Joint Attention through the Hands: Investigating the Timing of Object Labeling in Dyadic Social Interaction

Martin E. Rickert and Chen Yu
Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405

***Abstract*—Previous studies on joint attention and its role in language learning focus on eye gaze cue. The goal of the present paper is to discover fine-grained patterns of joint hand activities in child-parent social interaction that lead to successful word learning. To this end, we focus on the following three topics: 1) quantifying joint manual actions between the parent and the child and in particular how the child follows the parent's bid of attention through manual actions; 2) discovering the timings between joint manual actions and object naming events; and 3) linking those timings with language learning results. Multiple high-resolution data streams were examined for episodes involving object-labeling events that either preceded or followed joint attentional focus as established through the hand actions of the dyad. Our findings suggest that the success of word learning through social interaction depends on the specific timing between follow-in joint hand activities and naming events.**

***Index Terms*—joint attention, embodied cognition, social interaction, lexical acquisition, visual data mining.**

I. INTRODUCTION

Chilren learn about their world – about objects, actions, other social beings, and language – through their second-by-second, minute-by-minute sensorimotor interactions. Everyday social activities such as toy play with parents are the context for learning as it unfolds in real time. A well coordinated child-caregiver interaction seems likely to lead to better learning while a decoupled or non-coordinated interaction may disrupt learning and development. Indeed, a diagnostic marker for autism is a failure to engage in normal social interactions with others [1]; these disrupted interactions may not only be a manifestation of this developmental disorder, but also causally contribute to the broader behavioral and cognitive, social, and behavioral delays often observed. At the macro behavioral level, various methods have been developed to access early social communication, such as surveys and interviews from the caregiver, video observation of activities at home, and experimental studies to measure the child's sensitivity to social stimuli [2] and [3]. But we know very little about how macro-level behaviors work at the sensorimotor micro-level, in real time and in the cluttered context of everyday parent-child interactions. Moment-by-moment, the child's actions – head and eye movements, hand movements, picking up objects – create within the child dynamic *multimodal dependencies* of looking, seeing, touching and feeling, with each bodily action determining the next sensory event and each sensory event generating new momentary goals and the next action. Thus the child is a dynamic multimodal system. But the toddler is not alone in the physical environment. Instead, in his social environment, a mature partner – who is also a multimodal system – offers words, gestures and actions. Moreover, the experiences and behaviors in child-parent interaction -- the streams of touches, sights and sounds – are not independent for the two participants. Instead, dependencies and interaction patterns occur in two directions with the young child shaping the experiences and behaviors of the mature partner through his own bodily actions and his own sensorimotor experiences, and with the mature partner likewise directly influencing the sensorimotor experiences and actions of the young learner.

The goal of the present research is to seek to develop a set of novel data collection and analysis methods that allow us to document and describe the dependencies and fine-grained sensorimotor patterns in child-parent interaction, and discover how they organize early social communication in toddlers. In particular, we are interested in studying joint attention in child-parent interaction and how this joint social activity may lead to better learning. While most previous studies of joint attention focus on eye gaze as a primary cue to establish and maintain joint attention between two social partners, the present study investigates the role of manual actions of hands in the context of everyday toy play. We study naturalistic everyday interaction because that is where learning and interaction happen in the real world. We study hand activities because hands may generate stronger and more salient signals of one's attention (therefore it can be easily detected) as a more reliable and sustainable cue compared with more subtle and temporally brief cue such as eye gaze. As our first attempts to systematically study joint manual actions in lexical acquisition, this paper concentrates on the following three topics: 1) quantifying joint manual actions between the parent and the

Manuscript received March 6, 2010. This work was supported by NSF BCS 0924248 and AFOSR FA9550-09-1-0665.

Computational Cognition and Learning Laboratory, Dept. of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405 USA (phone: 812-856-0840; e-mail: rickertm@indiana.edu).

child in toy-play interaction; 2) discovering the timings between joint manual actions and object naming events; and 3) linking those timings with language learning results. In the following sections, we will first introduce our experimental setup and data collection methods, and then present our results from data-mining behavioral patterns from such data.

II. Multimodal Sensing System

A multimodal sensing system (Fig. 1) was developed to capture moment-by-moment changes in the sensorimotor information available in naturalistic parent-child interactions. This system simultaneously records streams of time-series data representing multiple visual perspectives, speech and vocal utterances, as well as body, head, and hand-motion and position-in-space. The room environment in which interactions were studied is described briefly in Section I-A. An overview of data acquisition components are described in Sections I-B, -C, and –D. An overview of signal and data pre-processing is provided in Section I-E. More complete descriptions of the system can be found in [4]- [6].

