Embodied Attention Resolves Visual Ambiguity to Support Infants' Real-Time Word Learning

Sara E Schroer and Chen Yu
University of Texas at Austin, Department of Psychology, Austin, TX 78712 USA

Abstract

The input for early language learning is often viewed as a landscape of ambiguity with the occasional high-quality naming event providing resources to resolve uncertainty. Word learning from ambiguous naming events is often studied using screen-based cross-situational learning tasks. Little is known, however, on how ambiguity impacts real-time word learning in free-flowing interactions. To explore this question, we asked parent-infant dyads to play in a home-like environment with unfamiliar objects while wearing head-mounted eye trackers. After the play session, we tested whether infants learned any of the object-label mappings and categorized individual words as learned or not learned. Dyadic behaviors and the visual information available to infants during the naming moments of learned and not learned words were analyzed. The results show that infants' embodied attention during ambiguous naming moments was the key to predicting learning outcomes. Specifically, infants held and looked at the target object longer in ambiguous instances that led to learning. Our results emphasize the importance of studying word learning in naturalistic environments to better understand the cues infants use to resolve ambiguity in everyday learning contexts.

Keywords: word learning; parent-infant interactions; eye tracking; embodied attention; sensorimotor development

Introduction

The language learning environment is frequently described in terms of the quantity and quality of input. This research often focuses on global measures of how much speech children hear and how rich parent speech is, using metrics such as diversity of vocabulary (e.g., Rowe, 2012). Quality and quantity can also be used at a smaller scale to study how individual words are learned or to measure language exposure in an interaction. Recently, Yu et al. (2021) studied the relationship between quality and quantity by analyzing how much infants attended to labeled objects during free play. They found a striking bimodal distribution. Although infants often only looked at one object, naming events were either "highly informative" with attention on the target or "misleading", with attention on a single competing distractor.

Cross-situational learning offers a mechanistic explanation of how infants could learn from both high- and low-quality naming moments. The hypothesis of cross-situational learning is that infants can integrate information across multiple naming events. Since correct object-label pairs are more likely to occur than incorrect pairings, mappings can eventually be determined by aggregating these statistics across naming events (e.g., Siskind, 1996; Zhang, Yurovsky, & Yu, 2021). Learners can accumulate statistics across high-quality and ambiguous naming events to discover object-label mappings (Yurovsky, Smith, & Yu, 2013). And while sheer exposure to many ambiguous naming events is enough for learning (Zhang et al., 2021), a few high-quality moments where an object-label mapping is easily inferred can scaffold the learning process (e.g., Clerkin & Smith, 2022).

Recent research on cross-situational learning in infants has explored the extent to which some uncertainty can support learning. In lab-based experiments, background noise, variability, and even referential ambiguity improved learning outcomes, as opposed to hurt them (Twomey, Ma, & Westermann, 2018; Bunce & Scott, 2017; Cheung, Hartley, & Monaghan, 2021). But too much ambiguity, or too many misleading naming events, creates bad statistics for cross-situational learning (Bunce & Scott 2017; Zhang & Yu, 2017). Luckily for infant learners, the labels being produced by social partners often co-occur with other behaviors.

Social cues, such as object handling and gaze, are an important means of resolving ambiguity for both infant and adult learners (Baldwin, 1993; MacDonald, Yurovsky, & Frank, 2017). In moments of high ambiguity, parents are more likely to generate gestural cues to support infant learning (Cheung, Hartley, & Monaghan, 2021). And infants can even differentiate between referential cues to a target object and passive holding of competitor objects (Baldwin, 1993). The role of hands in reducing uncertainty is not surprising, as object handling is a powerful tool for facilitating dyadic coordination in parent-infant interactions.

