Examining Real-time Attention Dynamics in Parent-infant Picture Book Reading

Yayun Zhang
yayunzhang@utexas.edu
Chen Yu
chen.yu@austin.utexas.edu

Department of Psychology
The University of Texas at Austin, USA

Abstract

Picture book reading is a common word-learning context from which parents repeatedly name objects to their child and it has been found to facilitate early word learning. To learn the correct word-object mappings in book-reading context, infants need to be able to link what they see with what they hear. However, given multiple objects on every book page, it is not clear how infants direct their attention to objects named by parents. The aim of the current study is to examine how infants mechanistically discover the correct word-object mappings during book-reading in real time. We used head-mounted eye-tracking during parent–infant picture book reading and measured infant's moment-by-moment visual attention to the named referent. We also examined how gesture cues may influence infants' attention at naming moments. We found that although parents provide many object labels during book reading, infants were not able to attend to the named object easily. However their abilities to follow and use gestures to direct the other social partner's attention increase the chance of looking at the named target during parent naming.

Keywords: picture book reading, word learning, visual attention, gesture

Introduction

Shared storybook reading is a naturalistic context in which parents read an illustrated book to their child (see e.g., Levy et al., 2006). It is one of the most common daily activities for young children and it has been found to have many long-term benefits including parent-child bonding (Barratt-Pugh and Rohl, 2015), reading and literacy skills (Sulzby and Teale, 1987), academic achievement (Sénéchal et al., 1998) and learning to sustain attention (Lawson, 2012). According to a large-scale survey study, many parents begin to read to their children shortly after birth and about 95% of parents of children ages 18 to 23 months report reading books to their infants at least once or twice a week and 50% of them reported reading books at least once a day (Young et al., 1998).

With overwhelming evidence supporting early shared book reading as a critical training ground for language learners, it is important to understand how learning happens during picture book reading. To learn the correct word-object mappings from picture books, young children need to be able to link what they see with what they hear. However, children's books usually portray complex scenes with multiple objects on each page. When parents label an object on a book page, it is not clear how children direct their attention to the named object given there are multiple potentially correct referents on a book page. Evans and Saint-Aubin (2013) investigated how eye movements in shared storybook reading are related to the time-locked spoken language input. They found that 4-year-olds did look at the target region of the illustration after the critical word was spoken by the reader. However, it took them 4–5 seconds on average to do so. Given that adults read aloud at a rate of almost 200 words per minute (about three words per second; Ashby, Yang, Evans, and Rayner, 2012), children's eye movements may be too slow to keep up efficiently with the reader's spoken language output. Children are facing a real-time challenge of mapping the heard label with the right object in view during book-reading interactions.

Despite this word learning challenge, shared storybook reading is a dynamic interaction involving more than just the audio-visual input. There are multiple factors simultaneously influencing what the child is seeing and hearing at the moment of naming. For example, parents often also offer concurrent cues while naming an object to engage the child and help reduce referential uncertainty. Previous studies have explored the many possible pathways through the use of cues from social (Baldwin, 1993; Bloom, 2002; Tomasello, 2000), linguistic (Gleitman, 1990), attentional (Smith, 2000), and conceptual (Gentner, 1982) constraints. One of such cues that has been studied extensively and found to support word learning is hand gesture. Deictic gestures, such as pointing, can highlight the correct referent ostensively and offer crucial clues for infants to locate the intended referent when facing referential uncertainty (Rowe et al., 2008). Because deictic gestures provide an easier pathway for infants to identify and integrate the audio-visual information (Cook et al., 2008), they have been found to facilitate language comprehension (Morford and Goldin-Meadow, 1992).

