The Effects of Deictic Pointing in Word Learning

Hilary Kalagher
Indiana University
Bloomington, IN 47405
hkalaghe@indiana.edu

Chen Yu
Indiana University
Bloomington, IN 47405
chenyu@indiana.edu

Abstract

Previous research suggested that eye gaze as a social cue plays a crucial role in early word learning. In light of this, we investigated another kind of embodied social cue, pointing, and asked how it relates to word learning in young children as it is ubiquitous in day – to – day parent- child interactions. Parents were asked to narrate a story book displayed on a computer screen. Each page of the story contains the pictures of multiple objects and the novel spoken names of those objects were introduced during the narration. Word learning was measured at the end of the story. The three learning conditions were, (1) pointing to the correct object while labeling it, (2) no pointing, and (3) general pointing to the center of the screen but not to a specific object. The results showed embodied pointing actions significantly increase word learning. Moreover, a touch screen panel placed over the computer screen was used to record the time and location of each pointing action. We developed and implemented various approaches to measure the spatial and temporal correlations of parental speech and pointing actions. The results of detailed analyses suggest that exact synchrony and degree of overlap of speech and pointing streams of action are not directly relevant to learning efficiency. Overall, this work suggests both that social cues, such as pointing, are embedded in a system of correlations relating the speech stream to the physical world of objects and events and that the human word-learning system is robust.

Index Terms: word learning, social cues, language learning, and intermodal synchrony

I. Introduction

Children learn language similarly to the way that they put individual pieces into a puzzle. They perceive multiple spoken words (pieces), while simultaneously perceiving objects to which those words refer (places where the pieces belong). The task of word learning is then to put particular pieces into the appropriate places, and thus solve the puzzle.

Before children are able to appropriately place words in the correct locations in the puzzle they must first be able to accurately segment speech streams. That is, children must be able to parse a continuous speech stream into discrete pieces. Research on this topic has shown that infants as young as 8 months of age can accurately segment speech streams after only two minutes of exposure to an artificial language [1]. Further research suggests that slightly older American infants were able to successfully discern between phonetic units of Mandarin Chinese after only 5 hours of exposure [2]. This evidence implies that infants at a very young age are quite capable of completing this very basic first step of language acquisition.

Once they are able to segment speech properly (pick out the pieces), the next step is to discover the places that those pieces belong to. Linking spoken words to their intended referents is not an easy task. At any given moment when words are heard there are a myriad of co-occurring objects and events in the physical world to which those words could be referring. This problem is termed reference uncertainty [3]. Richards and Goldfarb (1986) proposed a solution wherein children add words into their vocabularies primarily by a statistical associative process. They learn to pair a word with its corresponding referent by repeatedly seeing the object while simultaneously hearing the word. The more times a word is heard while the referent is present, the more quickly and accurately the association between the two is made. Unfortunately, this approach does not provide a clear explanation as to how children select the correct referent of a word from all the co-occurring objects while the word is heard. If this were the only mechanism through which children attain language they would not be able to add words into their vocabulary smoothly and efficiently. Furthermore, they would make wrong associations by pairing irrelevant but co-occurring words and referents. Therefore, associative learning seems less likely to be the whole story of word learning.

Bloom (2000) proposed that there is another mechanism that children employ. He suggested that the majority of word learning is dependent on how well a child can understand and interpret the thoughts of a speaker. The ability to decipher the actions of a mature speaker through reading the speaker's thoughts would enable a child to narrow down the possible referents. Tomasello (2000) argued that children can determine adults' referential intent in complex situations through the understanding of the speaker's thoughts.

It is widely accepted that children are also sensitive to the actions of those speaking in their presence [7]. Actions which contain clues as to referential intent, or social cues, are signaled by parents and guide children's learning by linking the spoken words in the auditory stream with the correct referents in the visual stream. That is, parents, and other mature speakers of a language act as co-puzzle solvers, leading children in the correct direction. In this view, children would consult the speaker's actions when trying to discern the intended referent. Children then consider those actions when making decisions and narrowing down the possible referents. Coupling the consideration of speakers' actions with statistical regularities in the environment would facilitate the learning of a language by making it more manageable. The child would not have to understand the thoughts of the speaker but would only rather have to know that the speaker's actions while talking provide important clues as to the correct referent.

