Grounding Word Learning in Multimodal Sensorimotor Interaction

Chen Yu, Linda B. Smith and Alfredo F. Pereira (chenyu@indiana.edu)
Department of Psychological and Brain Sciences, and Cognitive Science Program, Indiana University
Bloomington, IN, 47405 USA

Abstract

A central problem in the study of language acquisition is word learning – how the child's mental representation of objects and events on the one hand is associated with the linguistic input on the other, and how young children acquire the vocabulary of their first language so effortlessly and smoothly, with virtually no errors along the way. This paper aims at understanding the mechanisms through which word learning is grounded in sensorimotor experience, in the physical regularities of the world, and in the time-locked and coupled multimodal interactions between the child's own actions and the actions of their caregivers. We designed and implemented a novel multimodal sensing environment consisting of two head-mounted mini cameras that are placed on both the child's and the parent's foreheads, motion tracking of head movements and recording of caregiver's speech. Using this new technology, we captured the dynamic visual information from both the learner's perspective and the parent's viewpoint while they were engaged in a naturalistic toy-naming interaction, to study the regularities and dynamic structure in the multimodal learning environment. To achieve this goal, we also implemented various data processing programs that can automatically extract visual, motion and speech information from raw sensory data. Our results show that a wide range of perceptual and motor patterns, such as the proportion of the named objects in both the child's and the caregiver's visual fields, the proportion of time that the child's hands are holding the named objects when those names are uttered, and as well as the child's head movements, are predictive of successful word learning through social interaction. In light of this, we suggest that high-level social-cognitive cues in word learning can be grounded in embodied perceptual and motor patterns that are part of a natural social interaction.

Keywords: language development, embodied cognition, learning, computational modeling

Introduction

A major recent advance in understanding word learning has been the documentation of the powerful role of social-interactional cues in guiding infants' attention and in linking the linguistic stream to objects and events in the world (Baldwin, 1993; Tomasello & Akhtar, 1995; Yu, Ballard & Aslin, 2005). There can be no doubt that young learners are highly sensitive to the social information in these interactions (e.g., Baldwin, 1993; Bloom, 2000; Woodward, 2004). However, the nature of this sensitivity and the relevant underlying processes are far from clear. Often in this literature, children's use of social cues is interpreted in terms of (and seen as diagnostic markers of) their ability to infer the intentions of the speaker. This kind of social cognition is called "mind reading" by Baron-Cohen (1995). Butterworth (1991) showed that even by 6 months of age, infants are sensitive to social cues, such as monitoring and following another's gaze, although infants' understanding of the implications of gaze or pointing does not emerge until approximately 12 months of age. Based on this evidence, some researchers (e.g. Bloom, 2000, Tomasello, 2000; Woodward, 2004) have suggested that children's word learning in the second year of life actually draws extensively on their understanding of the thoughts of speakers.

However, there is an alternative explanation to that of "mind-reading". Smith (2000) has suggested that these results may be understood in terms of the child's learning of correlations among actions, gestures and words of the mature speaker as predictors of attention and intended referents. Smith (2000) argued that construing the problem in this way does not "explain away" notions of "mind-reading" but rather grounds those notions in the perceptual cues available in the real-time task that infants must solve (see also Smith & Breazeal, 2007).

One problem with resolving (or integrating) these two ideas about the information available in social interactions is that most experiments are designed as demonstrations of children's sensitivity to social cues and as such, they focus on macro-level behaviors (such as pointing, or direction of eye gaze) in constrained contexts in which the experimenter presents one or two objects on an uncluttered table and where language is also uncluttered by remarks about anything other than those objects on the table (e.g. Baldwin, 1993). To truly understand mechanisms of learning, however, we may need to focus on more micro-level behaviors as they unfold in *real time* in the richly varying and dynamically complex interactions of children and their mature partners in more naturalistic tasks (such as toy play). Further, whereas the studies at the macro-level clearly demonstrate many intelligent behaviors in infant word learning, they have not as yet led to a formal account of the underlying mechanisms. Thus, we want to know not only that learners use social cues but also *how* they do so in terms of the real-time processes in the naturalistic tasks where everyday language learning must take place.

