Multimodal Attention Creates the Visual Input for Infant Word Learning

Sara E Schroer
Department of Psychology
University of Texas at Austin
Austin, Texas
saraschroer@utexas.edu

Chen Yu
Department of Psychology
University of Texas at Austin
Austin, Texas
chen.yu@austin.utexas.edu

Abstract—**Infant language acquisition is fundamentally an embodied process, relying on the body to select information from the learning environment. Infants show their attention to an object not merely by gazing at the object, but also through orienting their body towards the object and generating various types of manual actions on the object, such as holding, touching, and shaking. The goal of the present study was to examine how multimodal attention shapes infant word learning in real-time. Infants and their parents played in a home-like lab with unfamiliar objects with assigned labels. While playing, participants wore wireless head-mounted eye trackers to capture visual attention. Infants were then tested on their knowledge of the new words. We identified all the utterances in which parents labeled the learned or not learned objects and analyzed infant multimodal attention during and around labeling. We found that proportion of time spent in hand-eye coordination predicted learning outcomes. To understand the learning advantage hand-eye coordination creates, we compared the size of objects in the infant's field of view. Although there were no differences in object size between learned and not learned labeling utterances, hand-eye coordination created the most informative views. Together, these results suggest that in-the-moment word learning may be driven by the greater access to informative object views that hand-eye coordination affords.**

Keywords—Language acquisition, Multimodal perception, Sensorimotor development, Embodiment

I. Introduction

Infants' worlds are rich with multimodal information. A typically developing infant can hear, see, touch, and move around their environment. Critically, infants' bodies and motor abilities constrain the available information. Motor milestones (like sitting, crawling, and walking) and growing height change the infant's view of the world [1]. Shorter arms bring objects closer to the infant's face and thus bigger in their field of view (FOV) [2]. And the development of intrapersonal coordination gives infants the ability to attend to both their social partner's and their own hands as they manipulate objects [3]. Learning spoken languages relies on each of these sensory inputs and the constraints imposed by infants' bodies.

Much of early language learning research has focused on infants' auditory environment. The quantity of input, often measured as the total number of words infants hear, has been linked to later language abilities [4]. More recent research suggests, however, that it is the quality of input that matters most. The content of parent speech, such as the use of diverse vocabulary [5], infant-directed speech [6], and even conversational turns between parent and infant [7] are predictive of language outcomes and real-time word processing speed. When parents choose to talk and how they change the content of their speech in response to infant's behaviors is even more predictive of language milestones than the infant's own behaviors [8]. Nonetheless, speech is only one modality of input. Hearing words is largely meaningless unless infants can form mappings between words and their environment.

The solution to learning word-referent mappings likely depends on what infants see while they hear the label of an object or event. The use of head-mounted cameras and eye trackers have shown that infants have a view of the world that is unique from their parents'. There tend to be fewer objects in infant's FOV and those objects tend to be larger too [9]. A "visual signature" of successful naming moments has been identified – infants were more likely to learn the name of an object if their parent labeled it while it was big and centered in the infant's FOV [9]. Additionally, successful naming events have another "advantage" in that the labeled object is bigger in infant's FOV than the mean size of distractors [10]. The visual properties available in the infant's first-person view benefit not only infant learners, but other learners as well. Adult participants learned novel words better when stimuli were captured from a head-mounted camera on a toddler, as opposed to a traditional 3rd-person view [11]. Similarly, computer vision research has found that object recognition models perform better when trained on the toddler's view than an adult's [12]. Subsequent analyses suggest that the greater variability in object views created by infants played a more important role in the model's performance, not just the size of objects in the FOV [12].

These visual properties are not static features of the environment, but are created by active infants, most often through infants' hands. Infants often hold the labeled object during successful (but not unsuccessful) labeling events [9]. Additionally, infant holding of an object was more likely to create a dominant object view than when their parents held it [9]. Home observations saw that parents are more likely to label objects when their infant is in hand-eye coordination with the object (holding and looking at the object at the same time) [13]. The real-time effects of holding and manipulating objects have a cascading impact beyond a single interaction. The variability in object views created by 15-month-old infants' own hands during an experiment was predictive of their vocabulary growth over the next 6-months, but the variability generated by their parents was not [14]. In addition to changing the visual input to include more informative object views, hands themselves may act as a frame for attention and minimize the interference of distractors in the environment [15].

