Cognition 125 (2012) 244–262

Contents lists available at SciVerse ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Embodied attention and word learning by toddlers

Chen Yu *, Linda B. Smith

Department of Psychological and Brain Sciences, Cognitive Science Program, Indiana University, United States

ARTICLE INFO

Article history:
Received 18 August 2010
Revised 26 June 2012
Accepted 29 June 2012
Available online 9 August 2012

Keywords:
Embodied cognition
Language learning
Perception and action

ABSTRACT

Many theories of early word learning begin with the uncertainty inherent to learning a word from its co-occurrence with a visual scene. However, the relevant visual scene for infant word learning is neither from the adult theorist's view nor the mature partner's view, but is rather from the learner's personal view. Here we show that when 18-month old infants interacted with objects in play with their parents, they created moments in which a single object was visually dominant. If parents named the object during these moments of bottom-up selectivity, later forced-choice tests showed that infants learned the name, but did not when naming occurred during a less visually selective moment. The momentary visual input for parents and toddlers was captured via head cameras placed low on each participant's forehead as parents played with and named objects for their infant. Frame-by-frame analyses of the head camera images at and around naming moments were conducted to determine the visual properties at input that were associated with learning. The analyses indicated that learning occurred when bottom-up visual information was clean and uncluttered. The sensory-motor behaviors of infants and parents were also analyzed to determine how their actions on the objects may have created these optimal visual moments for learning. The results are discussed with respect to early word learning, embodied attention, and the social role of parents in early word learning.

1. Introduction

Infants learn their first words through the co-occurrence of a heard word and a visual scene. By many analyses (Frank, Goodman, & Tenenbaum, 2009; Quine, 1964; Smith & Yu, 2008; Snedeker & Gleitman, 2004; Waxman & Booth, 2001), the central theoretical problem in explaining how infants break into word learning is the ambiguity inherent in everyday scenes with their many potential referents. In this view, it seems unlikely that an infant would, for example, hear the word "train" when then named object was the only object in view; instead, it seems that the infant would more often hear the label when the intended referent, a toy train perhaps, was part of a visual jumble of many things, for example, with a toy car, a ball and a cup on the floor. This, then, is the theoretical problem: Given the ambiguity inherent in such everyday scenes and a learner who may as yet know none of the names of the things in that scene, how can that learner determine the intended referent?

Contemporary solutions endow infants with remarkable cognitive skills, including prior knowledge about the kinds of concepts that are lexicalized by languages (Waxman & Booth, 2001), the ability to make inferences about the thoughts and intentions of the speaker (Baldwin, 1993), and powerful statistical mechanisms that evaluate data across many word-scene experiences (Frank et al., 2009; Smith & Yu, 2008; Yu & Smith, 2007). These are all *internal* cognitive solutions that accept the premise of referential ambiguity. Here we consider an *external* sensory-motor solution and the possibility that the premise of referential ambiguity is exaggerated. Early word learning often takes place in the context of infants' active exploration of objects: infants do not simply look passively at the jumble of toys on the floor but rather use their body – head, hands, and eyes – to select and potentially visually

* Corresponding author.
E-mail address: chenyu@indiana.edu (C. Yu).

0010-0277/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cognition.2012.06.016

isolate objects of interest, thereby reducing ambiguity at the sensory level. These bodily movements also create overt cues that might be exploited by the mature social partner. If infants through their own actions on objects create possible optimal visual moments – with minimal clutter – and if parents are congenial enough to name objects at those moments, then the degree of referential ambiguity may be reduced at the level of the sensory input itself.

This hypothesis was suggested by several recent studies that used head-cameras to capture infants' egocentric views during interactions with objects. The findings suggest that during active play with multiple objects, infants create clean one-object-at-a-time views as a byproduct of their own manual engagement with the objects (Smith, Yu, & Pereira, 2011; Yoshida & Smith, 2008; Yu, Smith, Shen, Pereira, & Smith, 2009). In the contexts used in these studies, there were always multiple objects close together in the play area but analyses of the head-camera images indicated that the infant's view often contained a single object that was close to the infant's body and head, and thus visually larger than the other objects. The specific empirical question for the present study is whether these visually selective moments – observed in contexts of toy play – are also optimal moments for object name learning. If so, it would suggest a bottom-up sensory solution to referential uncertainty.

As in the previous studies, we used head cameras to record the first-person views of toddlers and parents as they jointly played with toys. However, in the present study, all toys were novel and parents were asked to name them with experimenter-supplied novel names. At the end of the play session, infants' knowledge of the object names was tested via a preferential looking measure. Parents were *not* explicitly told to teach the object names, and were not told that their infants would be tested at the end of the play session. In this sense, the task was an incidental learning task embedded in the context of toy play, in which parents named and infants heard those names alongside of other activities such as stacking, rotating, exploring, and playing with objects, similar to the free-flowing play contexts in which everyday word learning is assumed to take place (Hart & Risley, 1995; Hirsh-Pasek, Golinkoff, Berk, & Singer, 2009; Ruff & Rothbart, 2001). We took this approach – not explicitly telling parents to teach the names – because we did not want parents to exaggerate or alter their behaviors in response to perceived demands characteristic of the laboratory setting.

The head-camera images were analyzed frame-by-frame to extract the properties and dynamics of the first-person views of both toddlers and parents during play and naming moments. In addition, the participants' holding of the toys and their head movements were measured. The number of participating child–parent dyads was small ($n = 6$) but the number of data points per subject was extremely large. The small number of participants with a large number of data points per participant is consistent with contemporary approaches to the study of sensory and motor systems (Blake, Tadin, Sobel, Raissian, & Chong, 2006; Jovancevic-Misic & Hayhoe, 2009; Najemnik & Geisler, 2005; Thelen et al., 1993). The key analyses center on the visual properties of the child head-camera images during naming events that were and were not associated with learned object names as measured at test. In addition, we examined both participants' moment-to-moment motor behaviors around naming events in an effort to better understand how optimal visual moments for object name learning are created.

2. Method

2.1. Participants

Six parent–infant dyads participated (three male and three female infants). Three additional infants began the study but did not contribute to data because of refusals to wear the measuring equipment. The mean age of the infants was 18.5 mo (range 17–20 mo).

2.2. Stimuli

There were nine unique novel "toys", organized into three sets of three. Each toy was a simple shape with a uniform color made from plastic, hardened clay, aggregated stones, or cloth. All objects were similar in size, on average 288 cm^3. Fig. 1 shows three toy objects on the table top during play as well as all of the nine objects and their associated names.

2.3. Experimental room

Parents and infants sat across from each other at a small table (61 cm × 91 cm × 64 cm) that was painted white. The infant's seat was 32. 4 cm above the floor (the average distance of eye to the center of the table was 43.2 cm). Parents sat on the floor such that their eyes, heads and head cameras were at approximately the same distance from the tabletop as those of the infants (the average distance of eye to the table center for parents sitting on the floor was 44.5 cm). A previous head camera study of object play (Smith et al., 2011) explicitly compared parent and infant head camera images when parents were sitting naturally in a chair or on the floor and found no differences in any aspects of infant or parent behavior as a function of the task geometry (see also Yoshida & Smith, 2008, who used a somewhat slightly different geometry and observed the same results as Smith, Yu, & Pereira, 2010). To aid in the automatic image analysis, both participants wore white clothing. There were also white curtains from floor to ceiling and a white floor such that everything in the head-camera images was white with the exception of heads, faces, hands and the toys.

2.4. Apparatus

The toddler and participating parent wore identical head cameras, each embedded in a sports headband. The cameras were Supercircuits (PC207XP) miniature color video cameras and weighed approximately 20 g. The focal length of the lens was f3.6 mm. The number of effective pixels were 512 (horizontal) × 492 (vertical) (NTSC). The resolution (horizontal) was 350 lines. The camera's visual

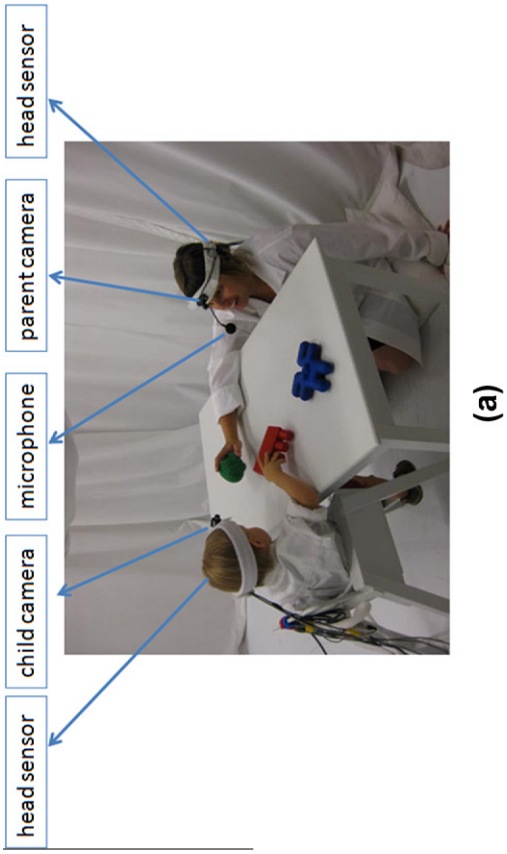

Fig. 1. (a) Parent and child at the play table with head-mounted sensors, including head-mounted cameras, head motion tracking sensors and a microphone to record parental speech. (b) The nine toys and their names as organized into the three play sets of three toys each.

field was 70° and provided a broad view of objects in the head-centered visual field that was less than the full visual field (approximately 170°). The recording rate was 10 frames per second. The direction of the camera lens when embedded in the sports band was adjustable. Input power and video output went through a camera cable connected to a wall socket, via a pulley, so as to not hinder movement. The head cameras were connected via standard RCA cables to a digital video capture card in a computer in an adjacent room. The headband was tight enough that the camera did not move once set on the child. The multi-channel video capture card in the recording computer adjacent to the experiment room simultaneously recorded the video signal from the cameras. The head camera moved with head movements but not with eye movements and therefore provided a head-centered view of events that may be momentarily misaligned with the direction of eye gaze. In a prior calibration study using a similar tabletop geometry, Yoshida and Smith (2008) independently measured eye gaze direction (frame by frame via a camera fixated on the infant's eyes) and head direction and found that eye and head directions were highly correlated such that 87% of head camera frames coincided with independently coded directions of eye gaze. Moments of non-correspondence between head and eye directions in that study were generally brief (less than 500 ms). Thus, although head and eye movements can be decoupled, the tendency of toddlers to align the head and eyes when actively reaching for and interacting with objects suggests that the head camera provides a reasonable measure of the toddler's first person view.

A high-resolution camera (recording rate 30 frames per second) was mounted above the table providing a bird's eye view aligned with the table edges. This camera provided visual information about the events that was independent of participants' movements and was used to resolve any ambiguities in the head-camera images. In addition, for the object name-learning test, a small camera was mounted on the table (in front of the experimenter doing the testing) and was centered on the infant's face and eyes so as to record the direction of eye gaze during the testing procedure.

A Liberty motion tracking system (www.polhemus.com) was used with two sensors embedded in the infant's and the parent's headbands respectively to measure their head movements. Each sensor generated 6 degree-of-freedom data – 3D coordinates (x, y, z) and 3D orientations (heading, pitch and roll) of the participant's head relative to the source transmitter centered above the table. Sampling rate was 240 Hz.

The parent's voice during the interaction was recorded with a standard headset with a noise reduction microphone.

2.5. Procedure

Prior to entering the experimental room and while the infant played with an experimenter, the parent was given a sheet with the pictures and names of the nine novel objects. The parent was asked to use these names when playing with the infants. Parents were not told that the purpose of the study was for them to teach the infant these names

but rather that the goal of the study was simply to observe how they and their infant interacted with a set of novel toys and that they should try to play as naturally as possible. The parent and infant were then fitted with white smocks.