Despite a large amount of observational studies on book reading, limited experimental work has been done on examining real-time visual attention during shared book reading with very young children (less than two years of age). The goal of the current study was to provide a mechanistic account of how correct word-object mappings may be established in the context of book reading by quantifying word learning input provided by parents. Towards this goal, we recruited 18-to-24 month-old children and their parents and fitted them with head-mounted eye trackers to capture both the children's and the parents' first-person views while parents read several storybooks to their children for 15 minutes. Using linguistic input from the parent and sensory-motor level behaviors of eyes and hands from both the parent and the child, we examined how likely the child was attending to the named object when

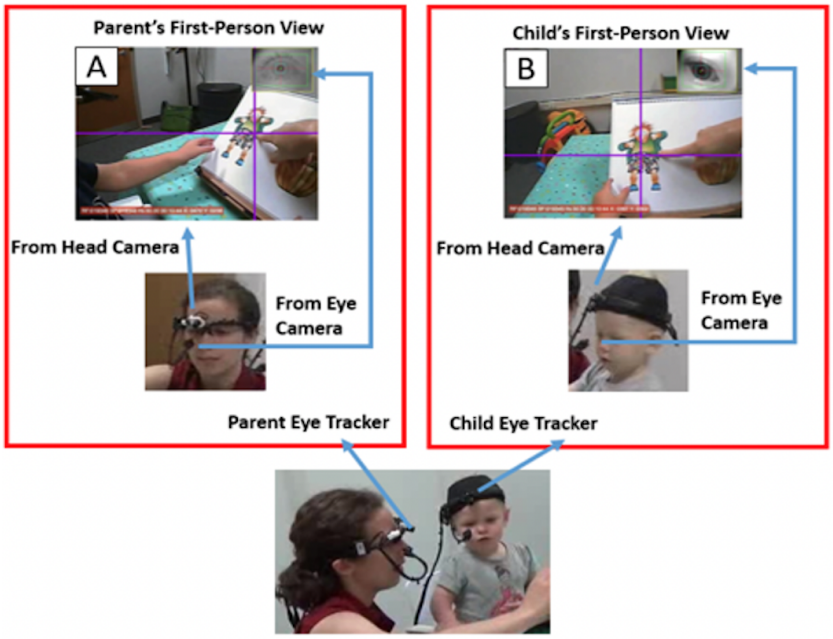

Figure 1: Experimental setup. Child's (A) and parent's (B) first-person views with eye images superimposed on the upper-right corner. Crosshair indicates where each agent looks.

naming occurred. Specifically, we focused on three sets of analyses: 1) how frequently parents provide object names; 2) where do infants visually attend when hearing object names; 3) how gesture cues impact infants' attention when hearing object names.

We hypothesized that 1)parents produce many naming instances during book reading, providing a lot of word learning opportunities. However, 2)infants may not be able to attend to the named object easily given there are many potentially correct referent on every book page. 3) Their abilities to follow parent's gesture and use gestures to direct parent's attention could increase the chance that infants look at the right target during parent naming.

Method

Participant

Participants were 16 parent-child dyads who resided in Midwest, U.S.A. All children (12 female) were between the ages of 18 and 24 months (M = 19.03, SD = 1.6, Min = 18, Max = 24.4). Twelve additional dyads participated in the book reading but contributed no or limited eye-tracking data due to children' unwillingness to wear the head camera equipment.Procedures in this study were approved by the Human Subjects and Institutional Review Boards at xx University.

Materials

We used 5 commercially available storybooks: I Went Walking (1989), Goodnight, Gorilla (1994), Let's go visiting (1997), Sammy the Seal (2005), I am a Little Lion (1994). The selected books vary in their story contents and illustration styles, but all have clear story-lines centering around one main character. Because these books are intended for beginner-level readers, some have very few lines of text and some have no text. To be consistent, we removed written texts from all books and asked parents to come up with their own stories based on the printed images. This manipulation would elicit more diverse linguistic input that allows us to examine spontaneous interactions between parent and child and potentially compare individual differences in the future.