In naturalistic interactions, parents' and infants' hands can be crucial in resolving referential ambiguity and creating high-quality naming moments. Dyads often attend to each other's hands during free play and can use object handling as a cue to infer what their social partner is attending to (Yu & Smith, 2017). Attention to each other hands' matters because hands create scaffolding to support infant learning. When infants hold objects, the held object becomes big and centered in their field of view (FOV) and naming during these moments supports learning (Yu & Smith, 2012). Hand-eye coordination often elicits naming from parents (West & Iverson, 2017) and increased hand-eye coordination is predictive of real-time learning (Schroer & Yu, 2022). Missing from this body of literature, however, is an understanding of the cognitive mechanisms through which infant object handling is so important for learning. We hypothesize that the dominant object views described by Yu & Smith (2012) hold the key to linking manual action with word learning.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

The goal of the present study was to quantify the information infants have access to during free-flowing interactions and examine how real-time word learning is affected by ambiguous information. We hypothesized that infants' embodied attention would play a key role in resolving ambiguity. To test this hypothesis, we invited parent-infant dyads into a home-like lab environment to study the patterns of naturalistic interactions that may create or resolve visual ambiguity, when the intended label of a target is ambiguous based on the information available in infant's FOV. In contrast to Yu & Smith (2012), we used head-mounted eye trackers to get a direct measure of infant gaze. We compared the visual attention and object handling of parents and infants during naming instances that varied in information quality and linked these real-time behaviors to word learning outcomes. We assessed the information that is available to infants during naming, as well as whether they are actually using that information to learn.

Methods

Participants

Twenty-nine 12- to 26-months-old infants (mean age = 17.2, 12 F) and their parents were recruited from a primarily White, non-Hispanic community of working- and middle-class families in the Midwest of the United States. Most parents spoke to their infants exclusively in English during the study, but no data was collected on at-home language use. The data used in this paper were previously described in Schroer & Yu (2022), but all presented analyses are unique.

Data Collection

Parent-infant dyads played in a home-like lab with 10 unfamiliar toys for 10 minutes (or until the infant became fussy, mean =7.12min [range=2.22-11.26min]) (**Figure 1**). The 10 toys were selected as objects that infants are unlikely to know the names of as they are not included on the MacArthur-Bates Communicative Development Inventory (Fenson et al., 1993). Parents were instructed to play as they would at home. Parents were asked to use an assigned label for each object, but were not given instructions on how often to label objects and were not told that there was a word learning test after the play session.

While playing, dyads wore wireless head-mounted eye trackers (Pupil Labs Core). The eye trackers were equipped with a camera that records the participant's eye movement and a scene camera that captures their FOV. The eye trackers were connected to an Android smart phone that transmitted data to a nearby computer. Participants wore jackets with a pocket sewn onto the back to hold the phone. This set-up allowed participants to move around freely during the experiment, since they were not tethered to any computer. The eye tracking videos were calibrated after data collection to determine where participants were looking in their FOV.

After the play session, dyads were brought into a smaller testing room. Infants sat on their parents' laps facing a computer screen and screen-based eye tracker (SMI REDn Scientific Eye Tracker). The word learning test consisted of 20 7s-long trials, divided into two blocks of 10 trials. In each trial, two objects were presented on the left- and right-hand side of a white screen. After 2s of silence, a 1s-long labeling utterance was played ("where's the X?"), followed by 3s of silence. Infants' knowledge of each object was tested twice, with a different distractor in the two trials and the target appearing on each side of the screen.

The tests were scored offline. Trials were only scored if infants looked at the screen for more than one third of the 3s window after the label. A trial was "correct" if infants looked at the target (labeled) object for a greater proportion of the 3s