Three experimenters worked together in this experimental setup. Upon entering the experiment room, the infant was seated in the chair and a push-button pop-up toy was placed on the table. One experimenter played with the infant while the second experimenter placed the headband low on the forehead of the infant at a moment when the child was engaged with the toy. The first experimenter then directed the infant to push a button on the pop-up toy while the second experimenter adjusted the camera such that the button being pushed by the infant was near to the center of the head-camera image (as viewed by a third experimenter in the control room). To calibrate the parent's camera, the experimenter asked the parent to look at one of the objects on the table, placed close to the infant. During both the infant's and the parent's head-camera calibration, the third experimenter in the control room confirmed that the object was at the center of the image and if not small adjustments were made to the camera.

2.5.1. Play session

The containers holding the objects had the written names of the objects as reminders to the parents of the names. Parents were told to take all three objects from one set, place them on the table, and engage the infant with the toys. These toys were removed and replaced with the next set of three toys given an audio command from the experimenters. In this manner, the parent cycled through each set of three toys twice for six play trials, each approximately lasting 1 min. The whole interaction was about 9 min in total with a brief break between trials for switching toy sets.

2.5.2. Object-name test

Immediately, at the end of the play session, an experimenter tested the child in a name comprehension task. In a 3-alternative forced choice, each of the nine names was tested twice. On each trial, the foils were two randomly selected objects from the set of nine objects. The experimenter sat across the table from the child. One camera was directed at the child's face and eyes, and a second camera was directed at the experimenter to ensure that the experimenter provided no social cues – by look, posture, or other behavior as to the requested object. On each trial, the experimenter put three objects – 40 cm apart – onto a tray out of view of the child. The experimenter then brought the tray into view and said "look at the x, where is the x, look at the x". The trial lasted approximately 40 s. During this testing, the parent sat behind the child and was explicitly asked not to interact with her child. The order of the 18 testing trials was randomly determined in two blocks of 9 with which each object name tested once in a block and thus twice overall. Testing took 5–10 min. Naïve coders who knew *when* the name was mentioned but did not know the target object coded the video for the direction of infant eye-gaze to the three objects. The main dependent measures were looks immediately following the naming event and total looking time to each object during the testing event, with looks to an object interpreted as indicating the child's answer to the comprehension question (Hirsh-Pasek & Golinkoff, 1996). A second coder scored a randomly selected 25% of the test trials; the level of agreement exceeded 90%. In addition, naïve coders also coded a portion of the video recordings of the experimenter's behavior during testing to ensure no unconscious prompting; these coders watched the experimenter, with the sound off, and then guessed which object of the three the experimenter was asking for.

2.6. Data processing

2.6.1. Visual images

The main dependent measures for the head camera images were the sizes and numbers of objects in the images for each of the approximately 3600 frames contributed by each participant. These two measures were automatically coded, frame by frame, via a machine vision program (Yu et al., 2009). See Appendix A for technical details and Yu et al. (2009) and Smith et al. (2010) for comparison to frame-by-frame hand coding. Holding behaviors (who and which object) were coded manually, frame-by-frame, from the images captured by the overhead camera. The two coders independently coded the same randomly selected 25% of the frames (checking head camera images to resolve any ambiguities) with 100% agreement.

2.6.2. Motion data processing

The three-dimensional head position data were reduced to one dimension and the three-dimensional orientation data were reduced to a second dimension by aggregating across three dimensions. Each one-dimensional signal was smoothed with high-frequency components removed using a standard Kalman filter (Haykin, 2001) with a single set of parameters estimated for the Kalman filter and used for all signals. In addition, these speed time series were down-sampled to be at 60 Hz. After these three steps, each parent–child dyad generated four time series corresponding to the parent's head position and orientation and the infant's head position and orientation. For all analyses, position and orientation movements were converted into two binary categories: moving and not moving using the threshold for position of 3 cm and for rotation of 15°. The main dependent measure used in the analyses is percentage of time that the head was moving.

2.6.3. Speech processing

A silence duration of more than 0.4 s was used to mark the boundaries of utterances. Human coders listened and transcribed these speech segments to determine which were naming events. A naming event was defined as a parent utterance containing the name of a novel toy. The duration of each naming event was defined by the onset and offset of the spoken utterance in which the name was included. For example, "Can you get the dax?" and "look at the dax" were two naming events, and the onset and offset of an entire utterance was marked to define the temporal duration of the naming event. The average length of naming events was 1.86 s. All other moments were designated

as non-naming events. Two coders transcribed the same randomly selected set of utterances with 90% agreement; all disagreements were resolved by re-listening to the audio recordings.

2.6.4. Statistical analyses

The main statistical analyses are based on linear mixed-effects models (Bates & Sarkar, 2007) – using the lmer function of the R package lme4 (Doran, Bates, Bliese, & Dowling, 2007). Unless specified otherwise, each of sensorimotor patterns extracted from raw data for each participant was treated as a dependent variable (e.g. size of object in the child's view, proportion of time holding an object), along with participant (child versus parent) and event (naming versus non-naming) were fixed factors within the analyses. Random effects for subjects, trials, instances of events, and objects were also included to account for any non-independence among different participants, behaviors, objects, words, and trials (Baayen, Davidson, & Bates, 2008). All p-values and confidence intervals reported in mixed-model analyses were derived from posterior simulation using the R language package (Baayen, 2008) to yield standard p-value statistical significance. Some of the analyses are also trial based (six dyads $*$ six play trials $*$ two participants, child and parent) and as such provide a description of first-person views that are grosser than that of an individual image frames (as there are 3600 frames per participant) but finer than that of all naming moments aggregated within a single dyad. We believe this to be an appropriate level to capture the variance in these sensory-motor measures.

3. Results

3.1. Visual selection

We first present analyses pertinent to visual selection without regard to whether an object was or was not being named. This is necessary to show that the present context replicates previous findings of one-object visual dominance in the infant's view when playing with multiple objects. These initial analyses also provide baseline measures against which to consider the visual properties of naming moments from the infant's view.

3.1.1. Head camera images

On each experimental trial, there were three objects on the table and thus three objects that could be in the infant's and the parent's views. Further, if the infant and parent were to sit back and take a broad view of the table, not moving their heads, all three objects would be in view and would all have approximately the same image size. However, the sizes of the objects in the head-camera images change as their distance to the viewer changes. Fig. 2a shows several examples of head camera images from both parents' and toddlers' views at simultaneous moments. The sizes provided (% of image pixels) indicate the image size of the largest object in the infant view. As is apparent, an object that took up even just 5% of the image was very large and visually dominating. These images from the child, parent and overhead cameras also illustrate how objects are generally larger in the infant's view than in the parent's view and that this was because the objects were closer to the child. These three views – child, parent, and overhead – also show how, despite there being three objects on the table in relative close proximity to each other, there was often just one dominating object in the infant's view.

The first two panels in Fig. 2b shows the mean image size for all objects in view and the mean number of objects in view for parent and child head-camera images calculated across the entire dataset. The means and standard errors are based on frame-by-frame measures averaged within each 60-s trial. The proportion of image pixels taken up by all the objects together was greater in the infants' than parents' views ($\beta = -5.43$; $p < 0.001$); however, there were fewer objects in the infants' than parents' views ($\beta = 0.91$, $p < 0.001$) as the infant's head camera images often contained one or two visually large objects. These group differences also characterized individual dyads: the *difference* between the average image size for an infant and parent ranged from 5% to 7% across dyads and this difference was individually reliable for each dyad as determined by frame-by-frame comparisons of parent and infant images within a dyad (minimal β value among all of the dyads $\beta = -0.74$, $p < .0001$). The *difference* in the mean number of objects in view between infants and parents ranged from 0.75 to 1.12 and was also individually reliable for each dyad (with a minimal β value among all of the dyads $\beta = 0.43$, $p < .0005$).

The third panel in Fig. 2b shows how much the view was dominated by a single object. Visual dominance by a single object was defined using both a more conservative and a more liberal criterion. Both criteria took into account the absolute size of the object *and* its relative size with respect to other objects in the view. By the more conservative standard, an object was considered dominating if it comprised at least 5% of the image and if it was greater than 50% of the size of all objects in view, thus if it was by itself at least as big as all other in-view objects combined. By the more liberal criterion, an object was considered dominating if it comprised at least 3% of the image and if it was greater than 50% of the size of all objects in the image. The 3% criterion is roughly comparable to the size of the fovea. By the more liberal definition, more than 60% of the infant images within a trial met the criterion for a visually dominating object whereas by this same definition only 12% of the parent images did ($\beta = -1.36$, $p < 0.001$). By the stricter size criterion, almost 40% of infant images within a trial contained a dominating object whereas only 5% of the parents' images met this criterion ($\beta = -1.46$, $p < 0.001$). When each dyad's head-camera images were considered separately, the mean number of images for each infant with a dominating object ranged from 43% to 81% for the liberal measure and from 25% to 63% for the conservative measure; for the individual parents, these means ranged from 8% to 15% for the liberal measure and 3–7% for the conservative measure. These findings replicate those of earlier studies (Smith et al., 2010; Yu et al., 2009): the infant's egocentric view is often characterized by a single object close to the infant and thus large and dominating in the infant's visual field. Before

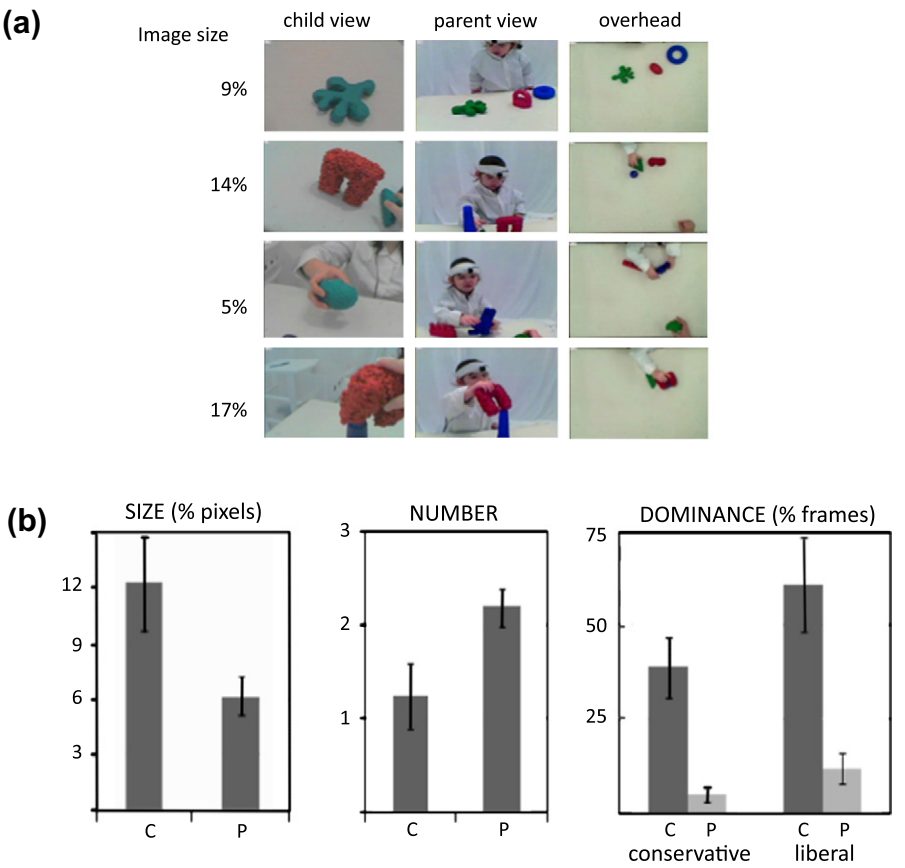

Fig. 2. (a) Simultaneous views from the child's head camera, the parent's head camera, and the overhead camera. Object size indicates the percentage of head camera image taken up by the largest objects in the child's view. (b) Differences between child (C) and parent (P) head camera images: mean image size of objects in each frame; mean number of objects in view; and percentage of frames with one dominating object by both a conservative and liberal criterion for dominance (see text). Standard errors of the mean are calculated with respect to each trial (six trials $*$ six participants).

considering the main question of how these moments of visual selection in the input may matter to object name learning, we consider how these one-dominating object views relate to hand and head movements of the infants and also those of the parents.

