A Computational Model of Early Word Learning from the Infant's Point of View.

Md Alimoor Reza
Indiana University Bloomington
21 PUBLICATIONS **73** CITATIONS
SEE PROFILE

Chen Yu
Indiana University Bloomington
179 PUBLICATIONS **4,204** CITATIONS
SEE PROFILE

Satoshi Tsutsui
Indiana University Bloomington
13 PUBLICATIONS **28** CITATIONS
SEE PROFILE

A Computational Model of Early Word Learning from the Infant's Point of View

Satoshi Tsutsui[1], **Arjun Chandrasekaran**[3], **Md Alimoor Reza**[1], **David Crandall**[1], **Chen Yu**[2]
{stsutsui,mdreza,djcran,chenyu}@iu.edu, arjun.chandrasekaran@tuebingen.mpg.de
[1] Luddy School of Informatics, Computing, and Engineering, Indiana University, USA
[2] Department of Psychological and Brain Sciences, Indiana University, USA
[3] Max Planck Institute for Intelligent Systems, Germany

Abstract

Human infants have the remarkable ability to learn the associations between object names and visual objects from inherently ambiguous experiences. Researchers in cognitive science and developmental psychology have built formal models that implement in-principle learning algorithms, and then used preselected and pre-cleaned datasets to test the abilities of the models to find statistical regularities in the input data. In contrast to previous modeling approaches, the present study used egocentric video and gaze data collected from infant learners during natural toy play with their parents. This allowed us to capture the learning environment from the perspective of the learner's own point of view. We then used a Convolutional Neural Network (CNN) model to process sensory data from the infant's point of view and learn name-object associations from scratch. As the first model that takes raw egocentric video to simulate infant word learning, the present study provides a proof of principle that the problem of early word learning can be solved, using actual visual data perceived by infant learners. Moreover, we conducted simulation experiments to systematically determine how visual, perceptual, and attentional properties of infants' sensory experiences may affect word learning.

Keywords: Word learning, Computational Modeling, Eye Tracking and Visual Attention, Parent-Child Social Interaction

Introduction

Infants show knowledge of their first words as early as 6 months old and produce their first words at around a year. Learning object names — a major component of their early vocabularies — in everyday contexts requires young learners to not only find and recognize visual objects in view but also to map them with heard names. In such a context, infants seem to be able to learn from a sea of data relevant to object names and their referents because parents interact with and talk to their infants in various occasions — from toy play, to picture book reading, to family meal time (Yu & Smith, 2012).

However, if we take the young learner's point of view, we see that the task of word learning is quite challenging. Imagine an infant and parent playing with several toys jumbled together as shown in Figure 1. When the parent names a particular toy at a particular moment, the infant perceives 2-dimensional images on the retina from a first-person point of view, as shown in Figure 2. These images usually contain multiple objects in view. Since the learner does not yet know the name of the toy, how do they recognize all the toys in view and then infer the target to which the parent is referring? This *referential uncertainty* (Quine, 1960) is the classic puzzle of early word learning: because real-life learning situations are replete with objects and events, a challenge for young word learners is to recognize and identify the correct referent from many possible candidates at a given naming moment. Despite many experimental studies on infants (Golinkoff et al., 2000) and much computational work on simulating early word learning (Yu & Ballard, 2007; Frank, Goodman, & Tenenbaum, 2009), how young children solve this problem remains an open question.

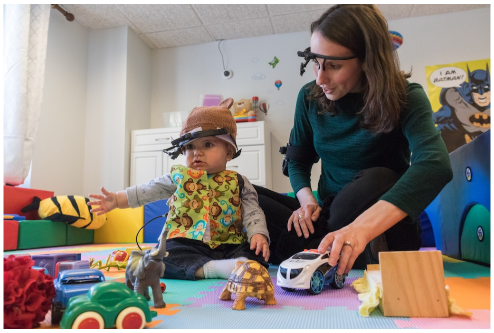


Figure 1: An infant and parent play with a set of toys in a free-flowing joint play session. Both participants wore head-mounted cameras and eye trackers to record egocentric video and gaze data from their own perspectives.