To this end, we sought to study the dynamics of social cues to word learning at the sensorimotor levels – in the bodily gestures and as well as momentary visual and auditory perception of the participants. The study presents a new design and implementation of a sensing system for recording multisensory data from both the child and the caregiver. With this new methodology, we compare and analyze the dynamic structure of natural parent-child interaction in the context of language learning, and further discover perceptual and motor patterns that are

informatively time-locked to words and their intended referents and *predictive* of word learning.

Multimodal Sensing Environment

As shown in Figure 1, the naturalistic interaction of parents and toddlers in the task of table-top toy play is recorded by three cameras from different perspectives: one head-mounted camera provides information about the scene from the child's point of view; a second head-mounted camera provides the parent's viewpoint; and one from a top-down third-person viewpoint allows a clear observation of exactly what was on the table at any given moment (mostly the participants' hands and the objects being played with). In addition, our multimodal system also records participants' body movements through a motion tracking system and as well as parents' speech through a headset.

Interaction room setup. Parents and children sat across each other on a small table (61cm × 91cm × 64cm) that was painted white. The child sat on a high chair and the parent sat on the floor – which places their head-cameras about equal distance from the tabletop. Both participants were asked to wear white outfits. White curtains from floor to ceiling surrounded the table. The experimental room was setup in such a way that everything was white as seen from the vantage point of the participants – with the exception of heads, faces, hands and objects on the table. This greatly simplified automatic visual object segmentation since any white areas of an image could be considered as background.

Head-mounted cameras. The head-mounted cameras are a lightweight mini camera attached to a sports headband. This allowed us to place the camera on the forehead close to the participants' eyes. A small plastic encasing supported rotation of the camera in order to adjust the camera such that during calibration an object to which the participant was attending was near the center of the field. The headband was tight enough that the camera did not move (unless the child pulled at the band during the experiment – an event that caused the data from that child beyond that point to be excluded unless centering could be re-achieved). The visual angle recorded by the camera was 70°. The cabling was long and lightweight enough not to push down on the participant's head or get tangled during movement. A digital video recorder card in a computer adjacent to the experiment room simultaneously recorded the video signal from these two cameras.

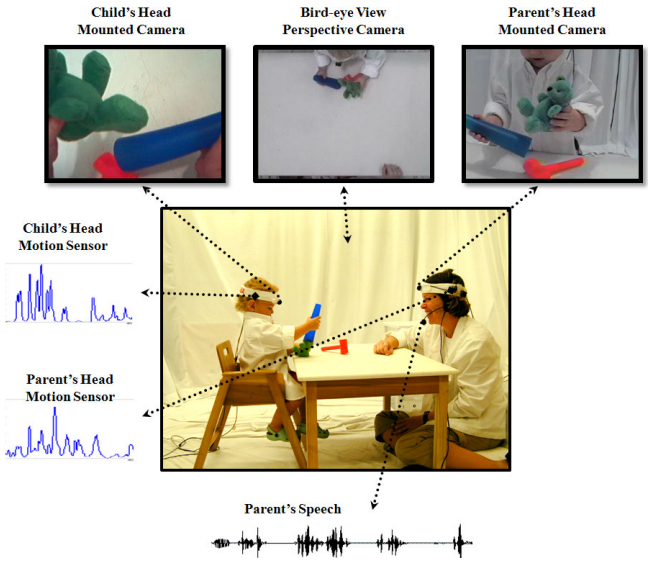

Figure 1: Multimodal sensing system. The child and the mother play with a set of toys at a table. Two mini cameras were placed onto the child's and the mother's heads respectively to collect visual information from two first-person views. A third camera mounted on the top of the table records the bird-eye view of the whole interaction. They also wore motion sensors to track their head movements. A headset was used to record the caregiver's speech.

Bird-eye view camera. A high-resolution camera was mounted right above the table and the table edges aligned with edges of the bird-eye image. This view provided visual information that was independent of gaze and head movements of a participant and therefore it recorded the whole interaction from a third-person static view. An additional benefit of this camera lay in the high-quality video, which made our following image segmentation and object tracking software work more robustly compared with two head-mounted mini cameras. Those two were light-weight but with a limited resolution and video quality due to their small size.