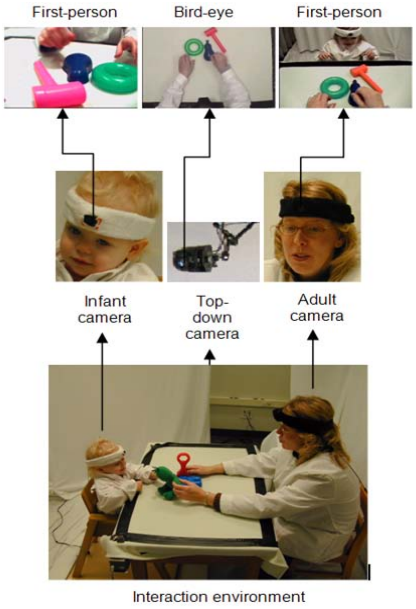

Fig. 1. Multimodal sensing system. The child and the parent play with a set of novel toys at a table. A mini-camera is worn by each participant to collect visual information from a first-person view. A third, high-resolution, camera located above the table records the bird's-eye view of the parent-child interaction.

A. Dyad-Interaction Environment

A large 6.5m x 3.3m room was partitioned into two spaces using portable sound-treated wall dividers. One side of this room contained all recording equipment including data acquisition computers, multiple monitors, as well other digital and analog devices. Recording of the parent-child interactions were monitored in real-time by one or more research assistants seated in this control-room area. The other side of the room was designed as an interaction environment for the parent-child dyads. The walls and floor of the 3.3m x 3.1m interaction space were covered with all white fabrics. A small 61cm x 91cm x 64cm table painted a nonglossy white was positioned in the center of this space and provided the surface for table-top play and object interaction. In addition, both participants wore white shirts. This uniform white 'background' facilitates feature extraction based on detection of nonwhite pixels (i.e., objects and hands) during post-processing of the recorded video data streams.

B. Multi-Camera Setup

Dynamic visual scene information was recorded from both the parent and child point of view. Each participant wore a lightweight head-mounted camera fastened to a sports headband. The camera mount piece allowed the camera to rotate on the horizontal axis perpendicular to the camera's line of sight. We used this to adjust the angle between the camera line of sight and the participant's head orientation. Each camera covered a visual field of approximately 90°, both horizontally and vertically. Static visual scene information was recorded using a high resolution digital camera mounted above the table-top in the interaction space. As shown in Fig. 1, this bird's-eye view provides a source of visual information that is independent of participants' gaze direction and body motion and position. This high-resolution data stream is used to improve automated visual segmentation of objects from the low-resolution head-mounted cameras data streams via information fusion techniques.

C. Speech Recording and Motion Tracking

The data stream for parents' speech was recorded through a headset microphone at a sampling rate of 44.1 kHz and with 16-bits A-D resolution. Head position and orientation was recorded using a Polhemus motion tracking sensor attached to headband worn by both the parent and child.

D. Object-in-Hand Coding

Two human coders were asked to manually code the information whether the child or the parent is holding one of three visual objects. An agent is coded as holding or touching a visual object as far as his hands made contact with an object. This coding was done by a home-made program which simultaneously displays three video streams (the child's head camera, the parent's head camera and the bird's eye view) and therefore allows the coders to reliably detect object-in-hands events reliably even in case the visual information from one camera might not be clear. The result from this coding consists of 12 binary time series indicating whether either of agents is holding one of the three objects using either his left or right hands (2x3x2), which will be used in the following data analyses.

III. Experimental Procedures

A. Participants

The data reported here are for a sample of eighteen parent-child dyads. The target age period for children in this study was 18 to 24 months, a period of large developmental changes in early word learning and visual object recognition (Smith and Pereira, in press). Children's mean age was 21.3 month, anging from 19.1 to 23.4. All participants were residents from the Bloomington, IN area, white, and middle class.

