Developmental changes in children's visual access to faces during early word learning

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Kaia Simmons
kaias@stanford.edu
Program in Human Biology
Stanford University

Daniel Yurovsky
dyurovsky@stanford.edu
Department of Psychology
Stanford University

Guido Pusiol
pusiol@stanford.edu
Department of Psychology
Stanford University

Abstract

The faces of other people are a critical information source for young word learners. Yet during the period of early word learning, children are also undergoing significant postural and locomotor development, changing from lying and sitting infants to toddlers. We used a head-mounted camera in conjunction with a face-detection system to explore the effects of these changes on children's visual access to their caregivers' faces during an in-lab play session. In a cross-sectional sample of 4–20 month old children playing with their caregivers, we found substantial changes in face accessibility based on age and posture. These changes translate into changes in the accessibility of social information during word learning.

Keywords: Social development; face processing; word learning; head-camera

Introduction

A father offers his young daughter two toys. One—a ball—is familiar, while the other—a bright yellow feather duster—is not. After she accepts the toys, he remarks, "Isnt the zem funny?" To learn the word "zem," the child must determine his mothers intended referent from a wide range of possible targets, including the two salient objects. Many factors can be useful in making this inference, including the contrast with a known object (Markman & Wachtel, 1988; Clark, 1988). But in many cases, the simplest solution may be for the child to look to her father for a sign of what he is talking about (Baldwin, 1991; Vaish, Demir, & Baldwin, 2011).

The ability to follow social signals like eye-gaze is a strong predictor of children's early vocabulary growth. For example, Brooks and Meltzoff (2005) found that children who followed an experimenter's gaze better before their first birthday had larger vocabularies at 18 months. Similarly, Carpenter, Nagell, and Tomasello (1998) found that children's level of joint engagement (as well as the degree to which mothers followed the child's focus of attention in their labeling) predicted vocabulary growth in both language production and comprehension. These studies suggest that children's social environment plays a powerful supportive role in language learning.

At the same time as children are beginning to learn their first words, their view of the world is changing radically (Adolph & Berger, 2007). As speechless infants, they are unable to locomote independently and are moved from place to place and set in particular postures by their caregivers. Before their first birthday, they begin crawling; soon after, they begin to walk independently. These changes may have a profound effect on what children see.

A recent study suggests the possibility of links between motor milestones, social cognition, and language. Walle and Campos (under review) noted robust correlations between children's ability to walk and their vocabulary, both receptive and productive. On the basis of an observational study of parent input, they speculated that the emergence of walking may change the ability of the child to use and appreciate a number of social cues. One possible explanation for this pattern is that children's posture may mediate their access to social cues, and seeing more social information may in turn allow children to discover word meanings more effectively.

Recent methodological developments provide data that allow this hypothesis to be tested. The rise of head-mounted cameras and eye-trackers allow for measurement of children's naturalistic environment in a way that was not previously possible. Yoshida and Smith (2008) gave the first demonstration of the radical differences between toddler and adult perspectives on the social world, with toddlers' visual field being dominated by hands and objects much more than adults'. More recent work has used head-mounted eye-tracking methods to measure young toddlers' fixations (Franchak, Kretch, Soska, & Adolph, 2011), also finding that children look relatively infrequently at their mothers' faces in naturalistic play.

These methods are now being applied to understand inputs to language acquisition. Work by Yu, Smith, and colleagues has suggested that when parents or children create moments in which the visual field is dominated by a single object and that object is named, word learning and retention is facilitated for toddlers (Smith, Yu, & Pereira, 2011; Yu & Smith, in press). Some data even suggest that young children's restricted viewpoint may be more effective for learning words than the comparable adult perspective(?, ?). Together, this broader body of evidence suggests that understanding infants' perspective during language use—and how it interacts with their development—is a critical part of understanding language learning.

In the current study we took a developmental approach to understanding the relationship between perspective and language input. We recorded head-camera data from a group of infants and children across a broad development range (4 – 20 months) as they played with their caregivers during a brief laboratory visit. We then hand annotated these data for both the child's posture and the naming instances provided by the parents and used face-detection algorithms to measure the frequency and size of faces in the child's visual field. The resulting dataset allows us to analyze both changes in access to faces according to age and posture and whether access to faces is related to naming of objects during play.