Decades of research in developmental psychology and cognitive science have attempted to resolve this mystery. Researchers have designed human laboratory experiments by creating experimental training datasets and testing the abilities of human learners to learn from them (Golinkoff et al., 2000). In computational studies, researchers have built models that implement in-principle learning algorithms, and created training sets to test the abilities of the models to find statistical regularities in the input data. Some work in modeling word learning has used sensory data collected from adult learners or robots (Roy & Pentland, 2002; Yu & Ballard, 2007; Rasanen & Khorrami, 2019), while many models take symbolic data or simplified inputs (Frank et al., 2009; Kachergis & Yu, 2017; K. Smith, Smith, & Blythe, 2011; Fazly, Alishahi, & Stevenson, 2010; Yu & Ballard, 2007). Little is known about whether these models can scale up to address the same problems faced by infants in real-world learning. As recently pointed out in (Dupoux, 2018), the research field of cognitive modeling needs to move toward using realistic data as input because all the learning processes in human cognitive systems are sensitive to the input signals (L. B. Smith, Jayaraman, Clerkin, & Yu, 2018). If our ultimate goal is to understand how infants learn language in the real world — not in laboratories or in simulated environment — we should model internal learning processes with natural statistics of the learning environment. This paper takes a step towards this goal and uses data collected by infants as they naturally play with toys and interact with parents.

Recent advances in computational and sensing techniques (deep learning, wearable sensors, etc.) could revolutionize the study of cognitive modeling. In the field of machine learning, Convolutional Neural Networks (CNNs) have achieved impressive learning results and even outperform humans on some specific tasks (Silver et al., 2016; He, Zhang, Ren, & Sun, 2015). In the field of computer vision, small wearable cameras have been used to capture an approximation of the visual field of their human wearer. Video from this egocentric point of view provides a unique perspective of the visual world that is inherently human-centric, giving a level of detail and ubiquity that may well exceed what is possible from environmental cameras in a third-person point-of-view. Recently, head-mounted cameras and eye trackers have been used in developmental psychology to collect fine-grained information about what infants are seeing and doing in real time (He et al., 2015; Silver et al., 2016). These new technologies make it feasible to build computational models using inputs that are very close to infants' actual sensory experiences, in order to understand the rich complexity of infants' sensory experiences available for word learning.

In the present study, we collect egocentric video and gaze data from infant learners as they and their parents naturally play with a set of toys. This allows us to capture the learning environment from the perspective of the learner's own point of view. We then build a computational system that processes this infant sensory data to learn name-object associations from scratch. As the first model taking raw egocentric video to simulate infant word learning, the present study has two primary goals. The first aim is to provide a proof of principle that the problem of early word learning can be solved using raw data. The second aim is to systematically determine the computational roles of visual, perceptual, and attentional properties that may influence word learning. This examination allows us to generate quantitative predictions which can be further tested in future experimental studies.

Method

Data Collection

To closely approximate the input perceived by infants, we collected visual and audio data from everyday toy play — a context in which infants naturally learn about objects and their names. We developed and used an experimental setup in which we placed a camera on the infant's head to collect egocentric video of their field of view, as shown in Figure 1. We also used a head-mounted eye gaze tracker to record their visual attention. Additionally, we collected synchronized video and gaze data from the parent during the same play session.

Thirty-four child-parent dyads participated in our study. Each dyad was brought into a room with 24 toys (the same as in (Bambach, Crandall, Smith, & Yu, 2018)) scattered on the floor. Children and parents were told to play with the toys, without more specific directions. The children ranged in age from 15.2 to 24.2 months (μ=19.4 months, σ=2.2 months). We collected five synchronized videos per dyad (head camera and eye camera for child, head camera and eye camera for parent, and a third-person view camera – see Figure 1). The final dataset contains 212 minutes of synchronized video, with each dyad contributing different amounts of data ranging from 3.4 minutes to 11.6 minutes (μ=7.5 minutes, σ=2.3 minutes). The head-mounted eye trackers recorded video at 30 frames per second and 480 × 640 pixels per frame, with a horizontal field of view of about 70 degrees. We followed validated best practices for mounting the head cameras so as to best approximate participants' actual first-person views, and for calibrating the eye trackers (Slone et al., 2018).

Training Data

Parents' speech during toy play was fully transcribed and divided into spoken utterances, each defined as a string of speech between two periods of silence lasting at least 400ms (Yu & Smith, 2012). Spoken utterances containing the name of one of the objects were marked as "naming utterances" (e.g. "that's a helmet"). For each naming utterance, trained coders annotated the intended referent object. On average, parents produced 15.51 utterances per minute (σ=4.56), 4.82 of which were referential (σ=2.09). In total, the entire training dataset contains 1,459 naming utterances.

