Examining Referential Uncertainty in Naturalistic Contexts from the Child's View: Evidence from an Eye-Tracking Study with Infants

Yayun Zhang and Chen Yu
yayzhang@indiana.edu, chenyu@indiana.edu
Department of Psychology & Brain Science, Indiana University, 1101 E. 10th Street, Bloomington, IN 47405 USA

Abstract

Young infants are prolific word learners even though they are facing the challenge of referential uncertainty (Quine, 1960). Many laboratory studies have shown that human infants are skilled at inferring the correct referent of an object from ambiguous contexts (Swingley, 2009). However, little is known regarding how children visually attend to and select the target object among many other objects in view when parents name it during free play interactions. In the current study, we explored the looking pattern of 12-month-old infants using naturalistic first person images with varying degrees of referential ambiguity. Our data suggest that infants' attention is selective and they tend to only select a small subset of objects to attend to at each learning instance despite the complexity of the data existed in the real world. This work allows us to better understand how perceptual properties of objects in infants' view influence their visual attention, which is also related to how they select candidate objects to build word-object mappings.

Keywords: statistical learning; word-referent mapping; learning mechanisms

Introduction

Infants encounter words in complex environment and one challenge in early word learning is that of referential uncertainty: how infants manage to find the right word-referent pairs in the noise (Quine, 1960). Many studies have shown that human infants are able to infer the correct referent of an object from ambiguous contexts (Swingley, 2009; Waxman & Booth, 2001). Using the cross-situational word-learning task, Smith and Yu (2008) have found that infants can learn word-referent pairs by computing distributional statistics across the co-occurrences of words and referents at multiple naming moments, suggesting that infants attend to and systematically store the co-occurrence information during training. Additional evidence demonstrates that infants keep track of not only the strongest available associations but also low-frequency information, which further supports the notion that infants are sensitive to the co-occurrence statistics between words and referents and they keep track of a system of associations (Vouloumanous & Werker, 2009).

While laboratory tests have led to significant advancement in our understanding of the underlying word-learning mechanism, cognitive scientists have also started to investigate word learning using more naturalistic data. One interesting ongoing discussion in the literature is centered on the question of how noisy our daily environment is. Medina, Snedeker, Trueswell, and Gleitman (2011) argue that learners encounter words in complex environments where infinite referents might be treated as the label's correct referent, therefore co-occurrences in the real world are too noisy to be effectively learned by human learners. In their study using the "Human Simulation Paradigm" (HSP). They showed adults video clips of parent interacting with an infant. The original sound of the video was muted and a beep was inserted at the onset of the label when parent named an object. Adult learners were asked to watch these videos and guess the intended referent by the parent at the moment of the beep. They found that participants were not able to aggregate information and learn the correct word-referent mapping across trials. The researchers concluded that because there are potentially too many candidate referents which could be mapped on to a label, it is impossible for learners to continuously store and update the word-object co-occurrences across word learning moments and make appropriate decisions based on aggregated statistics.

However, other investigators reached different conclusions by using variants of the HSP method. Yurovsky, Smith and Yu (2013) used training videos from both the observers' view (captured by a tripod-mounted camera) and the child's view (captured by a head-mounted camera) to study how uncertain participants were when asked to make explicit hypothesis regarding the intended referent by the parent. They found that about 50% of the naming episodes by mothers to toddlers were not ambiguous to the adults, who could accurately guess the target referent. They then investigated whether participants were able to learn artificial language labels by integrating statistics across the most ambiguous naming events, which were instances that most adults could not guess the correct target referent. Significant learning was found only from the child's perspective but not from the observer's perspective, suggesting that the kind of input children experience may facilitate statistical aggregation. In a related follow-up study done by Zhang, Yurovsky and Yu (2015), participants were presented with a mixture of ambiguous and unambiguous first person videos and were asked to make guesses about the correct referent on a trial-to-trial basis. Their results suggest that word-learning is a continuous process that learners make progress gradually by integrating previous knowledge. Being able to remember and carry over partial knowledge, despite the uncertainty of the information at a moment, could facilitate learning and partial knowledge can be especially helpful when the learning situations are ambiguous.