B. Stimuli and Procedure

Parents were given a maximum of six sets of toys (three toys per set) in a free-play task. The toys were either rigid plastic objects or plush toys with simple shapes and a single color or an overall main color to facilitate computer-based visual processing. The protocol used in this study required a team of three experimenters. Two had specific roles involving the parent-child dyad; one provided instructions to the parent and placed equipment on the child; the other distracted the child until recording of the interaction began; the third provided feedback during calibration and monitored the quality of the video recording in real-time. As soon as the child seemed well distracted, the headband with the camera and motion sensor was placed on the child's head so that the camera was in a good position re: the child's eyes. Instructions for the parent were phrased in a conversational tone and included a brief overview of the study's goal, i.e., to investigate how parent and children interact with the each other. In addition, parents were to take three toys from a drawer and engage in table-top play with the child for a short period of time and then switch to the next set of toys when cued by an auditory signal. The main function of the second experimenter was to keep the child distracted

C. Trial structure and timing

After this calibration phase, the experimenters removed all objects from the table, asked the parent to start the experiment, and left the room. Both experimenters left the interaction space following successful calibration of the head-mounted cameras. Parents listened for a specified sound and when alerted by it, began a new trial. There were a total of six 90-second trials. The entire study, including initial setup, lasted for 10-15 minutes.

D. Object-Label Learning

Following completion of the free-play interaction trials, we obtained a learning score for each of the six toys using a 3-AFC procedure and two trials per *object*. There were two foils and a single target object in each testing trial; learners indicated their preference by pointing to or gazing at the target object. If the learner correctly selected the target object twice in both testing trials for that object, we scored the learning of that word-object pairing as 2. If the learner correctly selected the target object in one of the two trials, the score is 1. Otherwise, the score is zero. Thus, a perfect learner would score 12 if s/he successfully learned all of the six objects. This learning result serves as an important dependent measure in our analyses; specifically, we correlate these scores with measures quantifying sensorimotor patterns discovered from interaction to identify which patterns lead to better word learning.

IV. Results

Section IV-A begins with a brief description of our approach for extracting object-in-hand timing relations for purposes of histogram calculation. Overall patterns of in-hand events observed during parent-child dyad interactions are presented. Frequency distributions presented in Section IV-B illustrate the range of potential patterns in the timing relation of referent-object labeling by the parent to object-in-hand event by child; we assume that children's attentional focus is on the object in-hand (i.e., consistent with extant definitions of follow-in labeling). Finally, quantitative measures including global and conditional counts, estimates temporal location and dispersion, and entropy were derived from the individual histograms. Although the majority of the derived measures show only a weak association with object-label learning scores, there is evidence that learning scores are correlated positively with mean temporal correlation between the naming events and the onset of an object in-hand event by the child.

A. Parent-Initiated Object Exchanges

We operationally defined two types of parent-initiated events involving object exchanges with the child. The schematic timelines shown in Fig. 2 demonstrate the qualitative difference in motor coupling for these two cases. In the first case, our algorithm scanned the parent and child object-in-hand event stream data for occurrences when the same referent object was jointly held by the dyad under the constraint of a minimum temporal overlap of 1 s. In the other case, the algorithm scanned the event stream data for object exchanges involving a temporal gap between parent and child in-hand

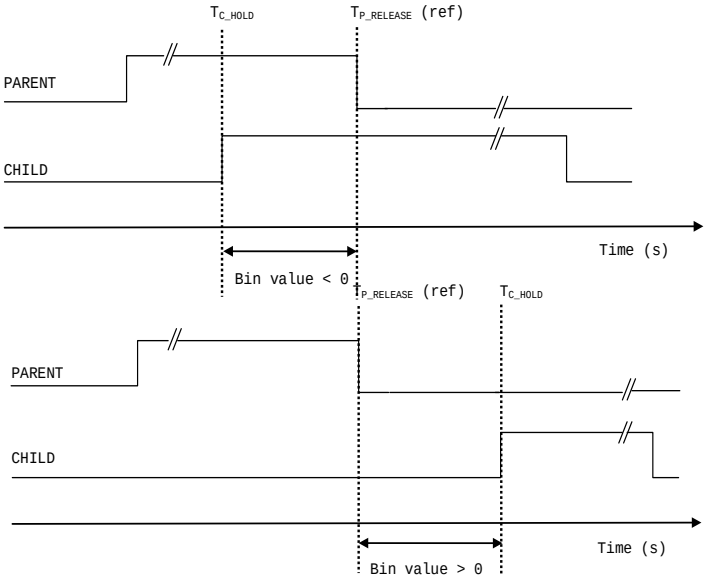

Fig. 2. Illustration of the relative timing of parent and child holding event streams. (Top) Follow-in with temporal overlap. Object-specific holding by child is initiated at time T_{C_HOLD} while parent is holding the object but before parent releases the object at time $T_{P\_RELEASE}$. (Bottom) Follow-in events with a temporal gap. Object-specific holding by child begins at time T_{C_HOLD} and occurs after the object is released by the parent at time $T_{P\_RELEASE}$. Histograms of follow-in events were constructed using a fixed bin width of ΔT = 300 ms referenced to $T_{P\_RELEASE}$.

actions, i.e., gap episodes occur when the parent places the referent object on the table top and the child picks it up and holds it for a minimum duration of 1 s.

