was set at 500ms, the High Learners tended to have more fixations than the other two groups, especially in the second half of training. When the threshold was set at 1000ms, there did not seem to be group differences.

The patterns observed above were confirmed by statistical analyses. We compared the numbers of Short (100ms-500ms), Mid-length (500ms-1000ms), and Long fixations (<1000ms) of different groups of learners. With regard to Short fixations, trial-by-trial ANOVAs showed that group differences were significant between Trial 38 and Trial 42 (*p*s <.05). Pair-wise comparisons showed that the High Learners generated more Short fixations than the Low Learners (*p*s <.05). For Mid-length fixations, Trial-by-Trial ANOVAs revealed that significant group differences occurred between Trial 31 and Trial 39 at *p* level of .05. Pair-wise comparisons showed that the High Learners generated more Mid-length fixations than the Mid and Low Learners (*p*s < .05). In addition, the Mid Learners also generated more Mid-length fixations than the Low Learners

in Trial 13, 16, 39 and 40. When the threshold was raised to 1000ms, all three groups had about equal numbers of fixations across trials. Significant group differences were only found at Trial 26, in which the High Learners generated more fixations than the Mid and Low Learners (*p*s<.05).

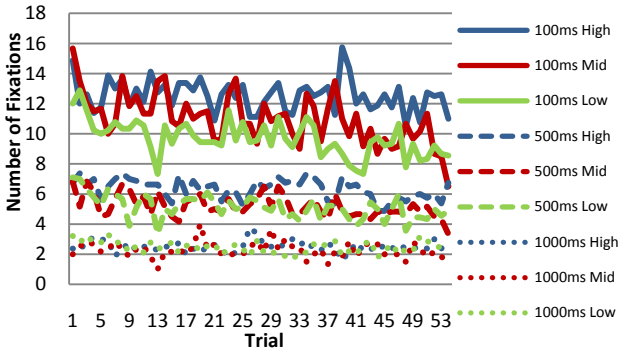

Figure 4 Number of Fixations of High, Mid, and Low Learners. The number of fixations was counted separately with 100ms, 500ms, and 1000ms as thresholds of minimal eye fixation length.

To summarize, the major differences between the High, Mid, and Low Learners were caused by the decreasing Short and Mid-length fixations of the Mid and Low Learners. The High Learners had more Short and Mid-length fixations than the other two groups, especially in the second half of training. The Mid learners also generated more Mid-length fixations than the Low learners.

Proportion Looking Time

Proportion Looking Time By Trial We first looked at the dynamics of attention allocation during the course of statistical learning. For ease of comparison, Figure 5 to Figure 7 present the normalized Proportion Looking Time of the High, Mid, and Low Learners across training trials. The Proportion Looking Time to a certain type of object is normalized so that the chance level is 25%. As can be seen from Figure 5, there was a drastic increase in the High Learners' Proportion Looking Time to the Correct Object. There was also a decreasing trend in their looking at the Between-Category Distractors. Starting from Trial 34, the High Learners looked at the Correct Object significantly more than expected by chance (*p*s < .05). They also looked at the Between-Category Distracters significantly less than chance from Trial 35 on (*p*s < .05). As to the Mid Learners in Figure 6, even though there was an increasing trend in their Proportion Looking Time to the Correct Object, it did not reach statistical significance. As can be seen in Figure 7, the Low Learners had chance level performance across the training. Though they had above- or below-chance performance in a few trials, the patterns were not reliable.

We also conducted trial-by-trial ANOVAs to compare group performance. Starting from Trial 38, the High Learners looked at the Correct Object more than the Mid and Low Learners (at *p*s <.05). The pattern can be seen in Figure 8. There was also a trend that the Mid Learners looked at the Correct Object more than the Low Learners at the last third of training. But the pattern was not reliable. As to Within-Category Distractors, there were significant group differences in a few trials in which the High and Mid Learners looked at the Within-Category Distractors more than the Low Learners. But the patterns were not reliable either. With regard to Between-Category Distractors, there were significant group differences starting from Trial 24. Compared to the High Learners, the Low Learners looked more at the Between-Category Distractors in the second half of training. Additionally, they looked more at the Between-Category Distractors than the Mid Learners in the last third of training.