Several recent studies have investigated word learning from learners' own perspective by placing lightweight head-cameras and eye-trackers on children while they interact with their parents. For example, researchers have found that referential uncertainty in 1½ year olds infants' own visual field is significantly reduced at the sensory level. The clutter and distraction in child's visual field are effectively reduced when objects are close to their eyes and head as close objects are visually large and can block the view of potential distracters (Yu & Smith, 2012). In addition, when parents played with and talked about novel objects with their toddlers, the visual properties (e.g. object's image size or centeredness relative to other objects) of the target object during naming predicted children's later novel object-name learning (Pereira, Smith & Yu, 2014). These perceptual cues available in children's view may play an important role in children's internal statistical computations. These findings are quite informative considering the previous assumption that cluttered everyday environment can cause a high degree of referential uncertainty. There is a need to take the learners' view into account and to study the visual input directly perceived by the learners, because ultimately the statistical information that makes contact with children's learning system matters the most. Even though low referential ambiguity facilitates word learning and parents do create relatively clean and unambiguous naming moments (Frank, Tenebaum & Fernald, 2013), naturalistic learning situations vary in their quality with some being more ambiguous than others, and some may facilitate learning and some may not.

In addition, reserachers have also started to investigate whether visual attention plays a role in infants' word learning process (e.g. Smith & Yu, 2013). Although past research on visual attention indicates that statistical word-learning is constrained by infants' developing attention system (Yu & Smith, 2011), little is known regarding how children visually attend to and select the target object among many other objects in view when parents name it during everyday interactions. Given that infants' attention is selective as they voluntarily direct their attention to certain aspects of the environment moment by moment, will they select a subset of information to attend to at each naming instance and aggregate their knowledge over time? Is it the case that when the adults utter a new word during interaction, children pay attention to a lot of objects, happenings, and properties that can possibly be a match, therefore word learning by aggregating information is so hard and impossible? In the current study, we explored the looking pattern of 12-month-old infants using naturalistic images with varying degrees of referential ambiguity in order to examine whether perceptual properties of objects in children's own view during naming moments would influence how young infants select candidate objects to build word-object mappings.

Experiment 1

The paradigm of presenting dynamic natural first-person scenes obtained from an infant's first person perspective to another age-matched infant while gathering on-line eye-tracking data has been done successfully in other studies (e.g. Aslin, 2009). Different from the original paradigm that used dynamic videos, we used still images in the current study. Our goal of Experiment 1 was to measure 12-month-old's looking behaviors during free viewing of natural word-learning scenes. Specifically, we wanted to compare and examine how different visual properties of target objects at naming moments influence the way infants allocate their attention. No spoken label was provided in this condition.

Participants. Twenty-five 12-month-old infants (10 female, ages ranged from 11.7 to 13.2 months, M_{age} = 12.28, SD_{age} = .43) participated the study. Parental consent was obtained for all participants in compliance with the IRB of Indiana University. All children received a gift for their participation.

Materials. Forty-four images were selected from a set of naming moment vignettes collected by Yurovsky et al. (2013) for their original study. This set of vignettes included play sessions from eight parent-child dyads. All vignettes were captured from children's first person view using head-mounted cameras during toy play with their parents and each vignette was 5 seconds long with the target name's onset occurred at the third second. As shown in Figure 1, we selected one frame from each 5-second naming window and systematically varied both the size (big vs. small) and location (centered vs. off-centered) of the target toy to create 4 experimental conditions (Figure 1). The four frames of the same object were selected from different naming moments.

Figure 1: Sample images from target object "ball" for four experimental conditions.

As shown in table 1: 1) there were many objects in view (ranging from 10 to 15) to which infants could direct their attention, which suggests that overall there was a high degree of ambiguity and uncertainty in all 4 experimental conditions; 2) there were also distinct differences in visual complexity and uncertainty among the four conditions. The target objects seem to be more visually dominant and salient in the big/centered condition and therefore more likely to attract infants' attention while the targets in small/off-centered

Table 1: Visual property details averaged across all 11 images in each condition.