The present study investigates the effects of deictic pointing actions in early word learning. Our research is different from previous studies in several important ways. First, the role of eye gaze in word learning has been systematically examined. However, there are naturally occurring environments where eye gaze cannot be useful for the infant trying to make associations. In situations such as story book narration, a child does not track eye gaze, as the child and parent generally face the same direction. Instead manual pointing is most often used to signal referential intent. In light of this, we studied how deictic pointing highlights aspects of the visual environment and therefore facilitates language learning. Second, previous studies on eye gaze were conducted in constrained learning environments. In the present study, there were multiple words and multiple pictures temporally co- occurring in a trial, this attempts to simulate ambiguous learning situations that young children deal with in their everyday life. Third, previous studies measured social cues at a macro- behavioral level. In contrast, the present research recorded parents' actions at the sensory level and systematically analysed the synchrony between their speech and social actions.

II. EXPERIMENT: WORD LEARNING

This experiment was designed to determine the degree to which children take advantage of social cues during word learning. To understand the use of social cues as a learning mechanism, multi-sensory parental input was analysed in this present experiment as they unfolded in real time in the naturalistic situation of story book narration.

Parents sat in front of a computer screen with their children on their laps. A story created by the investigators titled; 'Mr. Squirrel,' was read by the parents to their children. Six novel objects were introduced in the story and parents were asked to repeatedly label those objects. The six objects were, ring, canoe, yo-yo, rhino, pin, and desk. These objects were chosen through consideration of the MacArthur Communicative Development Inventory. Each object appeared three times in different positions on the page and with different objects each time. The written labels of the objects also appeared on the computer screen simultaneously with the objects. The written labels were displayed at the bottom of the screen and did not correspond spatially to the objects as they were positioned on the screen. Each object appeared a total of three times and each page consisted of three randomly selected objects. Thus, on a learning trial, there were three objects and three labels temporally co-occurring. Without any social cues, it would be unlikely that young learners would make the correct association between spoken words and objects. Pages of the story appeared one by one on the computer screen. Parents were instructed to press the spacebar to move from page to page so that they could read at a natural rate.

There were three conditions in which we varied the amount of deictic pointing. In Condition 1, parents pointed explicitly to objects during narration. A touch panel placed over the computer screen recorded when and where parents pointed as they labeled the objects. In Condition 2, parents were instructed not to point as they narrated. Therefore, there was no information about which word labeled which object. In Condition 3, parents pointed to an area on the screen that was equidistant from the three objects while labeling them. In this condition, pointing could not be used to tackle reference uncertainty but may have made learners more engaged in the interaction and therefore more engaged in word learning.

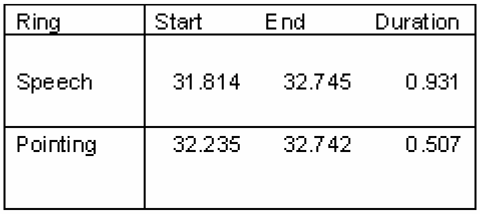
During the testing phase, parents asked their children to find (by responding in the most comfortable way) one of two objects displayed on the screen. There were no written words on the screen during the testing phase. Each of the six objects were tested for once. Responses were captured via a video camera located below the computer monitor and through the verbal responses of both the parents and their children.

A. Method

Participants. Children ranged in age from 18 to 30 months with an average age of 23 months, and came from the community surrounding Indiana University in Bloomington. They were recruited through letters sent in the mail.

Data recording. We placed a touch screen panel over the computer monitor. The touch panel recorded the time and location of each pointing action in milliseconds. Below the monitor we placed a camera (which was mostly hidden from the child's view with a piece of felt fabric) allowing us to record where each child was looking through the narration. The parents were asked to wear a microphone that recorded their speech stream at the frequency of 8000Hz. These recording devices allowed us to obtain information similar to that shown in table 1 for each time the parent labeled an object. Each parent labeled each of the six objects three times giving us 18 blocks of information similar to that shown in table 1 for each parent. Fig 1 is a picture of the set- up, in particular, showing the touch screen panel and the location of the camera.

Table 1. Information block, showing times for one labeling phase out of 18 for one parent.