Experimental setup

During the experiment, child and parent sat next to each other at a table (61cm x 91cm x 64cm). Infants sat in a customized highchair that supported sitting stability and parents sat on the floor. A bookstand is used to hold the book at a consistent 60° angle and roughly 10cm away from the edge of the table. This setup allows parents to freely interact with their children while avoiding displacement of the eye-tracking devices due to voluntary head-movements. Both participants wore head-mounted eye trackers from Positive Science, LLC (Franchak et al., 2010; Yu and Smith, 2013). As shown in Figure 1A and 1B, each eye-tracking system includes an infrared eye camera and a scene camera. The eye camera is mounted on the head and pointed to the right eye of the participant that records eye images, and the scene camera captures the first-person view from the participant's perspective. The scene camera's captures a 90° visual field. Although less than the approximately 170° full visual field of natural human vision, it captures area on the book-page space that is critical to determine gaze location. Each eye tracking system records both the egocentric-view video and gaze direction in that view, with a sampling rate of 30 Hz. Parent speech was recorded from a microphone

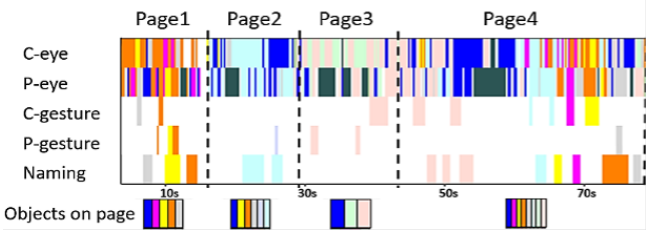

Figure 2: Data visualization for 5 coded variables: child eye gaze, parent eye gaze, child gesture, parent gesture, parent naming.

built in the parent eye tracker. We also added three additional high-resolution cameras on two walls and the ceiling to capture the interaction from three third-person views.

Procedure

We first fit parent with the eye-tracking gear. After both the parent's and the child's eye-tracking gears are placed properly, we collect calibration points for eye tracking. We place a letter-sized sheet with 5 points (4 at corners and 1 at center) on the bookstand. The experimenter randomly points to one of the five points using a laser pointer and makes sure both the parent and the child's attention is directed to that point. This procedure is repeated at least 15 times with the calibration points placed in various locations on the sheet. Parents are then instructed to read books to their children as they naturally would for 15 minutes. They do not need to follow any order or finish reading all the books. They are told to put the book on the bookstand when reading it and to keep the original sitting configuration as much as possible. They are not aware that the study is about word learning nor are they instructed to name the objects.

Corpus In total, we collected 45 book-reading sessions with good eye-tracking data from 16 parent-child dyads. This equals to 157 minutes of usable video data. On average, each book was read 9 times. Each dyad contributed 2-5 books. On average, parents spent about 3.49 minutes on each book, with the shortest single-book interaction lasted 1.19 minutes and the longest one lasted 9.38 minutes. To process data for analyses, we synchronized and calibrated first-person view videos from both parent and child. Using calibrated videos with crosshairs superimposed on the videos indicating gaze directions, we manually annotated five variables listed below: child and parent gaze, child and parent gestures and parent speech using the following coding scheme.

Data processing

Gaze data. We first identified a list of region-of-interest (ROIs) for each book. All ROIs are whole objects on the page that can be named using concrete nouns. The number of objects varies page by page. The average number of objects on a page is 5.45 (SD = 2.83, Min = 2, Max = 15). This shows that book reading creates word learning moments that are referentially uncertain as there are always multiple objects on a page when naming happens.

Coders watched the calibrated first-person view videos and coded these ROIs frame by frames by using an in-house program. Together, the whole 2.5 hours interactions yield 572,190 frames extracted from both social partners. Within these interactions, roughly 31% of frames from the infants and 30% of frames from the parents were not codable either due to loss of tracking or participants being off task (not looking at the book at all). Gesture data. Deictic gestures from both social partners were coded. Deictic gesture is used for referent identification. In the context of book reading, the most common deictic gestures used is pointing at a referent using ones' hands or fingers. Using videos from all views, coders identify segments of the video in which parent or child points at objects. The duration of each gesture covers the time period in which the object that parent or child intends to point is clearly identifiable from the any of the videos.