Frequency histograms characterizing the global pattern of object-directed joint hand actions for the entire sample of N=18 dyads are shown in Fig 3. These distributions were constructed using a fixed bin-width of 300 ms and constraining

the event search to fall within ±5 s of the object release by the parent. Each color represents a different dyad. The top panel is for the overlap condition (negative bin values) wherein both the parent and the child simultaneously held the same object during exchange and the bottom panel is for the gap condition (positive bin values) wherein the parent released the object on the table and the child picked it up. There are two notable features in the aggregate distributions. First, although there is considerable and asymmetric dispersion over the entire analysis window, there is little or no bias in the total number of in-hand events for overlap (53%) versus gap (47%). Second, the majority of exchanges occur within a relatively narrow temporal window (1-2 s) of parent's release of an object.

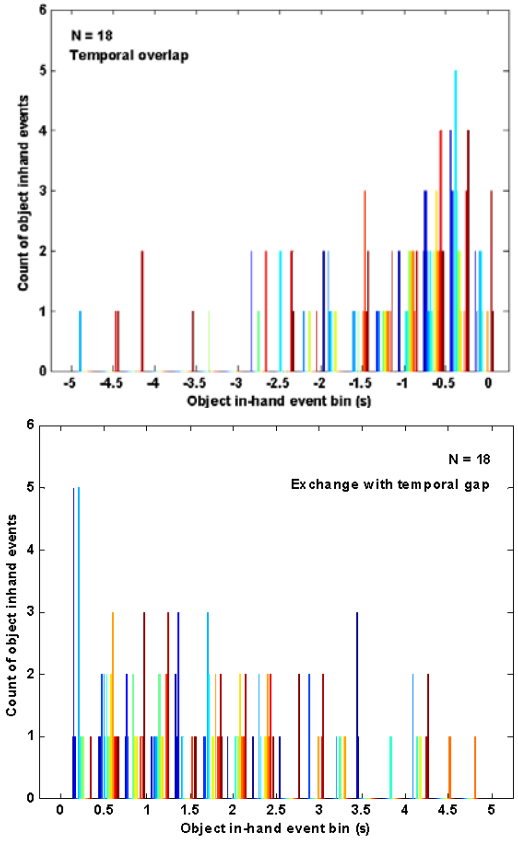

Fig. 3. Parent-initiated object-exchange histograms for N=18 dyads. Negative event bins (top panel) correspond to parent and child in-hand. Positive event bins (right panel) correspond to temporal gap, i.e., parent releases an object before child picks it up.

Potentially significant sources of variation in the in-hand event data can be identified by plotting cumulative distributions for individual subjects collapsed across all ten second of the analysis window. As evident in the asymptotic count values of Fig. 4 (left panel), there is considerable individual variation in the total number of event for these eighteen dyads. The median number of counts is 21 with a normal-consistent estimate of the standard deviation equal to 5.93. The slopes in the cumulative relative frequency distributions shown in the right panel of Fig. 4 are reasonably consistent; only a few dyads exhibit an obvious bias in the number of in-hand exchanges for either overlap (olive green and turquoise traces) or gap (black trace) exchanges.

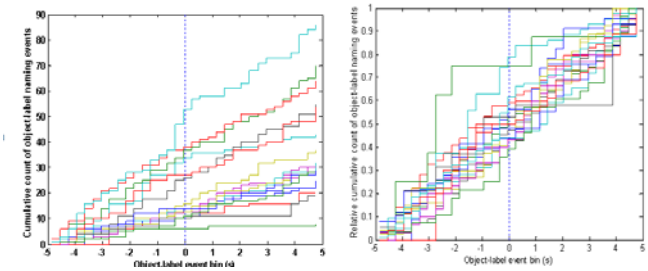

Fig.4. Cumulative in-hand event count histograms for individual dyads (each color represents a different dyad). Raw and relative count distributions are plotted in the left and right panels, respectively.