Ring	Start	End	Duration
Speech	31.814	32.745	0.931
Pointing	32.235	32.742	0.507

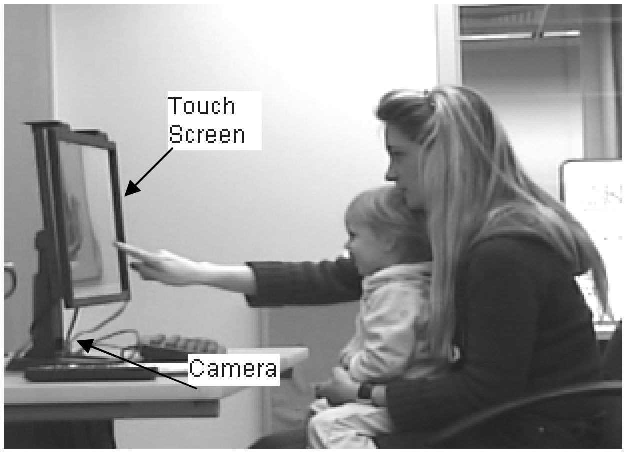


Fig. 1 A parent pointing to an object while labeling it.

B.Results

There were 12 children in all three conditions. Males and females were evenly distributed between the conditions. One testing trial for one child in Condition 1 was excluded because of parental input during the testing phase. Explicit parental pointing significantly increased the number of words that children learned. The average number of correct responses for children in Condition 1 (pointing) was (m= 5, SD= .85), for Condition 2 (no pointing) (m= 3.17, SD= 1.99), and for Condition 3 (non-directional pointing) (m= 3.33, SD= .985). There was a significant difference between the groups, with F (2, 33) = 6.529 p< .01. As shown in figure 2, a direct comparison between results in Condition 1 and Condition 2 indicated a significant difference between pointing and no pointing conditions, t=2.9 p=.01. We also found that there is no significant difference between Condition 2 and Condition 3, t=-.26 p=.797. Relevant research suggested that social cues might serve as a temporal "social gateway" to engage learners [8]. However, our results did not show this effect, if they had we would expect the mean scores of the point condition and the general point condition to be closer to each other. A plausible reason is that even if children were more engaged in learning in the general point condition compared to the no point condition, there was still a high degree of ambiguity in each trial so that they cannot discover correct word- to- object associations. Thus, to investigate the temporal role of social cues we may need to design a new condition in which word- to- object ambiguity is removed as a factor that may influence learning results.

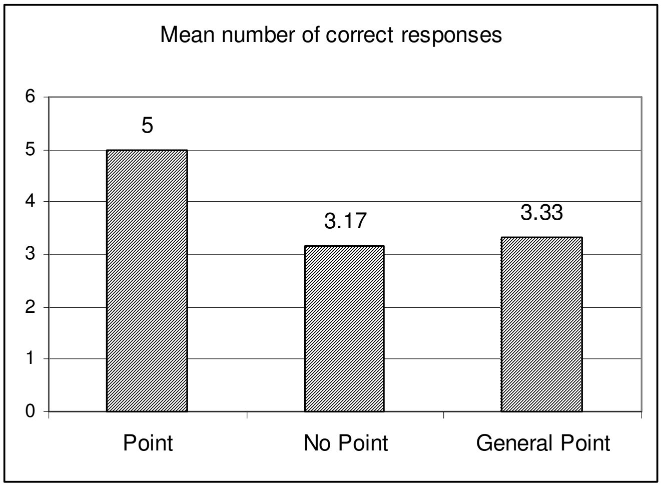


Fig. 2. The mean number of correct responses out of 6 trials.

III. DATA ANALYSIS AT THE BEHAVIORAL LEVEL

In addition to measuring the number of words learned for each child, we were also able to measure the synchrony between the speech stream and pointing actions of the parents. One of the aspects of social cues that we were interested in centers around the correlations that are present between the statistical regularities of the speech stream and those of the pointing action, and how this relates to words being learned. Our working hypothesis is that the more overlap (correlation) between the two streams of information would result in better word learning. Through multimodal data that we recorded we can then determine the degree of temporal synchrony that is necessary in order for word learning to be more effective. Figure 3 demonstrates how the speech stream and pointing action are recorded. From this information we can calculate the synchrony between these two streams.

