Recent studies have suggested that both that infants use extensive knowledge about the world in language learning and that much of that knowledge is communicated through the intentions of the mother. Furthermore those intentions are embodied through a body language consisting of a repertoire of cues, the principal cue being eye gaze. While experiments show that such cues are used, they have not quantified their value. We show in a series of three related studies that intentional cues encoded in body movement can provide very specific gains to language learning. A computational model is developed based on machine learning techniques, such as expectation-maximization, which can identify sound patterns of individual words from continuous speech using non-linguistic contextual information and employ body movements as deictic references to discover word-meaning associations.
It is quite obvious that thinking without a living body is impossible. But it has been more difficult to appreciate that our bodies are an interface that represents the world and influences all the ways we have of thinking about it. The modern statement of this view is due to Merleau-Ponty (1968).

If we started with sheet music, everything that Mozart ever wrote would fit on one compact disc, but of course it could never be played without the instruments. They serve as a ‘body’ to interpret the musical code. In the same way human bodies are remarkable computational devices, shaped by evolution to handle enormous amounts of computation. We usually are unaware of this computation as it is manifested as the ability to direct our eyes to objects of interest in the world or to direct our body to approach one of those objects and pick it up. Such skills seem effortless, yet so far no robot has come close to being able to duplicate them satisfactorily.

If the musculoskeletal system is the orchestra in our metaphor, vision is the conductor. Eye movements are tirelessly made at the rate of an average of three per second to potential targets in the visual surround. In making tea we pre-fixate the objects in the tea-making plan before each step (Land, Mennie, & Rusted, 1999). In a racquet sport we fixated the bounce point of the ball to plan our return shot (Land & Mcleod, 2000). Since the good resolution in the human eye resides in a central one degree, our ability to maneuver this small acuity ball throughout a volume of over a million potential fixation points in a three dimensional world is all the more impressive. Nonetheless we do it. We have to do it for our muscuoskeletal system is designed with springs and the successful manipulation of objects is dependent on the ability to preset the tension and damping of those springs just before they are needed. This interplay of fixation and manipulation is a central feature of primate behavior and preceded language, but it contains its own implicit syntax. The ‘I’ is the agent. The body’s motions are the verbs and the fixations of objects are nouns. In infant language learning, the ‘you’ is the caregiver, a source of approval and other reward. Given this introduction, you are primed for our central premise, namely that the body’s natural ‘language’ can serve and in fact did serve as a scaffold for the development of spoken language.
1 Early Language Learning

Infant language learning is a marvelous achievement. Starting from scratch, infants gradually acquire a vocabulary and grammar. Although this process develops throughout childhood, the crucial steps occur early in development. By the age of three, most children have incorporated the rudiments of grammar and are rapidly growing their vocabulary. Perhaps most impressive, they are able to do this from the unprocessed audio stream which is rife with ambiguity. Exactly how they accomplish this remains uncertain. It has been conjectured that it may be possible to do this by bootstrapping from correlations in the audio stream and indeed recent experimental evidence demonstrates that the cognitive system is sensitive to features of the input (e.g., occurrence statistics). Among others, Saffran, Newport, and Aslin (1996) showed that 8-month-old infants are able to find word boundaries in an artificial language only based on statistical regularities. Later studies (Saffran, Johnson, Aslin, & Newport, 1999) demonstrated that infants are also sensitive to transitional probabilities over tone sequences, suggesting that this statistical learning mechanism is more general than the one dedicated solely to processing linguistic data. The mechanisms may include not only associative processes but also algebraic-like computations to learn grammatical structures (rules). The recent work in Pena, Bonatti, Nespor, and Mehler (2002) showed that silent gaps in a continuous speech stream can cause language learners to switch from one computation to another.

In addition to word segmentation and syntax, the other important issue in language acquisition is how humans learn the meanings of words to establish a word-to-world mapping. A common conjecture of lexical learning is that children map sounds to meanings by seeing an object while hearing an auditory word-form. The most popular mechanism of this word learning process is associationism. Richards and Goldfarb (1986) proposed that children come to know the meaning of a word through repeatedly associating the verbal label with their experience at the time that the label is used. Smith (2000) argued that word learning is initially a process in which children’s attention is captured by objects or actions that are the most salient in their environment, and then they associate it with some acoustic pattern spoken by an adult. This approach has been criticized on the grounds that it does not provide a clear explanation about how infants maps a word to a potential infinity of referents when the word is heard, which is termed reference uncertainty by Quine (1960). Quine presented the following puzzle to theorists of language learning: Imagine that you are a stranger in a strange land with no knowledge
of the language or customs. A native says "Gavagai" while pointing at a rabbit in the distance. How can you determine the intended referent? Quine offered this puzzle as an example of the indeterminacy of translation. Given any word-event pairing, there are, in fact, an infinite number of possible intended meanings – ranging from the rabbit as a whole, to its color, fur, parts, or activity. But Quine’s example also includes a powerful psychological link that does rule out at least some possible meanings – pointing. The native through his body’s disposition in space narrows the range of relevant perceptual information. Although not solving the indeterminacy problem, pointing (1) provides an explicit link between the word and location in space and in so doing (2) constrains the range of intended meanings. Thus, auditory correlations in themselves are unlikely be the whole story of language learning, as studies show that children use prodigious amounts of information about the world in the language process and indeed this knowledge develops in a way that is coupled to the development of grammar (Gleitman, 1990). A large portion of this knowledge about the world is communicated through the mother’s animated social interactions with the child. The mother uses many different signaling cues such as hand signals, touching, eye gaze and intonation to emphasize language aspects. Furthermore we know that infants are sensitive to such cues from studies such as Baldwin et al., 1996; Bloom, 2000; Tomasello, 2000, but can we quantify the advantages that they offer? In this chapter, we report on three computational and experimental studies that show a striking advantage of social cues as communicated by the body. First, a computational analysis of CHILDES database is presented. This experiment not only introduces a formal statistical model of word-to-world mapping but also shows the role of non-linguistic cues in word learning. The second experiment uses adults learning a second language to study gaze and head cues in both speech segmentation and word-meaning association. In the third experiment, we propose and implement a computational model that is able to discover spoken words from continuous speech and associate them with their perceptually grounded meaning. Similar to infants, the simulated learner spots word-meaning pairs from unprocessed multisensory signals collected in everyday contexts and utilizes body cues as deictic (pointing) reference to address the reference uncertainty problem.