Figure 1: Our light-weight, low-cost head-mounted camera with fisheye lens.

Methods

Participants

Participants were 20 infants and children (N=4 each at 4, 8, 12, 16, and 20 months, 9 females total), recruited from the surrounding community via state birth records. All participants had no documented disabilities and were reported to hear at least 80% English at home. Overall, XYZ children visited the lab, with XYZ contributing significant data. Of these, XYZ% wore the camera successfully (with success rates at different ages varying from 100% at 4 months to approximately XYZ% at 20 months). Our current sample of participants were selected randomly to comprise a an age-balanced sample from the total group of participants in the ongoing study.

Head-mounted camera

Our head-mounted camera ("headcam") is composed of a small, inexpensive MD80 camera attached to a soft elastic headband from a camping headlamp. An aftermarket fisheye lens intended for iPhones and other Apple devices is attached to increase view angle. The total cost of each camera is approximately $60. The camera captures 720x480 pixel images at approx. 25 frames per second, and has battery life of 60–90 minutes. With the fisheye lens affixed, it has a viewing angle of approx. 60 degrees of visual angle in the vertical axis and approx. XYZ degrees in the horizontal. The device is pictured in Figure 1.

The vertical field of view of the camera was smaller than the child's approximate vertical field of view, which—even at 6-7 months—spans around 100–120 degrees (Mayer, Fulton, & Cummings, 1988; Cummings, Van Hof-Van Duin, Mayer, Hansen, & Fulton, 1988). We were therefore faced with a choice in the orientation of the camera. If we chose a lower or higher orientation, we would be at risk of truncating either the child's own hands and physically proximate objects, or the faces of the adults around the child. Yet if we chose the middle orientation, we would still be at risk of underestimating the proportion of faces viewed by the child. Thus, for the purposes of the current study—measuring visual access to faces—we chose to orient the camera towards the upper part of the visual field.[1] While this orientation decreased our chances of recording the objects being manipulated by the child, it nevertheless allowed us to capture the majority of the faces in the child's visual field.

Procedure

After coming to the lab, families were seated in our waiting room where they signed consent documents and where children were fitted with the headcam. After a short period of play, they were escorted to a playroom in the lab where the free-play session (the focus of the current study) was conducted.

In the waiting room, the experimenter placed the headcam on children's heads after they had time to adjust to the environment. For children who resisted wearing the headcam, the experimenter used "distractor" techniques (presenting stimulating toys or engaging the children in hand-occupying activities) intended to keep children's focus elsewhere and prevent them from taking off the camera (Yoshida & Smith, 2008). Once the child was wearing the camera comfortably for a period of time, child and caregiver or caregivers (in two cases, there were two adults present) were escorted to the playroom.

In the playroom, the experimenter presented the child's parent with a box containing three labeled pairs of objects, each consisting of a familiar and a novel object: a ball and a feather duster ("zem"), a toy car and an XYZ (blah) and an XYZ and a (blah). Parents confirmed that the child had not previously seen the novel toys. Parents were instructed to play with the object pairs with the child, one at a time, "as they typically would" and to use the novel labels to refer to the three toys. After giving these instructions, the experimenter left the room for a period of approximately 15 minutes. During this time, a tripod-mounted camera recorded video from a corner of the room and the headcam also captured video from the child's perspective.

Data Processing and Annotation

All headcam videos were cropped to exclude the period of entry to the playroom and any points at which the camera needed to be adjusted and were automatically synchronized with the tripod-mounted videos using FinalCut Pro Software. The final sample was approx. 5 hours of headcam video (M=12m, range: 2–12m), for a total of roughly 400,000 frames.