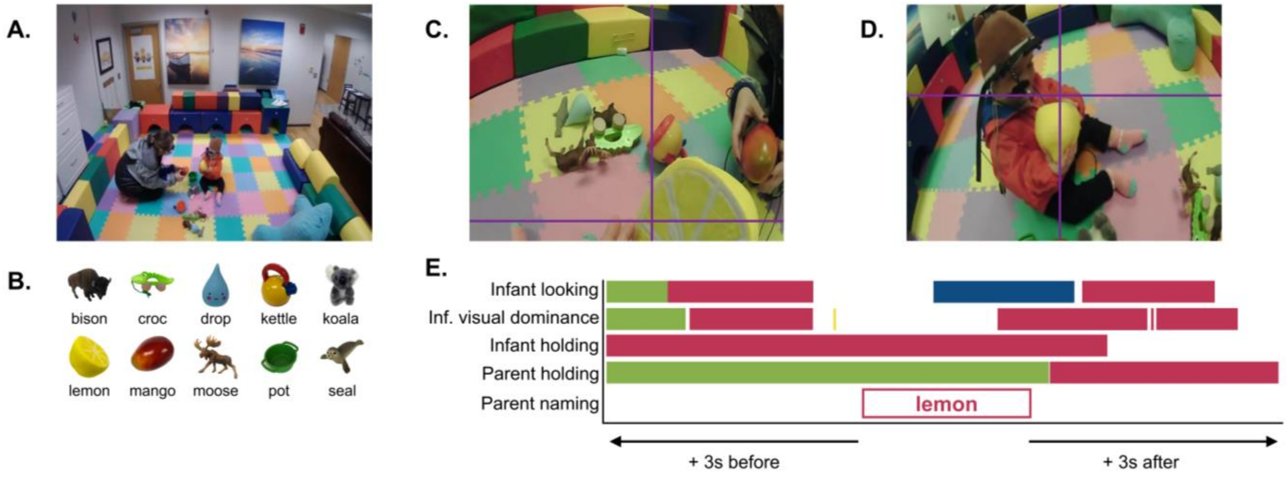

Figure 1. On the left side is the play space used in the study (A) and the 10 unfamiliar object-label mappings (B). The right shows the infant's (C) and parent's (D) view during a frame taken from a naming moment of the lemon. The purple crosshair indicates their gaze. (E) shows the infant's visual attention, the dominance of objects in the infant's FOV, and infant's and parent's object handling during the temporal window used in our analyses. Each rectangle represents the onset and offset of a behavior with the color indicating the region of interest being attended to (magenta is lemon, green and yellow are other objects, and blue is social partner's face).

window than they looked at the distractor and "incorrect" if the infant looked more at the distractor. If the infant got both trials for an object correct, the object-label mapping was coded as "learned". If both trials were incorrect the object was considered "not learned". Across all 29 infants, 66 object instances were learned (mean=2.3 objects) and 56 were not learned (mean=1.9 objects). Objects that had one correct and one incorrect trial (mean=3.6 objects) or were missing at least one test trial (mean=2.2 objects) were excluded from the presented analyses. Despite the wide age range in our experiment, the number of object-label mappings infants learned or did not learn was not correlated with age (*p*s > 0.094) or concurrent vocabulary size (measured with the MCDI, *p*s > 0.382).

Behavioral Coding

Visual attention was coded frame-by-frame using an in-house coding program. We annotated when participants looked at the 10 toys and their social partner's face. We also coded the objects participants touched frame-by-frame. Any contact with an object was considered touch. Lastly, we transcribed parent speech at the utterance level in Audacity. We defined utterances as any parent talk including non-word vocal play (such as saying "vroom vroom"). Consecutive utterances had to be separated by at least 400ms of silence (following Yu & Smith, 2012). We then identified any time the parent labeled an object that the infant learned (N = 256) or did not learn the name of (N = 248). Learned and not learned objects were labeled a similar amount (learned = 4.0 times, not learned = 4.6 times; *p* = 0.479).

We also annotated the frames collected from the infant's scene camera to quantify properties of their visual field. Using methods similar to Bambach et al. (2018), detectron2 object detection (Wu et al., 2019) was used to automatically detect the size and location of objects in the infant's FOV. We report object size as the proportion of pixels the object occupied in the FOV. To better understand the quality of naming instances, we defined "dominance" as moments when an object occupied at least 5% of infant's FOV and was 2 times larger than all other objects in view. Only a single object could be dominant at a time and there were times when no object met the dominance criteria. The dominance measure allowed us to see how big a labeled object was in the infant's FOV relative to all other in-view objects.

Data Analysis

To understand the behaviors that support word learning, we analyzed attention to the labeled object around naming moments that did or did not lead to learning. A temporal window was defined that started 3s before onset of naming and ended 3s after offset of naming. Naming utterances lasted an average of 1.25s, resulting in a mean temporal window of 7.23s. Within this window, we measured the proportion of time infants looked at and held the intended target object (as well as time engaging with other objects) and when parents held the target or other objects (**Figure 1**).