Speech data. Coders transcribed parent' speech using only the audio recordings from the interactions. Parental speech was then divided into utterances, which is defined as strings of speech between two periods of silence lasting at least 400msec. Among those spoken utterances, ones that contain at least one labeling of an object printed on the page (e.g., "What is the duck doing over there?") were then coded as "naming utterances." As shown in a data visualization in Figure 2, all coded variables are represented as n categorical (n = number of Region-Of-Interest defined) temporal data streams with different colors indicating different ROIs in each moment. Data analyses were carried out using these five data variables.

Results

Quantifying linguistic input

Parents produced 2690 speech utterances in total. Among these speech utterances, about 50% are naming utterances (1360), which is defined as an utterance containing at least one object label. Infants on average hear 17.54 utterances per minute (SD = 3.25) and 8.95 naming utterances per minute (SD = 2.26). This finding suggests that book reading is a very fast-paced interaction in which parents provide a lot of labels.

In addition, we observed large individual differences across different dyads. As shown in Figure 3, some parents (highlighted blue dot) have a high speech rate, but not many speech utterances are naming utterances, whereas other parents (highlighted black dot) have a relatively low speech rate, but almost all speech utterances are naming utterances. On average, parents mentioned about 13 unique object names (M = 12.73, SD = 5.08), which is about 56% (SD = 13%) of all unique objects printed on the book.

Because picture books are designed in a way that the same object appears many times across different pages, we next measured how often parents name the same object and whether naming frequency of an object is associated with its occurrences on the book. We found that parents labeled some

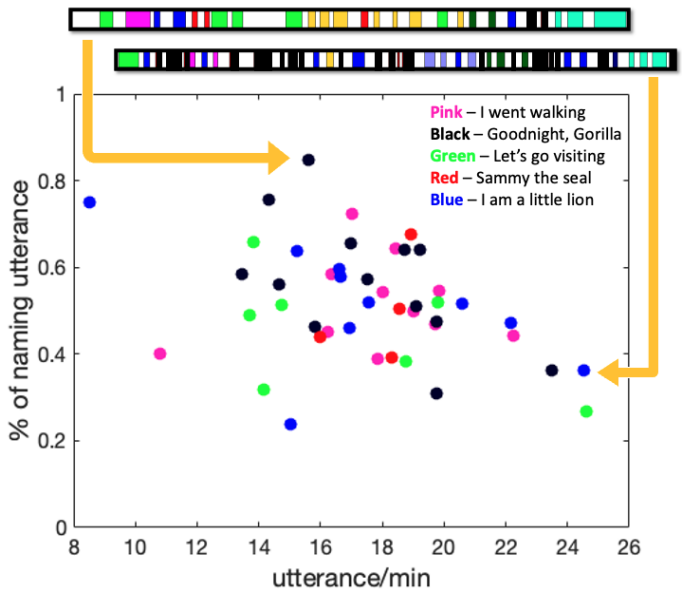

Figure 3: A scatter plot showing linguistic input from different dyads. Two color-coded speech streams from different parents are plotted on the top of the figure. Different colors indicate different objects being labeled, black indicate non-naming utterances.

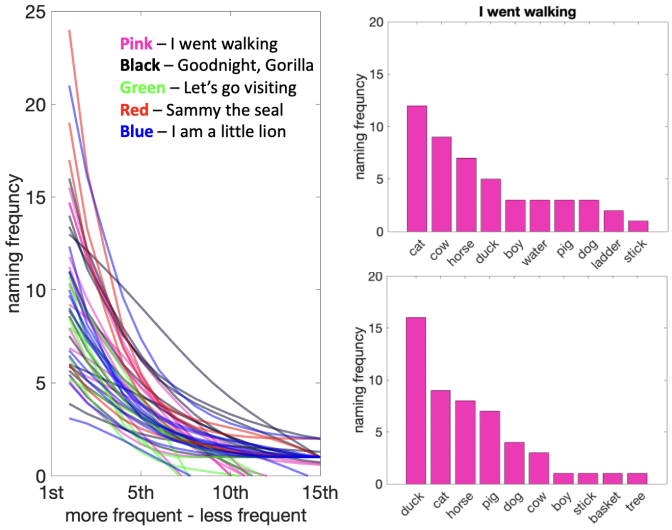

Figure 4: Left: Frequency distribution from 45 book sessions; Right: histograms showing naming frequency from two book sessions.

objects more than others (Figure 4, left panel). More than half (56%) of the naming instances are about the top 3 named objects. In addition, we found that the set of most named objects tend to be different for different dyads. For example, as shown in Figure 4 (right panel), duck was the most named object for one dyad and was named 16 times, but it was only the fourth named object for another dyad and was named 5 times. This seems to suggest that parents create their own linguistic input that is not entirely tied to the printed pictures.