Of course, infants are more than just ears, eyes, and hands. Infants' whole bodies shape their learning opportunities. Most notably, there is ample research demonstrating that motor development changes parent-infant social interactions. Once infants make the transition from crawling to walking, they start "bidding" for parent attention differently (i.e., how they show objects to their parents) and, as a result, hear new types of information from their parents [16]. The development of independent locomotion grants infants the ability to explore their environment. Together, infants' increased ability to select their own learning experiences and the changes in parent speech may support the increased vocabulary growth that accompanies the transition from crawling to walking [17].

Despite the interconnectedness of these sensory inputs and evidence supporting theories of embodied cognition, few studies of language learning consider all four of the discussed modalities. The goal of the current study was to measure the multimodal information available to infants as they hear new words and how it affects learning outcomes. To do so, we invited parent-infant dyads into a home-like laboratory and asked them to play with unfamiliar toys. Although parents were not told that this was a study on word learning, we tested infant's knowledge of the unfamiliar object-label mappings after a short play session. While they were playing, dyads wore wireless head-mounted eye trackers. In contrast to our previous work [14,18], we used wireless eye trackers so that participants could freely move around the lab, allowing infants to select where they wanted to play and with which toys. Using this data, we were able to measure both the sensory input and infant actions during play, and link those with infant word learning acquired at test.

We chose to analyze sensory input in two ways. First, we compared infant visual and manual attention within the labeling utterances, as well as in the seconds leading up to and following a labeling event. We studied visual attention before the labeling event to see whether parents were following infant attention to label the attended object or shifting their infant's attention to a different object. By studying attention after labeling, we can see whether labeling extends infant attention to that object, perhaps creating more information for learning. Then, we analyzed the visual properties of infant's FOV during successful and unsuccessful labeling and tested how different types of visual and manual attention change these visual properties. In particular, we measured the number of objects in view and the size of the labeled object and distractors. We hypothesized that hand-eye coordination would be a better predictor of infant learning than only looking at or only holding the labeled object, because of the rich multimodal information hand-eye coordination creates. We also hypothesized that this learning advantage exists because hand-eye coordination creates the most informative object views during labeling utterances.

II. Methods

A. Data Collection

29 parent-infant dyads (infants aged 12-26 months, average = 17.2; 12 F) played in a home-like laboratory environment for 10 minutes (or until the infant no longer wanted to wear the eye tracker, M = 7.12 minutes [range = 2.22-11.26 min]). Dyads were given 10 unfamiliar toys to play with and parents were asked to use a specific label for each toy (Fig. 1a). Parents were asked to play as they would at home and were not told that we were interested in word learning. During the play session, parents and infants both wore wireless head-mounted eye trackers (Pupil Labs, Fig. 1b). The eye trackers consist of two cameras – one that is centered on the participant's forehead to capture the view in front of them and a second that is pointed towards their eye to record eye movements. After the experiment, the eye tracking videos were calibrated to determine where the participant was looking in their FOV throughout the experiment (Yarbus). Using wireless eye trackers allowed participants to move freely around the lab space, promoting a more naturalistic, free-flowing interaction.

