14
LINKING WORDS TO WORLD
An embodiment perspective

Chen Yu

Introduction

Children begin to comprehend words at 9 months. They say their first word at around 12 months. The pace with which children add new words to receptive and productive vocabulary then accelerates such that by 24 to 30 months, children add words at the staggering rate of 5 to 9 new words a day (Bloom, 2000). Just as in many other cognitive learning tasks, a critical problem in word learning is the uncertainty and ambiguity in the learning environment – young word learners need to discover correct word-referent mappings among many possible candidate words and many possible candidate referents from potentially many objects that are simultaneously available. For example, Quine (1960) famously presented the core problem for learning word meanings from their co-occurrence with perceived events in the world. He imagined an anthropologist who observes a speaker saying "gavagai" while pointing in the general direction of a field. The intended referent (rabbit, grass, the field, or rabbit ears, etc.) is indeterminate from this example.

Past work has shown that the social context plays a key role in young learners' prowess in disambiguating cluttered learning situations. The literature provides many powerful demonstrations of how social-interactional cues guide infants' word learning, and in many cases, these seem essential to successful learning (Baldwin, 1993; Tomasello and Akhtar, 1995; Woodward and Guajardo, 2002; Bloom, 2000). Often the importance of social cues is interpreted in terms of children's understanding that words are used with an "intent to refer." Thus children's early dependence on social cues is seen as a diagnostic marker of their ability to infer the intentions of the speaker. This kind of social cognition is called "mindreading" by Baron-Cohen (1997) or more generally "theory of mind" (Wellman and Liu, 2004). Consistent with these ideas, studies have shown that very young learners map nouns to objects only when the speaker is intentionally attending to the named object and not, for example, when there is an accidental co-occurrence of object and name (Tomasello, 2000). Such results point to the importance of understanding the social structure of learning experiences. However, there is much that is still not understood:

(1) At the behavioral level, most studies have examined early word learning in constrained experimental tasks with only one or two objects in view. The adult partner (usually the

experimenter) is focused on the child, on effective teaching, and provides clear and repeated signals of her attention to the object being named. In this way, the attentional task is simple, and easily described in discrete and categorical terms (the attended object vs. the distractor). These contexts are not at all like the real world in which word learning is embedded in *a stream of activity* – in which parents both react to and attempt to control toddler behaviors and in which toddlers react to, direct, and sometimes ignore parents as they pursue their own goals. If we are going to understand the role of social cues in real-world word learning, we need to study social interactions and learning as they unfold in real time in dynamically complex and cluttered contexts.

(2) At the theoretical level, the focus on macro-level behaviors and folk-psychological constructs does not connect easily to the real-time events in which learning happens. Current theoretical models explain the role of social cues in word learning via *internal* computations – mental models about the intentional states of the social partner and inferences about the goals and plans of the other (Breazeal and Scassellati, 2000). It is not at all clear that such abstract logic-like inferences about the internal states of others can happen *fast enough* to explain the exquisite real-time "dance" of social interactions in which effective adjustments within the dyad happen in fractions of seconds.

(3) At the computational level, the analysis of the learning task has been based on an *adult's* – and *third person's* – view of the learning environment. Experimenters and theorists of children's word learning are adults with a mature and developed view of the structure of the learning task and of the attentional and intentional states of the learner. As observers, we watch interactions between child and parent and we interpret these interactions from a vantage point that sees both the causes and effects of each action on each participant. It is seductively easy to describe such events in folk-psychological terms that sound like explanations: "the mother tried to elicit the child's attention by waving the toy," "the child wanted the toy and so reached for it." There are many philosophical, methodological, and theoretical problems with this (Pfeifer and Scheier, 1999). One straightforward problem is that the third-person observer's view (the experimenter's view, etc.) of the learning task is not the learner's view. Instead, what the learner sees – moment to moment – is a dynamic event that depends on the learner's own momentary interests and *bodily orientation*. Recent studies using head-mounted cameras indicate that the adult view of the learning task *does not align at all with the dynamic first-person view of toddlers*, and is, therefore, a poor basis for theorizing about underlying processes (Pereira, Smith, and Yu, in press; Smith, Yu, and Pereira, 2011; Yu, Smith, Shen, Pereira, and Smith, 2009).