Head motion tracking. To measure the activity of each partner's head we used an electromagnetic motion tracking solution, the Liberty system from Polhemus (Polhemus, Colchester, Vermont, USA). This tracker uses passive electromagnetic sensors and a source that emits a electromagnetic field. The source was placed above the table. The sensors consist of electromagnetic coils in a plastic casing, assembled as small cubes measuring 22.9 mm x 27.9 mm x 15.2 mm and weighing 23g. A wire connects each sensor to the base and multiple sensors can be acquired simultaneously with high sampling rates and precision. When tracking, the system provides 6DOF data -- 3D coordinates (x, y, z) and 3D orientation (heading, pitch and roll) relative to the source position.

Parent's speech. To record the parent's voice we used a standard headset with a noise reduction microphone. The parent wore the headset while interacting with her child.

Word Learning through Social Interaction

The task is a common one in the everyday lives of children and parents – to take turns in jointly acting on, attending to, and naming objects. This is a common context in which children learn names for things. The toys used in this experiment were novel things. The child and parent sat opposite each other at a small table and the parent was instructed to interact naturally with the child, engaging their attention with the toys while teaching the words for them.

Participants. The target age period for this study was 18 to 20 months. We invited parents in the Bloomington, Indiana area to participate in the experiment. 5 dyads of parent and child were part of the study (2 male and 3 female). 3 additional children were not included because of fussiness

before the experiment started or failure to keep the head camera on. For the child participants included, the mean age was 18.5, ranging from 17 to 20 months. All participants were white and middle-class.

Stimuli. Parents were given three sets, with three toys in each set, in this free-play task. The toys were rigid plastic objects of simple shapes and were painted with one primary color. Each set had a red, a green and a blue object.

Procedure. The study was conducted by three experimenters: one to distract the child, another to place the head-mounted cameras and a third one to control the quality of video recording. Parents were told that the goal of the study was simply to observe how they interacted with their child while playing with toys and that they should try to interact as naturally as possible. Upon entering the experiment room, the child was quickly seated in the high chair and several attractive toys were placed on top of the table. One experimenter played with the child while the second experimenter placed a sports headband with the mini-camera onto the forehead of the child at a moment that he appeared to be well distracted. Our success rate in placing sensors on children is now at over 60%. After this, the second experimenter placed the second head-mounted camera onto the parent's forehead and close to her eyes.

Calibration of head-mounted cameras. To calibrate the horizontal camera position in the forehead and the angle of the camera relative to the head, the experimenter asked the parent to look into one of the objects on the table, placed close to the child. The third experimenter controlling the recording in another room confirmed if the object was at the center of the image and if not small adjustments were made on the head-mounted camera gear. The same procedure was repeated for the child, with an object close to the child's hands.

Parent-child free play session. The instructions given to the parent were to take all three objects from one set, place them on the table, play with the child and after hearing a command from the experimenters, remove the objects in this trial and move to the next set to start the next trial. Parents were given the names of the objects that they were to use and were instructed to teach the children those object names. However, there was no special instruction as to what the parents had to say or what they had to perform, just that they were to engage their child. All the names were artificial words. There were a total of three trials, each about 1.5 minute long. The interaction between parent and child lasted between 4 and 7 minutes and was free-flowing in form.

Name-comprehension test. After the period of free interaction, the experimenter tested the child's comprehension of the object name for each of the 9 objects. This was done by placing three objects out of reach of the child about 30 inches apart, one to the left of the child one in the middle and one to the right. Then the experimenter looked directly into the child's eyes, said the name of one of the objects and asked for it (e.g. "I want the dax! The grizzly! Get me the grizzly!"). For this portion of the experiment, a camera was focused on the child's eyes. Direction of eye gaze – looking to the named object when named – was scored as indicating comprehension. These recorded eye movements were coded (with the sound off) by a scorer naïve to the purpose of the experiment.

Unimodal Data Processing and Results

The multisensory data recorded include three video sequences from three views, head motion of two participants, and parental speech. This section presents both the methods and the results of processing sensory data for each individual modality. The next section presents the results from an integration of this unimodal data processing.