B. Naming Events While Objects are In-Hand

In this section, we examine object-labeling timing relative to the child-inhand event given that parent-initated joint activity and object-exchange. As in the previous section, our approach involves histogram estimation conditional on detection of parent in-hand event. Previous reserch has demonstarted that object-labeling provides critical input for early lexical acquisition [3]. Object labels that follow-in to the child's attention correlate positively with lexical development, whereas redirective (i.e., "lead-in") object labels correlate negatively with child language. However, positive correlations are observed when lead-in labels are supported by a combination of gesture and vocabulary size. In the current analysis, we assume that parent-initated joint hand interectaion may be sufficient to establish a (momentary) state of joint attention and holding an object is a valid, albeit low-dimensional, indicator of latent attentional focus derived frorm the sensorimotor stream.

Only a subset of the eighteen individual results are presented in Fig 5. The data for four subjects shown here do, however, capture the main features and patterns of variation in the timing event histograms for the full sample. Each of the six objects has been color rendered consistently across panels. The abscissa represents the onset of labeling events referenced temporally to when the child began holding the referent object. Counts have been aggregated over all events involving each specific object. In addition, each panel legend provide dyad-specific information including the total number of labeling events (in parentheses) for all episodes involving the referent object as well as object-label learning scores for that object. As an example, for SUBID 65, there were a total of 14 labeling events in the analysis window while object O2.1 (cyan bars) was in-hand and, as is readily apparent, most of these events preceded the child in-hand episode. At test, object O2.1 was correctly selected by the learner one out of two times.

Overall, we note that there is considerable variation in the object-label timing re: object in-hand by the child both across dyads and within dyads across objects. More important, the number of correct learning responses is not correlated with the number of labeling events either preceding or following the in-hand event. It is also evident that perfect (2/2) learning scores for some objects and dyads are associated with relatively few labeling events. Such cases may reflect a high degree of efficiency or consistency in the interaction.

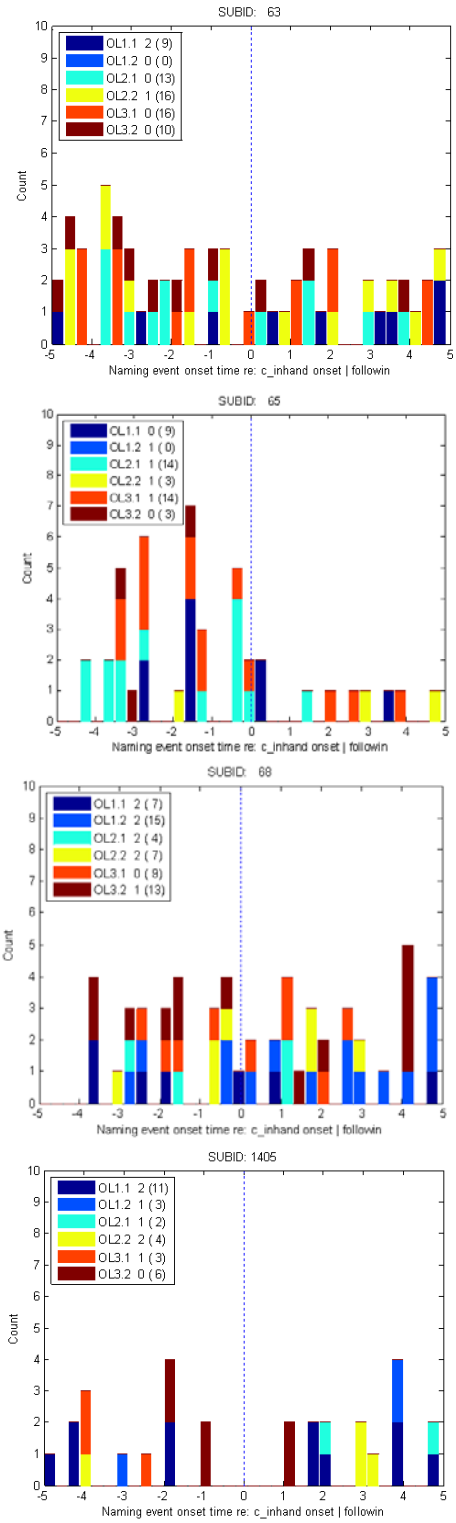

Fig. 5(a)-(d). Object-labeling histograms triggered on child in-hand event for four dyads. The object-label in the legend is followed by (1) the number of correct object-labeling responses at test, and (2) the total number of naming events counted within the ± 5 s analysis window re: parent release of an object. Thus negative bin values correspond to object labels that anticipate a child in-hand event whereas positive event bins correspond to labeling that is concurrent with a child in-hand event.