3.1.2. Hands

Infants were holding at least one object on 68% of the frames and parents were doing so on 61% of the frames ($\beta = -0.07$; $p = 0.08$; range for individual infants 49–81%, for individual parents 42–76%). Dyads differed on who held objects more overall: within two dyads, the parent held objects more frequently than the infant did, and within four dyads, the infant held objects much more than did the parent. However, for frames in which one object was visually dominant in the infant's view (by the conservative criterion), the dominating object was in the infant's hands reliably more often than it was in the parent's hands (52% of the time versus 20% of the time, $\beta = -2.86$, $p < .001$). This direction of difference characterized all six dyads with the smallest difference between parents and infants in who was holding the visually dominating object being 19%. On 28% of the infant head-camera frames, the dominating object was not being held by anyone but was sitting on the table close to the child. These results also replicate previous findings, suggesting that the infant's one-object views are associated with the infant's holding of the visually selected object (Yu et al., 2009).

To better understand the child's and parent's behaviors leading up to these one-object views, we determined the frame in which an object first became dominant in the infant's view by the more conservative definition and the frame at which it ceased to be dominant by this criterion. We then determined (frame by frame) whose hands (if any) were holding that target object or other objects for the 5 s preceding dominance and for the 5 s after the target object ceased to be dominant in the infant's head camera image. This yields a trajectory of the likelihood that the visually selected object was being held prior and after its dominance by the parent or child as shown in Fig. 3 (see Allopenna, Magnuson, and Tanenhaus (1998), for use of this approach in time-course analyses). The probability that the infant was holding the to-be-visually-dominant object, the target, shows a clear and dramatic increase as a function of temporal proximity to visual dominance. The target object was more likely to be held by the infant than other objects by 4.9 s ($\beta = -0.15$; $p < 0.005$) prior to becoming dominant and the once-dominant object was still more likely to be held by the infant than other objects up to 3.5 s ($\beta = -0.23$; $p < 0.005$) after no longer meeting

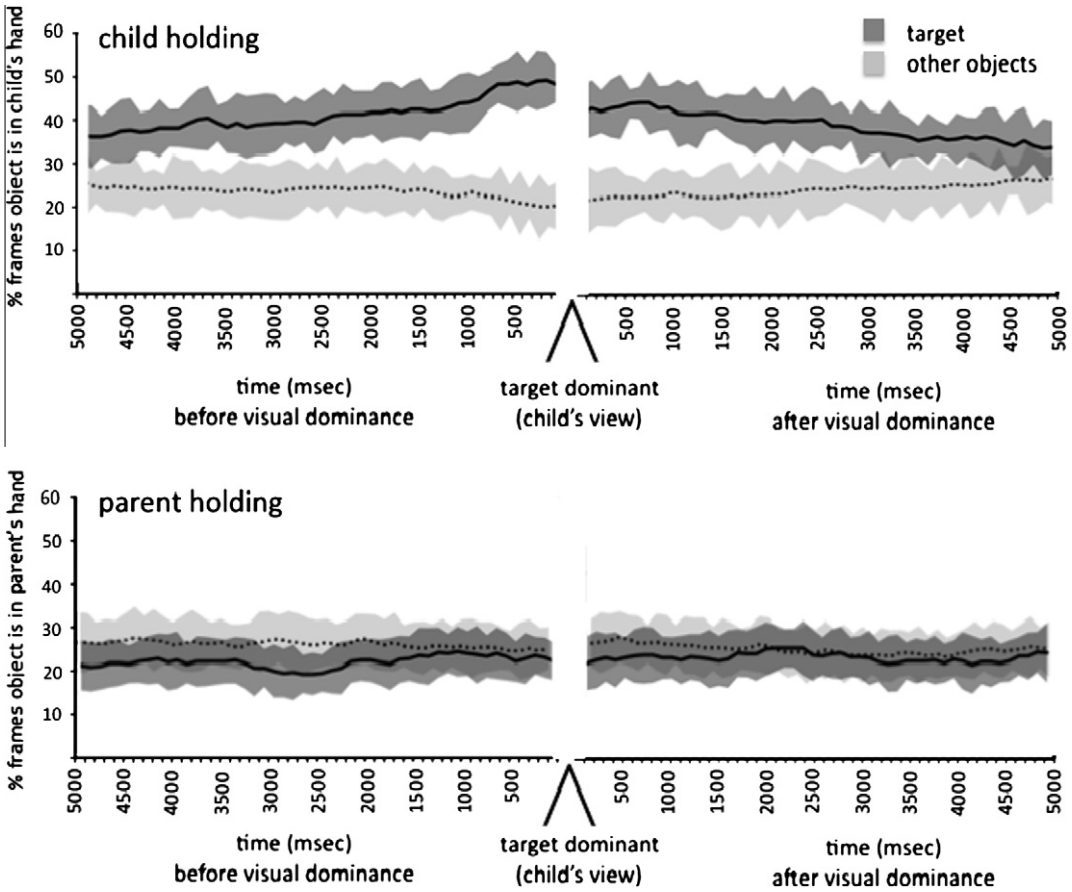

Fig. 3. Mean percentage of frames 5 s prior to and 5 s after an object (the target, etc.) becomes dominant (using the conservative criterion) in the child's head camera image that has being held by the child (solid line top) or parent (solid line bottom) and the percentage of frames that the child or parent was holding some other object (dotted lines). The indicated regions around the means are the standard error, with mean and standard error calculated with respect to trials.

the criterion for visual dominance. Statistical comparisons within a parent–child dyad of the mean likelihood of holding for the 5 s window before and the 5 s window after visual dominance indicates that within a dyad, and for both before and after, the infant within each dyad was more likely to be holding the dominant object than the parent ($\beta = -0.76$; $p < 0.005$). Not only does the pattern in Fig. 3 show that infant holding behavior and not parent holding behavior is associated with one-object views, the pattern also suggests that visual selection emerges in a temporal profile of motor behavior that is sustained for some time before and some time after visual dominance. These behaviors are potentially important clues to the infants' interest that may be exploited by the parent.

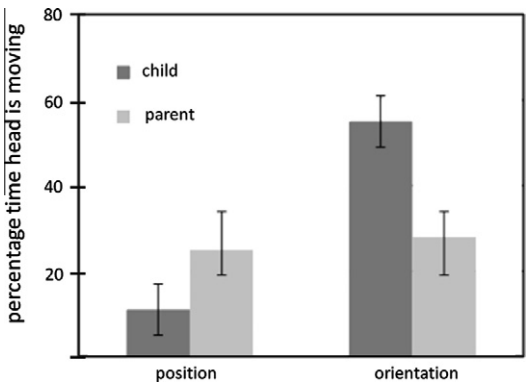

Fig. 4. Mean percentage time with position and orientation head movements (see text on page 12 for threshold measurement of movements) for children and parents. Standard errors calculated with respect to trials.

3.1.3. Heads

Head motion directly determines the head camera images. Changes in head position move the head (and also the eyes and head camera) closer or farther away from objects in view while changes in head orientation shift the head direction in the 3D environment. Fig. 4 shows the percentage of time that heads were moving. Overall, infants moved their heads more often than did parents ($\beta = -1.16$, $p < 0.001$). However, parents moved their heads more often positionally than did infants ($\beta = -2.51$; $p < 0.005$) whereas infants rotated their heads much more often ($\beta = -12.21$; $p < 0.001$) than did parents. Within all

six dyads, the percentage of time that the head was moving (combination of position and orientation) was greater by at least 1.5 times in the infant than the parent. This greater head movement, and particularly the greater changes in head orientation, means that overall the infants' views (and head camera images) were less stable than those of the adults, a fact which could make stabilizing attention more difficult and may mark moments of holding an object and head stabilization as a critical component of effective visual attention by toddlers.

In summary, the results reported in this section replicate previous findings (Smith et al., 2010; Yu et al., 2009) about object selection from the toddler's first-person view and they replicate the link between such selection at input and the infant's holding of the visually selected object.

3.2. Object name learning

During the play session, parents uttered on average 365 words (tokens). Each of the nine object names was produced by the parents on average only 5.32 times (SD = 1.12). An object name was categorized as learned for an infant if his looking behavior at test indicated learning on 2 out of the 2 testing trials for that object; all other object names were considered as unlearned. By this measure, infants learned on average 5.5 of the nine object names (range 3–8). The number of times parents named each object was negatively correlated with the likelihood that the infant learned the object name: 4.5 naming events for learned names and 6.5 per name for unlearned names, $r(52) = -0.35$; $p < 0.001$. This may be due to parents' use of the name in attempts to engage children with specific objects that were not of interest to the child. At any rate, the lack of correlation reminds that learning may depend on more than the mere frequency of heard names and more critically on the frequency with which naming coincides with the infant's visual selection of the named object.

All parent naming events associated with learned object names were designated as **successful** ($n = 149$). All other object- naming events were designated as **unsuccessful** ($n = 136$). Recall that objects were presented in 3 sets of 3. Successful and unsuccessful naming events did not differ in duration ($\beta = -0.07$; $p = 0.67$) nor any other noticeable property. Note, however, that if a parent named one object five times during play and the infant was judged to know that object name at test, all five naming events were considered "successful". Thus, there is noise in this categorization of naming events as successful and unsuccessful. Nonetheless, if we can discern a systematic relation between the visual dominance of the named object and object name learning – despite this noise – then we would have evidence for the hypothesis that toddlers may solve the referential uncertainty problem at a sensory level. To test this hypothesis, we measured the size of the named target and the size of other *distracter* objects in the head camera images. This provides a measure of the relative dominance of the referent of the object name and its visual competitors. We also computed the same measures – for a randomly designated "target" object – for all of the moments when no object was being named (non-naming events) as a baseline for comparison. The sizes of the target and other objects in both the infant and the parent head-camera views during naming events are shown in Fig. 5 and the average measure for the non-named "target" during non-naming events is indicated by the dotted line.

Consider first the pattern from the child's head camera images. The image sizes of the named target in the child head camera during *successful* naming events differed from non-naming events ($M_{\text{successful}} = 6.28\%$, $\beta = -1.75$, $p < 0.001$) but the target object sizes for unsuccessful naming events did not ($M_{\text{unsuccessful}} = 4.07\%$; $\beta = -0.05$, $p = 0.58$). This provides direct support for the hypothesis that referential selection at *input*, at the sensory level, matters to successful object name learning by infants. However, parent naming versus not naming was not strongly associated with the visual dominance of the target object in the child's view. Parents produced nearly as many unsuccessful naming events as successful ones, and only successful naming events show the visual signature of target objects in the child's view. Notice also that the named target object was larger in the child's head-camera view for successful than for unsuccessful naming events ($M_{\text{successful}} = 6.28\%$; $M_{\text{unsuccessful}} = 3.88\%$; $\beta = 2.62$, $p < 0.001$). We also examined whether these differences changed over the course of the play session: That is, it could be that infants learned some words early in the session and because they knew these words, they might interact with the objects differently or parents might name objects differently early versus later in play. Comparisons of the relative dominance of the named object for the first three versus second three play trials did not differ for either successful or unsuccessful naming events ($\beta = -0.08$, $p = 0.31$; $\beta = -0.09$, $p = 0.69$). These analyses provide strong support for the relevance of visual information at the moment an object name was heard for the learning of that name by 18-month old infants.