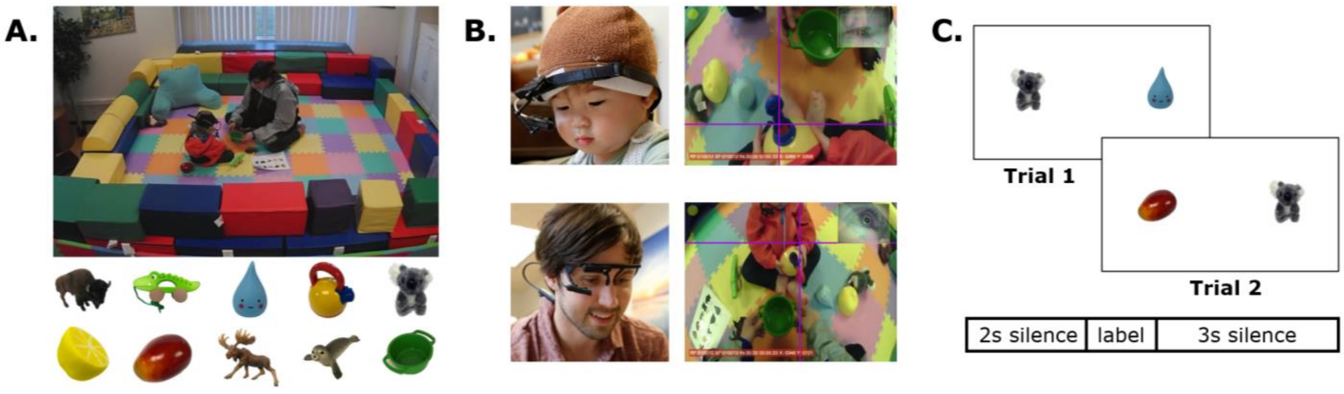

Fig.1. A. Top is the lab space with a dyad participating and below are the 10 objects used in the study. Parents were given a piece of paper that provides the labels to use for each toy. B. Top left and bottom left show an infant and parent wearing the eye tracker. The top right and bottom right images are a still from the infant's and parent's eye trackers during the same moment shown in (A). C. Infants were tested on their knowledge of object-label mappings after the experiment. Shown are the two trials used to test "koala". Each trial began with 2s silence, then the labeling utterance for 1s, followed by 3s silence.

After the play session, infants were tested on their knowledge of the object-label mappings (Fig. 1c). Dyads were brought into a smaller room equipped with a screen-based eye tracker (SMI REDn Scientific Eye Tracker). Infants sat on their parent's lap facing the screen and parents were asked to keep their eyes closed. The test was composed of 20 trials, so that each object could be tested twice. During each trial, two objects, a target and distractor, were presented on a white screen for 7 seconds. After 2s of silence, infants heard a 1s-long labeling utterance for the target object ("where's the X?"), which was then followed by 3s of silence. The order and presentation of trials was pseudo-random – the 20 trials were divided into two sets of 10 (so each object was tested once before any object was repeated) and target-distractor pairings were modified as needed so that different objects from the experiment served as the distractor in the two trials. Tests were then scored off-line. Only trials in which the infant looked at the screen for more than 1s of the 3s of silence after the naming event were scored. A trial was scored as "correct" if the participant looked at the target for a greater proportion of this 3s window than the distractor. Conversely, a trial was "incorrect" if the infant looked more at the distractor. Infant were considered to have "learned" an object-label mapping if they got both test trials correct and did "not learn" an object-label mapping if both trials were incorrect. Objects that did not have usable trials or that had one correct and one incorrect trial were not included in the analyses.

B. *Behavioral Coding and Analyses*

After the experiment, participants' visual attention and manual attention, as well as parent speech, were coded frame-by-frame. Visual attention was coded by identifying where the participant's gaze fell. Fixations could be to one of the provided toys, their social partner's face, or elsewhere. For manual attention, coders annotated each hand separately to determine whether a participant's left, right, or both hands were touching one of the toys. All instances of object touching were considered holding in these analyses. Parent speech was transcribed at the utterance-level. If parent speech was separated by more than 400ms of silence, it was divided into two utterances regardless of semantic content (e.g., a single sentence could be split into two utterances). We then identified the utterances in which parents labeled one of the objects their infant did or did not learn, also referred to as successful and unsuccessful naming events.

To specifically answer how infant's visual attention varies when learned or not learned objects are labeled, we measured infant attention in 3 temporal windows: 3s before the onset of the labeling utterances, during the labeling utterance, and the 3s after the offset of the labeling utterance. Three modalities of attention were defined for this first set of analyses: infants holding but not looking at the labeled object (holding only), infants looking at but not holding the labeled object (looking only), or infants holding and looking at the labeled object at the same time (hand-eye coordination, Fig. 2). We then created an event-level corpus dataset of each instance of successful and unsuccessful naming, collapsing across subjects and objects. We used logistic regressions to test if proportion of infant attention during or around labeling could predict learning at test [19]. We included a random effect of subject in the model. A separate regression was used for each type of attention (modality and temporal window). All models were compared to a null model (random effect only) using a chi square test.