Video Processing and Results

The recording rate for each of the three cameras is 10 frames per second. There were 3 trials in the interaction, each lasting about 90 seconds. In total, we have collected approximately 8100 ($10 \times 90 \times 3 \times 3$) image frames from each interaction. The resolution of image frame is 320×240. Figure 2 illustrates the procedure of image processing and results. The technical details can be found in (Yu, Smith, Christensen & Pereira, 2007). The relevant information we extracted from three image streams are where objects are in each view at each moment, and what are the sizes of those objects at each moment. In addition, we also calculated, from the bird-eye view, which objects were held by the child's and the caregiver's hands. As illustrated in an example shown in Figure 2, a direct comparison between the child's and the caregiver's views replicated our finding in our previous studies (Yu et al., 2007) – in the same interaction objects occupy about 11% of the child's head-camera field but less than 5% of the parent's visual field. This is because parents and children move the child-attended objects close to the child's head and because children also move their head close to attended objects.

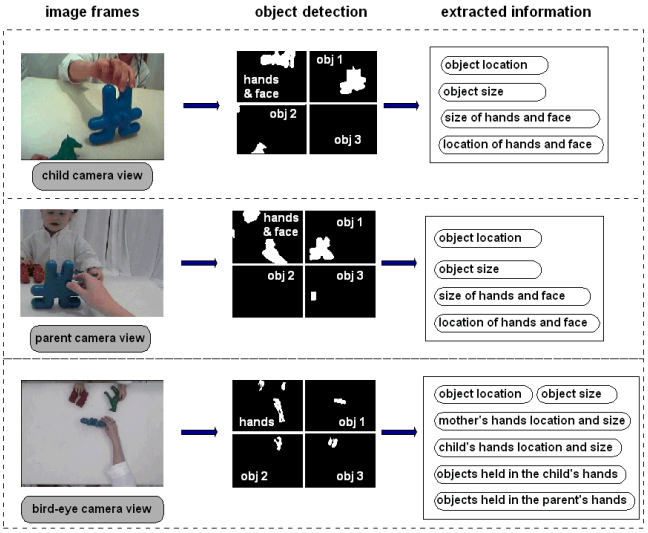

Figure 2: The overview of data processing using computer vision techniques. Our program can detect three objects on the table and participants' hands and faces automatically based on pre-trained object models and skin models. The extracted information from three video streams will be used in subsequent data analyses.

Motion data Processing and Results

Two motion tracking sensors on participants' heads recorded 6 DOF of their head movement at the frequency of 240 Hz. Given the raw motion data {x, y, z, h, p, and r} from each sensor, the primary interest in the current work is the overall dynamics of the head. We grouped the 6 DOF data vector into position {x, y, z} and orientation data {h, p, r}, and then we developed a motion detection program that computes the magnitudes of both head position movements and orientation movements. Figure 3 shows the proportion of time that either children or parents were moving their heads. Head position movements are equally frequent in children and parents. However, children rotate their heads much more frequently than adults do, in the same interaction. This result indicates that young children are more likely to switch their visual attention through head rotation while adults may rely more on gaze shifting. This measure supports our head-camera setup as a means of capturing the child's more dynamic view.

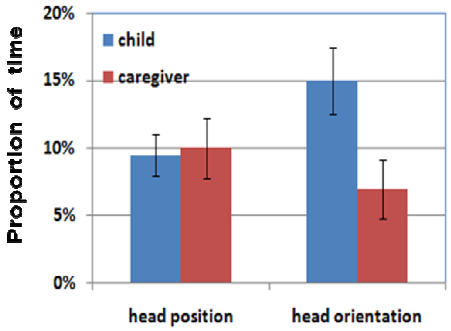

Figure 3: The proportion of time that the child's and the caregiver's heads are moving in the interaction.

Speech Processing and Results

We first segmented the continuous speech stream into multiple spoken utterances based on speech silence. Next, we asked human coders to listen to the recording and transcribe the speech segments. The statistics from the transcriptions show that on average, parents uttered 365 words in each interaction and each spoken utterance consists of 3.5 words. The average size of vocabulary for each interaction is 120. Moreover, nine target object names were produced 32 times in total in each interaction. In the whole dataset of 5 dyads, those object names occurred 161 times in spoken language. We extracted the onset and offset timestamps wherein an object name occurred in transcription and used them to define a naming event in time. We will use those naming events to data mine the patterns in visual and motion data streams.