Recent studies on infant word learning show that the moments during and after hearing a word are critical for young learners to associate seen objects with heard words (Yu & Smith, 2012). In light of this, we temporally aligned speech data with video data, and used a 3-sec temporal window starting from the onset of each naming utterance. Given that each naming utterance lasted about 1.5 to 2 seconds, a 3-sec window captured both the moments that infants heard the target name in parent speech and also the moments after hearing the name. For each temporal window, a total of 90 image frames (30 frames per second) were extracted. To summarize, the final training dataset consists of all the naming instances in parent-child joint play, with each instance containing a target name and a set of 90 image frames from the child's first-person camera that co-occur with the naming utterance. As shown in Figure 2, each image typically contains multiple visual objects and the named object may or may not be in view.

Testing Data and Evaluation Metrics

To evaluate the result of word learning, we prepared a separate set of clean canonical images for each of the 24 objects varying in camera view and object size and orientation in a similar manner to previous work (Bambach, Crandall, Smith, & Yu, 2016). In particular, we took pictures of each toy from eight different points of view (45 degree rotations around the vertical

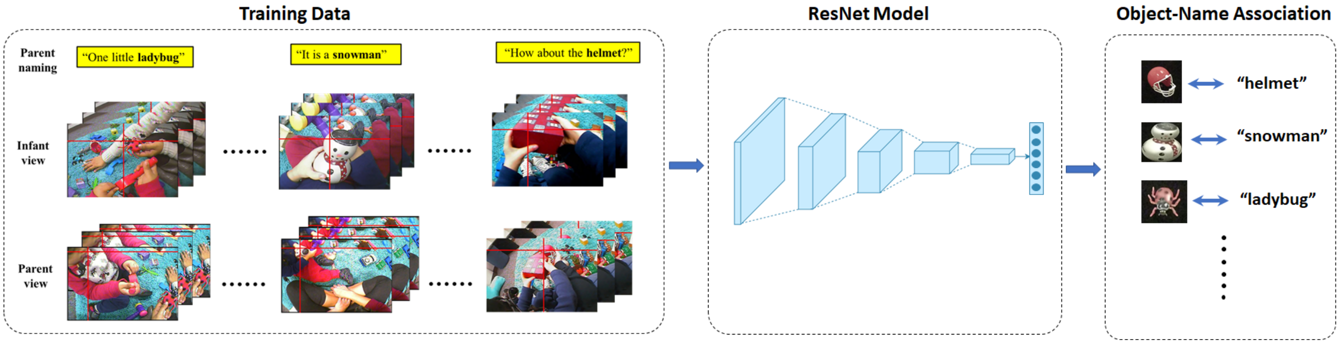


Figure 2: Overview of our approach. The training data were created by extracting egocentric image frames around the moments when parents named objects in free-flowing interaction. The data was fed into deep learning (ResNet) models to find and associate visual objects in view with names in parent speech. As a result, the models built the associations between heard labels and visual presentations of target objects.

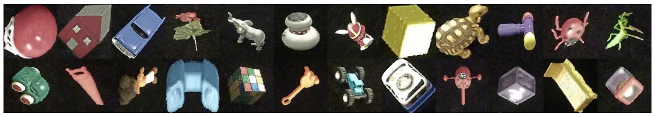
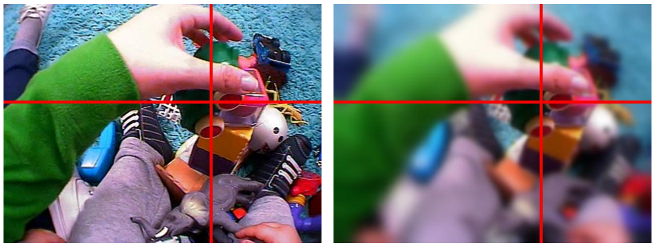


Figure 3: Testing images. We evaluate the models trained from egocentric images using systematically captured images from various views with a clean background.

axis), totaling 3,072 images (see Fig 3). This test set allowed us to examine whether the models generalized the learned names to new visual instances never seen before. During test, we presented one image at a time to a trained model and checked whether the model generated the correct label. We compute mean accuracy (i.e., the number of correctly classified images over the total number of test images) as the evaluation metric.