Now consider these same measures for the parent head-camera images, also shown in Fig. 5. The image size of the objects was always smaller (because the objects tend to be farther away) in the parent's than in the infant's head camera images. However, the pattern of image size for the named object for successful versus unsuccessful naming events *is the same for parents and infants*. More specifically, for the parent head-camera images, the named target was larger in the parents' head camera image during successful than unsuccessful naming moments ($M_{\text{successful}} = 3.46\%$; $M_{\text{unsuccessful}} = 2.29\%$; $\beta = 1.53$, $p < 0.001$) and differed reliably from the comparison measure for non-naming events ($M_{\text{non-naming}} = 2.36\%$, $\beta = -1.17$, $p < 0.001$). Considering that the target object was closer to the child (as established in the analyses of the child head-camera images), this pattern can happen *only* if parents move their head toward the named target (and child) during the naming event thereby reducing the distance between the object and the head (and the head camera). In brief, the target object was more visually dominant in *both* the infant's and the parent's view during successful but not unsuccessful naming events, indicating coordinated and joint attention during successful naming events. This result also suggests that parent behavior, as well as infant behavior, distinguished successful and unsuccessful naming events: the child may signal interest (as well as reduce the visual ambiguity) by holding

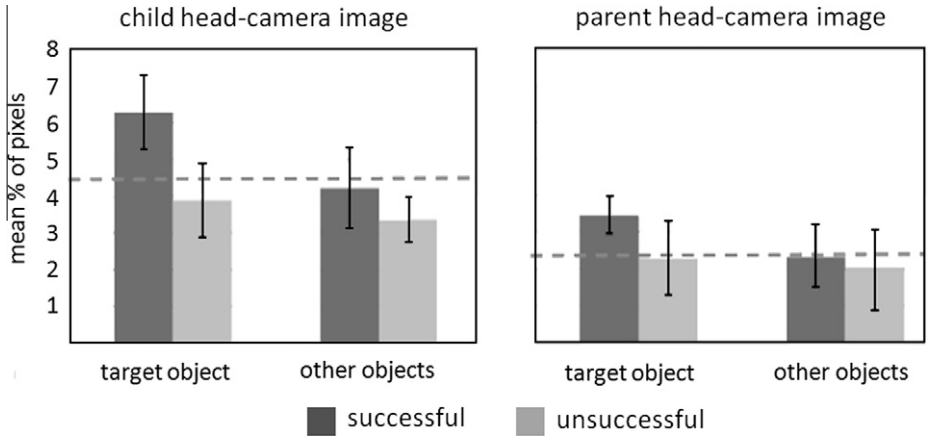

Fig. 5. Mean object size (% of pixels in image) for the named target and for other objects in child's and parent's head-camera images during the naming event, for successful naming events that led to learning at post-test and for unsuccessful naming events that did not lead to learning as measured at test. Means and standard errors were calculated with respect to trials. Dashed line indicates the mean object size during non-naming moments.

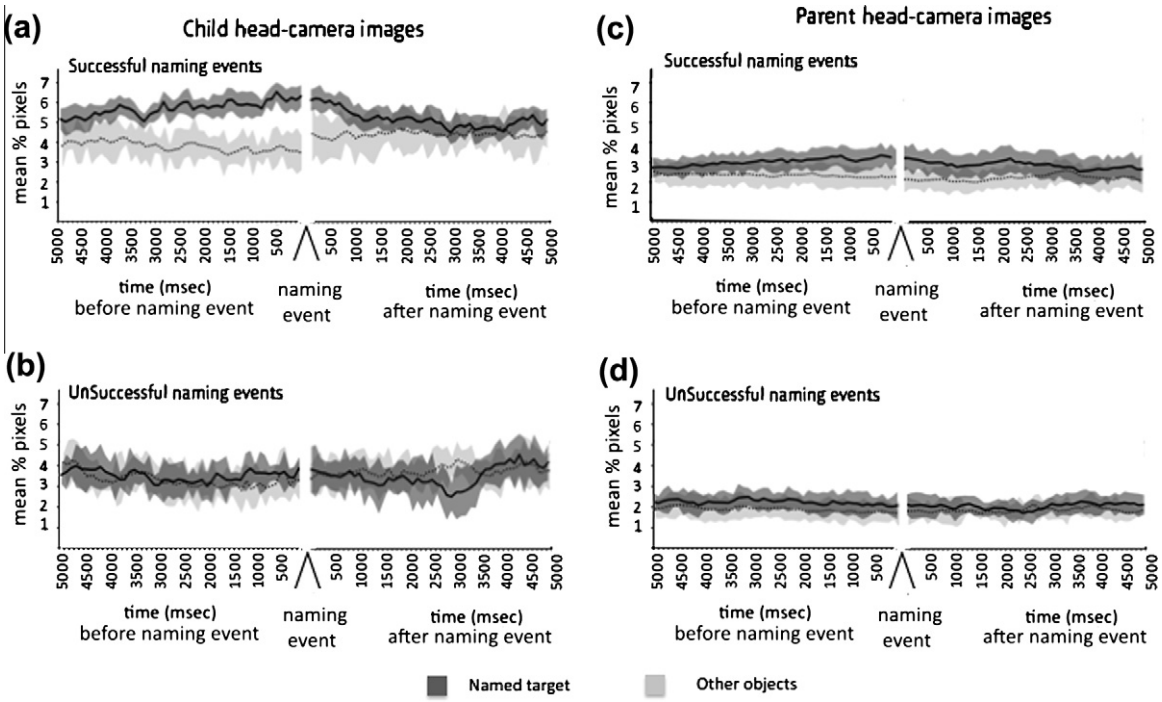

Fig. 6. Size of named target and other objects (% of pixels) in child and parent head-camera images for the 5 s before and after successful and unsuccessful name events. Solid line and dark grey shading indicate named target; and dotted line and light shading indicate other objects. Shading indicates standard error around the mean calculated with respect to variance across trials.

and moving the object close to the head and eyes, and in successful naming moments, the parent may signal a naming event by moving her head slightly toward the child and the named object.

The dynamics of visual selection suggest that this coordination emerges because the parent follows the infant's attentional lead. Fig. 6 shows the mean image size of named versus distracter objects from the child's head camera images (panels a and b) and from the parent's head-camera images (panels c and d) for the 5 s before and right after successful and unsuccessful naming events. For successful naming events, the image size of the named target diverges from the competitor objects in these temporal profiles much earlier (and to a much greater degree) in the images from the child than from the parent head camera. We defined the point of divergence as the first significant difference in a series of temporally ordered pairwise tests over time (Allopenna et al., 1998;

Gershkoff-Stowe & Smith, 1997). For successful naming events, the target object was reliably larger than other objects in the infant head-camera image beginning at 3.2 s before the naming event ($\beta = -0.49$; $p < 0.005$) and remained visually dominant for 900 ms after naming ($\beta = -0.78$; $p < 0.005$). For the parents, there was a reliable advantage in the head-camera image size of the named target over others only for successful naming events and only 2.5 s prior to the naming event itself ($\beta = -0.73$; $p < 0.005$), and only 500 ms after naming ($\beta = -0.48$; $p < 0.005$).

For unsuccessful naming events, there was no advantage for the named target over other objects in either the child or parent head-camera images. Thus, for infants and for parents, the advantage of the named over un-named objects in visual size characterized successful but not unsuccessful naming. However, for infants the visual selection and isolation of the object that led to successful object name learning began long before the naming event and lasted for some time after naming. The increased visual size of the target in the parent's view was temporally after selection made in the infant's view and was not maintained after the naming event in the parent's view as long as in the infant's view. Thus, the infant's selection may begin with general interest in the object, which then creates optimal moments for learning but the visual selection on the part of the parent appears more localized to the naming event itself. The adult pattern is thus consistent with effective naming that follows-in on the child's sustained interest in an object (Masur, 1997; Tamis-LeMonda & Bornstein, 1994; Tomasello & Farrar, 1986).

Further, the observed patterns of successful and unsuccessful naming events might be expected to differ across dyads with some parents being more and some less sensitive to the signals of visual selection and optimal naming moments (Tamis-LeMonda, Bornstein, & Baumwell, 2001). Analyses of individual dyad data in this small sample suggest that the general pattern characterizes all dyads. All dyads contributed both successful and unsuccessful naming events and for all dyads, the visual dominance of the named target over other objects was greater for successful than unsuccessful naming events (minimal β, $\beta_{child} = 1.08$, $p < 0.001$; $\beta_{parent} = 0.90$, $p < 0.005$). This is not to say that a larger sample would not show critical differences in some parents' abilities to select optimal moments for naming, but rather, in the present small sample, for all parents, naming sometimes led to learning and sometimes did not and for all parents in this sample, naming that led to learning was naming that occurred when the named object was visually dominating in the infant's view. These data are correlational and thus in and of themselves do not show that reduced visual clutter was the *reason* for the better learning at those naming moments, but they do show (1) that there are moments when the referential ambiguity believed to characterize the early word-learning context is significantly reduced and (2) that this reduction in ambiguity in associated with the learning of the object name.

3.2.1. Hands and heads

Visual selection and the reduction of referential ambiguity at the sensory level, at input, must be accomplished by changing the physical relation between the potential visual targets and the eyes. Hand actions that move the object close to the head and eyes and the quieting of head movements that stabilize the view are thus potentially important components of visual selection. The left side of Fig. 7 shows that infants were more likely to be holding the named object than other objects during both successful and unsuccessful naming events ($\beta = 0.52$, $p < 0.001$; $\beta = 0.42$, $p < 0.001$) but holding was more strongly associated with successful than unsuccessful naming events ($\beta = 0.32$, $p < 0.005$). The object-holding behavior of parents, shown on the right side of Fig. 7, was not reliably

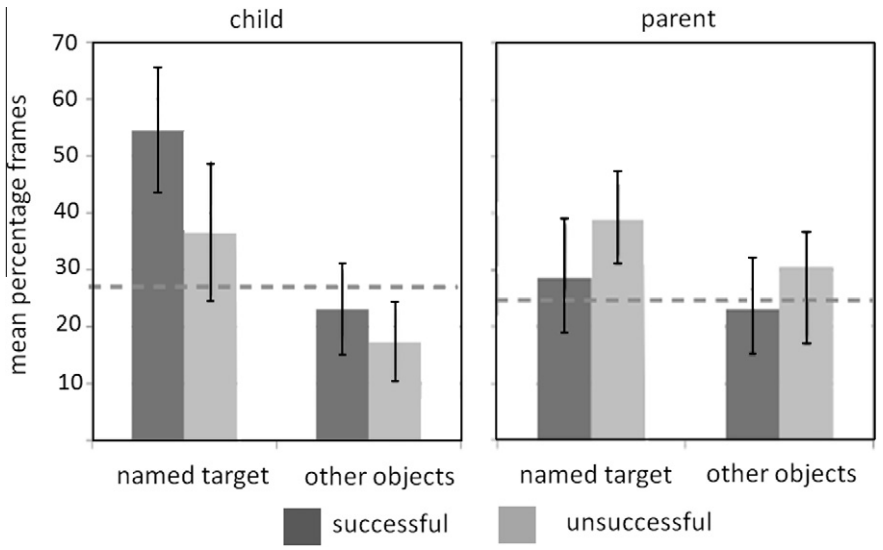

Fig. 7. Mean percentage of frames in which the parent or child has been holding the named object or another object for successful and non-successful naming events. Dashed line indicates mean holding for children and parents during non-naming events. Means and standard errors are calculated with respect to trials.