Results of Word Learning and Naming Events

We correlated the number of naming events for each object name with the learning results at testing and found these two (r=-0.3; p<0.001) at best weakly and negatively correlated. The average of naming events for learned object names is 2.45 per name and 3.5 per name for unlearned names. Thus, object names not learned through interaction were actually uttered more than those learned names. For example, some object names that were provided just once or twice were actually learned and others labelled by caregivers five or six times were not learned. This suggests that what matters are the specific contexts where those object names were named, what both caregivers and children visually attended to at those moments, and what they were doing at that time. We report those behavioral measurements in the next section.

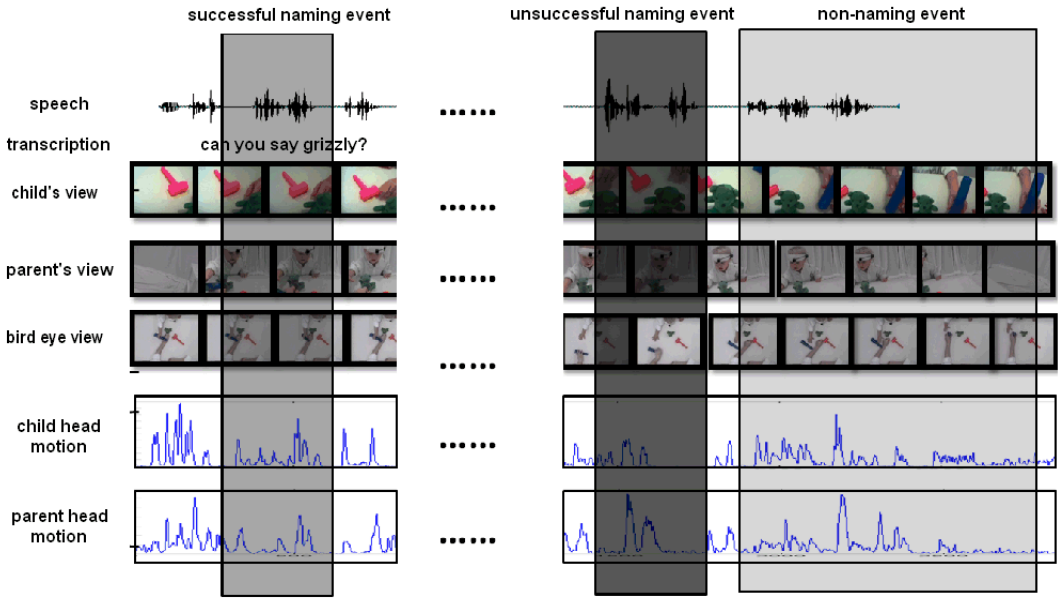

Figure 4: Continuous data were segmented and grouped into three categories: successful naming events, unsuccessful naming events and (other) non-naming events. A comparison of visual data and motion data was made based on these three event categories.

Multimodal Data Analysis and Results

Given the complex multimodal multi-streaming data collected from two participants, we opted to use the learning results collected in testing from young learners as teaching signals to guide us in data mining this fine-grained multimodal data. This method is different from most modeling approaches which build a simulated model first to make predictions about results and then correlate the predictions with actual experimental results. Instead we use here experimental results as supervisory information to search for reliable patterns from this complex multidimensional dataset. From a technical viewpoint, this approach is also different from standard unsupervised data mining approaches because we take advantage of behavioral information to facilitate data mining and pattern detection.