	Mean number of objects in view	Mean proportional size of target object	Mean distance of target from center (max=1)
Big/Centered	10	22.30%	0.008
Big/Off-centered	11	16.30%	0.124
Small/Centered	14	4.70%	0.013
Small/Off-centered	15	5.20%	0.305

condition were embedded in a set of objects in view, therefore less noticeable. These learning scenes with varying degrees of uncertainty allow us to examine how infants direct their attention in those different contexts.

Apparatus. The learners' eye gaze was measured by a Tobii 1750 eye tracker. 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 corneal 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. E-prime software was used to present the stimuli and to automate the recording of eye location with the eye tracker software.

Procedure. Infants were seated on their caregivers' laps approximately 60cm from the monitor in a quiet room. Parents were instructed to keep their child seated, facing forward and refrain from talking to them or direct their attention. We also told parents to either look down or close their eyes throughout the entire procedure so as to not to influence their infant's behavior.

The point of gaze was calibrated with a toy animation that appeared randomly at five locations (four corners and center) across the screen, one at a time. After successful calibration, the first trial began with the centered presentation of an animation to orient infants' attention to the screen. As soon as infants looked at the center, pre-selected first person view images would be presented full-screen. In total, 44 images (11 toys, each has 4 conditions) were displayed for 7 seconds each. The temporal order of images was pseudorandomized so that images of the same object and images of the same condition do not appear consecutively. The first attention grabbing slide was interspersed every 4 trials to maintain child's attention. While infants were attending to the images, they also heard soft music played in the background. The entire testing session was about 6 minutes long.

Results and Discussion. Because perfect tracking in a continuous mode is not possible due to technical limitation of the eye-tracker, involuntary head movement or loss of attention, we only included trials with more than 50% of gaze data points. Included trials have an average of 80% of gaze data points. For data analysis, we fit linear mixed effect model to the data by using size or location as the fixed factor, and subject and item as random factors. In Experiment 1, we focused on analyzing gaze data to 1) quantify the degree of uncertainty; and 2) to measure how much time infants attend to the target objects.

Quantifying Degree of Uncertainty. To investigate whether the number of objects attended by infants differed when target size or location changes, we first measured the total number of objects attended and found that even given that there were more than 10 objects in view, and also given plenty of viewing time (7 seconds per image) to attend to many objects, infants only selectively attended 3-6 objects per trial ($M_{big/centered}$=3.74; $M_{big/off\text{-}centered}$=5.24; $M_{small/centered}$= 5.11; $M_{small/off\text{-}centered}$=5.94). As shown in Figure 2A, we did not find a significant main effect for size or location (size: β=1.13, p=.05; location: β=.75, p=.21).

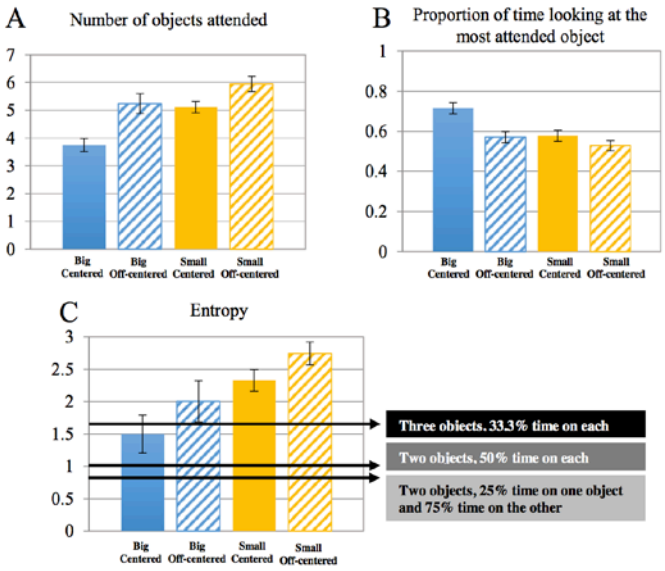

Figure 2: A. Mean number of objects attended; B. Mean proportion of time infants look at the most attended object; C. Mean entropy (averaged across trials in each condition).