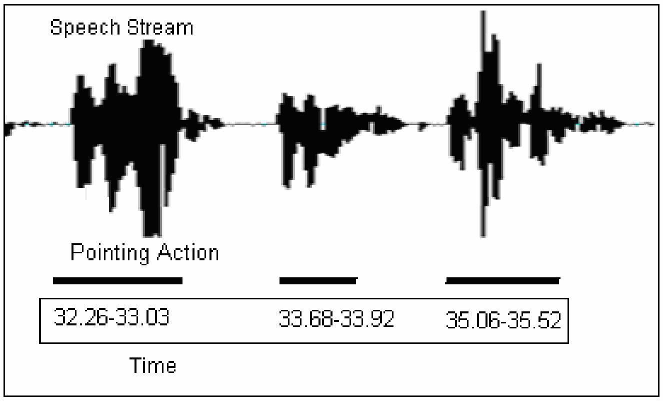


Fig. 3 The visualization of the synchrony between speech stream an pointing action.

A. Synchrony of the Onsets of Pointing and Speech Streams

The touch screen panel placed over the computer screen allowed us to measure the time of the pointing action to the millisecond. We were able to measure both when the action began, and also when it ended. We have also obtained equally precise audio information. Each parent labeled each of the 6 objects 3 times during the narration; this means that there were 18 labeling events for each parent. Out of the 12 participates we could only use 11 for touch screen labeling information because one of the parents failed to make contact with the screen when pointing. Therefore we obtained 198 labeling events. Furthermore, there were an additional 12 incidences where the parent failed to make contact with the screen for individual labeling events. Therefore, we were left with 186 events that we could analyze.

From the information obtained from both streams we could calculate the difference between the beginning of the pointing action and the beginning of the speech stream. This tells us which happened first, and the exact amount of time that lapsed before the second stream began. Since the time of the pointing action was subtracted from the time the speech stream began, a negative number indicated that the speech stream began before the pointing action. The results of this calculation are illustrated in figure 4.

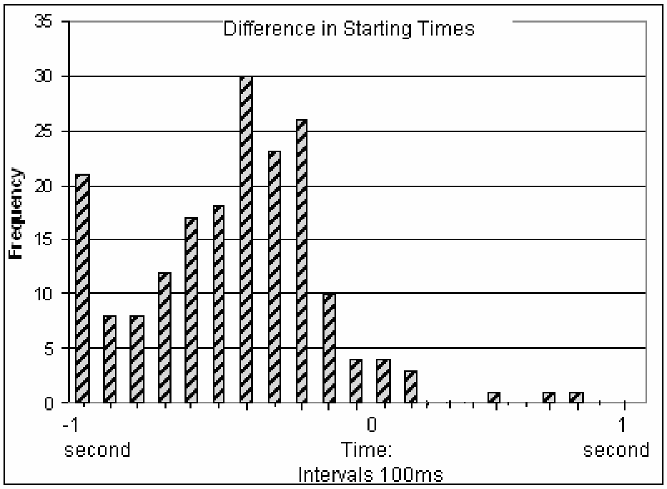


Fig. 4 The frequency of the difference in starting times between the speech stream and the pointing action.

As is evident from figure 4, the majority of the times are negative, signifying that the speech stream occurred first before the pointing action more times than the pointing action began. It also shows that there is no perfect synchrony between speech production and pointing action. Nonetheless, young children were able to utilize this kind of information to guide their word learning.

B. Overlap of Speech and Pointing Streams

The next metric we calculated was the amount of time that the speech stream and pointing action overlapped with each other. This reveals how long of a period the child received information from both sources. The more they overlap, the longer the child observes both events simultaneously and the easier the child can make the correct association. If the overlap only occurred for a very brief period of time, the child might miss the simultaneous actions. Figure 5 shows the number of times that specific periods of overlap occurred.