2 Experiment 1: Statistical Word Learning

The first of our studies uses mother-infant interactions from the CHILDES database (MacWhinney & Snow, 1985). These tapes contain simultaneous audio and video data wherein a mother
introduces her child to a succession of toys stored in a nearby box. The following transcript from the database is representative of the mother’s descriptions of one of the toys, in this case Big Bird from the television series Sesame Street:

– hey look over here see the birdie
– see the birdie
– oh yes yes I know
– let’s see
– you want to hold the bird you want hold this

The usefulness of non-linguistic cues has its critics, and this example shows why. In this kind of natural interaction, the vocabulary is rich and varied and the central item, Big Bird, is far from the most frequent word. Furthermore the tape shows numerous body language cues that are not coincident with the Big Bird utterance. This complex but perfectly natural situation can be easily quantified by plotting a histogram of word frequency for an extended sequence that includes several toys as shown in the first column of Figure 1. None of the key toy items make it into the top 15 items of the list. An elementary idea for improving the ranking of key words assumes that the infants are able to weight the toy utterances more by taking advantage of the approximately coincident body cues. For instance, the utterances that were generated when the infant’s gaze was fixated on the toys by following the mom’s gaze have more weights than the ones the young child just looked at when not paying attention to what the mother said. We examined the transcript and re-weighted the words according to how much they were emphasized by such cues, but as the second column in Figure 1 shows this strategy does little to help.

What is helpful is to partition the toy sequences (contextual information when the speech was produced) into intervals where within each interval a single toy or small number of co-occurring toys is the central subject or meaning, and then categorize spoken utterances using the contextual bins labeled by different toys. Associating meanings (toys, etc.) with words (toy names, etc.) can be viewed as the problem of identifying word correspondences between English and a “meaning language”, given the data of these two languages in parallel. With this perspective, a technique from machine translation can address the correspondence problem (Brown, Pietra, Pietra, & Mercer, 1993). We apply the idea of Expectation-Maximization (EM) (Dempster, Laird, & Rubin, 1977) as the learning algorithm. Briefly speaking, the algorithm
assumes that word-meaning pairs are some hidden factors underneath the observations which consist of spoken words and extralinguistic contexts. Thus, association probabilities are not directly observable, but they somehow determine the observations because spoken language are produced based on caregiver’s lexical knowledge. Therefore, the objective of language learners or computational models is then to figure out the values of association probabilities so that they can increase the chance of obtaining the observations. Correct word-meaning pairs are those which can maximize the likelihood of the observations in natural interactions. We argue that this strategy is an effective one that young language learners may apply during early word learning. They tend to guess most reasonable and most co-occurring word-meaning pairs based on the observations from different contexts.

The general setting is as follows: suppose we have a word set \( X = \{w_1, w_2, \ldots, w_N\} \) and a meaning set \( Y = \{m_1, m_2, \ldots, m_M\} \), where \( N \) is the number of words and \( M \) is the number of meanings (toys, etc.). Let \( S \) be the number of spoken utterances. All word data are in a set \( \chi = \{(S_w^{(s)}, S_m^{(s)}), 1 \leq s \leq S\} \), where each spoken utterance \( S_w^{(s)} \) consists of \( r \) words \( w_u^{(1)}, w_u^{(2)}, \ldots, w_u^{(r)} \), and \( u(i) \) can be selected from 1 to \( N \). Similarly, the corresponding contextual information \( S_m^{(s)} \) include \( l \) possible meanings \( m_v^{(1)}, m_v^{(2)}, \ldots, m_v^{(l)} \) and the value of \( v(j) \) is from 1 to \( M \). Assume that every word \( w_n \) can be associated with a meaning \( m_m \). Given a data set \( \chi \), the task is to maximize the likelihood of generating the “meaning” streams given English descriptions:

\[
P(S_m^{(1)}, S_m^{(2)}, \ldots, S_m^{(S)} | S_w^{(1)}, S_w^{(2)}, \ldots, S_w^{(S)}) = \prod_{s=1}^{S} P(S_m^{(s)} | S_w^{(s)})
\]  

The technical descriptions can be found in (Yu & Ballard, 2004). Figure 1 shows that this algorithm strikingly improves the probability of the toy vocabulary. 75% of words are associated with correct meanings, such as the word “hat” paired with the meaning “hat” (third column) and the word “book” paired with the meaning “book” (fourth column). In addition, all the toy words are in the top 3 of the corresponding objects (columns). Note that the object “ring” (the fifth column) seems to relate to multiple words. That is because in the video clips, the mothers introduced to the kids to a set of rings with different colors. Therefore, they spent significantly more time on the object “ring” and consequently many words co-occur more frequently with the meaning “ring” compared with other meanings.