In brief, traditional theories of learning and intelligence (and many contemporary theories of development, learning, and social interaction) concentrate on internal representations and inferences from those representations, paying little attention to the body and to the ways intelligence is affected by and affects the physical world. More recently, there has been a shift toward ideas of embodiment, that intelligence emerges in the interaction of an agent *and its body* with an environment (Brooks and Stein, 1994; Clark, 2008; Ballard, Hayhoe, Pook, and Rao, 1997; Pfeifer and Scheier, 1999; Gibbs, 2006; Shapiro 2011; Spivey, 2007). In these analyses, the body – its morphology and its own intrinsic dynamics – plays just as important role as the internal cognitive system and physical environment. Beer (1995) provided a principled theoretical analysis of these ideas in which behavior and cognition are understood as arising from the dynamical interaction between a brain (or cognitive system), body and environment which critically includes other brain-body-environment systems as shown in Figure 14.1 (left). From this perspective, the behavior and cognition *of an individual* may be conceptualized as arising from the

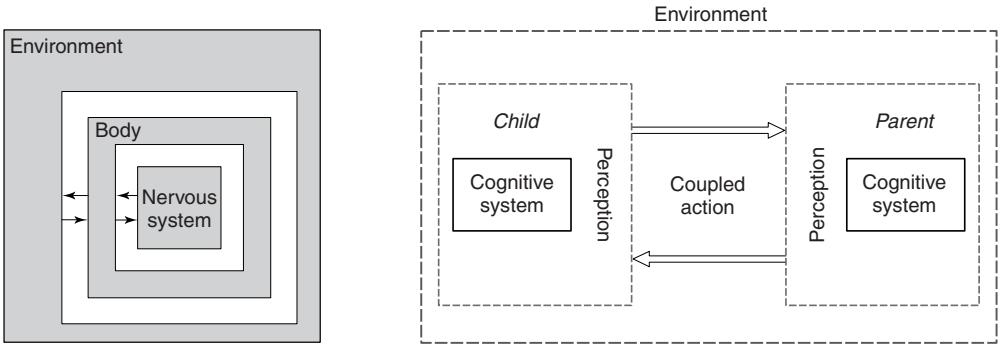

Figure 14.1 Overview of embodied social interaction. (Left) The brain-body-environment system in Beer (1995). (Right) Our proposed coupled embodied framework of child-parent interaction and communication.

closed-loop interaction of the cognitive system with the body and environment in which it is embedded, rather than as the sole product of any one component of this coupled system, such as the brain or internal representations. The behavior and collaboration of several individuals – for instance, word learning from child-parent social interaction – may be conceptualized as the *coupling* of these two systems as illustrated in Figure 14.1 (right).

Further, the critical role of embodiment has been demonstrated empirically and computationally in various behavioral science fields. For example, Ballard *et al.* (1997) proposed a model of "embodied cognition" that operates at timescales of approximately one-third of a second and uses subtle orienting movements of the body during a variety of cognitive tasks as input to a computational model. At this "embodiment" level, the constraints of the body determine the nature of cognitive operations, and the body's pointing movements are used as deictic (pointing) references to bind objects in the physical environment to variables in cognitive programs of the brain. In our studies of child word learning, we emphasize the dependencies between the learner's *own actions* and the learner's internal cognitive state. Accordingly, understanding how the sensorimotor dependencies in the child affect cognition and learning – how, for example, looking at an object, or holding it – may engage and maintain attention is viewed as critical. These sensorimotor-cognition couplings also mean that the momentary sensorimotor actions of the learner are likely to be indicative of the learner's internal state (Yu, Ballard, and Aslin, 2005). The basic idea behind our research is that the body – and its momentary actions – are crucial to social collaboration. Toward this goal, we seek to simultaneously measure multiple streams of behaviors and then to use data mining and machine learning techniques to *discover* patterns that support smooth interactions and word learning. We concentrate on measuring multiple streams of sensorimotor data because ultimately it is these coupled real-time behaviors that create the learning input and the learning environment.