Posture and Orientation Annotation One major goal of our study was to understand the relationship between children's posture and their access to information from the faces of their caregiver. To investigate this relationship, we created a set of hand annotations for the child's physical posture (e.g. standing, sitting) and orientation relative to the caregiver (e.g. in front of, behind, close, far away). For each headcam video,

[1] Previous studies have shown that children's head movements in the horizontal dimension are approximated by (though are slightly lagged by) their head movements (Yoshida & Smith, 2008). Our own experience with the current apparatus ratifies these conclusions for the horizontal field but suggests that head movements in the vertical field are less reliable.

Figure 2: Sample frames from the headcam videos for a child from each age group, selected because they featured successful face detections (green squares).

a coder used OpenSHAPA software to annotate both orientation and posture (Adolph, Gilmore, Freeman, Sanderson, & Millman, 2012).

Orientation was initially categorized as being in front of the caregiver, to the side, or behind, and close (defined informally as within arm's reach) or farther away. Because of data sparsity, we consolidated this scheme into three categories: close to the caregiver either in front or on the side, farther from the caregiver either in front or to the side, and a global category of behind the caregiver. Posture was categorized as being held/carried, lying down, sitting, crawling, standing, or other. Data from when the child was out of view of the tripod camera was marked as uncodable and excluded from these annotations. A second coder coded XYZ videos; their categorizations were reliable at XYZ.

Labeling Annotation We were also interested in the availability of social information proximate to naming events in the caregivers' speech to children. Accordingly, a human coder also marked the instance when the name of any of the six objects in the object set was used. Overall, caregivers produced a median of 35 labels in a highly skewed distribution across participants (range: 9 – 131), distributed across novel and familiar objects.

Face Detection A third goal was to measure the presence of caregivers' faces in the child's field of view (as approximated by the headcam). In order to avoid hand-annotating the size and position of faces in every frame of video, we used an in-house face-detection framework based on freely available tools (Bradski & Kaehler, 2008). This system is described in depth in our previous work (Frank, 2012) but we review it briefly here.

Our system has two parts. The first is the application of a set of four Haar-style face detection filters from the OpenCV library (Viola & Jones, 2004) to each frame of the videos independently. These detectors each provide information about whether a face is present in the frame as well as size and position for any detections. In a second step, these detections are then combined via a hidden Markov model (HMM), trained via annotated data. The intuition behind the HMM model is that a frame is much more likely to contain a face if the previous frame also contained one. The HMM model (which performed nearly as well as the more complex and computationally-intensive Conditional Random Field model used in our previous work) attempted to estimate whether a face was truly present in each frame of the videos, using as its input the number of Haar detectors that were active in any given frame. For training we annotated whether the individual face detectors were correct for XYZ XYZ.

Table 1: Model performance on gold standard generalization training set dataset.

	Precision	Recall	F-score
Baseline			
HMM	0.74	0.78	0.76
DLT			

We additionally computed a baseline model which simply assumed that the largest detector ...

Results for these models are reported in Table 1. The HMM model obtained a relatively high level of performance. ... Overall, the goal of our use of face-detection algorithms was to provide a measurement technique that eliminated tedious and expensive hand-coding and provided acceptable results. We therefore selected the highest performing algorithm (the XYZ) and use detections from this algorithm as an estimate of face presence in all further analyses. Sample frames from the video with overlaid detections are given in Figure 2.

Results

We report results from three different sets of analyses. First, we explore developmental changes in posture and orientation in our dataset. Next, we explore how these changes affect access to faces, as measured using our face-detection algorithm. Finally, we look at access to faces during labeling events. Although this rich dataset will support a large number of additional exploratory analyses, these analyses are motivated by theoretical hypotheses.