To quantify this observation, we measured how parent naming frequency is correlated with object occurrence in the book. In a hypothetical situation, if parent names every object printed on the book, we would see a perfect correlation and no individual differences between dyads. However, we only found a moderate correlation (r= 0.55, p< 0.001) between naming frequency and object occurrence (Figure 5). This suggests that parents do follow the general story line to some degree but they are certainly not labeling everything printed on each book page. Combining both results, we could argue that parents tend to name a small subset of objects very frequently. However, the subsets of objects they chose vary across dyads and are not completely tied to object occurrences printed on the pages.

Visual attention during naming

Given book reading is a linguistically rich environment in which children have many opportunities to learn object names, where do infants visually attend when parents provide object names? We analyzed the child's real-time gaze patterns during naming moments. We defined a window starting from the onset of each utterance and measured where the child looked moment by moment within the entire window. For multi-label utterances, such as 'duck, look at the duck!" (consist of 40% of all naming utterances), we equally split the utterance into n (n = number of labels) smaller labeling windows.

We plotted target look distribution in a normalized histogram (Figure 6) where x axis is proportion of time the child is looking at the named object, y axis is proportion of instances. A hundred percent on x axis means that when a naming event happens (i.e., mom is naming the object gorilla), the child is looking at gorilla 100% of the time within the naming window. Zero percent means when a naming event happens, the child is not looking at the correct target at all. Proportions between 0% and 100% mean the child at least spent some time looking at the target. We found that in over 50% of instances, infants completely missed the named object. The rest of the time, infants spent at least some time looking at the target. Only in about 15% of instances, infants attended to the named target 100% of time (Figure 6).

This pattern reflects different types of learning situations infants encounter in naturalistic storybook reading. Some naming moments are highly informative, from which children are able to find the correct word-object mapping very easily. Other naming moments are highly ambiguous that labeled object and attended object do not match. In these instances, the child is uncertain which object parent is naming, creating a word-object mapping challenge.

The effect of gesture on attention during naming

Both parents and children gesture often during book reading. We found that parents (M = 9.52 times/min, SD = 5.51 times/min) gestured significantly more than children (M=

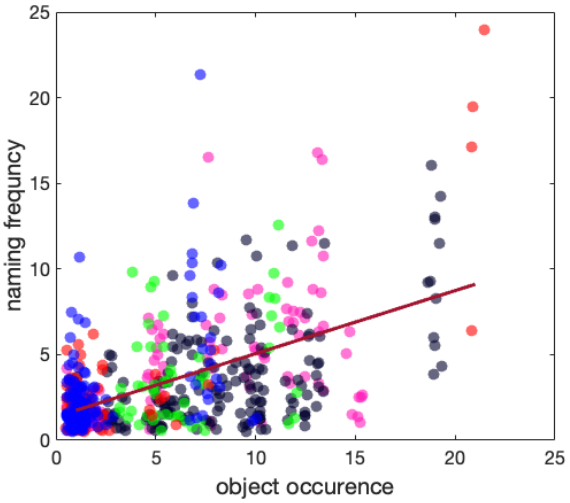

Figure 5: Moderate correlation between naming frequency and object occurrence.