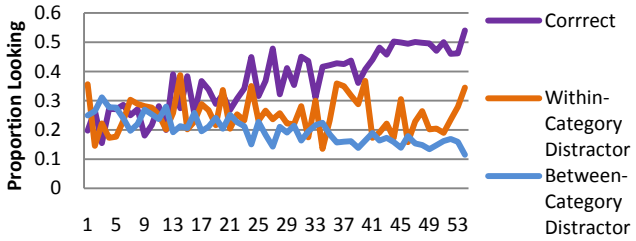

Figure 5 Proportion Looking Time of High Learners

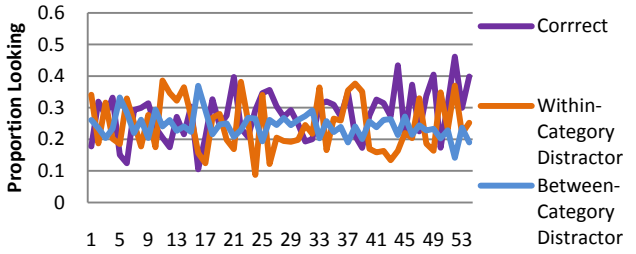

Figure 6 Proportion Looking Time of Mid Learners

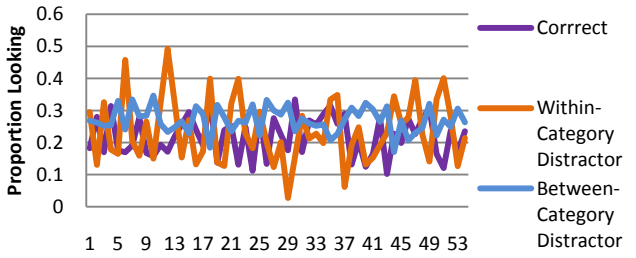

Figure 7 Proportion Looking Time of Low Learners

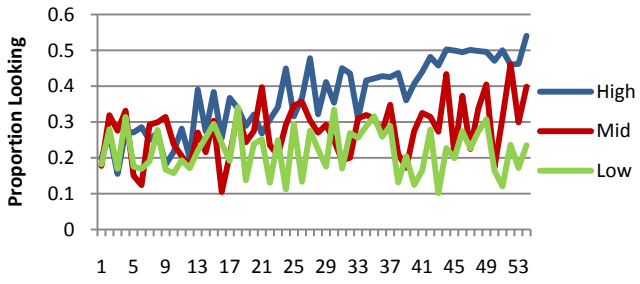

Figure 8 Proportion Looking Time to the Correct Object

Proportion Looking Time By Occurrences Across the Training session, each word-object pair occurred 12 times. For each participant, we calculated the Proportion Looking Time by word-object occurrences. For example, we took their Proportion Looking Time at the first occurrence of individual objects and averaged it across objects to get the Proportion Looking Time at Occurrence 1. This gave us 12 values for each participant. We then compared the High, Mid, and Low Learners' Proportion Looking Time to the Correct Object by occurrence.

Figure 9 illustrates that at about the third time the High Learners heard a word, they looked more at the Correct Objects than the Mid and Low Learners. Trial-by-trial analyses showed that group differences became significant at the third occurrence of a word (*p*s < .05). Except for the 6th occurrence, the High Learners were more likely to look at the Correct Object than the other two groups. The Mid Learners looked more at the Correct Objects than the Low Learners from Occurrence 10 to Occurrence 12.

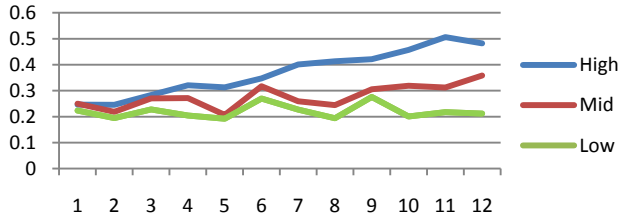

Figure 9 Proportion Looking Time to Correct Object by Occurrences

Compared to chance, the High Learners looked at the Correct Objects significantly above chance from the 7th to the last time they encountered a word (*p*s < .05). The Mid Learners looked at the Correct Objects significantly above chance from the 10th to the last time they heard a word (*p*s < .05). As for the Low Learners, they did not look at the Correct Objects more than chance. This indicates that it took only a few repetitions for the High Learners to detect the word-to-object co-occurrence regularities and that they could quickly tune their attention to the most probable referent of a word. However, it took longer for the Mid Learners to find the correct referent of a word.

Predictive Looking

Because the first syllable of a label indicated an object's membership, another question we were interested in was whether the participants made predictive looking and attended to objects from a relevant category even before the whole word was finished. For example, if the learners formed the association between the syllable *la-* and the spiral part, they might be able to use the syllable *la-* as a cue to rule out Between-Category Distractors even before the word "*lati*" was completed.

We calculated Proportion Looking Time to objects from a relevant category (i.e., the Correct Object and Within-category Distractor) and objects from irrelevant categories between 600ms and 900ms after the onset of a word. We chose the time between 600ms and 900ms based on the approximation that it took at least 200ms to generate stimulus-driven fixations and 600ms is about 200ms after the end of the first syllable while 900ms is about 200ms after the end of the word[1]. The Proportion Looking Time to object from a Relevant Category of the High, Mid, and Low Learners can be seen in Figure 10. For ease of comparison, the results were normalized, so that the chance value was .5. In the first half of training, all three groups had similar performance. In the second half of training, the Mid and the High Learners started to fixate on objects from a Relevant category even BEFORE the whole word was completed. However, for the Mid Learners, the trend was not as reliable as the High Learners.