Simulating acuity

Egocentric video captured by head-mounted cameras provides a good approximation of the field of view of the infant. However, the human visual system exhibits well-defined contrast sensitivity due to retinal eccentricity: the area centered around the gaze point (the fovea) captures a high-resolution image, while the imagery in the periphery is captured at dramatically lower resolution due to its lesser sensitivity to higher spatial frequencies. As a result, the human visual system does not process all "pixels" in the first-person image equally, but instead focuses more on the pixels around the fovea. To closely approximate the visual signals that are "input" to a learner's learning system, we implemented the method of (Perry & Geisler, 2002) to simulate the effect of foveated visual acuity on each frame. The basic idea is to preserve the original high-resolution image at the center of gaze while increasing blur progressively towards the periphery, as shown in Figure 4b. This technique applies a model of what is known about human visual acuity and has been validated with human psychophysical studies (Perry & Geisler, 2002).



(a) Original

(b) Acuity simulation

Figure 4: We simulated foveated vision by applying an acuity filter to the original egocentric image, based on the eye gaze position (red crosshairs).

Convolutional Neural Networks Models

We used a state-of-the-art CNN model, ResNet50 (He, Zhang, Ren, & Sun, 2016), trained with stochastic gradient descent (SGD). The network outputs a softmax probability distribution over 24 object labels, so the label with the highest probability is the predicted object. SGD optimizes the CNN parameters to minimize the cross entropy loss between the predicted distribution and the ground truth (one-hot) distribution. Before SGD, we initialized the parameters of ResNet50 with a model pretrained on ImageNet (Russakovsky et al., 2015). Thus, the model can reuse the visual filters learned on ImageNet to avoid having to learn the low-level visual filters from scratch. The training images were resized to 224 × 224 pixels with bilinear interpolation. We used SGD with batch size 128, momentum 0.9, and initial learning rate 0.01. We decreased the learning rate by a factor of 10 when the performance stopped improving, and ended training when the learning rate reached 0.0001. Because training was stochastic, there is natural variation across training runs; we thus ran each of our experiments 10 times and report means and standard deviations. Moreover, since our goal was to discover general principles that lead to successful word learning and not to analyze the results of individual objects, we applied a mixed-effect logistic regression with random effects of trial and object in each of our analyses.

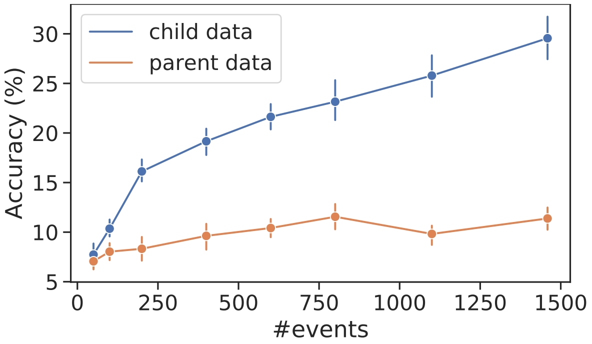


Figure 5: Results from models trained with infant data improve with more naming instances, while the models trained with the parent data show no improvement.

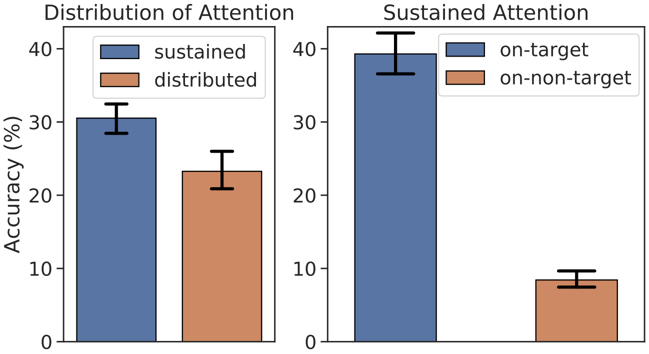


Figure 6: The model trained with sustained attention events outperformed the model trained with distributed attention events. Within the sustained attention events, the model trained with on-target instances outperformed the model trained with on-non-target instances.