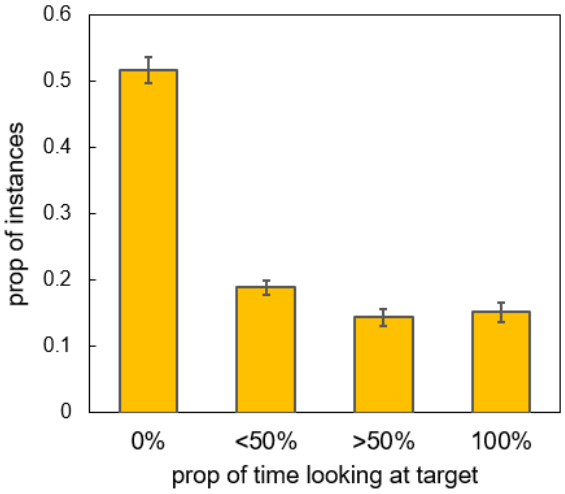

Figure 6: Target look distribution.

3.49 times/min, SD = 3.35 times/min, $t(44) = 5.63$, $p < .0001$). However, the duration of parent's (M =1.02 sec, SD = .52 sec) and child's gestures (M = 1.24 sec, SD = .77 sec) do not differ ($t(38) = 1.88$, ns).

Labeling and gesture are not only highly frequent events in book-reading context, they also tend to be coupled temporally. To quantify the coupling of these two types of events, we coded naming event as "naming with gesture" as long as there is one co-occurring gesture event. We found that 46% of naming instances are paired with a parent gesture and 18% of naming instance are paired with a child gestures (5% paired with both types of gestures), suggesting that gestures and labels are highly coupled. Knowing that the overall visual attention on target during naming is quite low, are infants more likely to attend to the target when gestures are also present?

We found that in naming instances with parent gestures

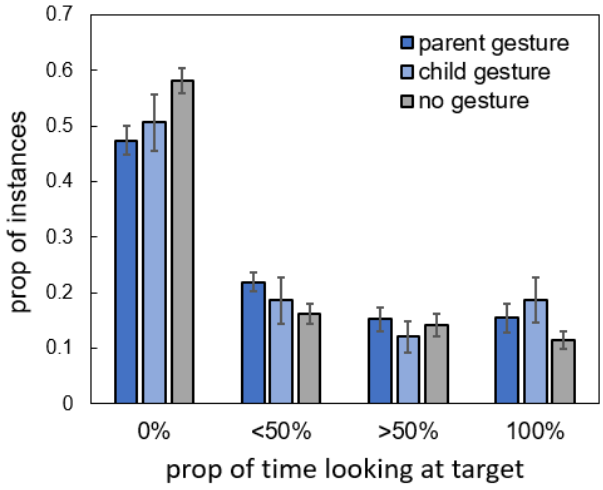

Figure 7: The effect of gesture on attention during naming

(Figure 7), there were fewer instances that the child never looked at the target. There was also an increase of target look compare to no gesture cases. We fit linear mixed effect models predicting proportion of target looking time from different types of naming instances with subject as the random factor. We first compared naming with parent gesture versus naming with no gesture and found the model to be statistically significant (model: target look $\sim$gesture + (1 | subject), $\beta = 0.04$, $p = 0.04$). We observed very similar patterns in naming instances with child gestures (Figure 7). We run a similar model as in the parent data set and found the model statistically significant: ($\beta = 0.12$, $p < .001$) for naming with child gesture versus naming with no gesture. In sum, we found that both parent and child gestures are effective in driving children's attention to the named target. Although children still may not solve the referential uncertainty problem at the moment, they are at least more likely to collect relevant information about the correct word-object mappings through increased visual attention.

Discussion

Understanding the learning input available to the child during word learning is critical for studying any learning mechanisms. In the current study, we focused on storybook reading interaction and found that parents provided a lot of different word labels in a short period of time. The fast-paced nature of book reading creates a big challenge for children who are acquiring new words through linking what they hear with what they see. By analyzing gaze data from their own perspective, we found that children are unsure which object is the correct referent been labeled or they may even mistakenly treat a wrong object as the correct referent. However, we found some evidence suggesting that deictic gestures from either parent or child help resolve the referential ambiguity problem by increasing the child's visual attention to target.