Knowing that the subset of objects infants attended to was quite small, we further examined how infants allocated their attention among the subset of objects they chose. Do infants attend to those objects equally frequently or do they only primarily attend to one or two objects? To answer this question, we measured infants' proportion of time looking at the most attended object and found that across all four conditions, as shown in Figure 2B, infants spent more than 50% of time looking at one selected object ($M_{big/centered}$=.71; $M_{big/off\text{-}centered}$=.57; $M_{small/centered}$= .58; $M_{small/off\text{-}centered}$=.53). By fitting lmer models, we found that if the target size was big, infants spent more time looking at their most attended object (β = -.13, p<.01). However, location does not have an impact (β = -.07, p=.13). This finding suggests that even infants focused on only a few objects per trial, they predominantly only look at one object at least half of the time.

Next, to further capture the uncertainty that infants faced within a trial, we calculated entropy based on their looking times. In information theory, Entropy can be used to describe the uncertainty given a distribution. In the present case, given

n objects in view, we calculated $I = \sum_{i=1}^{n} P_i log \frac{1}{P_i}$ where p_i is the proportion of time looking at object i. This Entropy measure captures the dynamics of attention as it takes into account not only the number of objects attended but also the looking duration on each object. For example, if one looked at two objects equally, entropy equals to 1. If one looked at two objects, but one look is much longer (75% of the time) than the other (25% of the time), then entropy value would get lower and equals to 0.81 as it was a less uncertain situation compare with looking at two objects equally. If one looked at three object equally, then entropy gets higher and equals to 1.56 (Figure 2C). Thus, both more looks and a more even distribution of looks will cause the increase of entropy with high uncertainty while fewer and uneven looks will cause the decrease of entropy with low uncertainty. As shown in Figure 2C, entropy measures in all four conditions were relatively low, suggesting low uncertainty based on infants' looking behavior ($M_{big/centered}$=1.50; $M_{big/off\text{-}centered}$=2.00; $M_{small/centered}$= 2.33; $M_{small/off\text{-}centered}$=2.75). We then assessed whether size or location of target objects influenced how uncertain learners were based on entropy value. We found that infants tended to be more uncertain (higher entropy) when target size was small (β=.42, p<.05), but location was not a significant factor (β=.26, p =.15).

These results are quite informative as they support the idea that the visual dynamics of children' visual field might not be as noisy as people previously believed if we consider statistical learning from an embodied view (Yu & Smith, 2012). The present results also show that children's selective attention may simplify the learning problem even more because they only look at a subset of objects in their visual field and spend most time attending to only one of the selected objects.

Target Look. Because all frames are taken from natural naming moments, there is a correct named target for each scene even though participants were not aware of which object was being named. We next explored whether learners attended to the correct target or not without labels and we used two different ways to quantify this measure: 1) proportion of time infants look at the correct target in a given trial; 2) if we treat the object that was attended the most by infants as the one selected by them as the target, then how likely the object they select is the correct target.

By examining whether size or location of target objects influences how long infants look at the target object ($M_{big/centered}$=.34; $M_{big/off\text{-}centered}$=.15; $M_{small/centered}$=.24; $M_{small/off\text{-}centered}$=.11, Figure 3A), we observed that infants looked at the target object significantly longer when it is big in view (β= -.90, p<.001) and when it is centered in view (β=-1.01, p<.01). These results suggest that if the target's size is big or if it is centered relative to other objects in the visual field, infants are more likely to pay attention to that object and treat it as a potential referent if naming occurs.