Different from previous models of cross-situational learning (e.g. Siskind, 1996) that are
Figure 1: **Word-like unit segmentation.** First column: the histogram of word frequency from Rollins’s video data in the CHILDES database shows that the most frequent words are not the central topic meanings. Second column: weighting the frequency count with cues improves the situation only slightly. The rest columns: The results of statistical word learning to build word-to-world mappings. The row is a list of words and the column is a list of meanings. Each cell is the association probability of a specific word-meaning pair. White color means low probability while dark means high probability. In our model, spoken utterances are categorized into several bins that correspond to temporally co-occurring attentional objects. The EM algorithm discounts words that appear in several bins, allowing the correct word-meaning associations to have high probability.
based inference rules and logic learning, our proposed model is based on probabilistic learning and is able to explicitly represent and estimate the association probabilities of all the co-occurring word-meaning pairs in the training data. Moreover, this formal model of statistical word learning provides a probabilistic framework to study the role of other factors and constraints in word learning, such as social cues and syntactic constraints. The results demonstrate the potential value of this mechanism – how multimodal correlations may be sufficient for learning words and their meanings. We also want to note two major assumptions in this computational study: (1) infants can segment words from continuous speech; and (2) they can partition the interaction intervals based on the focal toy. Our computational model described in Experiment 3 uses unprocessed multisensory data to associate spoken words with their perceptually grounded meanings and demonstrate that body cues play a key role in grounding language in sensorimotor experiences. In the following, we will first present an experimental study that provides empirical support for our argument of the role of body cues, and then describe the grounded model in Section 4 which provides a mechanistic explanation of how it works.

3 Experiment 2: Deictic Body Cues in Human Simulation

A major advance in recent developmental research has been the documentation of the powerful role of social-interactional cues in guiding the infants learning and in linking the linguistic stream to objects and events in the world. Studies (e.g., Baldwin, 1993; Baldwin et al., 1996; Tomasello, 2000; Tomasello, 2001; Bloom, 2000; Woodward & Guajardo, 2002) have shown that there is much information in social interaction and that young learners are highly sensitive to that information. Butterworth (1991) showed that even by 6 months of age, infants demonstrate sensitivities to social cues, such as monitoring and following another’s gaze, although infants’ understanding of the implications of gaze or pointing does not emerge until approximately 12 months of age. Based on this evidence, Bloom (2000) suggested that children’s word learning in the second year of life actually draws extensively on their understanding of the thoughts of speakers. Similarly, Tomasello (2000) showed that infants are able to determine adults’ referential intentions in complex interactive situations and he concluded that the understanding of intentions, as a key social cognitive skill, is the very foundation on which language acquisition is built. These claims have been supported by experiments in which young children were able to figure out what adults were intending to refer to by speech. For example, Baldwin
et al. (1996) proposed that 13-month-old infants give special weight to the cues of indexing the speaker’s gaze when determining the reference of a novel label. Their experiments showed that infants established a stable link between the novel label and the target toy only when that label was uttered by an adult who concurrently directed their attention (as indexed by gaze) toward the target. Such a stable mapping was not established when the label was uttered by a speaker who showed no signs of attention to the target toy, even if the object appeared at the same time that the label was uttered and the speaker was touching the object. However, there is an alternate understanding of these findings to the proposals of “mind-reading”. Smith (2000) has suggested that these results may be understood in terms of the child’s learning of correlations among actions, gestures and words of the mature speaker, and intended referents. Samuelson and Smith (2000) argued that construing the problem in this way does not much “explain away” notions of “mind-reading” but rather grounds those notions to the perceptual cues available in the real-time task that infants must solve. Further, grounding such notions as “referential intent” and “mind-reading” in correlations among words, objects and the co-ordinated actions of speakers and listeners provides a potential window into more conceptual understandings of referential intent. Relevant to this idea, Baldwin and Baird (2001) proposed that humans gradually develop the skill of mind reading so that ultimately they care little about the surface behaviors of others’ dynamic action but focus on discerning underlying intentions based on a generative knowledge system.

In light of this, our second experiment documents the power of the body’s disposition in space in helping language learning and attempts to ask more directly if body cues are in fact helpful for both speech segmentation and word-meaning association, which are two cruxes in early language learning. As in Quine’s example, the subjects are adults presented with a foreign word and a complex scene and the task is to determine the meaning of the word. The experiment uses eye gaze rather pointing as the explicit from word to world. Using adults is only an indirect way to explore infant language learning. The adults being exposed to a new language have explicit knowledge about English grammar that is unavailable to infants but at the same time do not have the plasticity of infant learners. Nonetheless, it has been argued that adult learning can still be a useful model (Gillette, Gleitman, Gleitman, & Lederer, 1999). Certainly if adults could not use body cues it would be a no vote for their use in the infant model, but, it turns out that the cues are very helpful.
Data. We use English-speaking adult subjects who are asked to listen to an experimenter reading a children storybook in Mandarin Chinese. The Mandarin is read in a natural tone similar to a caregiver describing the book to a child and with no attempts to partition the connected speech into segmented words as was done in the first study. The reader was a native speaker of Mandarin describing in his own words the story shown in a picture book entitled “I went walking” (Williams & Vivas, 1989). The book is for 1-3 year old children, and the story is about a young child who goes for a walk and encounters several familiar friendly animals. For each page of the book, the speaker saw a picture and uttered verbal descriptions. Figure 2 shows visual stimuli in three learning conditions. In one condition, Audio only, the speaker’s reading served as the stimulus training materials. In a second condition, Audio + Book, the Audio portion along with a video of the book as each page was turned served as the training material. In the third condition, Head and Eyes Cues, the audio portion, a video of the book as each page was turned, and a marker that showed where on the page the speaker was looking at each moment in time in the reading served as the training material. In the audio-visual condition, the video was recorded from a fixed camera behind the speaker to capture a view of the picture book while the auditory signal was also presented. In the eye-head-cued condition, the video was recorded from a head-mounted camera to provide a dynamic first-person view. Furthermore, an eye tracker was utilized to track the time-course of the speaker’s eye movements and gaze positions. These gaze positions were indicated by a cursor that was superimposed on the video of the book to indicate where the speaker was attending from moment to moment. Subjects were divided into three groups: audio-visual, eye-head-cued, and audio-only. The 27 subjects were randomly assigned to these three 3 training conditions. Each listened (watched) the training material 5 times.