Changes in Posture and Orientation

Our posture coding captured typical developmental milestones (Figure 3, left). Overall, sitting was the most common posture for interactions in the caregiver play session. The youngest infants in our sample mostly sat (with parental as-

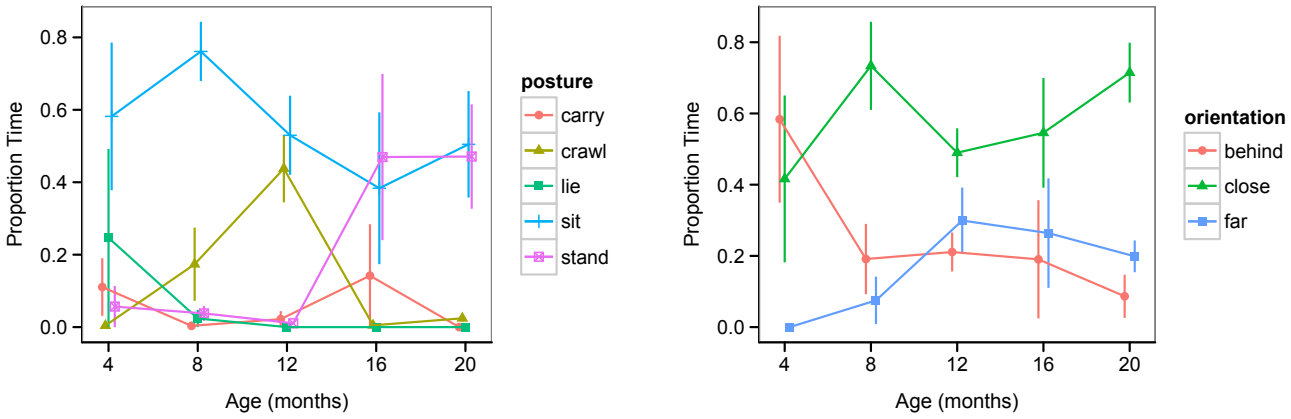

Figure 3: Left: Proportion time in each posture, plotted by child's age. Right: Proportion time in each orientation relative to the caregiver, again plotted by child's age. For clarity, the "other" code is not plotted in either figure. Error bars show standard error of the mean across participants.

sistance), but also lay down and were carried a significant proportion of the time. The 12-month-olds were the only group who spent a large amount of time crawling, and the 16- and 20-month-olds sat and stood in equal parts.

Similarly, our coding of orientation revealed some significant developmental changes (Figure 3, right). Younger children more frequently had the caregiver behind them, often because the caregiver was supporting the child's sitting posture (for the 4-month-olds especially). In contrast, the 12–20 month olds were able to locomote independently and so were able to spend more time further from the caregiver.

Access to Faces by Posture and Orientation

We next investigated the effects of the child's posture and orientation on the presence and size of the caregiver's face in the visual field. Figure 4 shows the proportion of frames with a positive face detection, plotted by the child's posture and orientation relative to the caregiver. Faces were seen most consistently when children were furthest from the caregiver, especially when the child was sitting or standing.[2] Since this combination of orientation and posture was primarily accessible to the toddlers, that meant that overall, older children tended to see more faces. In contrast, when young children were sitting close to (but not behind) their caregiver, they were often looking at the parent's hands and the objects they were manipulating. Crawling and lying down afforded almost no opportunity to see caregivers' face.

In addition to differences in the accessibility of faces, we also observed differences in the size of the faces that were observed based on the child's posture and orientation. The largest faces were seen when children were standing close to their caregivers (who were most often sitting on the floor during the play sessions), and faces were overall smaller when the children were further away.

Interaction between adults is often conducted at a greater distance than interaction between caregivers and children. This simple fact means that often young children have limited and non-canonical views of their caregivers' faces. It remains to be investigated whether young children can nevertheless extract information in these situations.

Access to Faces during Labeling

Our final analysis concerned the accessibility of caregivers' faces during labeling events. Franchak et al. (2011) found that referential speech was marginally more likely to draw toddlers' attention to mothers' faces. We were interested in whether looking at faces occurred during labeling, and especially whether this relationship was related to the type of labeling that occurred. Accordingly, we used the labeling annotations for each child to identify the 2s before and after each labeling event. [Why this window? Check other windows, plus add justification.] We then computed the proportion face detections within this window across ages, object types (familiar vs. novel), and whether this was the first instance of labeling for this object.

Figure 5 shows the results of this analysis. Overall we saw developmental increases in the amount of faces detected during labeling, consistent with the overall pattern of greater face accessibility for older children (who were after all more often sitting or standing further away from their caregiver). An interesting pattern emerged, however, when we broke down the data by label type and labeling instance: we found that there were more face detections around the first labeling instance for novel items [STATS].