Because infants were not aware that there was a potential target object in each scene, we were interested to see whether their most attended object during free viewing was likely to be the target, if so, whether the visual properties of the target toy influence their accuracy. Our results indicate that only location (β=-.32, p<.001) but not size (β=-.19, p=0.05) was a significant predictor ($M_{big/centered}$=.66; $M_{big/off\text{-}centered}$=.29; $M_{small/centered}$=.56; $M_{small/off\text{-}centered}$=.22, Figure 3B). This analysis provides evidence that when the target object is off-centered, infants are less likely to attend to it and treat it as a potential target even it is still big in view, which suggests that infants may have a center bias when free viewing natural scenes.

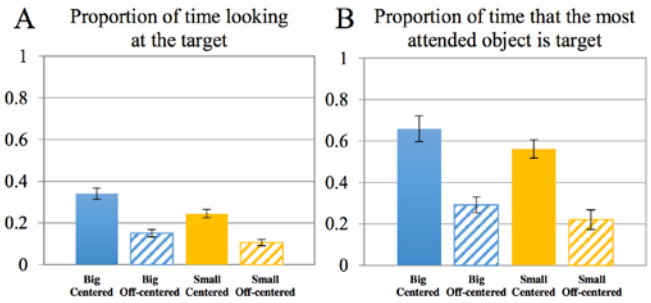

Figure 3: A. Mean proportion of time looking at the target object; B. Mean proportion of time that the most attended object is target (averaged across trials in each condition).

By analyzing free-viewing gaze data without naming in Experiment 1, we found that visual properties, such as size and location of objects in infants' own view can influence the way learners visually attend to those objects and the number of objects they are able to attend to is quite limited due to selective attention.

Experiment 2

Because we are interested in examining referential uncertainty during naming moments, in the second study, we investigated whether hearing a label during free-viewing would have an impact on how infants allocate their attention and whether their looking pattern changes after the label (e.g. look at more or fewer objects, stay longer or shorter on previously attended objects).

Participants. Twenty-three infants (11 females) between 11.4 and 12.6 months of age (M_{age} = 12.2, SD_{age} = .31) were recruited from the same population as in Experiment 1, none of these children participated in the previous experiment.

Materials. The same 44 images used in Experiment 1 were used in Experiment 2. A female native English speaker recorded the 44 labeling sentences that were infant directed. Toys' English labels were used. As shown in Figure 4, all labeling utterances were about 1 second long, with the onset of the utterance occurred at exactly the fourth second of each 7-second trial, so there were 3 seconds of silence both before and after the labeling sentences. To keep infants attentive, the same object was labeled using different sentence structures in different conditions, such as "Look at the __!" "There is a __!" See the__!" "It's a__!" and same sentence structure does

not occur consecutively. Same background music with lower volume was used.

Procedure. The procedure was the same as Experiment 1.

Results and Discussion. Mean percentage of gaze points contained across all usable trials is 83%. As shown in Figure 4, we are mainly interested in two types of comparisons: 1) compare looking behaviors happened in the last 3 seconds of the silence condition and the last 3 seconds of the label condition. This comparison controlled for the amount of visual experience infants received, the only difference between the two conditions was whether or not a label was presented; 2) calculate looking pattern changes between the first and the last 3 seconds of viewing for each condition, then compare the changes between silence and label conditions. In Experiment 2, we emphasized on analyzing gaze data to examine whether label would influence: 1) the number of objects infants select to attend; 2) the proportion of time infants attend to the correct target. For all subsequent analyses, we fit lmer models to the data and used label as the fixed factor and subject and item as random factors.

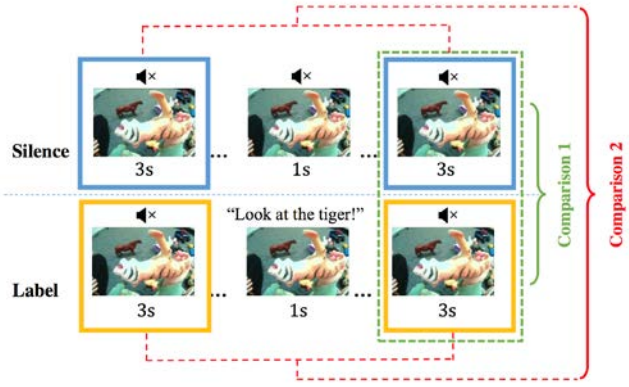

Figure 4: Types of comparisons implemented in Exp 2.