Testing. Testing differed somewhat for the three groups. All groups received a segmentation test: subjects heard two sounds and were asked to select one that they thought was a word but not a multi-word phrase or some subset of a word. They were given as much time as they wanted to answer each question. There were 18 trials. Only subjects in the audio-visual and eye-head-cued training conditions received the second test. The second test was used to evaluate knowledge of lexical items learned from the video (thus the audio-only group was excluded from this test). The images of 12 objects in the picture book were displayed on a computer monitor at the same time. Subjects heard one isolated spoken word for each question and were
asked to select an answer from 13 choices (12 objects and also an option for none of the above).

**Results.** Figure 3 shows the average percent correct on the two tests. In the speech segmentation test, a single-factor ANOVA revealed a significant main effect of the three conditions $F(2, 24) = 23.52; p < 0.001$. Post-hoc tests showed that subjects gave significantly more correct answers in the eye-head-cued condition ($M = 80.6\%; SD = 8.3\%$) than in the audio-visual condition ($M = 65.4\%; SD = 6.6\%; t(16) = 4.89, p < 0.001$). Performance in the audio-only condition did not differ from chance ($M = 51.1\%; SD = 11.7\%$). Subjects in this condition reported that they just guessed because they did not acquire any linguistic knowledge of Mandarin Chinese by listening to the fluent speech for 15 minutes without any visual context. Therefore, they were not asked to do the second test. For the word learning test, performance in the eye-head-cued condition was much better than in the audio-visual condition ($t(16) = 8.11, p < 0.0001$). Note also that performance in the audio-visual condition was above chance ($t(8) = 3.49, p < 0.005$, one-sample t tests).

![Figure 2: The snapshots when the speaker uttered “the cow is looking at the little boy” in Mandarin. Left: no non-speech information in audio-only condition. Center: a snapshot from the fixed camera. Right: a snapshot from a head-mounted camera with the current gaze position (the white cross).](image)

![Figure 3: The mean percentages of correct answers in tests. The results show the importance of explicit cues to the direction of attention of the speaker and suggest that this information importantly disambiguates potential meanings. This finding](image)
goes beyond the claims by Baldwin (1993) and Tomasello (2001) that referential intent as evidenced in gaze affects word learning. Our results suggest that information about the speaker’s attention, a social cue, not only plays a role in high-level learning and cognition but also influences the learning and the computation at the sensory level.

Figure 4: The level of synchrony between eye movement and speech production. Most spoken object names were produced after eye fixations and some of them were uttered before eye fixations. Occasionally, the speaker did not look at the objects at all when he referred to them in speech. Thus, there is no perfect synchrony between eye movement and speech production.

To quantitatively evaluate the difference between the information available in the audio-visual and eye-head-cued conditions, the eye-head-cued video record was analyzed on a frame-by-frame basis to obtain the time of initiation and termination of each eye movement, the location of the fixations, and the beginning and the end of spoken words. These detailed records formed the basis of the summary statistics described below. The total number of eye fixations was 612. Among them, 506 eye fixations were directed to the objects referred to in the speech stream (84.3% of all the fixations). Thus, the speaker looked almost exclusively at the objects that were being talked about while reading from the picture book. The speaker uttered 1019 spoken words, and 116 of them were object names of pictures in the book. A straightforward
hypothesis about the difference in information between the eye-head-cued and audio-visual conditions is that subjects had access to the fact that spoken words and eye movements are closely locked in time. If this temporal synchrony between words and body movements (eye gaze) were present in the eye-head-cued condition (but not the audio-visual condition), it could explain the superior performance on both tests in the eye-head-cued condition. For instance, if the onset of spoken words were always 300 msec after saccades, then subjects could simply find the words based on this delay interval. To analyze this possible correlation, we examined the time relationship of eye fixation and speech production. We first spotted the key words (object names) from transcripts and labeled the start times of these spoken words in the video record. Next, the eye fixations of the corresponding objects, which are closest in time to the onsets of those words, were found. Then for each word, we computed the time difference between the onset of each eye fixation and the start of the word. A histogram of this temporal relation is plotted to illustrate the level of synchrony between gaze on the target object and speech production. As shown in Figure 4, most eye movements preceded the corresponding onset of the word in the speech production, and occasionally (around 7%) the onset of the closest eye fixations occurred after speech production. Also, 9% of object names were produced when the speaker was not fixating on the corresponding objects. Thus, if the learner is sensitive to this predictive role for gaze-contingent co-occurrence between visual object and speech sound, it could account for the superior performance by subjects in the eye-head-cued condition on tests of both speech segmentation and word-meaning association. In the following study, we describe a computational model that is also able to use the information encoded by this dynamic correspondence to learn words. We also note here two important limitations of this experimental study: (1) the learners are adults and not children; (2) we marked the direction of eye-gaze on the page; the learner did not have to figure it out. Still, the study demonstrates the potential importance of these cues in real-time learning.