The current study focused on examining individual nam-

ing instances, but a lot of storybooks are designed in a way that objects repeated appear across pages. To make a coherent story, parent tend to repeatedly name the same object on and across pages, creating multiple opportunities for the child to learn words. The way storybook is structured is similar to a Cross-Situation Learning (CSL) paradigm used in many word learning studies. The logic of CSL is that when learners hear a word, they always see a set of potential candidate referents. Although learners are unable to identify the correct word-object mapping on a single exposure, but if they can combine information across multiple exposures, they are able to determine the most probable referent by integrating multiple mapping sets over time. In other words, hearing words in enough various contexts would allow learners to rule out incorrect associates and learn the most consistent mappings, which are likely be the correct ones.

In addition, deictic gesture is certainly not the only cue provided by parent during the entire interaction. Future work could also look at other gesture types, such as representational gestures. Those gestures are not only directive but also contain more complex information about a referent's size, shape, function, etc. (McNeill, 1992), which may offer additional clues to help infants identify the correct referent.

Together, we believe that in order to understand children's word-learning process, we need to first understand the learning input available to them during everyday word-learning moments and this critical learning input is jointly created by parents and children at the moment of learning (Cartmill et al., 2013; Hoff and Naigles, 2002; Weisleder and Fernald, 2013). It may be through multiple statistically sensitive processes of the input that learners gradually acquire the critical skills to solve the mapping problem in word learning.

Acknowledgments

This work is supported by R01HD093792.

References

Ashby, J., Yang, J., Evans, K. H., & Rayner, K. (2012). Eye movements and the perceptual span in silent and oral reading. *Attention, Perception, & Psychophysics*, *74*(4), 634–640.

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

Barratt-Pugh, C., & Rohl, M. (2015). 'better beginnings has made me make reading part of our everyday routine': Mothers' perceptions of a family literacy program over four years. *Australasian Journal of Early Childhood*, *40*(4), 4–12.

Bloom, P. (2002). Mindreading, communication and the learning of names for things. *Mind & Language*, *17*(1-2), 37–54.

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, *110*(28), 11278–11283.

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition*, *106*(2), 1047–1058.

Evans, M. A., & Saint-Aubin, J. (2013). Vocabulary acquisition without adult explanations in repeated shared book reading: An eye movement study. *Journal of Educational Psychology*, *105*(3), 596.

Franchak, J. M., Kretch, K. S., Soska, K. C., Babcock, J. S., & Adolph, K. E. (2010). Head-mounted eye-tracking of infants' natural interactions: A new method. *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, 21–27.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Center for the Study of Reading Technical Report; no. 257*.

Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition*, *1*(1), 3–55.

Hoff, E., & Naigles, L. (2002). How children use input to acquire a lexicon. *Child development*, *73*(2), 418–433.

Lawson, K. (2012). The real power of parental reading aloud: Exploring the affective and attentional dimensions. *Australian Journal of Education*, *56*(3), 257–272.

Levy, B. A., Gong, Z., Hessels, S., Evans, M. A., & Jared, D. (2006). " understanding print: Early reading development and the contributions of home literacy experiences"[journal of experimental child psychology 93 (2006) 63-93]: Erratum.

Morford, M., & Goldin-Meadow, S. (1992). Comprehension and production of gesture in combination with speech in one-word speakers. *Journal of child language*, *19*(3), 559–580.

Rowe, M. L., Özçalışkan, Ş., & Goldin-Meadow, S. (2008). Learning words by hand: Gesture's role in predicting vocabulary development. *First language*, *28*(2), 182–199.

Sénéchal, M., Lefevre, J.-A., Thomas, E. M., & Daley, K. E. (1998). Differential effects of home literacy experiences on the development of oral and written language. *Reading research quarterly*, *33*(1), 96–116.

Sulzby, E., & Teale, W. H. (1987). Young children's storybook reading: Longitudinal study of parent-child interaction and children's independent functioning. final report.

Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics*, *10*(4), 401–413.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological science*, *24*(11), 2143–2152.

Young, K. T., Davis, K., Schoen, C., & Parker, S. (1998). Listening to parents: A national survey of parents with

young children. *Archives of Pediatrics & Adolescent Medicine*, *152*(3), 255–262.

Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PloS one*, *8*(11), e79659659.