A case study: multimodal word learning

Experiment

Here we describe a general experimental context that has been developed and used to understand word learning through parent-child interaction (Yu and Smith, 2012; Smith *et al.*, 2011). The experimental task is unconstrained tabletop play between a toddler (children between 15 and 20

months of age) and a parent. We chose the context of parent and child playing with toys on a tabletop as our naturalistic context for three reasons: (1) it is a common everyday context in which parents and toddlers are jointly engaged and in which word learning takes place (Callanan,1985, 1990); (2) it is attentionally complex in that there can be many objects on the table, multiple and competing goals, and many shifts in attention; and (3) the geometry of tabletop play is sufficiently constrained that we can measure the first-person view and the head and hand movements of each participant. Figure 14.2 shows the basic set-up. The interaction is recorded

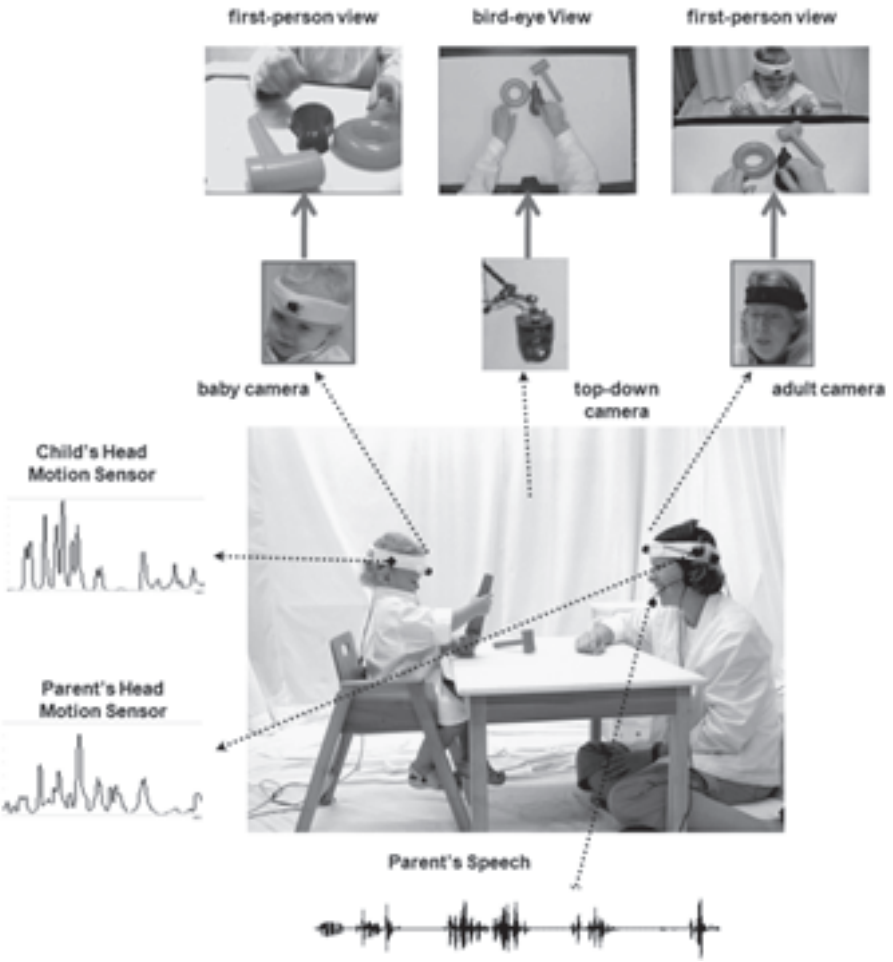

Figure 14.2 A multisensory system to collect multimodal data from child-parent interaction. The young word learner and the language teacher play with a set of objects at a table. Two mini cameras are placed onto the child's and the mother's heads respectively to collect visual information from two first-person views. Note that these two views are dramatically different. A third camera mounted high above the table records the bird's-eye view of the whole interaction. The participants also wore motion sensors to track their head and hand movements. A headset was used to record the caregiver's speech. In this way, we collected multimodal multistreaming data to analyze and detect interactive perception-action patterns from both the child and the parent that lead to successful word learning.