It is noteworthy that the High Learners' predictive looking could only be reliably observed in the last third of training, which occurred after their reliable above-chance looking at the Correct Objects. This indicates that prior to forming syllable-to-category associations, the learners needed to establish at least a few correct word-to-object mappings in order to extract the regularities across objects.

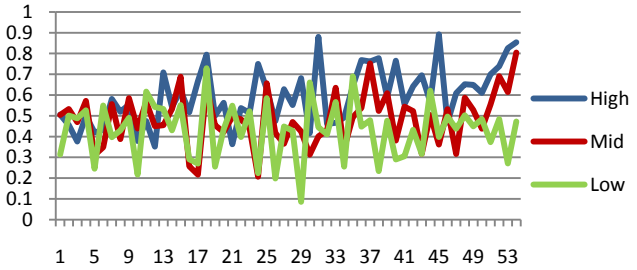

Figure 10 Proportion Looking between 600ms and 900ms after the onset of a word

Predictors of Behavioral Performance

As mentioned, the participants were grouped based on their performance in the Mapping task, which is a behavioral task administered after training. The above analyses showed that group differences could be observed from eye movement data during training. This suggests that eye gaze patterns during training might be used as predictors of behavioral performance.

Table 1: Correlations between Eye Gaze and Behavioral Measures.

		Mapping	Generalization
Number of Fixation	Short	.167	.107
	Mid-length	.339	.400*
	Long	.150	.118
Proportion Looking	Correct	.803**	.586*
	Within-category	.046	.278
	Between-category	-.749**	-.609**

* p < .05
** p < .001

[1] We also tried 500ms-800ms and 500ms-900ms. The trends are similar to the patterns observed here.

To find the best predictor of behavioral performance, multiple linear regression analyses were conducted. As can be seen from Table 1, there is a positive correlation between the number of Mid-length fixations and Generalization performance. The learners' Proportion Looking Time to the Correct Object is positively correlated with their Mapping and Generalization performance. In contrast, Proportion Looking Time to the Between-Category Distractors is negatively correlated with Mapping and Generalization performance. Stepwise regression showed that the best predictor of the Mapping performance is Proportion Looking Time to the Correct Objects during training. Consistent with the findings of previous studies, the more the learners looked at the correct features during training, namely the correct object, the better they performed in the following behavioral task. On the other hand, the best predictor of the Generalization performance is Proportion Looking Time to the Between-Category Distractors. The less the learners looked at the Between-Category Distractors, the better they did in the following Generalization task. This suggests that less looking at the Between-Category Distractors can be viewed as an indicator of category learning.

General Discussion

This study replicates previous findings that adult learners are able to use co-occurrence information to simultaneously learn word-to-object mappings and to form object categories. In addition, the current study shows that the learners' behavioral performance in the Mapping and Generalization tasks can be predicted from their looking patterns during the course of learning. Learners who generated more short- and mid-length fixations tended to perform better in the following behavioral tasks. However, there was no difference in the numbers of long fixations generated by different groups of learners. This indicates that more looking was not due to longer looking. Instead, the good learners tended to shift their attention back and forth among objects to check the possible referents of a word. Thus, rapid gaze shifts between several concurrent visual objects suggest a real time competition process which leads to better learning.

Patterns of attention allocation of the High, Mid, and Low Learners could be detected during the course of learning in addition. After accumulating certain statistical information, learners tended to shift their attention to objects containing relevant features. Moreover, at the third encounter with a word, the High Learners appear to have (partially) formed the association between a word and its referent. On the other hand, it took about 10 times for the Mid Learners to form correct mappings. This suggests that from eye movement data, we might be able to observe the accumulation of partial knowledge and how it leads to successful learning.

After forming a few individual word-to-object mappings, the High and Mid Learners shifted their attention to relevant categories BEFORE a word was completed. This suggests that after establishing syllable-to-category associations, they use the first syllable of a word to eliminate Between-Category Distractors as possible referents of the word. Together, the results of the present study reflect online interaction of word learning and category learning. It also provides evidence that word learning and category learning bootstrap each other.

References

Blair, M. R., Watson, M. R., Walshe, R. C., & Maj, F. (2009). Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1196-1209.

Chen, C., Yu, C, Wu, C.-Y., & Cheung, H. (2009). Statistical Word Learning and Object Categorization: A Cross-Linguistic Study in English and Mandarin. *Proceeding of the 31st Annual Conference of the Cognitive Science Society*.

Colunga, E. and Smith, L. B. (2008). Knowledge embedded in process: the self-organization of skilled noun learning. *Developmental Science, 11(2)*, 195-203.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12(6)*, 499-504.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70(2)*, 109-135.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82(3)*, B101-B111.

Rehder, B. & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, 51, 1-41.

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

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

Xu, F. & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition, 112*, 97-141.

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