Experiments and Results

Study 1: Learning object names from raw egocentric video

The aim of Study 1 is to demonstrate that a state-of-the-art machine learning model can be trained to associate object names with visual objects by using egocentric data closely approximating sensory experiences of infant learners. We also evaluated models learned with parent view data in order to compare the informativeness of these different views. Moreover, to examine the impact of properties of the training data, we created several simulation conditions by sub-sampling the whole set of 1459 into seven subsets with different numbers of naming events (50, 100, 200, 400, 600, 800, 1100). While we expected that more naming instances would lead to better learning, we sought to quantify this relationship.

Figure 5 reveals two noticeable patterns in the models trained on the infant data and the model trained on the parent data. First, when there are 200 or more naming events, models trained with infant data consistently outperformed the same models trained on parent data (e.g., for 200 naming events: $M_{infant} = 16.12\%, SE_{infant} = 1.73\%; M_{parent} = 8.32\%, SE_{parent} = 1.27\%; \beta = 0.34, t = 3.64, p < 0.001$). Second, as the quantity of training data increased, the models trained on infant data obtained better performance while the models trained on the parent data saturated. Taken together, these results provide convincing evidence that the model can solve the name-object mapping problem from raw video, and that the infant data contain certain properties leading to better word learning. The finding that infant data lead to better learning is consistent with recent results reported on another topic in early development – visual object recognition (Bambach et al., 2018).

Study 2: Examining the effects of different attentional strategies

Humans perform an average of approximately three eye movements per second because our visual systems actively select visual information which is then fed into internal cognitive and learning processes. Thus during the 3-second window during and after hearing a naming utterance, an infant learner may generate multiple looks on different objects in view, or, alternatively, may sustain their attention on one object. The aim of Study 2 is to investigate whether different attention strategies during naming events influence word learning, and if so, in which ways.

To answer this question, we first assigned each naming event into one of two categories: *sustained attention* if the infant attended to a single object for more than 60% of the frames in the naming event, and *distributed attention* otherwise. This split resulted in 750 sustained attention (SA) and 709 distributed attention (DA) events. In either case, the infant may or may not attend to the named object because the definition is based on the distribution of infant attention, *not* on which objects were attended in a naming event. We trained two identical models, one on SA instances and one on DA instances. The results in Figure 6 reveal that the model trained with sustained attention events ($M_{sustained} = 30.53\%, SE_{sustained} = 2.08\%$) outperformed the model trained with distributed attention events ($M_{distributed} = 23.26\%, SE_{distributed} = 1.78\%; \beta = 0.20, t = 2.65, p < 0.005$), suggesting that sustained attention on a single object while hearing a name leads to better learning.

Of course, infants may or may not show sustained attention on the object actually named in parent speech. In total, infants attended to the target in 452 out of 750 SA events, and attended to a non-target object in the other 298 SA events. Attending to the target object with sustained attention should help learning while sustained attention on a non-target object should hinder learning. To test this prediction, we sub-sampled 298 on-target events from 452 SA events, and compared them with the remaining 298 on-non-target events. As shown in Figure 6, the model trained with the on-target events ($M_{target} = 39.27\%, SE_{target} = 2.30\%$) achieved significantly higher accuracy than the model trained on on-non-target events ($M_{non-target} = 8.42\%, SE_{non-target} = 1.20\%; \beta =$

$0.98, t = 14.52, p < 0.001$).

In everyday learning contexts such as toy play, young learners do not passively perceive information from the environment; instead, the visual input to internal learning processes is highly selective moment-to-moment. The ability to sustain attention in such contexts is critical for early development and has been linked to healthy developmental outcomes (Ruff & Rothbart, 2001). The results from the present study suggest a pathway through which sustained attention during parent naming moments creates sensory experiences that facilitate word learning.

Study 3: Examining the effects of visual properties of attended objects

One effect of sustained attention during a naming moment is to consistently select a certain area in the egocentric view so that the learning system can process the visual information in that focused area to find the target object and link it with the heard label. Moving from the attentional level to the sensory level, we argue that associating object names with visual objects starts with visual information selected in the infant's egocentric view, and therefore the factors that matter to word learning may not just be attended objects but sensory information selected and processed in the naming moments. Study 3 seeks to determine how visual properties of attended objects influence word learning.