C. *Object Size Calculation and Analyses*

Our second set of analyses explored whether infants' visual scenes differed during successful and unsuccessful naming utterances. Previous work shows that infants are more likely to learn a word-object mapping if the object is big and visually salient when hearing the label [9,10]. Accordingly, we defined a pixel-level visual property of interest: the size of the labeled object in the infant's FOV. Similar to [12], YOLO object detection [20] was used to automatically detect the size and location of objects in the infant's FOV. For the purpose of these analyses, object size was defined as the proportion of pixels the object occupied.

The goal of the visual properties analyses was to see how object size differed during the successful and unsuccessful labeling events, specifically as a function of the type of attention infants engaged in. For these analyses, a fourth "modality" of attention was defined – no attention to the object. Each utterance was then assigned into an attention category. If infants never looked at or held the object during an utterance, the entire utterance was coded as "no attention". Proportion of time spent in hand-eye coordination, looking only, and holding only were compared in a winner-take-all approach – e.g., if infants spent the greatest proportion of the utterance in hand-eye

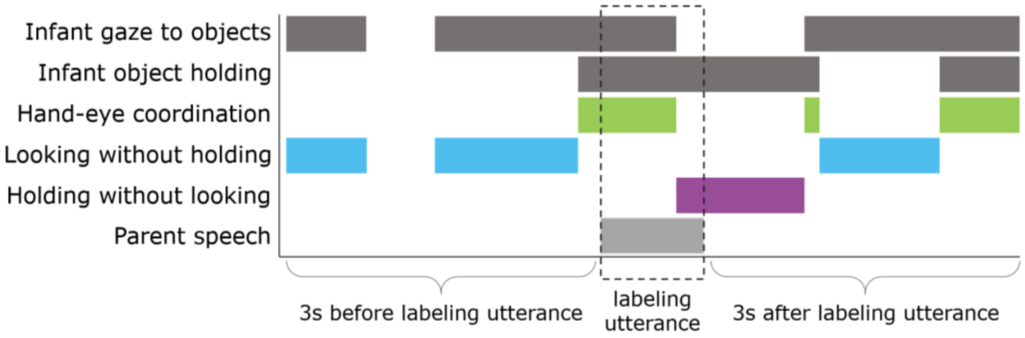

Fig.2. Example of a visualization stream showing a labeling utterance (a parent saying, "can you take the moose?") and infant's attention from 3s before the onset of the utterance to 3s after the offset of the utterance. Each rectangle represents the onset and offset of a behavior. In the stream, the top two rows in dark grey show the infant's "raw" looking at and holding the labeled object. The bottom row shows parent speech. We identified the moments when the infant was in hand-eye coordination with the target object (green), was just looking at the object but not holding it (blue), or just holding without looking (purple).

coordination, then the entire utterance was assigned to the “hand-eye coordination” category.

For each utterance we then calculated the mean size of the labeled object throughout the entire utterance, as well as the mean size of all of the distractor objects in view (referred to as “mean size of distractor”). We first compared the size of learned and not learned objects using a linear mixed effects model, with random effects for subject and object. We then used similar models to compare object size during the successful and unsuccessful utterances for each of the 4 types of attention. Then using separate ANOVAs for successful and unsuccessful utterances, we asked whether the size of the targets and mean size of distractors could be predicted by whether that object was named and the type of attention the infant was engaged in. We also compared the number of objects in view during labeling.

III. Results

Across all subjects, there were 66 learned objects (average of 2.3/subject) and 56 not learned objects (average of 1.9/subject). Learned and not learned objects were labeled a similar amount by parents (M learned = 3.88, M not learned = 4.43, p = 0.479), resulting in a corpus dataset of 256 successful labeling events of learned objects and 248 unsuccessful labeling events of not learned objects. The average utterance lasted 1.25s. Because amount of labeling alone did not differ between learned and not learned objects, we hypothesized that infant multimodal attention during and around the labeling utterances would be a greater predictor of learning.