C. Object-Labeling Learning Scores

A simple measure of object-label timing consistency was derived from the individual histograms and used as a predictor variable for object-label learning scores. Fig. 6 shows the distribution of object-label scores (median = 5.5 and a normal-consistent estimated standard deviation of 2.78). The regression of learning scores on mean object-label timing shown in Fig. 7 is statistically significant ($F_{(1,16)} = 4.786$, $p < 0.05$) and not dependent on data assumptions (i.e., using robust methods results in a nearly identical fit). Our results thus support this teaching strategy: parents should attract the child's attention to the visual object first and then release the target object to allow the child to pick it up; next parents should wait until the child has held an object for a little while (1 or 2 seconds) and then name the object. By contrast, those parents who named objects too early, even before the child's holding actions, may have failed to establish the best attentional context for teaching object labels. More generally, our results convincingly demonstrate that the timing of object labeling and joint manual activities may be a critical factor for successful learning from social interaction. The coupled hand activities between the child and the parent may establish a joint consensus in social engagement when object naming takes place right on time.

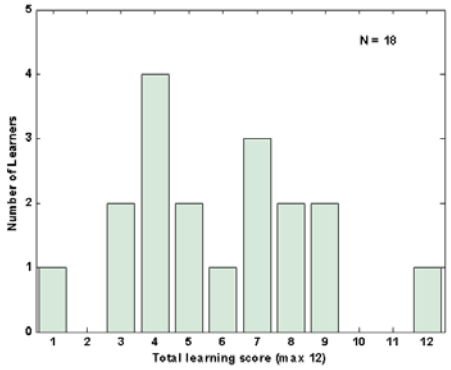

Fig. 6. Distribution of the number of learners with total learning score on the 3-AFC task as indicated on the abscissa. The test phase consisted of two repetitions with each of six novel toys; the maximum possible score of 12 was observed in only one learner.

V. Discussion

Joint attention is a well studied topic in both understanding human cognition and in building artificial intelligence systems. Most often eye gaze is treated as a primary cue in triggering and maintaining joint attention. The present study shows that both the child and the parent use their hands as a channel to indicate his/her communicative intent and also infer the other social partner's attention in the context of everyday toy play. We suggest joint activities through the hands might be more salient and therefore easy to detect from a pragmatic perspective, compared with other more subtle cues, such as eye gaze. The present study is our first effort toward understanding the role of the hands in dyadic social interaction and more importantly understanding how joint hand activities may lead to successful lexical acquisition from naturalistic interaction. Our approach is based on analyzing both the parent's and the child's momentary hand activities on objects. This moment-by-moment coding and data analysis lead to several interesting findings from this first effort. First, when a parent attempts to elicits the child's attention by using her hands to bring an

object to the child, it is critical that the child responses to this invitation within a temporal window of 2-2.5 seconds. Beyond this window, the chance that the child will follow the parent's elicitation is relatively low, suggesting that the parent should probably start a new attempt. Second, the number of naming events is not correlated with the success of word learning. More specifically, those words that were uttered more by the parent were not acquired better than those words uttered less

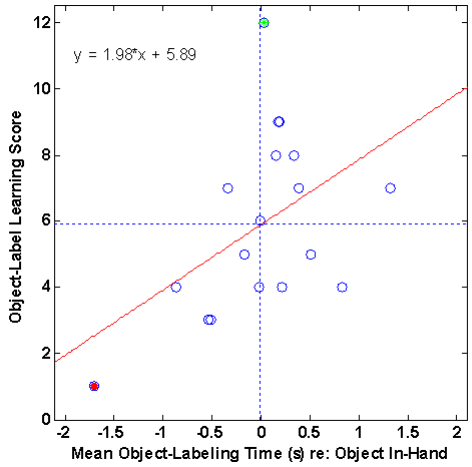

Fig. 7. Regression of learning score on mean naming event time (s) re: child in-hand event. Horizontal and vertical reference lines represent the arithmetic mean of each variable. Pearson correlation coefficient is equal to 0.4798