4 Grounding Spoken Language in Sensorimotor Experience

The Mandarin learning experiment shows conclusively that eye gaze is a big help in retaining vocabulary information in a new language but does not address the issue of the internal mechanism and provide a complete picture of early language learning. Thus, we want to know not only that learners use body cues but also how they do so in terms of the real-time processes
in the real-time tasks in which authentic language learning must take place. We want to study learners’ sensitivities of social cues that are conveyed through time-locked intentional body movements in natural contexts. In light of this, the last study introduces a computational model that learns lexical items from raw multisensory signals to closely resemble the difficulties infants face with in language acquisition, and attempts to show how gaze and body cues can be of help in discovering the words from the raw audio stream and associating them with their perceptually grounded meanings.

The value of this approach is highlighted by recent studies of adults performing visual-motor tasks in natural contexts. These results suggest that the detailed physical properties of the human body convey extremely important information (Ballard, Hayhoe, Pook, & Rao, 1997). They proposed a model of ”embodied cognition” that operates at time scales 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 references to bind objects in the physical environment to variables in cognitive programs of the brain. We apply the theory of embodied cognition in the context of early word learning. To do so, one needs to consider the role of embodiment from both the perspective of a speaker (language teacher) and that of a language learner. First of all, in the study of recent work (e.g. Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995, Meyer, Sleiderink, & Levelt, 1998, Griffin & Bock, 2000; for review, see Griffin, 2004), it has been shown that speech and eye movement are closely linked. Griffin and Bock (2000) demonstrated that speakers have a strong tendency to look toward objects referred to by speech and words begin roughly a second after speakers gaze at their referents. Meyer et al. (1998) found that the speakers’ eye movements are tightly linked to their speech output. They found that when speakers were asked to describe a set of objects from a picture, they usually looked at each new object before mentioning it, and their gaze remained on the object until they were about to say the last word about it. Additionally, from the perspective of a language learner, Baldwin (1993) showed that infants actively gathered social information to guide their inferences about word meanings and they systematically checked the speaker’s gaze to clarify his/her reference.

In our model, we attempt to show how social cues exhibited by the speaker (e.g., the mother) can play a crucial constraining role in the process of discovering words from the raw audio
stream and associating them with their perceptually grounded meanings. By implementing the specific mechanisms that derive from our underlying theories in explicit computer simulations, we can not only test the plausibility of the theories but also gain insights about both the nature of the model’s limitations and possible solutions to these problems.

Figure 5: The computational model shares multisensory information like a human language learner. This allows the association of coincident signals in different modalities.

To simulate how infants ground their semantic knowledge, our model of infant language learning needs to be embodied in the physical environment and sense this environment as a young child. To provide realistic inputs to the model, we attached multiple sensors to adult subjects who were asked to act as caregivers and perform some everyday activities, one of which was narrating the picture book (used in the preceding experiment) in English for a young child, thereby simulating natural infant-caregiver interactions. Those sensors included a head-mounted CCD camera to capture visual information about the physical environment, a microphone to sense acoustic signals, an eye tracker to monitor the course of the speaker’s eye movements, and position sensors attached to the head and hands of the caregiver. In this way, our computational model, as a simulated language learner, has access to multisensory data from the same visual environment as the caregiver, hears infant-directed speech uttered by the caregiver, and observes the body movements, such as eye and head movements, which can be used to infer what the caregiver refers to in speech. In this way, the computational model, as a simulated infant, is able to shared grounded lexical items with the teacher.

To learn words from caregivers’ spoken descriptions, three fundamental problems need to be addressed: (1) object categorization to identify grounded meanings of words from non-linguistic contextual information, (2) speech segmentation and word spotting to extract the sound patterns
of the individual words which might have grounded meanings, and (3) association between spoken words and their meanings. To address those problems, our model consists of the following components as shown in Figure 6:

- **Attention detection** finds where and when a caregiver looks at the objects in the visual scene based on his or her gaze and head movements. The speaker’s referential intentions can be directly inferred from their visual attention.

- **Visual processing** extracts perceptual features of the objects that the speaker is attending to at attentional points in time. Those visual features consist of color, shape and texture properties of visual objects and are used to categorize the objects into semantic groups.

- **Speech processing** includes two parts. One is to convert acoustic signals into discrete phoneme representations. The other part deals with the comparison of phoneme sequences to find similar substrings and cluster those subsequences.

- **Word discovery and word-meaning association** is the crucial step in which information from different modalities is integrated to discover isolated spoken words from fluent speech and map them to their perceptually grounded meanings extracted from visual perception.

The following paragraphs describe these components respectively. The technical details can be found in (Yu, Ballard, & Aslin, in press).