by three cameras: one head-mounted camera provides information about the scene from the child's point of view; a second head-mounted camera provides the parent's viewpoint; and a third from a top-down third-person viewpoint allows a clear observation of exactly what was on the table at any given moment (mostly the participants' hands and the objects being played with). We also measure both the child's and the parent's head and hand movements with a Polhemus 6 Degree-of-Freedom motion-tracking system and also the parent's speech through a headset. A particularly important and novel component of our method is the recording of visual information from the *learner's point of view* via a lightweight mini-camera mounted on a sports headband and placed low on the forehead. The angle of the camera is adjustable, and has a visual field of approximately 90 degrees, horizontally and vertically.

Parents were told that their goal was simply to engage the child with the toys and that they should interact as naturally as possible. The experimental objects were simple novel objects with novel names. Parents were taught the names prior to the experiment. Besides that, there were no constraints on what parents (or the children) had to say or what they had to do. Parents were told to engage their child with objects, to use the names we supplied *if* they named them, and that we were interested in the dynamics of parent-child play with toys. There were three toy objects in each of the three trials. At the end of interaction, we also tested each child's knowledge of the names of the nine objects that they played with, using a standard forced choice procedure. In this way, we used – as described above – completely novel methods of collecting multiple streams of sensorimotor data during the course of the learning experiences but we tied these measures to well-documented, standard, and highly reliable measures of word learning.

Video, motion-tracking, and speech data were coded to extract sensorimotor variables, such as object size in view, holding activities from both participants, and head stability. Technical details can be found in (Yu and Smith, 2012; Yu *et al.*, 2009). As a result, we have collected and extracted multiple time series that capture visual, motor, and speech behaviors moment by moment from both the child and the parent. Those derived data were further analyzed to discover various sensory and motor patterns from child-parent interactions.

Results

During the play session, parents uttered on average 365 words (tokens). Each of the 9 object names was produced by the parents on average only 5.32 times (standard deviation [SD] = 1.12). An object name was categorized as learned for an infant if his looking behavior at test indicated learning. 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. 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 us 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 "unsucessful" (n = 136). Recall that objects were presented in three sets of three. Successful and unsuccessful naming events did not differ in duration, nor in any other noticeable property.

Our first hypothesis was 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 *distractor* objects in the head-camera images. This provided a measure of the relative dominance of the referent of the object name and its visual competitors. The sizes of the target

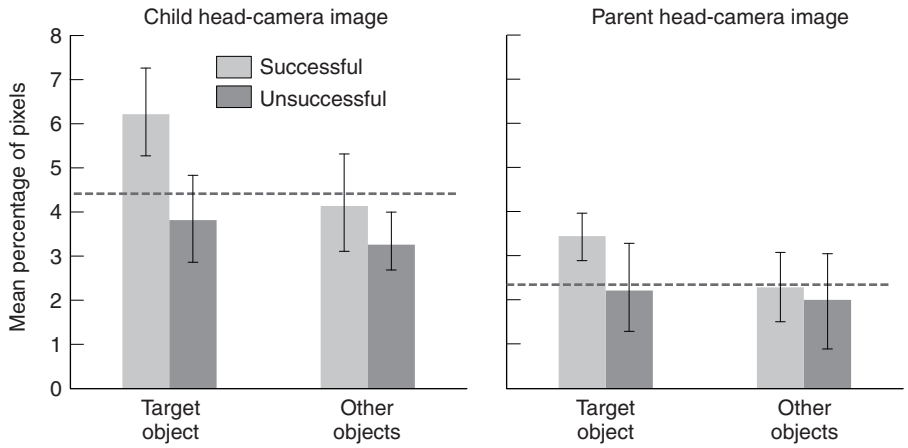

Figure 14.3 Mean object size (% pixels in image) for the named target, and for other objects in child's and parent's head-camera images during the naming event, for successful naming events that led to learning at post-test and for unsuccessful naming events that did not lead to learning as measured by test. Means and standard errors calculated with respect to trials. Dashed line indicates the mean object size during non-naming moments.