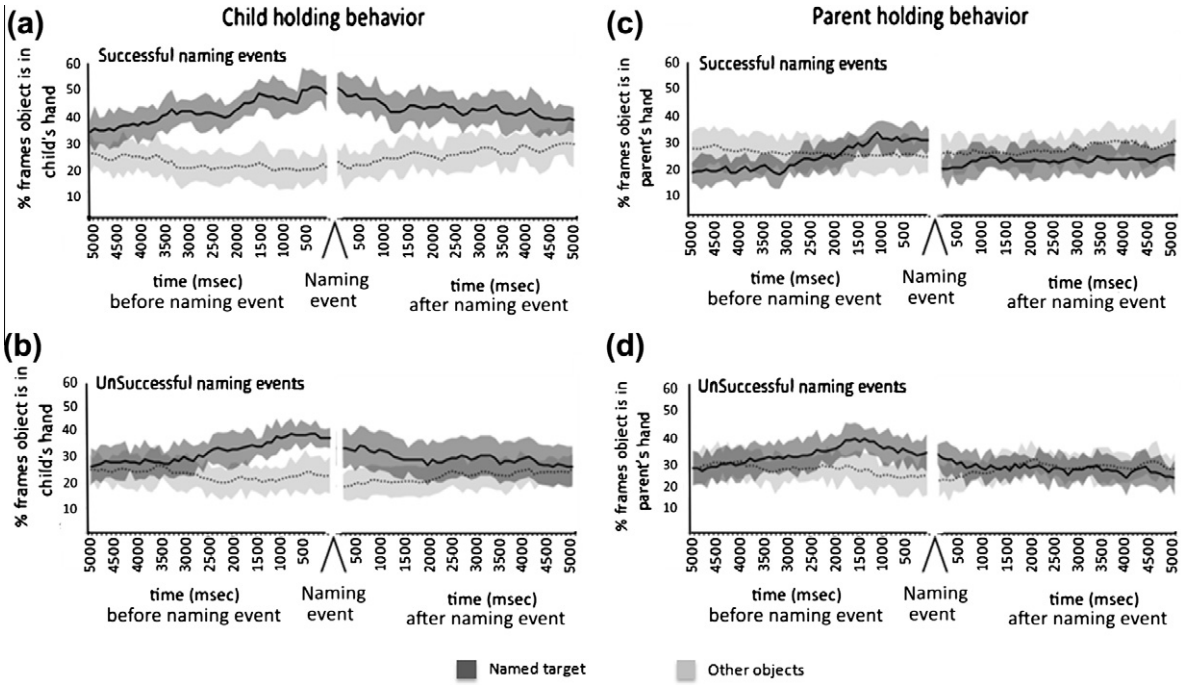

Fig. 8. Holding (mean percentage of frames) of named targets and other objects by child (a and b) and parent (c and d) for the 5 s before and after successful and unsuccessful name events. Solid line (mean) and dark grey shading (standard error) indicate the named target; and dotted line (mean) and light shading (standard error) indicate other objects. Means and standard errors are calculated in terms of trials.

related to naming or to the learning of the object name. But notice there was a slight tendency for parents to be holding the named object during *unsuccessful* naming events; in the present task, parents did not often jointly hold the object that the child was holding and thus parent-holding is associated with not-holding by the child, which, in turn is associated with less visual dominance for the named target and with a decreased likelihood of learning the object name.

Fig. 8 shows the dynamics of child and parent holding of the named and other objects for the 5 s before and after successful and unsuccessful naming events. For successful naming events, infants were more likely to be holding the target object than other objects 4.00 s prior to the naming event ($\beta = 0.21$, $p < 0.005$) and this likelihood increased steadily up to the naming moment. After successful naming events, infants continued to hold the named object more than other objects for a relatively long time, with the difference in the likelihood of holding the named versus other objects remaining statistically significant until 5.00 s after the naming event ($\beta = 0.67$, $p < 0.005$). Unsuccessful naming events showed a similar but weaker

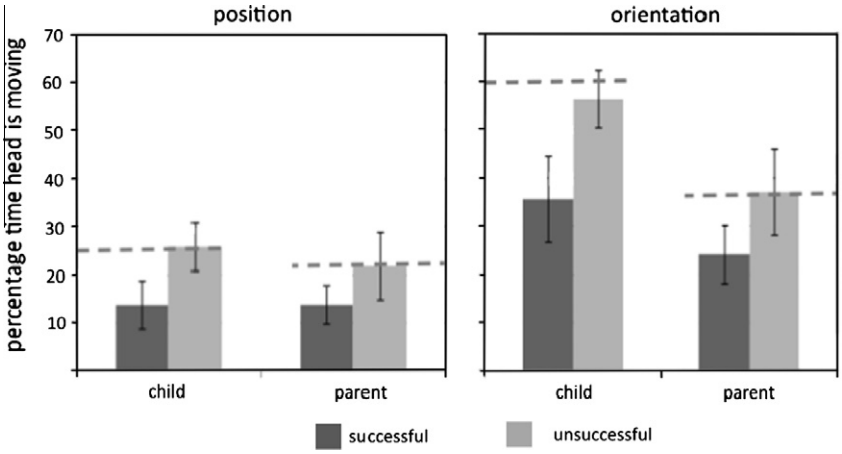

Fig. 9. Mean percentage time with position and orientation head movements during successful and unsuccessful naming events for children and parents. Means and standard errors are calculated with respect to trials. Dashed line indicates mean movements during non-naming moments.

pattern of differences; the likelihood that a named target was held more than others was reliable only 2.50 s prior to naming ($\beta = 0.11$, $p < 0.005$), and the likelihood that a named target was held more was reliable only 1.20 s after naming ($\beta = 0.56$, $p < 0.005$). Fig. 8 also shows the probability of the parent's holding behavior for the named target and for other objects for the 5 s prior to and after successful and unsuccessful naming events. In brief, the infant's sustained manual actions on objects are strongly indicative of optimal moments for learning object names.

If sustained visual selection is critical to infant learning, then learning may also depend on the quieting of head movements to stabilize the selected object in the visual field. Fig. 9 shows the percentage of time that infants and adults were moving their head during successful, non-successful, and non-naming events. For both head orientation and position and for both parents and infants, successful naming events are characterized by *less* head movement, suggesting the importance of stabilized visual attention ($\beta_{orientation} = 3.12$, $p < 0.001$; $\beta_{position} = 1.45$, $p < 0.001$). The fact that both parents and infants stabilized attention on the named object during successful but not unsuccessful naming events again points to coordinated or joint attention at the sensory-motor level. Considering the evidence on hands and heads together, successful naming events in the present context appear to have the following properties: During successful naming events, infants tend to hold the target object and visually isolate that object for some time before and after it is named, and in doing so, they stabilize head movements, maintaining this visual dominance of the selected object. During successful naming events, parent tend, immediately prior to the naming event, to move their head toward the named object and to hold the head steady at that moment, directed at the named object, but this increased visual dominance of the named object for the parent does not last and is localized to the naming event itself. Unsuccessful naming events have a different character, one in which both manual and visual attention on the part of the infant is more transient and one in which the visual field is more cluttered with other objects as large in the view as the named object. Both child's and parent's head movements may also reflect this greater clutter and more transient attention during non-successful naming events as infants and parents are less likely to move their head toward the target object and less likely to stabilize the head.

3.2.2. Summary

Table 1 summarizes sensory-motor patterns extracted from both successful and unsuccessful naming moments.

Table 1
Summary of sensory-motor patterns extracted from both successful and unsuccessful naming moments.

Data source	Descriptions of measurement	Agent	Results (S: successful; UnS: unsuccessful)
Head-camera images	Object size during naming (Fig. 5)	Child	S: The named object was much larger than other objects UnS: No difference
		Parent	S: The named object was much larger than other objects UnS: No difference
	At what moment the named object became visually dominant before naming (Fig. 6)	Child	S: 3.20 s UnS: Never happened
		Parent	S: 2.50 s UnS: Never happened
	How long the visual dominance of the named object lasted after naming (Fig. 6)	Child	S: 900 ms UnS: Never happened
		Parent	S: 500 ms UnS: Never happened
Holding actions	Holding the named versus other objects during naming (Fig. 7)	Child	S: Holding the named object more UnS: Holding the named object more
		Parent	S: No difference UnS: Holding other objects more
	How early holding the named object more before naming (Fig. 8)	Child	S: 4.00 s UnS: 2.50 s
		Parent	S: Never happened UnS: Never happened
	How long still holding the named object more after naming (Fig. 8)	Child	S: 5.00 s UnS: 1.20 s
		Parent	S: Never happened UnS: Never happened
Head movements	Positional movement during naming (Fig. 9)	Child	S: More stable UnS: No difference
		Parent	S: More stable UnS: No difference
	Orientational movement during naming (Fig. 9)	Child	S: More stable UnS: No difference
		Parent	S: More stable UnS: No difference

The main finding is that naming events that lead to learning have a visual signature, one in which the named object was visually dominant over possible competitor objects in the learner's view, and thus one in which there was minimal visual ambiguity as to the intended referent. These visually optimal moments for object-name learning were most closely associated with the infant's own actions – holding objects, bringing them close to the head, quieting of head movements. Parents named objects both when a single object dominated in the infant's view and when it did not; but, other aspects of parents' behaviors – moving and orienting the head toward the named object and stabilizing the head – were associated with successful naming, indicating that parents also distinguished these optimal moments for learning.

4. General discussion

The problem of referential uncertainty, a fundamental one for learners who must learn words from their co-occurrence with scenes, is reduced if object names are provided when there is but one dominating object in the learner's view. The present results show that infants often create these moments through their own actions and that object naming during these visually optimal moments is associated with learning. The present results are descriptive and correlational; therefore the implicated causal pathways must be considered as hypotheses in need of experimental test. However, the finding that 1½ year olds often visually isolate individual objects for extended periods and that they learn object names when naming coincides with such clean sensory data raises new questions and implications relevant to (1) theories of early word learning; (2) the embodiment of attention; (3) joint attention and social learning; and (4) the sensory-motor microstructure of cognition. We consider these in turn.

4.1. Early word learning

Most theoretical approaches (Frank et al., 2009; Smith & Yu, 2008; Snedeker & Gleitman, 2004; Waxman & Booth, 2001) to early word learning assume that mapping heard words to seen objects is fraught with referential uncertainty. However, the present results show that for 1½ year olds – infants who are in the midst of learning the names of everyday objects – there are moments within which referential uncertainty is significantly reduced at the sensory level. Learning during these moments would seem to require little cognitive work with respect to figuring out the intended referent. The evidence from many highly controlled experiments clearly indicate human infants have cognitive skills through which they can infer the intended referent given ambiguous data (Bloom, 2000; Smith & Yu, 2008; Swingley, 2009; Waxman & Booth, 2001). Nonetheless, early object-name learning may not depend solely on these advanced cognitive skills. Instead, the present results raise the possibility that during early stages of learning and outside of the laboratory in the dynamically complex and visually cluttered environment of everyday life, most object name learning may be the result of naming at optimal visual moments, when there is little referential ambiguity.

Indeed, it may be premature to conclude that clean sensory input is not necessary even in contexts in which children are inferring referential intent from the speaker's actions (Akhtar & Tomasello, 2000; Baldwin & Moses, 1996), when making inferences from linguistic and/or conceptual cues (Hall & Waxman, 2004; Markman, 1990) or making inferences based on statistical evidence across multiple encounters with the word (Smith & Yu, 2008). The evidence on these advanced skills in infants is derived mostly from laboratory experiments using discrete trials, uncluttered tabletops, and no measure of the personal view of the infants. Thus, it is possible that this inference making would not be robust in contexts of high visual clutter but might instead require that the *sensory* input to the cognitive machinery be unambiguous at least with respect to the object under consideration (for possibly related ideas, see Farzin, Rivera, & Whitney, 2010; Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005; Medina, Snedeker, Trueswell, & Gleitman, 2011). A recent finding reported by Yu and Smith (2011) in a study of statistical word-referent learning provides some support for this conjecture. Fifteen month olds were presented with a series of individually ambiguous learning trials, with two objects and two names presented per trial and no information about which object went with which name, a paradigm that had been used in a previous study to demonstrate infants' ability to aggregate and statistically evaluate word-referent co-occurrences (Smith & Yu, 2008). The newer study tracked infants' eye-gaze during the learning trials. The looking pattern by individual infants who did and did not learn suggested that eventual statistical learning required early trials in which the infant isolated the target object when the name was heard. Thus, selectivity at input, engendered by the infant's own actions of gaze direction, head movements, or – in active play contexts – hand movements, may be important to early word learning because these active movements effectively reduce the ambiguity in the input.