A. *What Behaviors Predict Word Learning?*

We first analyzed whether the three types of attention, multimodal attention (hand-eye coordination), unimodal visual attention (looking only), and unimodal manual attention (holding only) could predict learning outcomes. To gain a broader view of an object-labeling instance, we defined three temporal windows that were analyzed separately: before, during, and after the labeling utterance (Fig. 3, Table 1).

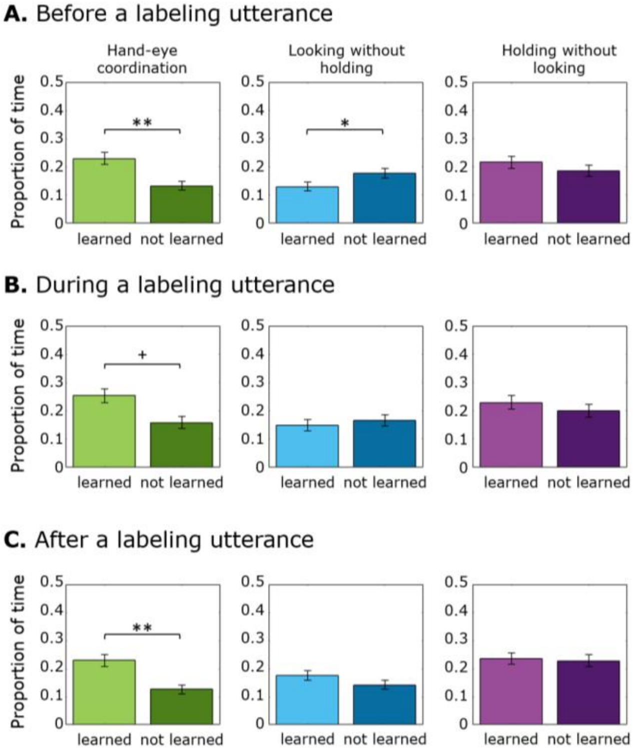

Fig.3. The average proportion of each attention window – before (A), during (B), or after the labeling utterance (C) – that the infant spent attending to the labeled object with hand-eye coordination (green), looking without holding (blue), and holding without looking (purple). For each type of attention, the light shade is attention to learned objects (left side of each plot) and the darker shade is attention to not learned objects (right). Error bars show standard error. + is trending, * is $p < 0.05$, ** is $p < 0.01$

Infant attention before the labeling utterance was predictive of learning outcomes. In particular, infants spent more time in hand-eye coordination with the labeled object before a successful labeling utterance (M = 0.230) than an unsuccessful utterance (M = 0.133, *p* = 0.004). Conversely, the proportion of time infants were only looking at the labeled object negatively predicted learning (learned = 0.131, not learned = 0.177, *p* = 0.014). The proportion of time only holding the object did not predict learning (learned = 0.217, not = 0.185, *p* = 0.545).

Infant attention during the labeling utterance was not as predictive of learning outcomes, perhaps because this temporal window was roughly half of the 3s windows before and after labeling utterances. Although infants spent a greater proportion of time in hand-eye coordination when the object’s label was

TABLE 1. Analyses on Attention

	Average proportion of time attending to labeled object		**learned? ~ attention + (1|subject)**		
	Learned	Not Learned	*β*	*p*	*null model comparison*
3s before labeling utterance					
Hand-eye coordination	0.230 (sd = 0.334)	0.133 (sd = 0.240)	1.071	0.004**	χ2(1)= 8.478, p = 0.004
Looking without holding	0.131 (sd = 0.240)	0.177 (sd = 0.270)	-1.020	0.014*	χ2(1)= 6.050, p = 0.014
Holding without looking	0.217 (sd = 0.336)	0.185 (sd = 0.314)	0.199	0.545	n.s.
During labeling utterance					
Hand-eye coordination	0.253 (sd = 0.402)	0.158 (sd = 0.335)	0.547	0.056†	χ2(1)= 3.663, p = 0.056
Looking without holding	0.150 (sd = 0.314)	0.167 (sd = 0.301)	-0.352	0.295	n.s.
Holding without looking	0.229 (sd = 0.388)	0.200 (sd = 0.368)	0.250	0.378	n.s.
3s after labeling utterance					
Hand-eye coordination	0.229 (sd = 0.338)	0.126 (sd = 0.255)	1.101	0.003**	χ2(1)= 9.343, p = 0.002
Looking without holding	0.175 (sd = 0.284)	0.142 (sd = 0.241)	0.446	0.269	n.s.
Holding without looking	0.236 (sd = 0.336)	0.228 (sd = 0.357)	0.005	0.987	n.s.