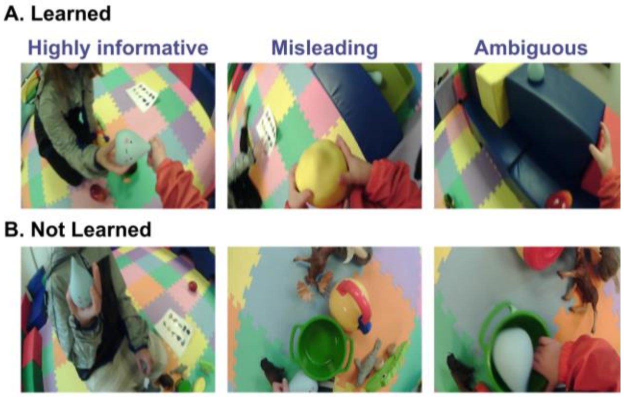

Figure 2. Frames taken from naming moments of an infant that learned the word "drop" (A) and an infant that did not (B). On the left is a Highly Informative naming moment, the middle is Misleading, and the right is Ambiguous.

To analyze visual properties, we calculated the mean size of the target and the number of objects in view during this temporal window. We also measured the proportion of time objects were dominant in infant's FOV during the temporal window around naming. We used object dominance to categorize naming moments into 4 different quality categories **(Figure 2)**: Highly Informative naming, when only the target object was dominant during the naming window; Misleading naming, when only a single distractor object was dominant during the naming window; Ambiguous naming, when multiple objects were dominant during the naming window; and naming instances that had no object dominance. We will not report on these "no dominance" instances since it often occurred when infants were moving and no objects were in view.

We report on three sets of analyses, done at the corpus-level, across all subjects and objects. First, we compared the visual properties of learned and not learned naming instances to see if **visual information** alone could predict learning outcomes. Second, we asked **whose hands create the visual information** in infant's FOV by comparing how much infants and parents held objects during Highly Informative, Misleading, and Ambiguous naming instances that did and did not lead to learning. Lastly, we analyzed **how infants use the visual information** across these different types of naming instances by comparing visual attention to the intended target and other objects in naming that does and does not lead to learning. For all analyses, we used binomial logistic regressions, with learning outcome as the dependent variable and included a random effect of subject (lmer Test package for R; Kuznetsova, Brockhoff, & Christensen, 2017). Separate analyses were run on attention to the intended target and attention to other objects. Models were compared to a null model with random effect only using a chi square test – all significant models were significantly improved from the null model.

Results

What information is available to infants?

We first analyzed the visual properties of learned and not learned naming instances at the corpus level. The visual scenes of learned naming instances may be slightly less cluttered, as there was a small difference in the number of objects in view during learned (m=7.906) and not learned naming (m=8.573, β = -0.147, *p* = 0.015). But unlike previous work (Yu & Smith, 2012), the size of the labeled object was not predictive of learning outcomes (*p*=0.065). We then compared the proportion of time the labeled object was dominant during the naming window, and again found no differences between learned (m=0.283) and not learned naming instances (m=0.233, *p*=0.711).

To better understand the quality of naming instances, we used object dominance to categorize naming. 66.1% of naming instances were Ambiguous, 16.9% were Highly Informative, 8.3% were Misleading, and 8.7% had no object dominance. The proportion of naming events classified as Ambiguous, Highly Informative, and Misleading were similar between learned and not learned words (**Table 1**). These results suggest that the information available to infants may be comparable during learned and not learned naming – so the difference that leads to learning may be due to how that information is made. To test these hypotheses, we turned to parent and infant object handling as it often creates the visual dominance of objects (e.g., Yu & Smith, 2012).

Table 1. Proportion of naming quality types

	Learned	Not learned
Ambiguous	0.641	0.681
Highly informative	0.164	0.173
Misleading	0.113	0.052
No object dominance	0.082	0.093

Whose hands create visual information?

We first compared the proportion of naming windows that infants held the target object (or other objects) during Ambiguous, Highly Informative, and Misleading naming. We found that infant holding during Ambiguous naming instances was the strongest predictor of whether or not an object was learned (**Figure 3**). Infants held the labeled object more in Ambiguous naming moments that led to learning (m_{learned}=0.541, $m_{\text{not-learned}}$=0.350; β=1.059, *p*=0.001) and held other objects more in Ambiguous naming that did not lead to learning (m_{learned}=0.434, $m_{\text{not-learned}}$=0.715; β = -1.353, *p* < 0.001). In Highly Informative naming, there was no difference in how much infants held the target (m_{learned}=0.598, $m_{\text{not-learned}}$=0.586; *p*=0.606) or other objects (m_{learned}=0.162, $m_{\text{not-learned}}$=0.118; *p*=0.267). There was also no difference in Misleading naming in how much infants held the intended target (m_{learned}=0.293, $m_{\text{not-learned}}$=0.222; *p*=0.815), but infants held other objects far more in Misleading naming that did not lead to learning (m_{learned}=0.422, $m_{\text{not-learned}}$=0.680; β=17.480, *p*=0.005).