Figure 4 shows our overall approach for multimodal information integration, which consisted of two steps. First, we started by grouping naming events (results from speech and language processing) into **successful naming events** (n = 65) and **unsuccessful naming events** (n= 96) based on the testing results measured at the end of each parent-child interaction. In addition, we grouped the remaining moments in the interaction as a third kind of event – **non-naming events** (n = 132). In this way, a whole temporal data stream can be decomposed and labeled by these three events. Next, we extracted various measures and statistics from visual and motion data, and compared those results across three event groups. Any differences on a certain measurement between successful and unsuccessful naming events will indicate the potential importance of this pattern in learning-oriented social interaction. In contrast, similar results across successful and nonsuccessful events suggest that the pattern under consideration may not play any major role in word learning. In addition, the third event group – non-naming event – provides a baseline. The differences between non-naming and two naming events will identify those behavioral patterns in a social interaction that caregivers generate when they teach object names, no matter whether the naming events themselves are successful or not. The following results will focus on three different measures: visual fields, hand movements and head movements, respectively.

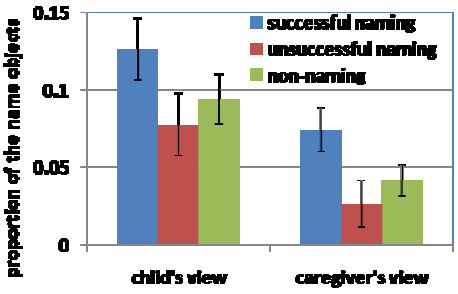

Figure 5: The proportion of the named objects in two views. In both views, the proportion of named objects is much bigger at the moments of successful naming events compared with either unsuccessful naming events or the baseline (other moments in the interaction).

Named objects in visual fields

The proportion of a visual field occupied by named objects may be viewed as a measure of the named objects' dominance over other objects in the viewers' attentional field. As shown in Figure 5, our analyses indicate that the named objects occupied a larger proportion of the child's visual field in successful naming events compared with that in unsuccessful naming events. The same trend holds with visual data from the parent's perspective. Putting together, the results suggest – not surprisingly – that object names are learned more effectively when the named object is visually salient in both the learner's view and the teacher's view, namely, when the child and the caregiver jointly attend to the same object.

Named objects in hands

The percentage of time in each event category that the named objects are either in the child's hands or in the caregiver's hands can also be viewed as a measure of attention to that object. As shown in Figure 6, more successful naming events are those in which the named object is in the child's but not the caregiver's hands. More specially, in about 45% of time when a successful naming event happened, the named object was in the child's hands. Meanwhile, the named objects were in the caregiver's hands in only 8% of time. Two implications follow from these results: First, those learning moments in which the parent correctly gauges the child's attention and then provides linguistic labels, may be most effective for word learning. Second, parents can infer the child's attention through the child's hand actions.

Head Movement and Word Learning

As shown in Figure 7, the third measure asks whether the child or the caregiver holds his/her head still during naming events. Our first finding is that both the child and the caregiver move their head more dramatically in unsuccessful naming events compared with successful

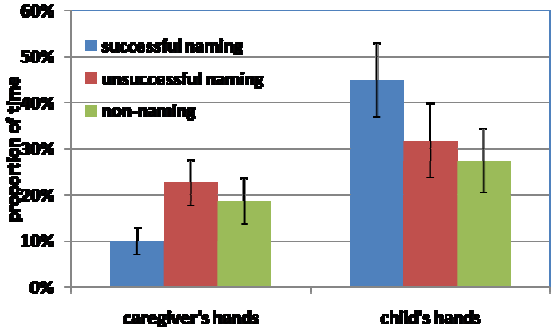

Figure 6: The proportion of time that participants' hands are holding an object. In successful naming events, the child's hands most often held the named objects, which happened less frequently in unsuccessful naming events. Indeed, the caregiver tended to hold the name objects in unsuccessful naming events even compared with non-naming moments.

naming events or the basic line. Second, the child's head is oriented more stably during successful naming events. This suggests that sustained attention is critical to learning object names.

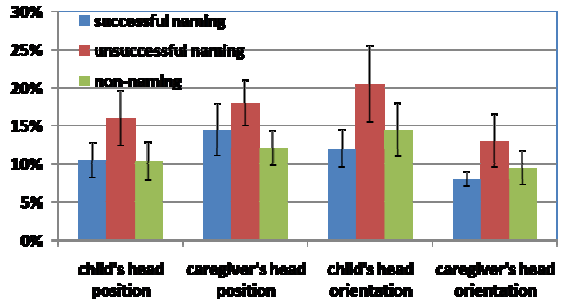

Figure 7: The proportion of time that the child's or the caregiver's head is moving. We found that the child's head tends to have a stable orientation in successful naming events. Also both the child and the caregiver move more dramatically in unsuccessful naming events.