learned, this result was only trending on significance (learned = 0.253, not learned = 0.158, p = 0.056). Additionally, the proportion of time only looking at (learned = 0.150, not learned = 0.167, p = 0.295) or only holding the labeled object (learned = 0.229, not learned = 0.200, p = 0.378) did not predict learning.

Infant attention after the labeling utterance followed a similar pattern. The proportion of time the infant was in hand-eye coordination with the just-labeled object positively predicted learning (learned = 0.229, not learned = 0.126, p = 0.002). In contrast, the proportion of time only looking at the object (learned = 0.175, not learned = 0.142, p = 0.269) or only holding the labeled object after it was labeled (learned = 0.236, not learned = 0.228, p = 0.987) did not predict learning.

These results suggest that not all attention is equal. Hand-eye coordination provides multimodal input and information-rich data to support learning. Compared to only looking or only holding an object, hand-eye coordination may be creating a "learning advantage." To begin understanding the nature of this learning advantage, we analyzed the visual properties of the labeled object in infant's FOV during labeling utterances.

B. Do Attention Types Provide a Learning Advantage?

To analyze the visual properties of labeling events we compared the number and size of objects in view during successful and unsuccessful labeling; as well as whether attention types created different types of information. To begin, we saw no difference in the number of objects in view as a function of learning outcomes (p = 0.086; using a Welch two sample t-test). Successful utterances had an average of 5.50 objects in view and the target was present for 74.22% of utterances. Unsuccessful utterances had an average of 5.88 objects in view and 76.31% of utterances had the target present.

We then analyzed whether size of the labeled object could be predicted by learning outcomes. There was no difference in the size of the labeled object when it was learned (0.070 of the FOV) or not learned (0.067, p = 0.742). Similarly, when comparing the size of learned and not learned objects during utterances of each type of attention, we saw no differences for no attention to the labeled object (learned = 0.019, not learned = 0.029, p = 0.604), only holding (learned = 0.052, not = 0.046, p = 0.308), only looking (learned = 0.088, not = 0.084, p = 0.946), or hand-eye coordination (learned = 0.134, not = 0.139, p = 0.150). Although the visual properties of learned and not learned objects did not significantly differ during naming utterances, there were noticeable differences in size of the labeled object across the four attention types.

In line with the "learning advantage" analyses in [10], we then compared the size of the labeled object to the mean size of the distractors. Successful and unsuccessful naming moments were analyzed separately (see section II.C). The labeled object was bigger than the mean size of distractors in view for both successful (F(1,504) = 82.36, p < 0.001) and unsuccessful object-labeling utterances (F(1,488) = 77.27, p < 0.001). These results suggest that naming utterances in general occur when the infant has an advantageous view of the labeled object. We also found a main effect of attention type for both successful (F(3,504) = 29.14, p < 0.001) and unsuccessful utterances (F(3,488) = 22.46, p < 0.001). As shown in Fig. 4, the size of the target increases across attention types: no attention has the smallest mean size and hand-eye coordination has the largest mean object size. The interaction between naming and attention type was also significant for both successful (F(3,504) = 26.75, p < 0.001) and unsuccessful utterances (F(3,488) = 28.12, p < 0.001). Most notably, for both learned and not learned objects, the differences in target and distractor size were significant for looking only and hand-eye coordination (ps < 0.001), but not for holding only and no attention.