Parent's manual activity had a different relationship with learning, with only holding during Misleading naming predicting learning outcomes (**Figure 4**). In Ambiguous naming there were no differences in how much parents held the target (m_{learned}=0.339, $m_{\text{not-learned}}$= 0.470; *p*=0.051) or other objects (m_{learned}=0.345,$m_{\text{not-learned}}$=0.360; *p*=0.877). Similarly, in Highly Informative naming, how much parents held the target (m_{learned}=0.376, $m_{\text{not-learned}}$=0.433; *p*=0.881) or other objects (m_{learned}=0.150, $m_{\text{not-learned}}$=0.244; *p*=0.198) did not predict learning. In Misleading naming, however, parents held the target more when it was not learned (m_{learned}=0.174, $m_{\text{not-learned}}$=0.305; β=21.407, *p*=0.007), but there was no difference in parents' holding other objects (m_{learned}=0.322, $m_{\text{not-learned}}$=0.326; *p*=0.992).

To summarize, we found that how much infants and parents held objects in Highly Informative naming does not predict learning – a high-quality naming moment will always be high quality. Holding objects in Misleading naming does predict learning, though: parents are more likely to hold the target and infants are far more likely to hold other objects in not learned naming moments. This suggests that who creates

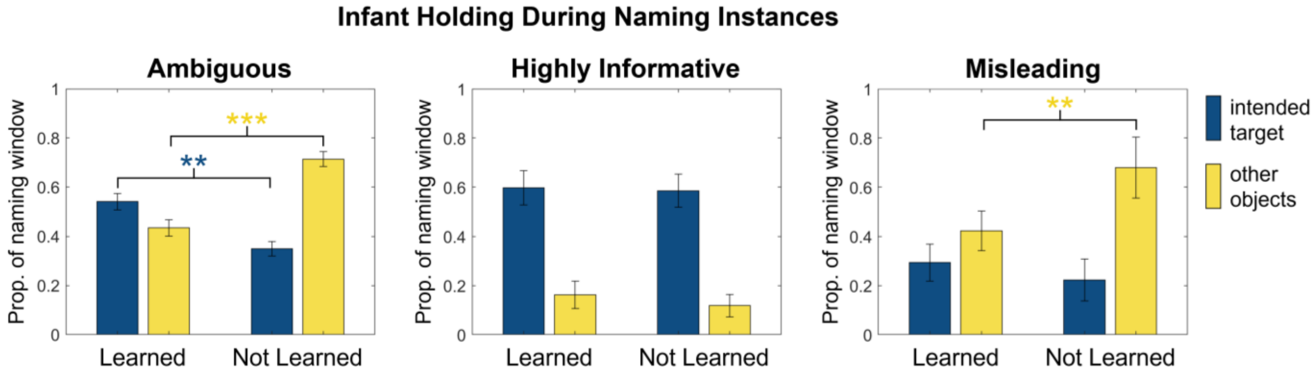

Figure 3. Proportion of time in Ambiguous (left), Highly Informative (middle), and Misleading (right) naming windows that infants held the labeled object (blue) and other objects (yellow). Plots compare holding for learned (left) and not learned (right) words. Error bars show standard error. ** *p < 0.01*, *** *p < 0.001*

