Cognitive Science 29 (2005) 961–1005


# The Role of Embodied Intention in Early Lexical Acquisition

Chen Yu[a], Dana H. Ballard[b], Richard N. Aslin[c]

[a]*Department of Psychology and Cognitive Science Program, Indiana University*
[b]*Department of Computer Science, University of Rochester*
[c]*Department of Brain and Cognitive Sciences, University of Rochester*

Received 9 July 2003; received in revised form 18 October 2004; accepted 18 January 2005

**Abstract**

We examine the influence of inferring interlocutors' referential intentions from their body movements at the early stage of lexical acquisition. By testing human participants and comparing their performances in different learning conditions, we find that those embodied intentions facilitate both word discovery and word-meaning association. In light of empirical findings, the main part of this article presents a computational model that can identify the sound patterns of individual words from continuous speech, using nonlinguistic contextual information, and employ body movements as deictic references to discover word-meaning associations. To our knowledge, this work is the first model of word learning that not only learns lexical items from raw multisensory signals to closely resemble infant language development from natural environments, but also explores the computational role of social cognitive skills in lexical acquisition.

*Keywords:* Language acquisition; Computational model; Machine learning; Embodied cognition; Cognitive development

## 1. Introduction

Children solve many complex learning problems during their first years of life. Perhaps the most remarkable and challenging of these tasks is learning language, which occurs both quickly and effortlessly. Language, of course, is a multileveled system of perception, production, and representation. This article focuses on three of the earliest problems that children need to solve as they acquire their native language: (a) segmenting the speech signal into lexical units, (b) identifying the meanings of words from their perceptual input, and (c) associating these meanings with lexical units.

Requests for reprints should be sent to Chen Yu, Department of Psychology and Cognitive Science Program, Indiana University, Bloomington, IN 47405. E-mail: chenyu@indiana.edu

The first problem is very difficult if attempted using only acoustic information. Before children can begin to map acoustic word forms onto objects in the world, they must determine which sound sequences are words. To do so, they must uncover at least some of the units that belong to their native language from a largely continuous stream of sounds. However, spoken language lacks the acoustic analog of blank spaces of written text, which makes the problem of discovering word boundaries in continuous speech quite challenging. In particular, if we assume that children start out with little or no knowledge of the inventory of words, the problem becomes much harder because segmentation of unknown stretches of speech cannot be achieved by first identifying known lexical items embedded in that speech.

Second, words usually refer to categories rather than single entities (the common noun, proper noun distinction). Therefore, young children must not only extract and recognize the possible meanings of words from their nonlinguistic perceptual input, but they must do so given the partially correlated cues that are present but which do not serve to define the category. Theories in which mental representations (concepts) must be acquired in advance of the labels that stand for them are well established in early word learning (Slobin, 1985). Those theories are generally accepted, although recent work shows that children can use linguistic labels to further constrain the category structure (Booth & Waxman, 2002; Sloutsky & Lo, 1999). Nevertheless, children have often formed rich conceptual categories prior to the development of language. For instance, children must have some understanding that a dog is furry and has four legs and two eyes, even if they do not know the linguistic labels of those concepts, such as *dog, leg,* and *eye.* Children's sensorimotor experiences are continually building up these embodied, prelinguistic concepts. If we assume that this conceptual machinery is already well established by the time the child's first words are learned, the word acquisition problem is simplified by directly associating a linguistic label with a category of sensorimotor experience that has already been established. For example, the appearances of objects could be obtained from visual perception and used to extract visual features of objects. Those visual features are then stored in the brain as the grounded meanings of object names and are ready to be associated with linguistic labels. This assumption has been supported by recent studies using a human simulation paradigm (Gillette, Gleitman, Gleitman, & Lederer, 1999; Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005; Snedeker & Gleitman, 2004) that is designed to investigate effects of linguistic input on the learning function by using adults (undergraduates) to model infants. The main finding from these studies suggests that the primary limiting factor on early vocabulary growth resides in the difficulty of solving the mapping problem (the third task described next) rather than in limitations in the early conceptual repertoire.

Third, young children need to associate sound patterns of words with meanings or concepts. Learning a word involves mapping a phonological form to a conceptual representation, such as associating the sound "dog" to the concept of *dog.* Quine (1960) pointed out that there is a potential infinity of referents when a word is heard, which is termed *reference uncertainty.* For instance, when a young child sees a dog and hears an isolated word "dog," the spoken word could refer to the whole dog, the dog and the ground it is on, a part of the dog, its color, shape, size, and so on. It might even refer to something that is not relevant to the dog at all (e.g., the brightness of the surroundings). In natural environments, infants hear spoken utterances that contain multiple words and are sensitive to aspects of the environment that have more than one possible target referent. They must determine which co-occurrences are relevant from a multitude of potential co-occurrences between words and things in the world.