and other objects in both the infant and the parent head-camera views during naming events are shown in Figure 14.3. 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 (mean object size, $M_{\text{Successful}}$ = 6.28% of pixels in the image; object size is measured as the total number of object pixels divided by the head-camera image size) but the target object sizes for unsuccessful naming events did not ($M_{\text{Unsuccessful}}$ = 4.07%). 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%). We also examined whether these differences changed over the course of the play session: 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. 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 Figure 14.3. 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%) and differed reliably from the comparison measure for non-naming

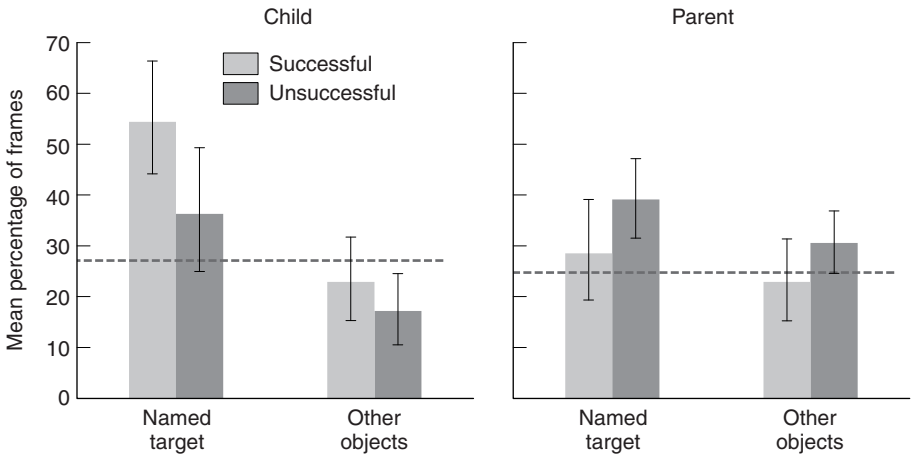

Figure 14.4 Mean percentage of frames in which the parent or child was holding the named object or another object for successful and unsuccessful naming events. Means and standard errors calculated with respect to trials. Dashed line indicates mean holding for children and parents during non-naming moments.

events ($M_{\text{Non-naming}}$ = 2.36%). 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.

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 Figure 14.4 shows that infants were more likely to be holding the named object than other objects during both successful and unsuccessful naming events but holding was more strongly associated with successful than unsuccessful naming events. The object-holding behavior of parents, shown on the right side of Figure 14.4, was not reliably 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.

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. Figure 14.5 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. 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 sensorimotor level. Considering the evidence on

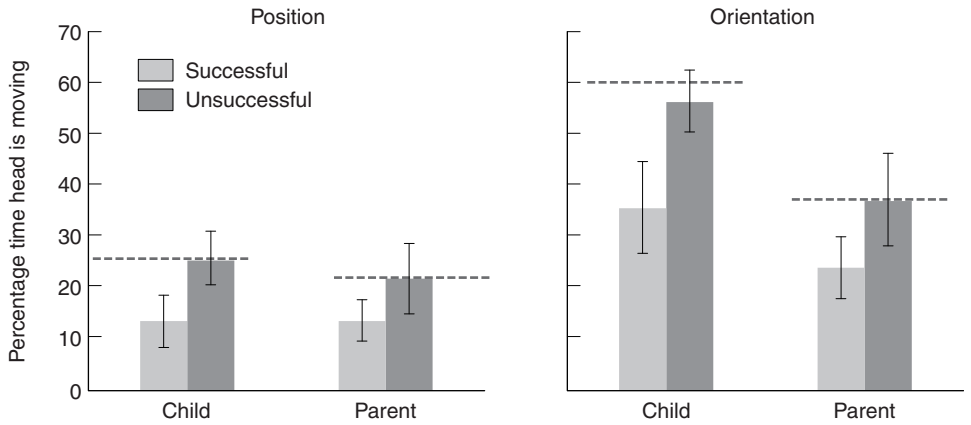

Figure 14.5 Mean percentage time with position and orientation head movements during successful and unsuccessful naming events for children and parents. Means and standard errors calculated with respect to trials. Dashed line indicates mean movements during non-naming moments.