frequently, suggesting that the quality but not the pure quantity of naming events is a more important factor in the context of learning from social interaction. Third, our results of correlating naming events and follow-in hand activities indicate potentially different ways of successful word learning. As showed in Figure 5, each naming event may differ from one dyad to another dyad and from one instance to another within a dyad. Nonetheless, one major finding in this paper is the correlation between the timing of follow-in events and naming events. Those parents who named a visual object consistently after the child was holding it created a better learning situation and therefore led to better learning. This result suggested two important components in successful learning from social interaction: 1) the role of the teacher is to attract the child's attention and name an object after the child takes the bid of the attention; and 2) whether a naming event is successful or not depends on the child's own activities, and the manual action of holding an object in hands can be a stronger commitment of embodied attention compared with using a brief glimpse by eye gaze. Moreover, the seamless coupling of those two sequential components can lead to better learning.

In the present work, there are two major motivations to link lexical learning with social interaction and communication. First, we are interested in how language learning happens in real-world scenarios wherein parents and children freely interact with each other. We believe that going beyond well-controlled experiments and moving towards studying naturalistic interaction may allow us to give complimentary insights how learning happens to advance our theoretical understanding with a more complete picture. Moreover, the findings from such research have a potential with applied utilities. For example, those results may provide some principled guidance to facilitate child-parent social interaction especially for young children with language learning deficits. Moreover, those findings from child-parent interaction can also be used to build intelligent robots that can learn from human teachers like young children. The other major reason to study lexical learning here is to use word learning results as a dependent measure which allows us to use this metric to evaluate interactive patterns in child-parent interaction. Due to the nature of this fine-grained dataset, as we show in Figs 3 and 4 we observed a range of temporal patterns in follow-in activities. Without a dependent measure as supervisory signals, we cannot tell which patterns are better (in terms of smooth communication) than others. Thus, more generally, we used word learning as a way to discover principles in human-human communication. Again, the findings from such research are also directly relevant to human-robot interaction [5]. The present paper reports the first effort towards this direction which raises a set of open questions for future research. For example, we are interested in further analyzing momentary actions between the parent and the child by incorporating visual information from both the parent's and the child's views. For another example, we are also interested in investigating both local momentary actions and global rhythmic micro-behavioral patterns in joint bodily activities and the roles of those two kinds of activities in communication and learning.

REFERENCES

[1] P. A. Filipek, P. J. Accardo, G. T. Baranek, E. H. Cook, G. Dawson, B. Gordon, *et al*, "The screening and diagnosis of autistic spectrum disorders," *Journal of Autism and Developmental Disorders*, *9*, pp. 439–483, 1999.

[2] D.A. Baldwin, "Understanding the link between joint attention and language," in C. Moore & P. J. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 131-158). Hillsdale, NJ: Lawrence Erlbaum, 1995

[3] M. Tomasello, and J. Todd, "Joint attention and lexical acquisition style," F*irst Language*, vol. 4, no. 12, pp. 197-212, 1983.

[4] H. Yoshida and L.B. Smith, "Hands in view: Using a head camera to study active vision in toddlers," *Infancy*, 2007

[5] C. Yu, L. B. Smith, H. Shen, A. F. Pereira, and T. Smith, "Active Information Selection: Visual Attention Through the Hands," *IEEE Transactions on Autonomous Mental Development*, Vol. 1 , No. 2, pp 1-11, 2009.

[6] C. Yu, M, Scheutz, M. and P. Schermerhorn (2010) "Investigating Multimodal Real-Time Patterns of Joint Attention in an HRI Word Learning Task," *5th ACM/IEEE International Conference on Human-Robot Interaction.*

[7] G. Deák, M. Bartlett, and T. Jebara, "New trends in Cognitive Science: Integrative approaches to learning and development," *Neurocomputing*, vol. 70, no. 13-15, pp. 2139-2147, 2007.

[8] C. Breazeal, and B. Scassellati, "Infant-like social interactions between a robot and a human caregiver," *Adaptive Behavior*, vol. 8, no. 1, pp. 49, 2000.

[9] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots," *Robotics and Autonomous System*, 37:185–193, 2001.

[10] M. Booth, A.E., McGregor, K.K. & Rohlfing, K. J. "Socio-pragmatics and attention: Contributions to gesturally guided word learning in toddlers," *Journal of Language Learning and Development* 4: 179-202, 2008.