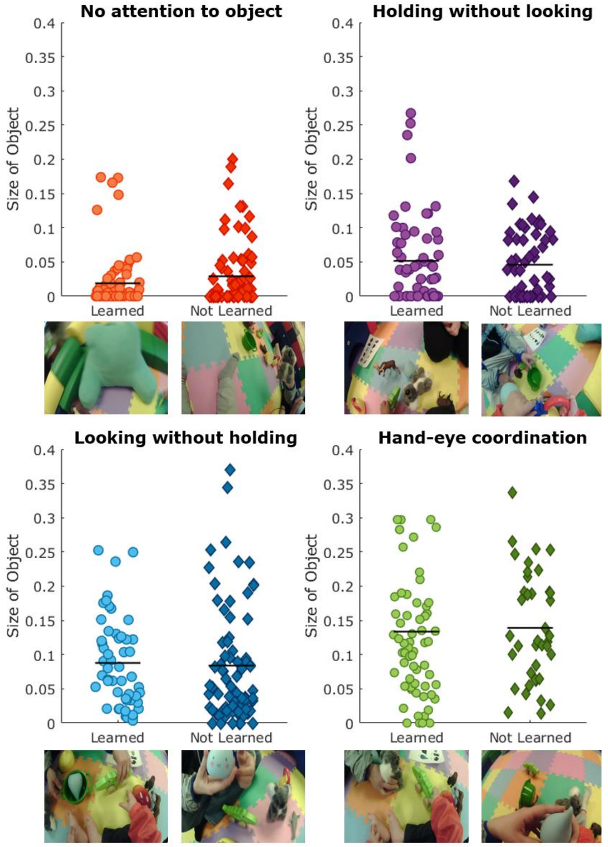

Fig.4. Mean size of the labeled object in proportion of pixels for each attention type when the label is learned (lighter shade) or not learned (darker shade). Black line shows overall mean. Below each scatter plot is an example image for each attention type. The learned object is the green crocodile and the not learned object is the blue drop (see Fig.1).

IV. Discussion

The goal of the current paper was to study the multimodal information that supports the real-time learning of new words. Infants and their parents played in a home-like environment while wearing wireless head-mounted eye trackers. Wireless sensing afforded unconstrained movement for the participants, as well as high-density data on visual attention. We first analyzed infant attention during and around the labeling utterances. We found that successful object naming instances are characterized by infants being in hand-eye coordination with the labeled objects. In particular, our data suggest that parents follow infant hand-eye coordination and that labeling the attended object may potentially extend that type of attention [18]. We hypothesized that hand-eye coordination is creating more informative object views than other attention types, such as only holding or only looking at the labeled object.

To test this hypothesis, we then analyzed the visual properties of successful and unsuccessful labeling utterances. Based on previous research [9,10], we chose to measure object size. Although the size of learned and not learned objects did not differ during naming, we found that the different types of infant attention generated different visual information. Hand-eye coordination produced higher quality information than the other types of attention. Supporting the importance of quality and quantity, learned objects had more hand-eye coordination and thus greater access to this high-quality information.

Our results support an embodied view of language acquisition and further define the information needed for word learning. We extended the findings in [9,10] to demonstrate that hand-eye coordination in particular, not just holding, creates informative views of object. We did not, however, find a difference in the visual properties of successful and unsuccessful naming, which may be due to a dramatic difference in experimental setting. [9, 10] were conducted in a small, all-white room with dyads seated at a table to play with 3 objects. The more naturalistic environment we used provides different affordances and changes the information available to the infant. Additionally, our analyses are also limited in scope. Future work will better match [9,10] by studying the frame-by-frame temporal profiles of infant multimodal attention in the moments before, during, and after object naming. We believe these analytical changes will better replicate [9,10] as attention leading up to and after naming were the most predictive windows in our behavioral analyses. Nonetheless, object holding may be providing a learning advantage that would not be captured in the visual scene (see [15]). Our more naturalistic environment may better highlight the importance of multisensory information and integration than a traditional lab setting. We are also beginning to collect parent-infant eye tracking data at home to see how well our findings extend to the actual playing field of word learning. The informative big-and-centered object views have been observed at home [21], suggesting that we would likely replicate our major findings.

To further test our hypothesis that hand-eye coordination creates the most informative object views for language learning, we are developing an ideal learner model. Based on [22], we will train an object recognition model on the object views created by participants in this experiment – and compare performance on no attention, holding only, looking only, and hand-eye coordination training sets. We are also beginning to analyze the variability in object views during successful and unsuccessful naming utterances [12]. More generally, our work is continuing to explore why and how infants create informative object views and how to use these findings to improve the training sets used in computer vision research.