Clearly, any theory of word learning must consider the input. But the only relevant input is that which makes contact with the learner's sensory system. Most theories of word learning also recognize that learning moments vary in their quality with some being more ambiguous than others, and some leading to learning and some not. Quality and effectiveness as dimensions of the input have been most systematically considered in terms of the linguistic (Mintz, Newport, & Bever, 2002), conceptual (Markman, 1990), and social information (Baldwin, 1993) in the learning context. The present results suggest that quality and effectiveness need to also be considered at the level of the sensory input and in terms of the infant's own personal view.

The real world is much more visually ambiguous than the present experimental context in which parents and infants interacted with just three objects at a time. But the present results suggest that this experimental simplification was not enough in and of itself to guarantee learning the intended object names. Instead, learning depended on the infants' own actions which further simplified and

cleaned up the sensory input. Thus, the critical reduction for learning in the present study appears to have been from three potential referents in the child's view to one visually dominant object, and this reduction was implemented by the infants' own actions. In the context of the even greater ambiguity that characterizes natural learning contexts, the infants' reduction of the number of potential referents at the sensory level and through their own action may be even more critical (see a recent study by Yurovsky, Smith, and Yu (2012)).

4.2. Embodied attention

When infants bring objects close to their eyes and head, they effectively reduce the clutter and distraction in the visual field as close objects are visually large and block the view of potential distracters. This is a form of externally rather internally accomplished visual selection and it highlights how the early control of attention may be tightly linked to sensory-motor behavior. This is a particularly interesting developmental idea because many cognitive developmental disorders involve attention and because there is considerable evidence of co-morbidity of these cognitive disorders with early usual sensory-motor patterns (Hartman, Houwen, Scherder, & Visscher, 2010).

Experimental studies of adults show that the mature system can select and sustain attention on a visual target solely through internal means, without moving any part of the body and while eye gaze is fixated elsewhere (e.g. Müller, Philiastides, & Newsome, 2005; Shepherd, Findlay, & Hockey, 1986). However, visual attention is also usually linked to eye movements to the attended object's location (Hayhoe & Ballard, 2005). Moreover, eye movements (Grosbras, Laird, & Paus, 2005; Rizzolatti, Riggio, Dascola, & Umiltá, 1987), head movements (Colby & Goldberg, 1992), and hand movements (Hagler Jr., Riecke, & Sereno, 2007; Thura, Hadj-Bouziane, Meunier, & Boussaoud, 2008) have been shown to bias visual attention – detection and depth of processing – in the direction of the movement. This link between the localization of action and the localization of visual attention may be revealing of the common mechanisms behind action and attention as indicated by growing neural evidence that motor planning regions play a role in cortical attentional networks (Hagler Jr. et al., 2007; Kelley, Serences, Giesbrecht, & Yantis, 2007; Knudsen, 2007). Perhaps for physically active toddlers, visual attention is more tightly tied to external action and with development these external mechanisms become more internalized.

In this context, we note a limitation of the present study, the lack of information on the momentary direction of eye gaze. The first-person view, moment-to-moment, is central to understanding attention and learning at the micro-level. This view changes with every shift in eye gaze, every head turn, and with hand actions on an object. Here, we have provided information about heads and hands, and show that the content of the *head-centered* visual field predicts word learning by toddlers. Since a stabilized head-centered view with a single dominant object predicts learning, it seems likely that the head and eyes were aligned during successful naming moments. But we do not have evidence on the finer-grained information of the infant's specific focus within that larger head-centered field nor fine-grained temporal information about how clutter in the visual field and nearby competitors, may lead to shifts in eye-gaze direction and then to shifts in head direction and in these ways destabilize attention.

We also do not have information on the role that eye-gaze direction plays in social cuing. Evidence from adults (Hanna & Brennan, 2007; Kreysa & Knoeferle, 2010; Richardson, Dale, & Tomlinson, 2009; Shockley, Richardson, & Dale, 2009) demonstrates that the momentary eye gaze direction of a social partner disambiguates potential referents for mature listeners rapidly, within the time frame of milliseconds, making the social partner's momentary eye gaze an important component of online word-referent mapping. Although the evidence indicates that infants follow eye gaze and that this relates to language learning (Brooks & Meltzoff, 2005), much less is known about the temporal dynamics of eye-gaze following in complex social contexts in which heads and bodies are continually moving. Most experiments on the following of eye gaze, in infants and adults, manipulate eye-gaze direction in a straight-on face (see Langton, Watt, and Bruce (2000), for a review). However, in natural contexts, heads and eyes can move together or independently (Einhäuser et al., 2007); adults, children and infants are known to have difficulty ignoring the direction of the head in judging eye gaze direction (Corkum & Moore, 1998; Doherty & Anderson, 1999, 2001; Doherty, Anderson, & Howieson, 2009; Langton et al., 2000; Loomis, Kelly, Pusch, Bailenson, & Beall, 2008; Moore & Corkum, 1994). The needed next step to understand the dynamics of eyes and heads in toddlers' embodied attention and to understand the roles of heads and eyes in parent–child social coordination requires head-mounted eye trackers on both participants so as to capture both the head-centered view and the dynamics of eye-gaze within that view. Ongoing but rapid advances in the development and use of head-mounted eye-trackers with active toddlers suggest that this is possible (Franchak, Kretch, Soska, & Adolph, 2011).

4.3. Social learning

Toddlers cannot learn object names by themselves (Baldwin, 1993; Baron-Cohen, 1997; Bloom, 2000; Woodward, 2004). The parents in the present study provided object names both at optimal sensory moments and also at less optimal moments. Analyses of infant and parent actions suggest that when parents supplied object names at optimal moments they were following their infant's lead and interest in the attended object, a pattern of responsivity on the part of the parent that has been linked to successful word learning in previous research (Bornstein, Tamis-LeMonda, Hahn, & Haynes, 2008; Gros-Louis, West, Goldstein, & King, 2006; Miller, Ables, King, & West, 2009). Parents may have provided object names at less optimal moments because they were trying to lead the infant's attention instead, or they were trying to "follow" the infant's lead but misread the degree to which the infant's interest would persist or because they were unable to "see" the visual clutter in the infant's view that led to more transient attention. By this account, parent head movements toward the named

object during successful moments could have emerged – not because parents knew in some way that these were optimal moments – but because they were dynamically tracking their infant's attentional shifts, and thus moving their eyes and heads moment by moment to the object to which their infant was attending. However, it is also possible that parents offered names for different reasons with different goals, and that these different goals were in part indicated by the different patterns of head movements at the moment of naming. Thus, parents may play a larger role in providing more and cleaner input for the infant's word learning processes than is apparent in the present data and infants may be cuing parents in ways not evident in the present analyses. These roles of children and parents in creating and exploiting optimal sensory moments may also change with development. Therefore, the present results demonstrate the value of richer analyses of parent and infant behavior with respect to these clean visual moments in which one object is dominant in the child's view.

The literature provides a long list of bodily actions that may be relevant to orchestrating these optimal visual moments. Evidence from discrete-trial and highly controlled laboratory experiments makes it clear that very young children are highly sensitive to momentary eye-gaze direction, points, and other manual gestures as cues to the intended referent (Akhtar & Tomasello, 1997, 2000; Baldwin, 1993; Baron-Cohen, 1997; Csibra & Gergely, 2006; Gergely, Nádasdy, Csibra, & Bíró, 1995; Woodward, 2004). However, in real-world social interactions, and in the social context of the present design, the interaction is not made up of discrete trials but unfolds in time, with each moment building on the past activity of the individual and the past activity of the social partner (de Barbaro, Chiba, & Deák, 2011; Goldstein & Schwade, 2008; Sebanz & Knoblich, 2009; Shockley et al., 2009). Studies of dynamic coordination between adults in these free-flowing contexts indicate a complex interplay between various bodily cues, including important roles for such subtle movements as bodily sway, mouth openings, posture, and very small head movements. Perhaps, more critically, these studies reveal a rhythm and entrainment of the social partner that is evident in the durations and amplitudes of speech rate, turn duration, and bodily movements (Sebanz, Bekkering, & Knoblich, 2006; Shockley, Santana, & Fowler, 2003). Recent studies of parent–toddler free-flowing interactions also suggest a role for a complex set of cues including head direction, vocal intensity, object holding behavior, the spatial segregation of objects in the play area, and rhythmic bodily movements (Amano, Kezuka, & Yamamoto, 2004; Liszkowski, Carpenter, Henning, Striano, & Tomasello, 2004; Yoshida & Smith, 2008; Yu et al., 2009). Thus in ways yet to be discovered, parent actions may show signs of sensitivity to and may also play a role in coordinating their infant's bodily orientation to objects and thus may foster these optimal visual moments for learning.

4.4. Going micro

Children learn the names of objects in which they are interested. Therefore, as shown in Fig. 10a, "interest", as a macro-level concept, may be viewed as a driving force behind learning (Bloom, Tinker, & Scholnick, 2001). Given this, what is the new contribution of the present study? One might argue that the main result is that infants learn object names when they are *interested* in those objects: that holding an object and a one-object view are merely indicators of the infant's interest in the object. That is, the cause of learning may not be the lack of visual clutter at the moment of object naming, but be the child's interest in the object which happens to be correlated with the not causally relevant one-object view. By this argument (as shown Fig. 10b), the results show only that infants learn the names of things in which they are interested more readily than the names of things for which they have little interest; visual selection at the sensory level is merely associated attributes but not essential to nor contributory to learning. From this perspective, the present study has gone to a lot of trouble and a lot of technology to demonstrate the obvious. Although we disagree with this view, the proposal that our measures of image size and holding are measures of infants' interest in the target object and that the results show that infants learn when they are interested in an object seems absolutely right to us. What the present results add to the macro-level construct of "interest" is two alternative explanations shown in Fig. 10c and d. First, the present study may provide a mechanistic explanation at a more micro-level of analysis of why "interest" matters to learning. As proposed in Fig. 10c, interest in an object by a toddler may often *create* a bottom-up sensory input that is clean, optimized on a single object and sustained. Interest may mechanistically yield better learning (at least in part) *because* of these sensory consequences. Therefore, at the macro-level, one may observe the correlation between learning and interest; at the micro-level, the effect of interest on learning may be implemented through clean sensory input, and through perceptual and action processes that directly connect to learning. Fig. 10d provides a more integrated version of these ideas: interest may initially drive learning (through a separate path); and interest may also drive the child's perception and action – which feed back onto interest with sustained attention to support learning. That is, interest may drive actions and the visual isolation of the object and thus increase interest. These sensory-motor behaviors may also directly influence learning by localizing and stabilizing attention and by limiting clutter and distraction. In brief, the micro-level analyses presented here are not in competition with macro-level accounts but offer new and testable hypotheses at a finer grain of mechanism – moving forward from Fig. 10a to d.