This result could be caused by children, caregivers, or a combination of the two. Older toddlers could be more interested in the novel objects and could seek out caregivers (who are initially holding the objects) to find out more; or caregivers could wait until the child can see what is being talked

[2] Since orientation was coded via body posture, faces seen while standing behind the caregiver are presumably due to caregivers turning their heads to look at children who have run behind them.

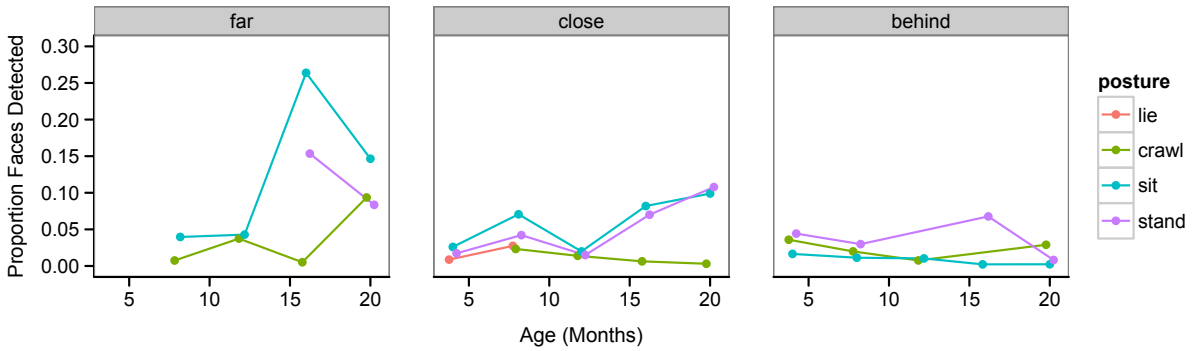

Figure 4: Proportion of frames with a face detected, plotted by child's age, posture (line and point color), and orientation with respect to caregiver (panel). No error bars are shown due to the small and uneven distribution of data across points.

about before introducing a new object. Most likely, both of these explanations is true in part. [SAY MORE].

General Discussion

Using a head-mounted camera, we explored the relationship between infants' postural and locomotor development and their visual access to social information. The use of automated annotation tools from computer vision allowed us to measure the prevalence and size of caregivers' faces in their children's visual field. We found systematic differences in the visual accessibility of faces based on the child's orientation and posture, variables that were further mediated by age. Put simply, older children, who could move themselves, got to choose where they were and when they saw their caregivers' face. Accordingly, they saw faces more around the introductory labeling of novel, engaging objects.

Our study complements a wide variety of work that has investigated the relationship between children's perspective and their access to social and referential information (Yoshida & Smith, 2008; Franchak et al., 2011; Smith et al., 2011; Yu & Smith, in press; Walle & Campos, under review).

The measures developed here have broad applicability to the study of individual and cultural differences. Since the physical circumstances of child rearing vary widely across households and across cultures with different material circumstances, there may be predictable differences in children's visual experience. As suggested by the correlations between walking and vocabulary (Walle & Campos, under review), shifts in how infants are placed in particular postures by strollers or carriers (Zeedyk, 2008) or how much exercise they are given (Bril & Sabatier, 1986).

While infants' visual field is often subject to the whims of their caregivers, toddlers determine their own input to a much greater degree. Toddlers live in a world populated by knees. [FINISH]

Acknowledgments

Thanks to Ally Kraus, Kathy Woo, Aditi Maliwal, and other members of the Language and Cognition Lab for help in recruitment, data collection, and annotation. This research was supported by a John Merck Scholars grant to MCF.

References

Adolph, K., & Berger, S. (2007). Motor development. *Handbook of child psychology*.

Adolph, K., Gilmore, R., Freeman, C., Sanderson, P., & Millman, D. (2012). Toward open behavioral science. *Psychological Inquiry*, *23*(3), 244–247.

Baldwin, D. (1991). Infants'contribution to the achievement of joint reference. *Child Development*, 875–890.