Previous studies using head-mounted cameras and head-mounted eye trackers showed that visual objects attended by infants tend to possess certain visual properties — e.g., they tend to be large in view, which provides a high-resolution image of the object (Yu & Smith, 2012). In light of this, the present simulation focused on object size. The naming events were grouped into two subsets by a median split of object size. The large subset contains naming instances in which named objects are larger than the median size (6%) whereas the small subset contains naming instances in which named objects are smaller than the median. The same model was separately trained on the large set and the small set. We found the model trained with large objects achieved significantly higher accuracy on the test dataset than that trained with small objects ($M_{large} = 30.50\%, SE_{large} = 2.20\%; M_{small} = 18.81\%, SE_{small} = 1.81\%; \beta = 0.29, t = 4.12, p < 0.001$).

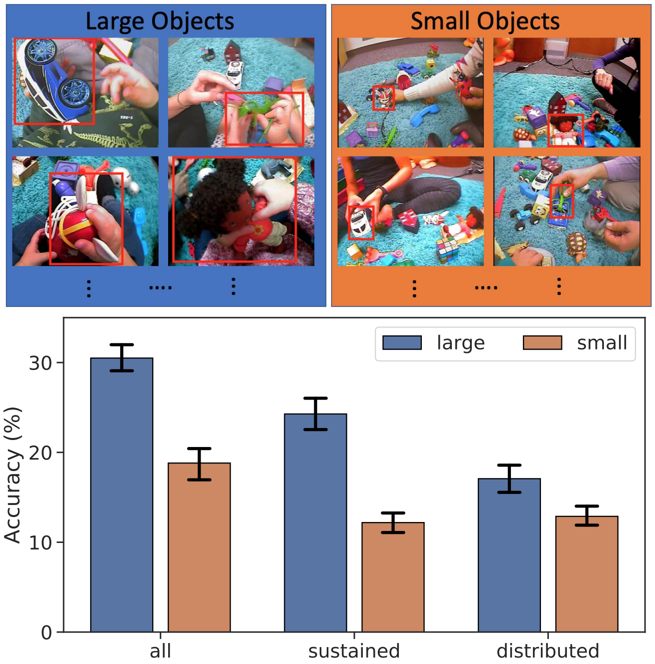


Figure 7: Effect of object size. Naming events were divided based on object size; instances in the large set contain visual instances of named objects large in view whereas the named objects are small in view in the small set.

If the target object in a naming event is large in view, that object is more likely to be attended by infants. Thus, infants' sustained attention on a target object is likely to co-vary with the size of the object. If so, the difference in the learning results described above could be due to sustained attention but not object size. To distinguish the effects on word learning between those two co-varying factors, we divided naming events into sustained attention and distributed attention as in Study 2, and examined the effects of object size in those two situations. In each case, we used a median split to further divide naming events into a large subset and a small subset. As shown in Figure 7, when infants showed sustained attention on named objects, the model trained based on large targets outperformed the same model trained with small targets ($M_{large} = 24.27\%, SE_{large} = 2.03\%; M_{small} = 12.18\%, SE_{small} = 1.5\%; \beta = 0.37, t = 4., p < 0.001$). In the cases of naming events with distributed attention, the model again favored events with large target objects over those with small targets ($M_{large} = 17.07\%, SE_{large} = 1.60\%; M_{small} = 12.88\%, SE_{small} = 1.33\%; \beta = 0.20, t = 2.02, p < 0.05$). Taken together, these results suggest that visual properties of the target object during a naming event have direct and unique influence on word learning.

General Discussions

Despite the fact that the referential uncertainty problem in word learning was originally proposed as a philosophical puzzle, infant learners need to solve this problem at the sensory level. From the infant's point of view, learning object names begins with hearing an object label while perceiving a visual scene having multiple objects in view. However, many computational models on language learning use simple data pre-selected and/or pre-cleaned to evaluate the theoretical ideas of learning mechanisms instantiated by the models. We argue that to obtain a complete understanding of learning mechanisms, we need to examine not only the mechanisms themselves but also the data on which those mechanisms operate. For infant learners, the data input to their internal processes are those that make contact with their sensory systems, so we capture the input data with egocentric video and head-mounted eye tracking. Moreover, compared to prior studies of word learning from third-person images (Chrupała, Kádár, & Alishahi, 2015), the present study is the first, to our knowledge, to use actual visual data from the infant's point of view to reconstruct infants' sensory experiences and to show how a computational model can solve the referential uncertainty problem with the information available to infant learners.