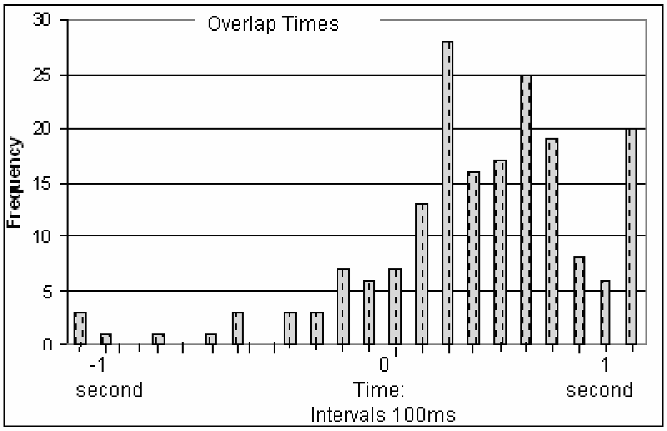


Fig. 5 The frequency of overlap interval times

The majority of overlap intervals were between 300 and 600 milliseconds long. A negative overlap time indicates a gap between the speech stream and the pointing action. This happened only 22 times out of 186 possible events. Table 2 shows the statistics 11 participants individually. We are particularly interested in individual standard deviation times. The results show that in general, overlapping times are centered around the means with relatively low standard deviations. We correlated standard deviations with the learning results of individual children. However, we did not detect a correlation between the overlapping time and learning results.

Table 2. Range of individual overlap between the speech and pointing action streams.

Subject#	Mean	Min	Max	Std
1	0.097	-1.870	1.969	0.917
2	0.578	0.010	1.150	0.304
3	0.285	-0.046	0.628	0.178
4	0.797	0.111	0.127	0.291
5	0.599	-0.288	0.907	0.272
6	0.556	0.021	1.708	0.417
7	0.956	0.171	1.942	0.419
8	0.235	-0.719	1.136	0.558
9	0.314	-0.451	0.744	0.333
10	0.298	-0.117	0.741	0.237
11	0.574	-0.416	1.145	0.413

We also note there was a significant difference in the overlap times when a comparison was done looking at the difference between the first object pointed at on a page and the second or third. As noted earlier, to move from page to page while reading the story the parents must push the spacebar on the computer keyboard. Therefore, when the parent goes to label the first object on the page their hand is down by the keyboard. However, when they then go on to label the second and third objects their hand is already up at the computer screen and therefore does not have to move as far. The mean overlap time for the first objects on the page was (m= .35, SD= .46), the average overlap for the second and the third combined was (m=.56, SD= .53), F (1,185) = 7.155 p< .01. This did not affect how well the words were learned. The only word that was not labeled first on the page was rhino, this was not the word least learned, canoe was although this was not a significant difference.

C. Stream Interaction

There are four different ways that the speech stream and pointing action can be related. The pointing action could take place within the speech stream; this is denoted as "Speech-Speech" in figure 6. The speech stream could begin the labeling phase and the pointing action could end after the speech stream; this is denoted as "Speech-Point" in figure 6. The pointing action could begin the labelling phase while the speech stream could end the phase; this is denoted as "Point-Speech" in figure 6. Finally, the speech stream could occur within the pointing action; this is denoted as "Point-Point" in figure 6.

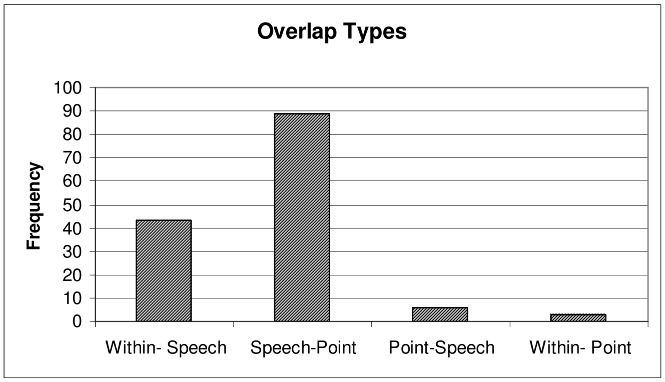


Fig. 6 The frequency of the ways that the pointing action and speech stream could be related.

The type of interaction that occurred most frequently was "Speech-Point". This occurred almost twice as many times as the combination of the other 3 different interactions.