One new hypothesis is that *visual* clutter itself may disrupt learning. A second hypothesis is that sustained sensory isolation of the named referent may be necessary for learning. That is, the visual dominance of the named object for a *short duration* just at the moment of naming may not be sufficient for toddlers to learn an object name; instead, sustained sensory isolation of the target some time prior to and after naming may be critical to bind the name to the object. These two hypotheses make clear the potential value of considering word learning at the sensory-motor level. The co-morbidity of motor and cognitive developmental disorders (Hartman et al., 2010; Iverson, 2010; Mostofsky

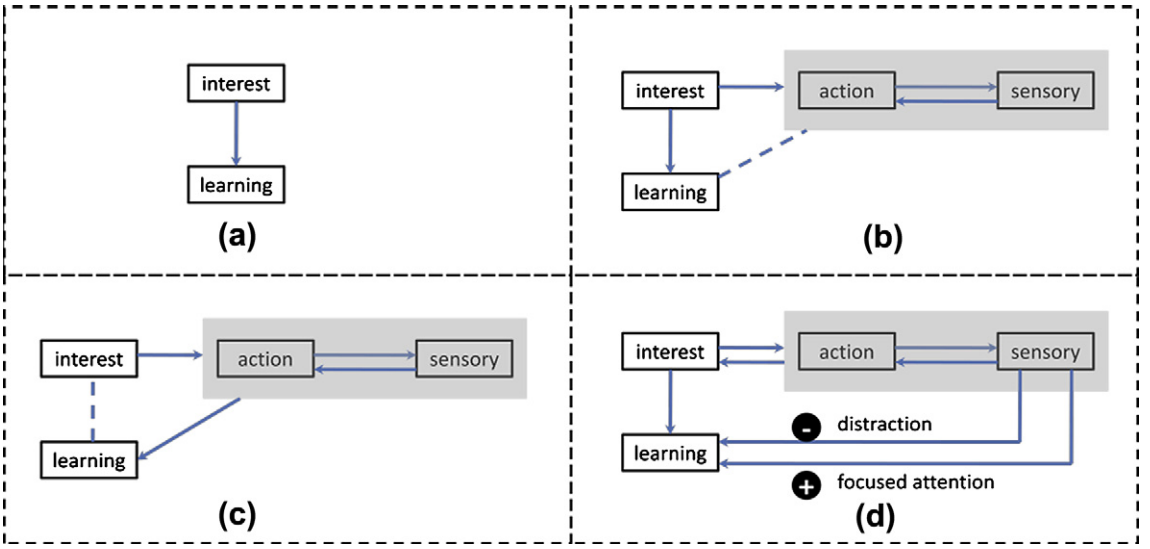

Fig. 10. Four hypotheses on child's interest, learning and sensory-motors behaviors. (a) Child's interest on target objects leads to learning. (b) Child's interest drives both learning and sensory-motors behaviors. Therefore, there are correlations between the two (the dotted line). (c) Child's interest leads to a sequence of actions on the interested object (e.g. holding and manipulating) which then lead to the visual dominance of that object. This clean visual input is fed into internal learning processes. In this way, child's interest is indirectly correlated to learning (then dot line) because interest is implemented through child's perception and action which directly connect to learning. (d) Initially, child's interest directly influences both learning and as well as sensory-motors behaviors. Thereafter, sensory-motors behaviors also directly influence learning (and maybe interest itself as well) as sustained attention on the target object may facilitate learning while distracting and messy sensory input may disrupt learning. In this way, both child's interest and sensory-motor behaviors jointly influence learning.

et al., 2009), and the link between abnormal movement patterns and poor attentional control in children (Mostofsky et al., 2006; Tillman, Thorell, Brocki, & Bohlin, 2007) are also well-known but not well-understood. Linking important macro-level achievements – such as mapping a name to an object – may be crucial in understanding the developmental dependencies between sensory-motor processes and early cognitive development. Research programs that attempt to cross and integrate the micro and macro might not only reveal these cross-level and cross-time scale dependencies but also provide translatable work-around solutions to the benefit of intervention. For example, if interest in an object were the primary driver of learning, but if interest by toddlers mechanistically benefited learning primarily through the sensory reduction of distractors, then we could focus intervention efforts not just on interest or motivational levels but also on artificially isolating the relevant object for the learner.

Thus, the findings also contribute by providing a more detailed and more quantitative description of sensory-motor behavior and its effects on word learning. One of our main conclusions is that the *visual* dominance of the named object matters. But we know more than that: we know the child creates moments of visual dominance through his bodily actions on objects; we also know the exact size of named objects in the child's egocentric view, and we know the temporal dynamics of object sizes before, during and after naming moments. Such detailed results open up new and potentially deep questions: for instance, is visual dominance with respect to word learning better understood in terms of absolute visual size or in terms of relative size with respect to competitors? If absolute size is critical, it may be an indicator of the processes in an immature system that are needed for the multisensory binding of a visual event to an auditory one. And, if it were absolute size it would predict that children would actively bring smaller objects closer to the eyes than larger ones. Alternatively, if relative dominance matters, then the key processes could concern competition in the visual system. Further, if objects compete for attention, then does dominance indicate a winner-take-all-like selection – such that at any moment, the largest object in view is considered as the only candidate referent, or is the process more probabilistic wherein each object gains a certain probability to be linked to the heard word and that probability is proportional to the visual saliency of that object? Are the dynamics of visual isolation of the target relative to naming – with the isolation (and thus potential representation of the object in memory prior to naming and sustained after naming) critical to binding the name and object? And, does it matter whether visual dominance is created by child or by parent? If visual dominance in the infant's view is necessary for word learning, it may not matter how one achieves that dominance. In interactive social play with equally sized objects, it may happen to be mostly the infant's manual actions that are the proximal cause. These are specific and answerable questions suggested by the present results and these questions are critical to a more complete understanding of the visual, attentional and memorial processes that support early word learning.

In conclusion, we began by noting that the challenge of the infant attempting to learn word-object mappings from

word-scene co-occurrences could be potentially solved at the sensory input level and that the assumption of referential ambiguity – an assumption that defines the major theoretical problem in early word learning – might be exaggerated. The results show that infants often create visual moments in which only one object is in their view and object name learning is strongly associated with naming events that occur during those moments. The main contribution of the present research, then, is that it suggests a bottom-up sensory solution to word-referent learning by toddlers. Toddlers, through their own actions, often create a personal view that consists of one dominating object. Parents often (but not always) name objects during these optimal sensory moments and when they do, toddlers learn the object name.

Acknowledgments

We thank Charlotte Wozniak, Amanda Favata, Alfredo Pereira, Amara Stuehling, and Andrew Filipowicz for collection of the data, Thomas Smith for developing data management and preprocessing software. We would also like to thank Scott Johnson and three anonymous reviewers for insightful comments and suggestions. This research was supported by National Science Foundation Grant 0924248 and AFOSR FA9550-09-1-0665.

Appendix A

A.1. Image processing

Our experimental setup significantly simplified the following image processing compared with other computer vision applications. However, there are two special challenges we faced with in our current setup. First, the quality of images captured from mini-cameras is limited due to the small size of the cameras. More specifically, the automatic gain control and the white balance functions in those cameras are always on which sometimes cause dramatic changes in "color temperature" frame by frame when participants moved their heads. The same object may look quite differently due to the automatic adjustment of the camera to compensate for "color temperature" changes caused by head movement. Second, the compositions of images from the two first person views (especially from the infant's camera) are quite different, compared from

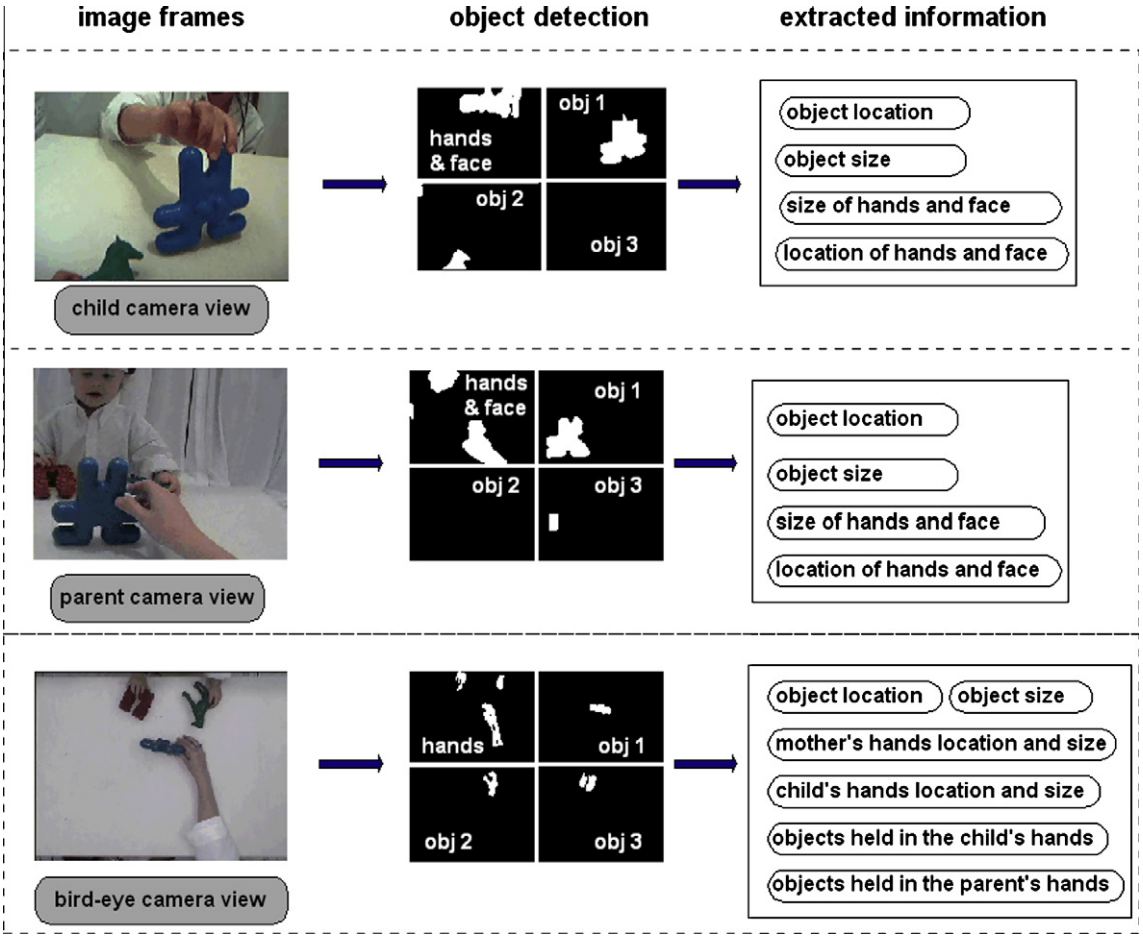

Fig. 11. The overview of data processing using computer vision techniques. Our program can detect three objects on the table and participants' hands and faces automatically based on pre-trained object models and skin models. The extracted information from three video streams will be used in subsequent data analyses.

images used in most other computer vision applications, in two unique ways: (1) Every object in the first-person camera was captured from a zoomed-in view. Therefore, the size of object was much bigger due to the close distance to the camera compared with standard images captured from a distance in most cases. (2) An object in head-camera images was always partitioned into several blobs due to the overlapping with other objects and hands.

The general image processing consists of two steps as shown in Fig. 11. We first pre-select 25 images per object and ask human coders to annotate those objects by clicking along the boundary of a desired object and then indicate its identity. We have developed a training program that takes the annotation information and builds a color histogram representation of each instance of an object.

Next, given a set of feature vectors based on color histogram, we cluster those vectors in a feature space and find a set of prototypes for each object. The Hierarchical clustering algorithm (Duda, Hart, & Stork, 2001) is used to group those vectors into a set of clusters. The center of each cluster is then calculated and used as a prototype. Moreover, we also assign a weight of each prototype based on the proportion of vectors that belong to this cluster. The outcome from training for each object is a set of vectors and weights. Next, given a new image frame, our image processing method is composed of two steps. First, the raw image is segmented into several blobs based on color constancy. Second, each blob is examined one by one and assigned to an object label based on the comparison with the color histogram representation extracted from a blob with the prototypes of objects from training. More specifically, we use earth-mover distance (Rubner, Tomasi, & Guibas, 2000) as a metric to compare two color histograms – a prototype from training and the color histogram extracted from the current blob. The central idea of the earth-mover distance is to take into account of the similarity between neighbor bins instead of treating them independently in histogram comparison. In this way, each blob extracted from image segmentation is assigned with either an object label or as background. More technical details can be found in (Yu et al., 2009).

References

Akhtar, N., & Tomasello, M. (1997). Young children's productivity with word order and verb morphology. *Developmental Psychology, 33*(6), 952–965.