There are three main contributions of the present paper as the first steps toward using authentic data to model infant word learning. First, our findings show that the available information from the infant's point of view is sufficient for a machine learning model to successfully associate object names with visual objects. Second, our findings here provide a sensory account of the role of sustained attention in early word learning. Previous research showed that infant sustained attention at naming moments during joint play is a strong predictor of later vocabulary (Yu, Suanda, & Smith, 2019). The results here offer a mechanistic explanation that the moments of sustained attention during parent naming provide better visual input for early word learning compared with the moments when infants show more distributed attention. Finally, our findings provide quantitative evidence on how in-moment properties of infants' visual input influence early word learning.

The present study used only naming utterances in parent speech (object names in those utterances, etc.), but we know that parent speech during parent-child interaction is more information-rich. For example, studies show that individual utterances in parent speech are usually inter-connected, forming episodes of coherent discourse that facilitate child language learning (Suanda, Smith, & Yu, 2016; Frank, Tenenbaum, & Fernald, 2013). To better approximate infants' learning experiences in our future work, we plan to include both object naming utterances and other referential and non-referential utterances as the speech input to computational models. Including the whole speech transcription will also allow us to examine how infants learn not only object names but also other types of words in their early vocabularies, such as action verbs. In addition, we know that social cues in parent-child interaction play a critical role in shaping the input to infant learners. With egocentric video and computational models, our future work will simulate and analyze how young learners detect and use various kinds of social cues from the infant's point of view.

Acknowledgment

This work was supported in part by the National Institute of Child Health and Human Development (R01HD074601 and R01HD093792), the National Science Foundation (CAREER IIS-1253549), and the Indiana University Office of the Vice Provost for Research, the College of Arts and Sciences, and the Luddy School of Informatics, Computing, and Engineering through the Emerging Areas of Research Project *Learning: Brains, Machines and Children*.

References

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

Bambach, S., Crandall, D. J., Smith, L. B., & Yu, C. (2016). Active viewing in toddlers facilitates visual object learning: An egocentric vision approach. In *Annual Conference of the Cognitive Science Society.*

Chrupała, G., Kádár, Á., & Alishahi, A. (2015). Learning language through pictures. In *Association for Computational Linguistics.*

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*(6), 1017–1063.

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

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), 1–24.

Golinkoff, R. M., Hirsh-Pasek, K., Bloom, L., Smith, L. B., Woodward, A. L., Akhtar, N., … Hollich, G. (2000). *Becoming a word learner: A debate on lexical acquisition*. Oxford University Press.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE Conference on Computer Vision and Pattern Recognition.*

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition.*

Kachergis, G., & Yu, C. (2017). Observing and modeling developing knowledge and uncertainty during cross-situational word learning. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(2), 227–236.

Perry, J. S., & Geisler, W. S. (2002). Gaze-contingent realtime simulation of arbitrary visual fields. In *Human Vision and Electronic Imaging.*

Quine, W. V. O. (1960). *Word and object*. MIT press.

Rasanen, O., & Khorrami, K. (2019). A computational model of early language acquisition from audiovisual experiences of young infants. In *INTERSPEECH.*

Roy, D. K., & Pentland, A. P. (2002, Jan). Learning words from sights and sounds: a computational model. *Cognitive Science*, *26*(1), 113–146.

Ruff, H. A., & Rothbart, M. K. (2001). *Attention in early development: Themes and variations*. Oxford University Press.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … others (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., … others (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484.

Slone, L. K., Abney, D. H., Borjon, J. I., Chen, C.-h., Franchak, J. M., Pearcy, D., ... Yu, C. (2018). Gaze in action: Head-mounted eye tracking of children's dynamic visual attention during naturalistic behavior. *Journal of Visualized Experiments*(141), e58496.

Smith, K., Smith, A. D., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*(3), 480–498.

Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences*, *22*(4), 325–336.

Suanda, S. H., Smith, L. B., & Yu, C. (2016). The multisensory nature of verbal discourse in parent–toddler interactions. *Developmental Neuropsychology*, *41*(5-8), 324–341.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13-15), 2149–2165.

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

Yu, C., Suanda, S. H., & Smith, L. B. (2019). Infant sustained attention but not joint attention to objects at 9 months predicts vocabulary at 12 and 15 months. *Developmental Science*, *22*(1), e12735.