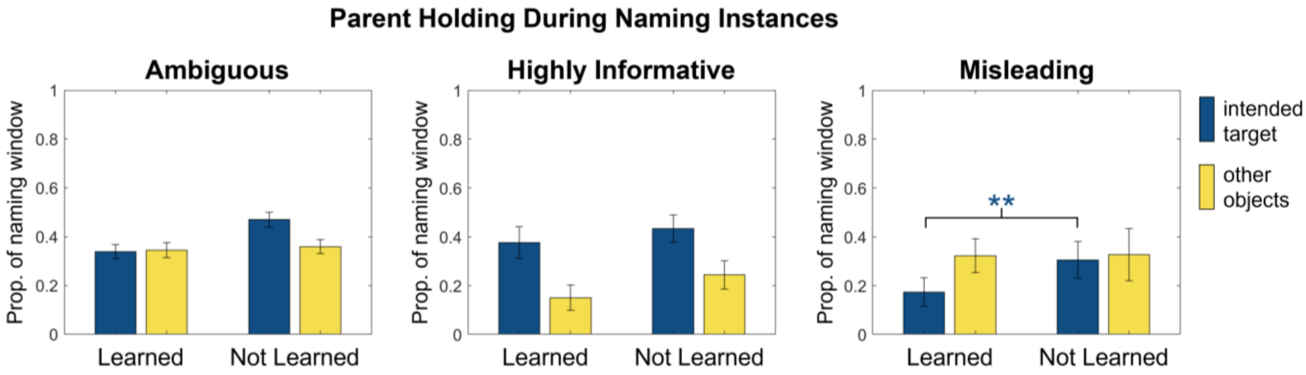

Figure 4. Proportion of time in Ambiguous (left), Highly Informative (middle), and Misleading (right) naming windows that parents held the labeled object (blue) and other objects (yellow). Plots compare holding for learned (left) and not learned (right) words. Error bars show standard error. ** *p < 0.01*

visual information does matter for learning. The most striking differences were in Ambiguous naming – infants held the target object more in moments that led to learning and held other objects more in not learned naming moments. We hypothesized that if the information is being created differently in moments that do and do not lead to learning, then infants might also attend to Ambiguous, Highly Informative, and Misleading naming moments differently.

How do infants visually select this information?

Infants' visual attention during naming moments mirrored the holding results, with attention during Ambiguous naming being the only predictor of learning outcomes (**Figure 5)**. During Ambiguous moments, infants looked more at a target that was learned ($m_{learned}$=0.417, $m_{not\text{-}learned}$=0.289; β=1.966, *p* < 0.001) and more at other objects during naming that did not lead to learning ($m_{learned}$=0.252, $m_{not\text{-}learned}$=0.362; β = -2.254, *p* < 0.001). We found no differences in Highly Informative naming in how often infants looked at the target ($m_{learned}$=0.563, $m_{not\text{-}learned}$=0.489; *p*=0.476) or other objects ($m_{learned}$=0.120, $m_{not\text{-}learned}$=0.127; *p*=0.436). Lastly, we observed no differences in Misleading naming in infant looking to target ($m_{learned}$=0.156, $m_{not\text{-}learned}$=0.100; *p*=0.810) or other objects ($m_{learned}$=0.444, $m_{not\text{-}learned}$=0.656; *p*=0.553).

Most naming happens in visually Ambiguous moments, so understanding how infants can use uncertain information is critical for predicting learning outcomes. In Ambiguous instances, we saw that infants' often look at the objects they hold. While we did not directly analyze the synchrony of infant visual attention and manual action in this study, previous work confirms the importance of hand-eye coordination in scaffolding learning and attention (Schroer & Yu, 2022; Yu & Smith, 2017; Yuan et al., 2019). Through embodied attention, visual ambiguity can be resolved to create more high-quality naming moments when referent-label mappings can be easily inferred. When parents take advantage of this clear signal, a word learning moment can be carved out of the noise.

Discussion

How learners handle misleading and ambiguous cases matters. In both screen-based tasks and everyday contexts, infants (and adults) cannot "just ignore" ambiguous information and may not even have enough knowledge to identify misleading naming moments. The goal of this study was to examine how real-time word learning is affected by ambiguous information – and how that uncertainty can be resolved through infants' embodied attention. We