Akhtar, N., & Tomasello, M. (2000). The social nature of words and word learning. In R. M. Golinkoff, K. Hirsh-Pasek, L. Bloom, L. Smith, A. Woodward, N. Akhtar, M. Tomasello & G. Hollich (Eds.), *Becoming a word learner*: *A debate on lexical acquisition* (pp. 115–135).

Allopenna, P., Magnuson, J., & Tanenhaus, M. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*(4), 419–439.

Amano, S., Kezuka, E., & Yamamoto, A. (2004). Infant shifting attention from an adult's face to an adult's hand: A precursor of joint attention. *Infant Behavior and Development, 27*(1), 64–80.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412.

Baldwin, D. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology, 29*(5), 832–843.

Baldwin, D., & Moses, L. (1996). The ontogeny of social information gathering. *Child Development, 67*(5), 1915–1939.

Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: The MIT Press.

Bates, D., & Sarkar, D. (2007). *Lme4: Linear mixed-effects models using S4 classes*. Madison: University of Wisconsin. Manuscript.

Blake, R., Tadin, D., Sobel, K. V., Raissian, T. A., & Chong, S. C. (2006). Strength of early visual adaptation depends on visual awareness. *Proceedings of the National Academy of Sciences of the United States of America, 103*(12), 4783–4788.

Bloom, L., Tinker, E., & Scholnick, E. K. (2001). *The intentionality model and language acquisition: Engagement, effort, and the essential tension in development*. Wiley-Blackwell.

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

Bornstein, M. H., Tamis-LeMonda, C. S., Hahn, C. S., & Haynes, O. M. (2008). Maternal responsiveness to young children at three ages: Longitudinal analysis of a multidimensional, modular, and specific parenting construct. *Developmental Psychology, 44*(3), 867–874.

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

Colby, C., & Goldberg, M. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. *Science, 255*(5040), 90.

Corkum, V., & Moore, C. (1998). Origins of joint visual attention in infants. *Developmental Psychology, 34*(1), 28.

Csibra, G., & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. *Processes of change in brain and cognitive development. Attention and performance, XXI*, 249–274.

de Barbaro, K., Chiba, A., & Deák, G. O. (2011). Micro-analysis of infant looking in a naturalistic social setting: Insights from biologically based models of attention. *Developmental Science, 14*(5), 1150–1160. http://dx.doi.org/10.1111/j.1467-7687.2011.01066.x.

Doherty, M. J., & Anderson, J. R. (1999). A new look at gaze: Preschool children's understanding of eye-direction. *Cognitive Development, 14*(4), 549–571.

Doherty, M. J., & Anderson, J. R. (2001). People don't keep their heads still when looking to one side, and other people can tell. *PERCEPTION-LONDON, 30*(6), 765–768.

Doherty, M. J., Anderson, J. R., & Howieson, L. (2009). The rapid development of explicit gaze judgment ability at 3 years. *Journal of Experimental Child Psychology, 104*(3), 296–312.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software, 20*(2), 1–18.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification*. Citeseer.

Einhäuser, W., Schumann, F., Bardins, S., Bartl, K., Böning, G., Schneider, E., et al. (2007). Human eye-head co-ordination in natural exploration. *Network: Computation in Neural Systems, 18*(3), 267–297.

Farzin, F., Rivera, S. M., & Whitney, D. (2010). Spatial resolution of conscious visual perception in infants. *Psychological Science, 21*(10), 1502–1509.

Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2011). Head-mounted eye tracking: A new method to describe infant looking. *Child Development, 82*(6), 1738–1750.

Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science, 20*(5), 578–585.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*(2), 165–193.

Gershkoff-Stowe, L., & Smith, L. (1997). A curvilinear trend in naming errors as a function of early vocabulary growth. *Cognitive Psychology, 34*(1), 37–71.

Gleitman, L., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. (2005). Hard words. *Language Learning and Development, 1*(1), 23–64.

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science, 19*(5), 515–523.

Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development, 30*(6), 509–516.

Grosbras, M. H., Laird, A. R., & Paus, T. (2005). Cortical regions involved in eye movements, shifts of attention, and gaze perception. *Human Brain Mapping, 25*(1), 140–154.

Hagler, D., Jr., Riecke, L., & Sereno, M. (2007). Parietal and superior frontal visuospatial maps activated by pointing and saccades. *Neuroimage, 35*(4), 1562–1577.

Hall, D., & Waxman, S. (2004). *Weaving a lexicon*. The MIT Press.

Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language, 57*(4), 596–615.

Hart, B., & Risley, T. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Brookes Publishing Company, Inc., PO Box 10614. 21285-0624 ($22).

Hartman, E., Houwen, S., Scherder, E., & Visscher, C. (2010). On the relationship between motor performance and executive functioning in children with intellectual disabilities. *Journal of Intellectual Disability Research, 54*(5), 468–477.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*(4), 188–194.

Haykin, S. (2001). *Kalman filtering and neural networks*. Wiley-Interscience.

Hirsh-Pasek, K., & Golinkoff, R. (1996). The intermodal preferential looking paradigm: A window onto emerging language comprehension. *Methods for Assessing Children's, Syntax, 1*, 105–124.

Hirsh-Pasek, K., Golinkoff, R., Berk, L., & Singer, D. (2009). *A mandate for playful learning in preschool: Presenting the evidence*. Oxford University Press.

Iverson, J. M. (2010). Developing language in a developing body: The relationship between motor development and language development. *Journal of Child Language, 37*(02), 229–261.

Jovancevic-Misic, J., & Hayhoe, M. (2009). Adaptive gaze control in natural environments. *The Journal of Neuroscience, 29*(19), 6234–6238.

Kelley, T., Serences, J., Giesbrecht, B., & Yantis, S. (2007). Cortical mechanisms for shifting and holding visuospatial attention. *Cerebral Cortex*.

Knudsen, E. (2007). Fundamental components of attention. *Annual Review of Neuroscience, 30*(1), 57–78.

Kreysa, H., & Knoeferle, P. (2010). Using speaker gaze for thematic role assignment in language comprehension.

Langton, S. R. H., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences, 4*(2), 50–59.

Liszkowski, U., Carpenter, M., Henning, A., Striano, T., & Tomasello, M. (2004). Twelve-month-olds point to share attention and interest. *Developmental Science, 7*(3), 297–307.

Loomis, J. M., Kelly, J. W., Pusch, M., Bailenson, J. N., & Beall, A. C. (2008). Psychophysics of perceiving eye-gaze and head direction with peripheral vision: Implications for the dynamics of eye-gaze behavior. *Perception, 37*(9), 1443–1457.

Markman, E. (1990). Constraints children place on word meanings. *Cognitive Science. A Multidisciplinary Journal, 14*(1), 57–77.

Masur, E. (1997). Maternal labelling of novel and familiar objects: Implications for children's development of lexical constraints. *Journal of Child Language, 24*(02), 427–439.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences, 108*(22), 9014.

Miller, J. L., Ables, E. M., King, A. P., & West, M. J. (2009). Different patterns of contingent stimulation differentially affect attention span in prelinguistic infants. *Infant Behavior and Development, 32*(3), 254–261.

Mintz, T., Newport, E., & Bever, T. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science. A Multidisciplinary Journal, 26*(4), 393–424.

Moore, C., & Corkum, V. (1994). Social understanding at the end of the first year of life. *Developmental Review*.

Mostofsky, S. H., Dubey, P., Jerath, V. K., Jansiewicz, E. M., Goldberg, M. C., & Denckla, M. B. (2006). Developmental dyspraxia is not limited to imitation in children with autism spectrum disorders. *Journal of the International Neuropsychological Society, 12*(03), 314–326.

Mostofsky, S. H., Powell, S. K., Simmonds, D. J., Goldberg, M. C., Caffo, B., & Pekar, J. J. (2009). Decreased connectivity and cerebellar activity in autism during motor task performance. *Brain, 132*(9), 2413–2425.

Müller, J., Philiastides, M., & Newsome, W. (2005). Microstimulation of the superior colliculus focuses attention without moving the eyes. *Proceedings of the National Academy of Sciences, 102*(3), 524.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature, 434*(7031), 387–391.

Quine, W. (1964). *Word and object*. MIT Press.

Richardson, D. C., Dale, R., & Tomlinson, J. M. (2009). Conversation, gaze coordination, and beliefs about visual context. *Cognitive Science, 33*(8), 1468–1482.

Rizzolatti, G., Riggio, L., Dascola, I., & Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia, 25*(1), 31–40.

Rubner, Y., Tomasi, C., & Guibas, L. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision, 40*(2), 99–121.

Ruff, H., & Rothbart, M. (2001). *Attention in early development: Themes and variations*. USA: Oxford University Press.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences, 10*(2), 70–76.

Sebanz, N., & Knoblich, G. (2009). Prediction in joint action: What, when, and where. *Topics in Cognitive Science, 1*(2), 353–367.

Shepherd, M., Findlay, J., & Hockey, R. (1986). The relationship between eye movements and spatial attention. *The Quarterly Journal of Experimental Psychology Section A, 38*(3), 475–491.

Shockley, K., Richardson, D. C., & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science, 1*(2), 305–319.

Shockley, K., Santana, M., & Fowler, C. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology, 29*(2), 326–332.

Smith, L., Yu, C., & Pereira, A. (2010). Not your mother's view: The dynamics of toddler visual experience. *Developmental Science*.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3), 1558–1568.

Smith, L. B., Yu, C., & Pereira, A. F. (2011). Not your mother's view: The dynamics of toddler visual experience. *Developmental Science, 14*(1), 9–17.

Snedeker, J., & Gleitman, L. (2004). Why it is hard to label our concepts. In D. G. Hall & S. R. Waxman (Eds.), *Weaving a lexicon* (pp. 255–293). MIT Press.

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1536), 3617–3632.

Tamis-LeMonda, C. S., & Bornstein, M. H. (1994). Specificity in mother–toddler language–play relations across the second year. *Developmental Psychology, 30*(2), 283–292.

Tamis-LeMonda, C. S., Bornstein, M. H., & Baumwell, L. (2001). Maternal responsiveness and children's achievement of language milestones. *Child Development, 72*(3), 748–767. http://dx.doi.org/10.1111/1467-8624.00313.

Thelen, E., Corbetta, D., Kamm, K., Spencer, J. P., Schneider, K., & Zernicke, R. F. (1993). The transition to reaching: Mapping intention and intrinsic dynamics. *Child Development, 64*(4), 1058–1098.

Thura, D., Hadj-Bouziane, F., Meunier, M., & Boussaoud, D. (2008). Hand position modulates saccadic activity in the frontal eye field. *Behavioural Brain Research, 186*(1), 148–153.

Tillman, C. M., Thorell, L. B., Brocki, K. C., & Bohlin, G. (2007). Motor response inhibition and execution in the stop-signal task: Development and relation to ADHD behaviors. *Child Neuropsychology, 14*(1), 42–59.

Tomasello, M., & Farrar, M. (1986). Joint attention and early language. *Child Development, 57*(6), 1454–1463.

Waxman, S., & Booth, A. (2001). Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. *Cognitive Psychology, 43*(3), 217–242.

Woodward, A. (2004). Infants' use of action knowledge to get a grasp on words. *Weaving a Lexicon*, 149–172.

Yoshida, H., & Smith, L. B. (2008). What's in view for toddlers? Using a head camera to study visual experience. *Infancy, 13*(3), 229–248.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*(5), 414–420.

Yu, C., & Smith, L. B. (2011). What you learn is what you see: Using eye movements to study infant cross-situational word learning. *Developmental Science, 16*(2), 165–180.

Yu, C., Smith, L. B., Shen, H., Pereira, A., & Smith, T. (2009). Active information selection: Visual attention through the hands. *IEEE Transactions on Autonomous Mental Development, 2*, 141–151.

Yurovsky, D., Smith, L. B., & Yu, C. (2012). Does Statistical Word Learning Scale? It's a Matter of Perspective. In *Paper presented at the proceedings of the 34th annual conference of the cognitive science society*.