IV. DISCUSSION

A. A Closer Look at Social Cues

Social cues range from conspicuous behaviors such as pointing, to less conspicuous behaviors such as eye gaze or slight shifts in body posture. Social cues could be thought of as packages of correlations that link regularities in language to regularities in the physical world of objects and actions. The physical embodiment that is a central aspect of social cues may be critical in connecting, in real time, the streams of information that are language, and the objects and events in the world that language is about. To fully determine the correlations among these co-occurring streams of information, they must be studied at the same time at a multi-sensory level instead of at the behavioral level as was the case in past studies.

B. Necessity of Examining Deictic Pointing

Past research regarding the use of social cues, specifically as it related to eye gaze, demonstrated that children are sensitive to the regularities of eye gaze. They disagree, however, as to the age at which children understand how those cues could help to solve the problem of reference uncertainty. Butterworth (1991) demonstrated that even by 6 months of age, infants display sensitivities to social cues, such as following eye gaze, however, not understanding the implications. Other research has shown that young children can not only follow the gaze successfully, but also use it in order to help them make the correct association between spoken utterance and object [7, 10]. In these studies, 13 month olds learned to link a novel label to an object by following the eye gaze of the speaker and then inferring the intent of the speaker. However, no such link was made when the speaker showed no signs of attention to the target object, even if the object appeared simultaneously with the label being spoken and the speaker was touching the object. Baldwin (1993) demonstrated that infants actively gathered social information to guide their inferences about word meanings and they checked the speaker's eye gaze to clarify the reference. Others have found evidence that suggests only infants 18 months or older understand the referentiality of seeing [11]. Caron, Butler, and Brooks (2002) used the same experimental design and obtained different results maintaining that infants as young as 14 months were influenced by eye gaze. Furthermore, there is additional evidence that suggests children as old as 3 years of age still do not completely understand the role of eye gaze. In these experiments, 3 year olds were shown to not understand that a line of vision must be straight in order for a person to see something [13].

These studies demonstrated that infants are sensitive to the actions of those speaking around them, but cannot determine the degree of the sensitivity. This is partially because they were not able to attain the proper measurements in order to definitively say what regularities are present in the daily lives of children. These studies only examined eye gaze behavior at the marco-level in a well- controlled laboratory condition; therefore, it is impossible to attain the kinds of information necessary in order to draw conclusions about statistical regularities in a more naturalistic learning environment. This present research demonstrated that through analyzing both what kinds of regularities are embedded in pointing actions and additionally how they relate to the speech stream we can attain detailed information about the degree to which children are sensitive to the information provided by their caregivers.

C. Future Conditions

Another condition will call for parents to only label the same three objects consistently through the story, instead of all six. This will be done in order to determine if only providing children with three labels will be supplying them with enough information (statistical regularities) to learn the remaining three objects. The same objects that were introduced in the first three conditions will be used in this condition. The testing phase of this condition will only present the child with either two objects that were both pointed at, or two that were both not pointed at. This will be done in an attempt to exclude the possibility of the child picking the correct answer simply because they have learned the name of one that was pointed to (i.e. knowing the answer through eliminated the incorrect answer).

The next condition will examine the role of embodied pointing. This experiment is designed to determine the degree to which the embodiment of the pointing is necessary in order for learning of the word. Instead of having the parent point throughout the story the object will be highlighted in another fashion. The parent will be instructed to label the object while it is being highlighted. Highlighting the object will draw the child's attention to it but since the parent is not pointing this highlighting cue cannot be considered embodied.

D. Deictic Pointing

Our results suggest that pointing, as a social cue, facilitates word learning. However, it is not enough for the pointing to simply draw attention to the page. In order for pointing to help children make correct associations it has to disambiguate between the objects presented on the page. There was not enough statistical regularity within this experiment for the children to be able to make the correct associations between object and spoken word label without the help of pointing. Smith (2000) and Yu, Ballard, and Aslin (2005) maintain that children are learning the correlations among actions, gestures and words of speakers in relation to their intended referents. A child would not have to understand what the speaker was thinking in order to make the correct pairing, but would only rather have to understand that the actions of speakers provide valuable information that facilitates finding the correct referent; our results come to a similar conclusion.