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, parents tend, immediately prior to the naming event, to move their heads 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 heads toward the target object and less likely to stabilize the head.

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.

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 distractors. 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 sensorimotor behavior. This is a particularly interesting developmental idea because many cognitive developmental disorders involve attention and because there is considerable evidence of comorbidity of these cognitive disorders with early usual sensorimotor patterns (Hartman, Houwen, Scherder, and 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, and Newsome, 2005; Shepherd and Findlay, 1986). However, visual attention is also usually linked to eye movements to the attended object's location (Hayhoe and Ballard, 2005). Moreover, eye movements (Grosbras, Laird, and Paus, 2005), head movements (Colby and Goldberg, 1992), and hand movements (Hagler, Riecke, and Sereno, 2007) 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, *et al.*, 2007). Perhaps for physically active toddlers, visual attention is more tightly tied to external action and with development these external mechanisms become more internalized.

The present study also raises a discussion on the level of understanding. Children learn the names of objects in which they are interested. Therefore, as shown in Figure 6(a), "interest," as a macro-level concept, may be viewed as a driving force behind learning (Bloom, Tinker, and Scholnick, 2001). Given this, what is the new contribution of the present study based on sensorimotor dynamics? 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 Figure 6(b)), 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 an associated attribute but not essential, 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" are two alternative explanations shown in Figure 6(c) 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 Figure 6(c), 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. Figure 6(d) provides a more integrated version of these ideas: interest may initially drive both learning (through a separate path); and interest may also drive the child's perception and action – which feed back onto interest and sustained attention and support learning. That is, interest may drive actions and the visual isolation of the object and thus increase interest. These sensorimotor behaviors may also directly influence learning by localizing and stabilizing attention and by limiting clutter and distraction. In brief, the micro-level embodied 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 Figure 6(a) to Figure 6(b), (c), and (d).

In conclusion, the main contribution of this research direction, then, is that it suggests a bottom-up embodied solution to word-referent learning by toddlers. Toddlers, through their

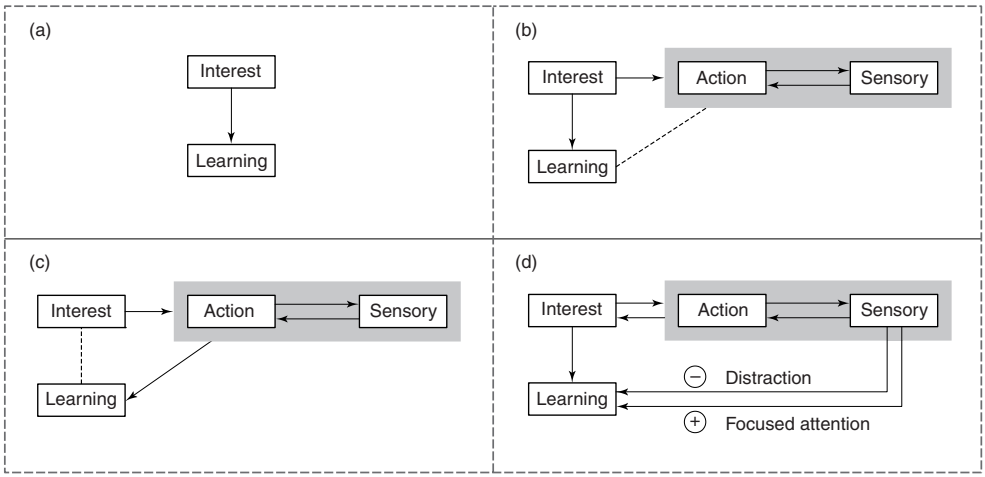

Figure 14.6 Four hypotheses on child's interest, learning and sensorimotor behaviors. (a) child's interest on target objects leads to learning. (b) Child's interest drives both learning and sensorimotor behaviors. Therefore, there are correlations between the two (the dashed line). (c) Child's interest 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 (the dashed 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 sensorimotor behaviors. Thereafter, sensorimotor behaviors also directly influence learning (and maybe interest itself as well) as sustained attention to the target object may facilitate learning while distracting and messy sensory input may disrupt learning. In this way, both child's interest and sensorimotor behaviors jointly influence learning.