Bradski, G., & Kaehler, A. (2008). *Learning opencv: Computer vision with the opencv library*. O'Reilly Media.

Bril, B., & Sabatier, C. (1986). The cultural context of motor development: Postural manipulations in the daily life of bambara babies (mali). *International Journal of Behavioral Development*, *9*(4), 439–453.

Brooks, R., & Meltzoff, A. (2005). The development of gaze following and its relation to language. *Developmental Science*, *8*(6), 535–543.

Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the society for research in child development*, *63*(4).

Clark, E. (1988). On the logic of contrast. *Journal of Child Language*, *15*, 317–335.

Cummings, M., Van Hof-Van Duin, J., Mayer, D., Hansen, R., & Fulton, A. (1988). Visual fields of young children. *Behavioural brain research*, *29*(1), 7–16.

Franchak, J., Kretch, K., Soska, K., & Adolph, K. (2011). Head-mounted eye tracking: A new method to describe infant looking. *Child development*.

Frank, M. C. (2012). Measuring children's visual access to social information using face detection. In *Proceedings of the 33nd Annual Meeting of the Cognitive Science Society*.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121–157.

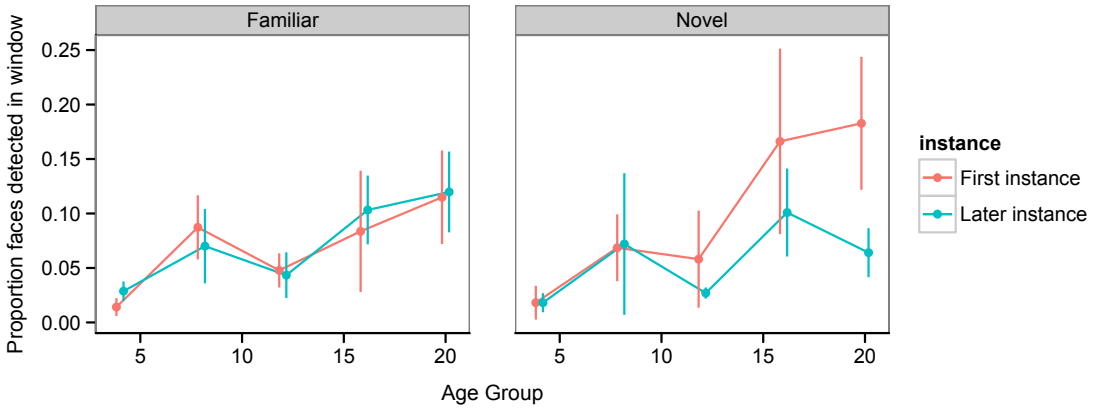

Figure 5: Proportion of faces detected in a 4s window of time centered around labeling events, split by labeling instance and whether the word was familiar or novel. Error bars show standard error of the mean.

Mayer, D., Fulton, A., & Cummings, M. (1988). Visual fields of infants assessed with a new perimetric technique. *Investigative ophthalmology & visual science*, *29*(3), 452–459.

Smith, L. B., Yu, C., & Pereira, A. F. (2011). Not your mother's view: The dynamics of toddler visual experience. *Developmental Science*, *14*, 9-17.

Vaish, A., Demir, Ö., & Baldwin, D. (2011). Thirteen-and 18-month-old infants recognize when they need referential information. *Social Development*, *20*(3), 431–449.

Viola, P., & Jones, D. H. (2004). Robust real-time face detection. *International Journal of Computer Vision*.

Walle, E. A., & Campos, J. J. (under review). Infant language development is related to the acquisition of walking.

Yoshida, H., & Smith, L. (2008). What's in view for toddlers? using a head camera to study visual experience. *Infancy*, *13*, 229–248.

Yu, C., & Smith, L. B. (in press). Embodied attention and word learning by toddlers. *Cognition*.

Zeedyk, M. (2008). *Whats life in a baby buggy like?: The impact of buggy orientation on parent-infant interaction and infant stress* (Tech. Rep.). National Literacy Trust.