Acknowledgment

This work was supported by the NIH (R01HD074601 and R01HD093792 to CY). SES was supported by the NSF GRFP (DGE-1610403) and the NIH (T32 Grant #HD07475). We thank the Computational Cognition and Learning Lab at Indiana University and the Developmental Intelligence Lab at UT Austin for their support in data collection and processing.

References

[1] Frank, M., Simmons, K., Yurovsky, D., & Pusiol, G. (2013). Developmental and postural changes in children's visual access to faces. In *Proc of the Annual Meeting of the Cognitive Science Society* (Vol. 35, No. 35).

[2] Smith, L. B., Yu, C., & Pereira, A. F. (2011). Not your mother's view: the dynamics of toddler visual experience. Dev Sci, 14(1), 9–17.

[3] De Barbaro, K., Johnson, C. M., & Deák, G. O. (2013). Twelve-month "social revolution" emerges from mother-infant sensorimotor coordination: A longitudinal investigation. *Hum Dev*, 56(4), 223-248.

[4] Hart, B., Risley, R. R. (1995). *Meaningful differences in the everyday experiences of young American children*. Baltimore, MD: Paul H. Brooks.

[5] Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Dev*, 83(5), 1762-1774.

[6] Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychol Sci*, 24(11), 2143-2152.

[7] Gilkerson, J., Richards, J. A., Warren, S. F., Oller, D. K., Russo, R., & Vohr, B. (2018). Language experience in the second year of life and language outcomes in late childhood. *Pediatrics*, *142*(4).

[8] Tamis-LeMonda, C. S., Bornstein, M. H., & Baumwell, L. (2001). Maternal responsiveness and children's achievement of language milestones. *Child Dev*, 72(3), 748-767.

[9] Yu, C., & Smith, L.B. (2012). Embodied attention and word learning by toddlers. *Cognition,* 125(2), 244-262.

[10] Pereira, A.P., Smith, L.B. & Yu, C. (2014). A bottom-up view of toddler word learning. *Psychol Bull & Review*.

[11] Yurovsky, D., Smith, L.B., & Yu, C. (2013). Statistical word learning at scale: the baby's view is better. *Dev Sci,* 16:6 (2013), 959-966.

[12] Bambach, S., Crandall, D. J., Smith, L. B. & Yu, C. (2018) Toddler-inspired visual object learning. *Advances in Neural Information Processing Systems (NeurIPS),* 31.

[13] West, K. L., & Iverson, J. M. (2017). Language learning is hands-on: Exploring links between infants' object manipulation and verbal input. *Coge Dev* 43, 190–200.

[14] Slone, L.K., Smith, L.B., & Yu, C. (2019). Self-Generated variability in object images predicts vocabulary growth. *Dev Sci.*

[15] Davoli, C. C., & Brockmole, J. R. (2012). The hands shield attention from visual interference. *Atten, Percept, & Psycho*, *74*(7), 1386-1390.

[16] Karasik, L. B., Tamis-LeMonda, C. S., & Adolph, K. E. (2014). Crawling and walking infants elicit different verbal responses from mothers. *Dev Sci*, *17*(3), 388-395.

[17] He, M., Walle, E. A., & Campos, J. J. (2015). A cross-national investigation of the relationship between infant walking and language development. *Infancy*, *20*(3), 283-305.

[18] Schroer, S., Smith, L., & Yu, C. (2019). Examining the multimodal effects of parent speech in parent-infant interactions. In *Procs of the 40th Annual Meeting of the Cognitive Science Society.*

[19] Kuznetsova A, Brockhoff PB, Christensen RHB (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." *J Stat Softw*, 82(13), 1–26.

[20] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proc of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

[21] Suanda, S.H., Barnhart, M., Smith, L.B. and Yu, C. (2019), The signal in the noise: The visual ecology of parents' object naming. *Infancy*, 24: 455-476.

[22] Tsutsui, S., Chandrasekaran, A., Reza, M.A., Crandall, D., & Yu, C. (2020). A computational model of early word learning from the infant's point of view. In *Proc of the Annual Meeting of the Cognitive Science Society.*