**Attention detection**  Our primary measure of attention is where and when the speaker directs gaze (via eye and head movements) to objects in the visual scene. Although there are several different types of eye movements, the two most important ones for interpreting the gaze of another person are saccades and fixations. Saccades are rapid eye movements that move the fovea to view a different portion of the visual scene. Fixations are stable gaze positions that follow a saccade and enable information about objects in the scene to be acquired. Our overall goal, therefore, is to determine the locations and timing of fixations from a continuous data stream of eye movements. Current fixation-finding methods (Salvucci & Goldberg, 2000) can be categorized into three types: velocity-based, dispersion-based and region-based. Velocity-based methods find fixations according to the velocities between consecutive samples of eye-position
data. Dispersion-based methods identify fixations as clusters of eye-position samples, under the assumption that fixation points generally occur near one another. Region-based methods identify fixation points as falling within a fixed area of interest (AOI) within the visual scene. We developed a velocity-based method to model eye movements using a Hidden Markov Model (HMM) representation that has been widely used in speech recognition with great success (Rabiner & Juang, 1989). A 2-state HMM was used in our system for eye-fixation finding. One state corresponds to the saccade and the other represents the fixation. The observations of the HMM are 2-dimensional vectors consisting of the magnitudes of the velocities of head rotations in three dimensions and the magnitudes of velocities of eye movements. We model the probability densities of the observations using a two-dimensional Gaussian. The parameters of the HMMs that need to be estimated comprise the observation and transition probabilities. The estimation problem concerns how to adjust the model $\lambda$ to maximize $P(O | \lambda)$ given an observation sequence $O$ of eye and head motions. We can initialize the model with flat probabilities, and then the forward-backward algorithm (Rabiner & Juang, 1989) allows us to evaluate the probabilities. As a result of the training, the saccade state contains an observation distribution centered around high velocities and the fixation state represents the data whose distribution is centered around low velocities. The transition probabilities for each state represent the likelihood of remaining in that state or making a transition to another state.
**Clustering Visually Grounded Meanings**  The non-linguistic inputs of the system consist of visual data from a head-mounted camera, head positions and gaze-in-head data. Those data provide the contexts in which spoken utterances are produced. Thus, the possible referents of spoken words that subjects utter are encoded in those contexts, and we need to extract those word meanings from raw sensory inputs. As a result, we will obtain a temporal sequence of possible referents depicted by the box labeled “intentional context” in Figure 7. Our method firstly utilizes eye and head movements as cues to estimate the subject’s focus of attention. Attention, as represented by eye fixation, is then used for spotting the target object of subject’s interest. Specifically, at every attentional point in time, we make use of eye gaze as a seed to find the attentional object from all the objects in a scene. The referential intentions are then directly inferred from attentional objects. We represent the objects by feature vectors consisting of color, shape and texture features. For further information see Yu, Ballard, and Zhu (2002). Next, since the feature vectors extracted from visual appearances of attentional objects do not occupy a discrete space, we vector quantize them into clusters by applying a hierarchical agglomerative clustering algorithm. Finally, for each cluster we select a prototype to represent perceptual features of this cluster.

**Comparing Phoneme Sequences**  We describe our methods of phoneme string comparison in this subsection. Detailed descriptions of algorithms can be obtained from Ballard and Yu (2003). First, the speaker independent phoneme recognition system is employed to convert spoken utterances into phoneme sequences. To fully simulate lexical learning, the phoneme recognizer does not encode any language model or word model. Therefore, the outputs are noisy phoneme strings that are different from phonetic transcriptions of text. Thus, the goal of phonetic string matching is to identify sequences that might be different actual strings, but have similar pronunciations. In our method, a phoneme is represented by a 15-dimensional binary vector in which every entry stands for a single articulatory feature called a distinctive feature. Those distinctive features are indispensable attributes of a phoneme that are required to differentiate one phoneme from another in English. We compute the distance between two individual phonemes as the Hamming distance. Based on this metric, a modified dynamic programming algorithm is developed to compare two phoneme strings by measuring their similarity.
Figure 7: **Overview of the method.** Spoken utterances are categorized into several bins that correspond to temporally co-occurring attentional objects. Then we compare any pair of spoken utterances in each bin to find the similar subsequences that are treated as word-like units. Next, those word-like units in each bin are clustered based on the similarities of their phoneme strings. The EM-algorithm is applied to find lexical items from hypothesized word-meaning pairs.

**Multimodal Word Learning** Figure 7 illustrates our approach to spotting words and establishing word-meaning associations, which consists of the following steps (see Yu et al., in press for detailed descriptions):

- Phoneme utterances are categorized into several bins based on their possibly associated meanings. For each meaning (an attentional object), we find the corresponding phoneme sequences uttered in temporal proximity, and then categorize them into the same bin labeled by that meaning.
- The similar substrings between any two phoneme sequences in each bin are found and treated as word-like units.
- The extracted phoneme substrings of word-like units are clustered by a hierarchical ag-
glomerative clustering algorithm. The centroids of clusters are associated with their possible grounded meanings to build hypothesized word-meaning pairs.

- To find correct lexical items from hypothesized lexical items, the probability of each word is represented as a mixture model that consists of the conditional probabilities of each word given its possible meanings. In this way, the same Expectation-Maximization (EM) algorithm described in Study 1 is employed to find the reliable associations of spoken words and their grounded meanings which maximize the likelihood function of observing the data.

Results. Six subjects, all native speakers of English, participated in the experiment. They were asked to narrate the picture book “I went walking” (used in the previous experiment) in English. They were also instructed to pretend that they were telling this story to a child so that they should keep verbal descriptions of pictures as simple and clear as possible. We collected multisensory data when they performed the task, which were used as training data for our computational model.

Table 1 shows the results for four measures. Semantic accuracy measures the categorization accuracy of clustering visual feature vectors of attentional objects into semantic groups. Speech segmentation accuracy measures whether the beginning and the end of phoneme strings of word-like units are word boundaries. Word-meaning association accuracy (precision) measures the percentage of successfully segmented words that are correctly associated with their meanings. Lexical spotting accuracy (recall) measures the percentage of word-meaning pairs