In this work, we present an implemented computational model of embodied word learning, which is able to associate spoken words with their perceptually grounded meanings in a completely unsupervised mode. The model not only addresses all three problems described previously but also collects and processes multisensory data to closely resemble the natural environments of infant language development. One of the central ideas is to make use of nonspeech contextual information to facilitate word spotting; the other is to use inference of speakers' referential intentions from their body movements, which we term *embodied intention* (Yu & Ballard, 2003; Yu, Ballard, & Aslin, 2003). Nonspeech contextual information and embodied intention can then be used as deictic references to discover temporal correlations of data in different modalities from which to build lexical items. The idea of embodied intention is derived from embodied cognition (Clark, 1997), and the theory of mind reading (Baron-Cohen, 1995).We argue that social cognitive skills can be grounded in sensorimotor-level behaviors (e.g., gaze and body position). In the context of language learning, we show that language learners are able to use others' intentional body movements in real-time natural interactions to facilitate word learning.

The remainder of this article is organized as follows. Section 2 gives a short overview of previous studies of infant language acquisition in both cognitive studies and computational modeling. Then we provide a framework for the theoretical arguments of embodied intention in Section 3. Section 4 describes an empirical study of the role of embodied intentions in learning by human participants (adults). In light of the findings from empirical studies, Section 5 presents both the theoretical model on which our studies are based and the implementation of the model. This model provides not only the machinery of early word learning but also a platform to investigate various kinds of cues that enable learning. The experimental setup and the results of a comparative study are reported in this section. In Section 6, we discuss several issues in word learning based on the results of empirical and computational studies. Section 7 concludes with a discussion of our future work.

Overall, our work renders the ideas of embodied intention as a formal model and measures its effectiveness for learning word meanings, using both experimental and computational approaches, suggesting that embodied intention plays an important role in both speech segmentation (the first task) and word-learning association (the third task). Note that we are not claiming that young children employ the exact method presented in this article. However, as a computational model, this work provides an existence proof for a machine learning technique that solves the lexical acquisition task. Furthermore, from empirical and computational studies, we provide a set of quantitative predictions for determining the role of infants' sensitivities to social cues conveyed through others' intentional body movements in early word learning. It leaves open for further empirical study the question of what techniques young children actually use to solve the problem. We hope that this work not only provides a computational account to supplement the existing related theories of language acquisition but also gives some useful hints for future research.

## 2. Related work

In the last 10 years, there has been tremendous progress in understanding infants' abilities to segment continuous speech, discover words, and learn their meanings. This section provides a brief overview of both empirical studies and computational modeling.

### 2.1. Experimental investigations

English-learning infants first display some ability to segment words at about 7.5 months (Jusczyk & Aslin, 1995). By 24 months, the speed and accuracy with which infants identify words in fluent speech is similar to that of native adult listeners. A number of relevant cues have been found that are correlated with the presence of word boundaries and can potentially signal word boundaries in continuous speech (see Jusczyk, 1997, for a review). Cutler and Butterfield (1992) argued that English-speaking infants appear to use prosodic cues, such as strong or weak stress, syllable units, and subsyllabic units, to parse continuous acoustic signals into words. Studies by Jusczyk and colleagues using streams of nonsense words (Johnson & Jusczyk, 2001) and fluent English sentences (Jusczyk, Hohne, & Bauman, 1999) suggest that prosodic cues (e.g., first-syllable stress) play an important role in word segmentation. There is also evidence that by 9 months of age, English learners have begun to determine the way that phonotactic sequences line up with word boundaries in their native language (Mattys & Jusczyk, 2001). *Phonotactics* refers to the constraints on the possible ordering of phoneme segments within morphemes, syllables, and words of a language. Similarly, different phoneme variants (allophones) of the same phoneme are often restricted in terms of the positions where they can appear within a word. Therefore, knowledge of the contexts in which such allophones appear could provide listeners with a clue to the location of word boundaries in fluent speech (Church, 1987). Another possible source of information for speech segmentation is the distributional statistics of phonemes or syllables. Saffran, Aslin, and Newport (1996) demonstrated that 8-month-old infants are able to find word boundaries in an artificial language based only on statistical regularities. And Thiessen and Saffran (2003) showed that statistical cues initially outweigh prosodic cues until 10 months of age. However, it still remains an open question as to how these constraints are actually applied by infant learners as they acquire their native language.