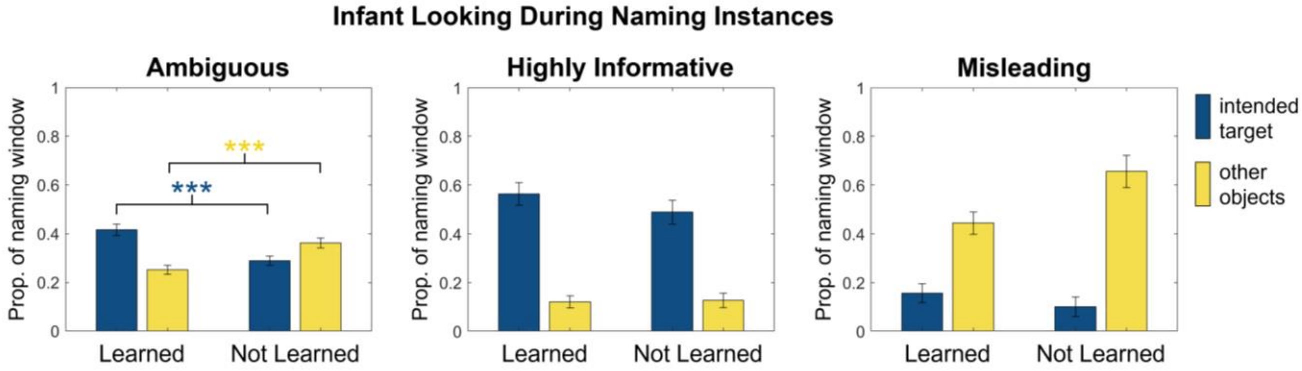

Figure 5. Proportion of time in Ambiguous (left), Highly Informative (middle), and Misleading (right) naming windows that infants looked at the labeled object (blue) and other objects (yellow). Plots compare holding for learned (left) and not learned (right) words. Error bars show standard error. *** *p < 0.001*

hypothesized that infants' actions would scaffold their information selection during ambiguous naming moments to reduce uncertainty. By splitting naming moments into Ambiguous, Highly Informative, and Misleading instances, we were able to compare how the visual information in infant's FOV is created and whether infants use that information to learn.

In naturalistic interactions, our results show that hands have the power to reduce ambiguity and create informative naming events, but they can also increase the noise of a naming moment by creating competitors that have a high likelihood of being a correct object-label mapping. When parents label an object that the infant is not visually and manually engaged with, ambiguity increases and creates a high-probability distractor (as in Bunce & Scott, 2017). Additionally, although parent's hands can provide the same visually dominant views, these moments were often categorized as Misleading. This supposedly similar visual information is likely used by the infant differently and, as a result, does not equally support learning outcomes. Infants' hands may uniquely serve as an attentional frame around the object in a way that their parent's hands could not (e.g., Davoli & Brockmole, 2012; Slone, Smith, & Yu, 2019). Embodied attention, when infants' eyes and hands are on the same object, resolves visual ambiguity and promotes learning.

In the real world, high-quality naming moments do occur (Suanda et al., 2019), but they are sprinkled into a sea of more ambiguous naming. Nonetheless, adults have learned in cross-situational learning experiments with only ambiguous naming events (Zhang et al., 2022) and some referential ambiguity may even boost toddler's cross-situational learning abilities (Cheung et al., 2021). But this robustness to ambiguity has a limit, and when faced with high-probability distractors, learnability does decrease (Bunce & Scott, 2017). How infants aggregate information across different quality naming events in free-flowing interactions is an open question.

Our corpus-level analyses consider each naming utterance as an independent event. In reality, our infant learners had access to the information across each naming utterance their parent provided. Future work could consider the accumulated statistics for each infant as an explanation for why some words are learned. Learned objects were still labeled in Misleading moments (and not learned objects in Highly Informative moments). These results invite many questions – do Misleading moments still provide ample information for learning or did infants also hear the object label in a few Ambiguous or Highly Informative moments? Conversely, could one Highly Informative naming moment be drowned out by many Misleading and Ambiguous ones? By considering all of the naming moments infants had access to, we may get a clearer picture of the information required for learning a word.

Additionally, one potential limitation of our study is the use of a screen-based paradigm to test infants' learning of real three-dimensional objects. The ability of infants to transfer their learning between 2D and 3D objects has often been studied by testing whether children can learn from a screen. This work has shown that there are reliable video deficit effects, but that infants are able to map a 2D image to the actual 3D version of an object (Krcmar, Grela, & Lin, 2007, discussed in Barr, 2010). As our infants learned about 3D objects and were tested using photographs (the opposite direction of video deficit research), it is probable they were able to use 3D-2D mappings during the test.

Notably, we were unable to replicate previous findings that holding creates dominant views of objects and that object size during naming predicts learning (Yu & Smith, 2012). There are several possible explanations. These original results were from a highly constrained experiment with only 3 toys that were equal in size and infants were unable to move around freely, limiting their ability to construct their FOV. In this clean environment, manual action can easily generate a size advantage that is likely to get infant attention to create a high-quality naming moment. In our study, we instead found that both learned and not learned objects have many visually Ambiguous naming moments, as well as a few High Informative and Misleading instances. In a more cluttered context, it is harder for manual actions to generate a size advantage when there are so many distractors. Having a direct measure of infant attention becomes more meaningful in a noisy environment, because we need to know what information infants are using – not just that it exists. We found that in learned naming moments, infants are still likely to selectively attend to the intended target, but it might not always be visually dominant. When we consider gaze data and see infants' information selection process, our results are in line with Yu & Smith (2012): infants' hands create the information that predicts learning.