<table>
<thead>
<tr>
<th>subjects</th>
<th>semantics</th>
<th>speech segmentation</th>
<th>word-meaning association</th>
<th>lexical spotting</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80.3%</td>
<td>72.6%</td>
<td>91.3%</td>
<td>70.3%</td>
</tr>
<tr>
<td>2</td>
<td>83.6%</td>
<td>73.3%</td>
<td>92.6%</td>
<td>73.2%</td>
</tr>
<tr>
<td>3</td>
<td>79.2%</td>
<td>71.9%</td>
<td>86.9%</td>
<td>76.5%</td>
</tr>
<tr>
<td>4</td>
<td>81.6%</td>
<td>69.8%</td>
<td>89.2%</td>
<td>72.9%</td>
</tr>
<tr>
<td>5</td>
<td>82.9%</td>
<td>69.6%</td>
<td>86.2%</td>
<td>72.6%</td>
</tr>
<tr>
<td>6</td>
<td>76.6%</td>
<td>66.2%</td>
<td>83.1%</td>
<td>72.8%</td>
</tr>
<tr>
<td>Average</td>
<td>80.6%</td>
<td>70.6%</td>
<td>88.2%</td>
<td>73.1%</td>
</tr>
</tbody>
</table>
that are spotted by the model. The mean Semantic accuracy of categorizing visual objects is 80.6%, which provides a good basis for the subsequent speech segmentation and word-meaning association metrics. It is important to note that the recognition rate of the phoneme recognizer we used is 75%. This rather poor performance is because it does not encode any language model or word model. Thus, the accuracy of the speech input to the model has a ceiling of 75%. Based on this constraint, the overall accuracy of speech Segmentation of 70.6%, is quite good. Naturally, an improved phoneme recognizer based on a language model would improve the overall results, but the intent here is to study the developmental learning procedure without pre-trained models. The measure of Word-Meaning Association, 88.2%, is also impressive, with most of the errors caused by a few words (e.g., “happy” and “look”) that frequently occur in some contexts but do not have visually grounded meanings. The overall accuracy of Lexical Spotting is 73.1%, which demonstrates that by inferring speakers’ referential intentions, the stable links between words and meanings could be easily spotted and established. Considering that the system processes raw sensory data, and our learning method works in an unsupervised mode without manually encoding any linguistic information, the accuracies for both speech segmentation and word meaning association are impressive.

To more directly demonstrate the role of body cues in language learning, we processed the data by another method in which the inputs of eye gaze and head movements were removed, and only audio-visual data were used for learning. Clearly, this approach reduces the amount of information available to the learner, and it forces the model to classify spoken utterances into the bins of all the objects in the scene instead of just the bins of attentional objects. In all other respects, this approach shares the same implemented components with the eye-head-cued approach. Figure 8 shows the comparison of these two methods. The eye-head-cued approach outperforms the audio-visual approach in both speech segmentation ($t(5) = 6.94, p < 0.0001$) and word-meaning association ($t(5) = 23.2, p < 0.0001$). The significant difference lies in the fact that there exist a multitude of co-occurring word-object pairs in natural environments that infants are situated in, and the inference of referential intentions through body movements plays a key role in discovering which co-occurrences are relevant.

**Significance.** To our knowledge, this work is the first model of word learning which not only learns lexical items from raw multisensory signals to closely resemble infant language develop-
development from natural environments, but also explores the computational role of social cognitive skills in lexical acquisition. In addition, the results obtained from this comparative study are very much in line with the results obtained from human subjects, suggesting that not only is our model cognitively plausible, but the role of multimodal interaction can be appreciated by both human learners and by the computational model.

5 General Discussion

5.1 The Role of Body Cues

Children do not hear spoken utterances in isolation. They hear them in a context. Ervin-Tripp (1973) found that normal children with deaf parents, who could access English only from radio or television, did not learn any speech. Macnamara (1982) argued that it is very difficult for a child to figure out what the silent actors in interactive materials (such as a video or a TV program) are talking about. By interacting with live human speakers, who tend to talk about things that are present in a shared context with children, the child can more effectively infer what the speaker might have meant. More recently, Kuhl, Tsao, and Liu (2003) showed that American 9-mo-old infants exposed to Mandarin Chinese under audio-videotape or auditory-only conditions did not show phoneme learning. Both studies indicate that learning is influenced by the presence of a live person generating body cues to attract infant attention and motivate learning. Recent experimental studies confirmed this idea and suggested that the existence of a theory of mind could play a central role in how children learn the meanings of certain words (Baldwin, 1993; Markson & Bloom, 1997; Tomasello & Farrar, 1986; Tomasello, 2001).

In this chapter, we focused on the ability of the young language learner to infer interlocu-
 tors’ referential intentions by observing their body movements, which may significantly facilitate early word learning. Clearly, this is the earliest and perhaps the lowest level of a theory of mind, and may not (at least for infants) involve any conscious knowledge that the speaker who is providing body-movement cues has explicit intentions. Nevertheless, if infants are sensitive to some of these body-movement cues, that may constrain the word-learning process sufficiently to enable it to function effectively and efficiently in early lexical development. Different from most other studies, our work explores the dynamic nature of body cues in language acquisition by closely resembling the natural environment of infant-caregiver interaction. In our preliminary experiment that simulated word learning using human adults, the experimenter narrated the story shown in the picture book naturally by using infant-directed speech. The adult learners were therefore presented with continuous speech and visual information as well as the dynamic movements of the speaker’s gaze and head. Similarly, in our computer simulation, the computational model we built of a young language learner received continuous sensory data from multiple modalities. As we pointed out in both of these situations (adult learning and model learning), the timing of speech productions and eye movements were not perfectly aligned in these complex natural contexts. Nevertheless, the results of empirical studies showed that adult language learners exposed to a second language in the eye-head-cued condition outperformed subjects in the audio-visual condition in both word discovery (segmentation) and word-meaning tests, indicating that human subjects can utilize dynamic information encoded in the continuous body movements of the speaker to improve the learning results. How do adults take advantage of the partial, imperfect temporal synchrony between sounds and object-directed gaze? Our computational model answered this question by simulating the underlying mechanism of using body cues.