E. Conclusion

The present experiment attempted to study the effects of deictic pointing in order to obtain the statistical information that those studies focusing on eye gaze could not obtain. In addition to the statistical regularities we could also access word learning. The speech stream and pointing action do not have to be simultaneous in order for word learning to occur.

One of the major aspects that we were most interested in studying during this experiment was determining the kinds of statistical regularities that are present in the environments of children. It was important that we measured the regularities that the parents themselves provided and not those of a co-experimenter because we wanted to ascertain the regularities children were frequently provided with in naturalistic environments. On the one hand, story book narration is a representative activity through which young children learn the names of objects in stories. On the other hand, the experiment is well controlled in terms of the number of objects on each page and in total. Moreover, we recorded detailed behaviors of parents while they narrated.

This experiment focused on documenting what regularities were present in a natural learning environment, as a next step, we then moved on and examined to what degree children attend to those regularities by analyzing the degree of synchrony that was necessary in order for word learning to occur. We found a range of overlap times that children are sensitive to suggesting that young language learners are quite competent at determining referential intent in relatively complex situations similar to the one we presented.

References

[1] R. Aslin, J. Saffran, and E. Newport, "Computation of conditional probability statistics by 8- month- old infants," *Psychological Science*, vol. 9, no. 4, pp. 321- 324, July 1996.

[2] P. Kuhl, F. Tsao, and H. Liu, "Foreign- language experience in infancy: Effects of short- term exposure and social interaction on phonetic learning," *PNAS*, vol. 100, no. 15, pp. 9096- 9101, July 2003.

[3] W.V.O. Quine, *Word and Object*, Cambridge, MA, MIT Press, 1960.

[4] D. Richards and J. Goldfarb, "The episodic memory model of conceptual development: An integrative viewpoint," *Cognitive Development*, vol. 1, no. 3, pp. 183- 219, July 1986.

[5] P. Bloom, *How Children Learn the Meaning of Words*, Cambridge MA: The MIT Press, 2000.

[6] M. Tomasello, " Perceiving intentions and learning words in the second year of life," in *Language acquisition and conceptual development*, M. Bowerman and S. Levinson, Eds, pp. 111- 128. Cambridge University Press, 2000.

[7] D. Baldwin, E. Markman, B. Bill, R. Desjardins, J. Irwin, and G. Tidball, "Infant's reliance on a social criterion for establishing word- object relations," *Child Development*, vol. 67, no. 6, pp. 3135-3153, December 1996.

[8] D. White, A. King, A. Cole, and M. West, "Opening the social gateway, Early social and vocal sensitivities in brown- headed cowbirds (Molothrus ater)," *Ethology*, vol 108, no. 1, pp. 23- 37, January 2002.

[9] G. Butterworth, "The ontogeny and phylogeny of joint visual attention," in *Natural theories of mind: Evolution, development, and simulation of everyday mind reading*, A. Whitten Eds, pp. 223- 232, Oxford, England; Blackwell, 1991.

[10] D. Baldwin, "Early referential understanding: Infants' ability to recognize referential acts for what they are," *Developmental Psychology*, vol. 29, no. 5, pp. 832- 843, February 1993.

[11] C. Moore, and V. Corkum, "Infants gave following based on eye direction" *British Journal of Developmental Psychology*, vol. 16, no. 4, pp. 495- 503, November 1998.

[12] A. Caron, S. Butler, and R. Brooks. "Gaze following at 12 and 14 months: Do the eyes matter?" *British Journal of Developmental Psychology*, vol. 20, no. 2, pp. 225- 240, June 2002.

[13] J. Flavell, F. Green, C. Herrera, E. Flavell. "Young children's knowledge about visual perception: Lines of sight must be straight" *British Journal of Development Psychology. Special Issue: Perspectives on the child's theory of mind: I.* vol 9. no. 1. pp. 73-87, March 1991.

[14] L. Smith, " How to learn words: An associative crane," in *Breaking the word learning barrier*, R. Golinkoff and K. Hirsh- Pasek, Eds, pp. 51-80. Oxford: Oxford University Press, 2000.

[15] C. Yu, D. Ballard, R. Aslin, "The role of embodied intention in early lexical acquisition," *Cognitive Science,* vol. 29, no. 6, pp. 961-1005, January 2005.