General Discussions and Conclusion

Most of children's word learning takes place in messy contexts – like the tabletop play task used here. There are multiple objects, multiple shifts in attention by both partners, and many object names that might be learned. In these contexts, very young children do not always learn the names of things but they must learn some. The goal of this work is to understand the qualities of real world interactions between young word learners and parents that organize that learning. The number of naming events is not the most important variable. Instead, naming needs to occur at the right moment in *time*, when both parent and child are attending to the same object. However, looking at an object, the metric of attention usually used in highly simplified artificial learning tasks, may not be the best real-world metric on attention. Instead, active engagement – that is, manual actions on the object –may be a better metric of the child's interest and thus readiness to learn the name. Finally, a quieting of head movements, an index of sustained orientation to the object, also predicts learning. These three dimensions of attention – shared visual attention, manual engagement, and sustained attention – fluctuate dynamically in the interactions between children and parents. Key questions for future work are whether dyads of parents and children differ in the dynamic qualities of these interactions with some modes of interaction being generally more effective than others. Also of interest is whether these individual differences in dyads emerge from children's attentional differences, from differences in parent sensitivity to the child's attention, or both.

Moving away from abstract and mechanistically ungrounded ideas such as "mind reading" and inferred intentions, and moving away from sparse experimental settings unlike the dynamic interactions of real world learning, may provide a leap forward in understanding natural word learning in humans (and in building computational devices that can learn words in the same contexts that children do). Further, inferences about the mental states of others must arise from their external bodily actions, bodily actions that in the real world are highly dynamic. The study reported here is a first step in understanding these dynamics.

In this paper, we use advanced sensing equipment and state-of-the-art experimental paradigms to collect multiple streams of real-time sensory data in parent-child interactions. A further strength of this research is the application of computational techniques to analyze these multisensory data to measure the statistical regularities in the learning environment. Thus, with more fine-grained data and advanced analysis tools, we have the opportunity to discover a more complete mechanistic explanation of early word learning.

Acknowledgments: We thank Amara Stuehling, Jillian Stansell, Saheun Kim, and Mimi Dubner for collection of the data. This research was supported by National Science Foundation Grant BCS0544995 and by NIH grant R21 EY017843.

References

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

Baron-Cohen, S. (1995). *Mindblindness: an essay on autism and theory of mind.* Cambridge: MIT Press.

Bloom, P. (2000). *How children learn the meanings of words.* Cambridge, MA: The MIT Press.

Butterworth, G. (1991). The ontogeny and phylogeny of joint visual attention. In A. Whiten (Ed.), *Natural theories of mind: Evolution, development, and simulation of everyday mindreading* (p. 223- 232). Oxford: Blackwell.

Smith, L. (2000). How to learn words: An associative crane. In R. Golinkoff & K. Hirsh-Pasek (Eds.), *Breaking the word learning barrier* (p. 51-80). Oxford: Oxford University Press.

Smith, L.B. & Breazeal, C. (2007) The dynamic lift of developmental process. *Developmental Science*, 10, 61-68.

Tomasello, M., & Akhtar, N. (1995). Two-year-olds use pragmatic cues to differentiate reference to objects and actions. *Cognitive Development*, *10*, 201-224.

Woodward, A. L. (2004). Infants' use of action knowledge to get a grasp on words. In D. G. Hall & S. R. Waxman (Eds.), *Weaving a lexicon.* MIT Press.

Yu, C., Ballard, D.H., & Aslin, R.N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science, 29* (6), 961–1005.

Yu, C. & Ballard (2004). A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perceptions* 1(1):57–80.

Yu, C., Smith, L. B., Christensen, M., & Pereira, A. F., (2007). Two Views of the World: Active Vision in Real-World Interaction. In McNamara & Trafton (Eds.), *Proceedings of the 29nd annual conference of the cognitive science society* (p 731-736). Mahwah.NJ: ErIBaum.