To compare looking behaviors observed in the last 3 seconds of the silence condition and the last 3 seconds of the labeled condition, we did not find a significant main effect of label on the number of objects attended (β=-.13, p=.46, Figure 5A), suggesting that label doses not influence how many objects infants choose to pay attention to. By comparing the number of objects attended before and after the label (first vs. last 3 seconds of the label condition), we found that the average number of new objects (the ones they did not attend to before the label) they chose to attend after the label was 1.33. The average number of new objects attended in the silence condition (first vs last 3 seconds of the silence) is 1.32, which is not significantly different from the label condition (β=.03, p=.77, Figure 5B). Our data suggest that infants do not change their looking patterns by selecting fewer or more objects to attend to because of the label. This is probably because if infants do not already know the referent's name, even with a label, there is still no clear indicator of which object might be the correct target. Although the information selected within a learning moment can be quite narrow that only a few objects are first attended and then stored in the memory, infant might still try to maintain more flexible visual attention and sample relatively broad co-occurrence data when the additional cues provided (e.g. label) could not help them narrow down the information selected further at the moment.

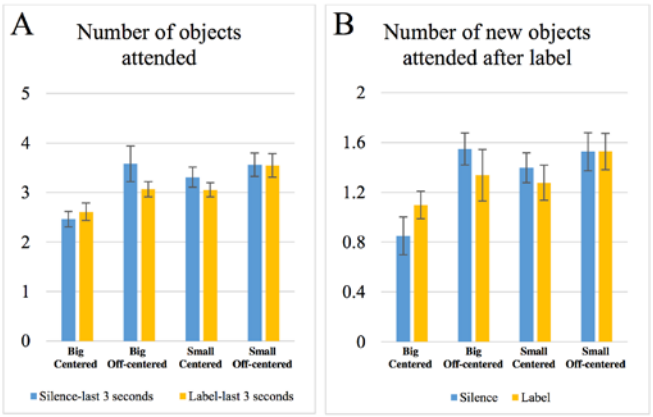

Figure 5: A. Mean number of object attended; B. Mean number of new objects attended after label in 4 conditions.

Because the labeling utterances are referring to the correct target in view, we further examined which object infants chose to pay attention to the longest and whether that object was the correct target. As shown in Figure 6, label does not influence the proportion of time infants look at the most attended object (β=.02, p=.21) nor the proportion of time that their most attended object is target (β=-.03, p=.32), suggesting that infants did not change their looking patterns dramatically after hearing the label.

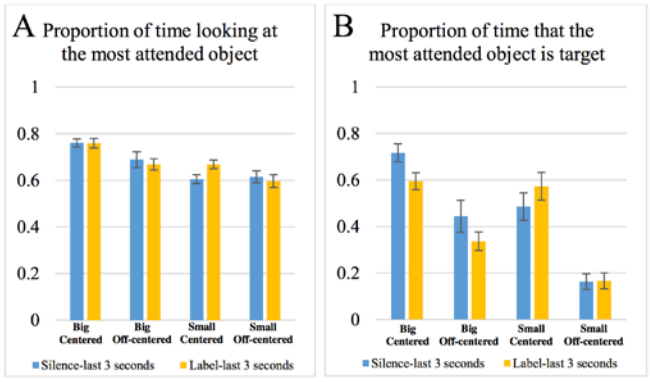

Figure 6: A. Mean proportion of time looking at the target object; B. Mean proportion of time that the most attended object is target in 4 experimental conditions.

The results in Experiment 2 are consistent with previous studies on parents' object labels in free play (e.g. Tomasello & Todd, 1983). They found that parents who use more follow-in labeling, which is the case the child is already attending to the object before labeling, have children with larger vocabulary, suggesting that just following and labeling what they have already attended to (instead of redirecting child's attention) would be quite effective because children would not switch their attention after hearing a label.