Body cues are referential in nature. In the computational model described in the previous section, a speaker’s referential intentions are estimated and utilized to facilitate word learning in two ways. First, the possible referential objects defined by gaze changes in real-time provide constraints for word spotting from a continuous speech stream. Second, a difficult task of word learning is to figure out which entities specific words refer to from a multitude of co-occurrences between words and things in the world. This is accomplished in our model by utilizing speakers’ intentional body movements as deictic references to establish associations between words and their visually grounded meanings. These two mechanisms not only provide
a formal account of the role of body cues in word learning, but also suggest an explanation of the experimental results obtained from adult learners of a second language in our human simulation. Furthermore, the combination of human simulation and computational modeling shows conclusively that body cues serve to facilitate, and may in fact be a necessary feature of, learning the vocabulary in a new language.

5.2 Modeling Embodied Word Learning

We are interested in not only what human language learners can achieve which is demonstrated in Experiment 2, but only how they do so. Theoretical simulation studies provide unique opportunities to explore the mechanistic nature of early word learning, to provide a quantitative computational account of the behavioral profile of language learners and to test hypotheses quickly (i.e., without requiring the collection of new data). Therefore, computational investigations of language acquisition have recently received considerable attention. Among others, MacWhinney (1989) applied the competition theory to build an associative network that was configured to learn which word among all possible candidates refers to a particular object. Plunkett, Sinha, Miller, and Strandsby (1992) built a connectionist model of word learning in which a process termed autoassociation maps preprocessed images with linguistic labels. The linguistic behavior of the network exhibited non-linear vocabulary growth (vocabulary spurt) that was similar to the pattern observed in young children. Siskind (1996) developed a mathematical model based on cross-situational learning and the principle of contrast, which learns word-meaning associations when presented with paired sequences of pre-segmented tokens and semantic representations. Regier’s work focused on grounding lexical items that describe spatial relations in visual perception (Regier, 1996). Bailey (1997) proposed a computational model that can not only learn to produce verb labels for actions but also carry out actions specified by verbs that it has learned. Tenenbaum and Xu (2000) developed a computational model based on Bayesian inference which can infer meanings from one or a few examples without encoding the constraint of mutual exclusion.

Computational models of development and cognition have changed radically in recent years. Many cognitive scientists have recognized that models which incorporate constraints from embodiment – that is, how mental and behavioral development depends on complex interactions among brain, body and environment (Clark, 1997) – are more successful than models which
ignore these factors. Language represents perhaps the most sophisticated cognitive system acquired by human learners, and it clearly involves complex interactions between a child’s innate capacities and the social, cognitive, and linguistic information provided by the environment (Gleitman & Newport, 1995). The model outlined in the present study focuses on the initial stages of language acquisition using the embodied cognition perspective – how are words extracted from fluent speech and attached to meanings? Most existing models of language acquisition have been evaluated by artificially derived data of speech and semantics (Brent & Cartwright, 1996; Siskind, 1996; Regier, 1996; Cohen, Oates, Adams, & Beal, 2001) (but also see Roy & Pentland, 2002). In those models, speech is represented by text or phonetic transcriptions and word meanings are usually encoded as symbols or data structures. In contrast, our model proved successful by taking advantage of recent advances in machine learning, speech processing and computer vision, and by suggesting that modeling word learning at the sensory level is not impossible and that embodiment has some advantages over symbolic simulations by closely resembling the natural environment in which infants develop. In both empirical and computational studies, we use storybook reading - a natural interaction between children and caregivers, to simulate the word learning in everyday life. Multisensory data (materials used by the model) are real and natural. To our knowledge, in the literature of language acquisition modeling, this experimental setup is the closest to the natural environment of early word learning that has been achieved.

Our model emphasizes the importance of embodied learning for two main reasons. First, the motivation behind this work is that language is grounded in sensorimotor experiences with the physical world. Thus, a fundamental aspect of language acquisition is that the learner can rely on associations between the movements of the body and the context in which words are spoken (Lakoff & Johnson, 1980). Second, because infants learn words by sensing the environment with their perceptual systems, they need to cope with several practical problems, such as the variability of spoken words in different contexts and by different talkers. To closely simulate infant vocabulary development, therefore, a computational model must have the ability to remove noise from raw signals and extract durable and generalizable representations instead of simplifying the problem by using consistent symbolic representations (e.g., text or phonetic transcriptions). Furthermore, our computational model addresses the problem of speech segmentation, meaning identification and word-meaning mapping in a general framework. It shows
the possible underlying mechanism by which linguistic processing, perceptual learning and social communication interact with each other in early word learning.

6 Conclusion

All three of our studies show quantitatively how body cues that signal intention can aid infant language learning. Such intentional body movements with accompanying visual information provide a natural learning environment for infants to facilitate linguistic processing. From a computational perspective, this work is the first model that explicitly includes social cognitive skills in language learning, such as inferring the mother’s referential intention from her body movements. The central ideas of our model are to identify the sound patterns of individual words from continuous speech using non-linguistic contextual information and employ body movements as deictic references to build grounded lexical items. By exploiting the constraints of social interaction and visual perception, probabilistic algorithms, such as expectation maximization, have the power to extract appropriate word-semantics associations even in the highly ambiguous situations that the infant normally encounters. Equally important is that the model suggests a framework for understanding the vocabulary explosion that occurs starting at age two. Beside providing a relatively limited number of the most probable lexical items, the EM model also generates a large amount of word-meaning pairs with uncertainty. This indicates that infants can potentially accumulate valuable information about many word-semantics associations long before these associations are unique. The rapid vocabulary expansion may be a product of this parallel accumulation process.

References and Notes