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.

Acknowledgements

I would like to thank Linda Smith as an incredible collaborator in the research discussed in this book chapter. I also thank Charlotte Wozniak, Amanda Favata, Alfredo Pereira, Amara Stuehling, and Andrew Filipowicz for collection of the data, Thomas Smith and Tian (Linger) Xu for developing data management and preprocessing software. This research was supported by National Science Foundation Grant 0924248 and AFOSR FA9550-09-1-0665.

References

Baldwin, D. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology, 29*, 832–43.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., and Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20*, 723–42.

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

Beer, R. D. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence, 72*, 173–215.

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

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

Breazeal, C., and Scassellati, B. (2000). Infant-like social interactions between a robot and a human caregiver. *Adaptive Behavior, 8*, 49.

Brooks, R. A., and Stein, L. A. (1994). Building brains for bodies. *Autonomous Robots, 1*, 7–25.

Callanan, M. A. (1985). How parents label objects for young children: The role of input in the acquisition of category hierarchies. *Child Development*, 508–23.

——(1990). Parents' descriptions of objects: Potential data for children's inferences about category principles. *Cognitive Development, 5*(1), 101–22.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford: Oxford University Press.

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

Gibbs, R. W. (2006). *Embodiment and cognitive science*. Cambridge: Cambridge University Press.

Grosbras, M. H., Laird, A. R., and Paus, T. (2005). Cortical regions involved in eye movements, shifts of attention, and gaze perception. *Human Brain Mapping, 25*, 140–54.

Hagler, D., Jr., Riecke, L., and Sereno, M. (2007). Parietal And superior frontal visuospatial maps activated by pointing and saccades. *Neuroimage, 35,* 1562–77.

Hartman, E., Houwen, S., Scherder, E., and Visscher, C. (2010). On the relationship between motor performance and executive functioning in children with intellectual disabilities. *Journal of Intellectual Disability Research, 54*, 468–77.

Hayhoe, M., and Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*(4), 188–94.

Müller, J. R., Philiastides, M. G., and Newsome, W. T. (2005). Microstimulation of the superior colliculus focuses attention without moving the eyes. *Proceedings of the National Academy of Sciences of the United States of America, 102*(3), 524–29.

Pereira, A. F., Smith, L. B., and Yu, C. (in press). A bottom-up view of toddler word learning. *Psychological Bulletin and Review*.

Pfeifer, R., and Scheier, C. (1999). *Understanding intelligence*. Cambridge, MA: MIT Press.

Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.

Shapiro, L. (2011). *Embodied cognition*. London: Taylor & Francis.

Shepherd, M., Findlay, J., and Hockey, R. (1986). The relationship between eye movements and spatial attention. *Quarterly Journal of Experimental Psychology Section A, 38*(3), 475–91.

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

Spivey, M. (2007). *The Continuity of Mind*. New York: Oxford University Press.

Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics, 10*, 401–13.

Tomasello, M., and Akhtar, N. (1995). Two-year-olds use pragmatic cues to differentiate reference to objects and actions. *Cognitive Development, 10*, 201–24.

Wellman, H. M., and Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, 75, 523–41.

Woodward, A. L., and Guajardo, J. J. (2002). Infants' understanding of the point gesture as an object-directed action. *Cognitive Development, 17*, 1061–84.

Yu, C., Ballard, D. H., and Aslin, R. N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science: A Multidisciplinary Journal, 29*, 961–1005.

Yu, C., and Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition, 125*, 244–62.

Yu, C., Smith, l. B., Shen, H., Pereira, A., and Smith, T. (2009). Active information selection: Visual attention through the hands. *IEEE Transactions on Autonomous Mental Development, 2*, 141–51.