Conclusion

By studying real-time word learning in naturalistic parent-infant interactions, we were able to explore how visual ambiguity may be created and resolved by an infant learner. Our results suggest that infant's embodied attention plays a critical role in reducing uncertainty to support learning.

Acknowledgements

This work was supported by NIH R01HD074601 and R01HD093792 to CY. SES was supported by the NSF GRFP (DGE-1610403). We thank Jane Yang, Spencer Caplan, the Computational Cognition and Learning Lab at Indiana University, and the Developmental Intelligence Lab at UT Austin, for their support in data collection, data annotation, and feedback on earlier versions of this work.

References

Baldwin, D. A. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental psychology*, 29(5), 832-843.

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

Barr, R. (2010). Transfer of learning between 2D and 3D sources during infancy: Informing theory and practice. *Developmental review*, 30(2), 128-154.

Bunce, J. P., & Scott, R. M. (2017). Finding meaning in a noisy world: exploring the effects of referential ambiguity and competition on 2.5-year-olds' cross-situational word learning. *Journal of child language*, 44(3), 650-676.

Cheung, R. W., Hartley, C., & Monaghan, P. (2021). Caregivers use gesture contingently to support word learning. *Developmental Science*, 24, e13098.

Clerkin, E. M., & Smith, L. B. (2022). Real-world statistics at two timescales and a mechanism for infant learning of object names. *Proceedings of the National Academy of Sciences,* 119(18), e2123239119.

Davoli, C. C., & Brockmole, J. R. (2012). The hands shield attention from visual interference. *Attention, perception, & psychophysics*, 74(7), 1386-1390.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., & Reilly, J. S. (1993). *MacArthur Communicative Development Inventories: User's guide and technical manual*. Baltimore, WV: Paul H. Brookes.

Krcmar M, Grela B, Lin K. (2007). Can toddlers learn vocabulary from television? An experimental approach. *Media Psychology* 10(1), 41–63.

Kuznetsova A., Brockhoff P.B., Christensen R.H.B. (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." *Journal of statistical software,* 82(13), 1–26.

MacDonald, K., Yurovsky, D., & Frank, M. C. (2017). Social cues modulate the representations underlying cross-situational learning. *Cognitive psychology*, 94, 67-84.

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

Schroer, S. E., & Yu, C. (2022). Looking is not enough: Multimodal attention supports the real-time learning of new words. *Developmental Science*.

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

Slone, L. K., Smith, L. B., & Yu, C. (2019). Self-generated variability in object images predicts vocabulary growth. *Developmental science*, 22(6), e12816.

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.

Twomey, K. E., Ma, L., & Westermann, G. (2018). All the right noises: Background variability helps early word learning. *Cognitive science*, 42, 413-438.

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

Wu,Y., Kirillov,A.,Lo,W.-Y.,Girshick,R.(2019). Detectron2, *https://github.com/facebookresearch/detectron2*.

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

Yu, C.,& Smith, L.B.(2017). Hand–eye coordination predicts joint attention. *Child development*, 88(6), 2060-2078.

Yu, C., Zhang, Y., Slone, L. K., & Smith, L. B. (2021). The infant's view redefines the problem of referential uncertainty in early word learning. *Proceedings of the National Academy of Sciences*, 118(52), e2107019118.

Yuan, L., Xu, T. L., Yu, C., & Smith, L. B. (2019). Sustained visual attention is more than seeing. *Journal of experimental child psychology*, 179, 324-336.

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

Zhang, Y., & Yu, C. (2017). How misleading cues influence referential uncertainty in statistical cross-situational learning. In *41st annual Boston University conference on language development* (pp. 820-833).

Zhang, Y., Yurovsky, D., & Yu, C. (2021). Cross-situational learning from ambiguous egocentric input is a continuous process: Evidence using the human simulation paradigm. *Cognitive science*, 45(7), e13010.