General Discussion

Our results show that perceptual properties of objects in infants' own view during naming moments dramatically influence how they select candidate objects to be considered to build word-object mappings. Experiment 1 results demonstrate that data available to statistical learners are not the data in the real world, but only a small subset of that data that is made into the learners' perceptual system at each learning moment. Such information is filtered through not only the dynamics of first person views, but also the learner's own developing attention system because it is not possible for infants to attend to everything in their own view. Thus, to address the question of whether natural learning moments are too complex for statistical learners to keep track of lots of information over time, we provide evidence to show that what the learners attend to at naming moments is not a large number of objects, but rather they attend to a small sample of available information in the world. This filtering process significantly simplifies the amount of information available for learners to carry over from one moment to the next, and to further process and integrate statistical evidence in their cognitive systems.

The quantitative results derived from gaze data can advance our understanding of the referential uncertainty problem encountered in real-life situations. At the same time, they are also in line with the previous results found using the cross-situational learning paradigm (Smith & Yu, 2008). The way infants learn word labels from real life learning moments might be similar to the way they learn words in cross-situational learning (CSL) tasks as in both cases they allocate their attention to only a few objects in view at a moment. Given that infants are able to learn the correct object-label mappings by aggregating information across trials in CSL tasks, it would be interesting to see whether they are also able to learn correct object names by collecting and accumulating information selected from first person scenes that resemble real-world learning situations. In addition, many adult studies using various paradigms (e.g. Yu & Smith, 2007; Zhang, Yurovsky & Yu, 2015) have shown that word-referent learning is a continuous statistical learning process and individual's ability to remember and carry over knowledge from past learning instances facilitates subsequent learning. One possible future direction along this line would be to design a word-learning experiment using first person view naming instances. By measuring learners' eye movement during training and comparing that with their learning outcome may allow us to understand real-time learning mechanisms, such as how statistical learners aggregate information moment by moment and whether the information learners select to attend to during training would link to what they learn at the end.

Despite the fact the environment young learners encounter is very complex and noisy, they are able to use selective attention to filter and clean up the inputs before processing them in their cognitive system. It is important to examine the underlying learning mechanisms by measuring and analyzing statistical information that is selected by, further stored and retained in the sensory, attentional, and memory processes as it is through the interactions of all these cognitive components in the learning system that young learners acquire the knowledge of solving word-learning problems and build their vocabularies.

Acknowledgments

This research was supported by NIH R01 HD074601. Special thanks to Lillian Hogan for data collection.

References

Aslin, R. N. (2009). How infants view natural scenes gathered from a head-mounted camera. *Optometry and vision science: official publication of the American Academy of Optometry*, *86*, 561.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, *9*, 1-24.

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

Pereira, A. F., Smith, L. B., & Yu, C. (2014). A bottom-up view of toddler word learning. *Psychonomic bulletin & review*, *21*, 178-185.

Quine, W. V. (1960). Word and Object. MIT Press. *Cambridge, MA*.

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

Smith, L. B., & Yu, C. (2013). Visual attention is not enough: Individual differences in statistical word-referent learning in infants. *Language Learning and Development*, *9*, 25-49.

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*, 3617-3632.

Tomasello, M., & Todd, J. (1983). Joint attention and lexical acquisition style. *First language*, 197-211.

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

Yu, C., & Smith, L. B. (2011). What you learn is what you see: using eye movements to study infant cross-situational word learning. *Developmental Science*, *14*, 165-180.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*, 244-262.

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science*, *16*, 959-966.

Vouloumanos, A., & Werker, J. F. (2009). Infants' learning of novel words in a stochastic environment. *Developmental psychology*, *45*, 1611.

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

Zhang, Y, Yurovsky, D., & Yu, C. (2015). Statistical Word Learning is a Continuous Process: Evidence from the Human Simulation Paradigm. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*.