Once infants have segmented auditory word forms from fluent speech, they must map those sounds onto meanings. One explanation of how infants discover one-to-one correspondences between multiple spoken words and their meanings, termed *cross-situational learning,* has been proposed by many authors, such as Pinker (1989) and Gleitman (1990). This scheme suggests that when a child hears a word, he or she can hypothesize all the potential meanings for that word from the nonlinguistic context of the utterance containing that word. On hearing that word in several different utterances, each of which is in a different context, he or she can intersect the corresponding sets to find those meanings that are consistent across the different occurrences of that word. Presumably, hearing words in enough different situations would enable the child to rule out all incorrect hypotheses and uniquely determine word meanings. However, as mentioned earlier, the precise mapping of sounds to meanings is complicated by the fact that many properties of the context could be shared (e.g., the presence of a white rabbit on a hard surface), despite many differences across contexts. In the absence of a single unique property across all contexts, it is not clear how infants disambiguate the meaning of a word (i.e., does *rabbit* refer to the rabbit or to the hard surface?).

A further challenge for the infant is to learn the meanings of verbs. There is an overwhelming preponderance of concrete nouns in children's early speech, not only in English but in most other languages, such as Italian (Caselli, Casadio, & Bates, 2000). Gentner

(1982) proposed a rationale that concrete nouns must precede verbs in early language development because of the conceptual limitations of young children. Nouns are easy to grasp based on relatively simple perceptual categorization of similar objects, whereas verbs reflect complex concepts, such as relations, events, or actions, which are harder to perceive. As the child develops, verbs are learned by mapping complex concepts that are easily perceived to those words that express them. However, recent cross-linguistic studies showed that children who learn Korean (Gopnik & Choi, 1995) and Mandarin Chinese (Tardif, 1996) do not display the early bias toward nouns like English learners. Tardif claimed that verbs may actually predominate statistically over nouns in many Chinese children. Gillette et al. (1999) offered a compelling explanation to account for which words (nouns or verbs) are acquired first. They provided strong evidence that learnability of a word is not primarily based on its lexical class but on the word's imageability or concreteness. The nouns are learned before most of the verbs because nouns are more observable than verbs. The imageability of a word is more important than the lexical class, and the most observable verbs are learned before the least observable nouns.

In summary, infants begin from a state at 4 months of age where they only appear to recognize highly frequent auditory word forms, such as their name (Mandel, Jusczyk, & Pisoni, 1995), and these word forms are often presented in isolation. By 6 months they begin to use statistical cues to extract candidate words from fluent speech, and this ability is quite robust by 8 months of age. By 9 months of age they have a bias to attend to the prototypical prosodic structure of their native language. And by 10 months they can use their sensitivity to language-specific phonotactics to assist in the word-segmentation task. Once the infant has extracted a small number of candidate word forms, these sounds are rapidly attached to meanings, and there is a strong bias to acquire the meanings of nouns before verbs because many nouns are more concrete and observable than verbs.

### *2.2. Computational modeling*

The foregoing experimental studies have yielded important insights into the linguistic abilities of infants and young children and have provided informative constraints for building computational models of language acquisition. On the other hand, modeling language acquisition can provide a quantitative computational account of the behavioral profile of language learners and test hypotheses quickly (i.e., without requiring the collection of new data). Therefore, computational investigations of language acquisition have recently received considerable attention. Generally, to simplify the problems that must be addressed, models of language acquisition are divided into several subtasks. For instance, Siskind (1996, 1999) presented a formal version of language acquisition by modeling it as involving three subtasks: (a) identifying the sequence of words in an utterance; (b) identifying a set of likely interpretations of the utterance based on the nonlinguistic context when the utterance is produced; and (c) inferring the meanings of words, given the results of the first two subtasks. Until now, most computational studies of how children learn their native language address only one particular subtask. This section gives an overview of modeling speech segmentation (the first subtask) and lexical acquisition (the third subtask). The studies of perceptual learning and categorization, reviewed in Section 1, focus on the second subtask and will be discussed in detail in Section 3.

*2.2.1. Speech segmentation*

To explore how young children discover the words embedded in a mostly continuous speech stream, several computational models of speech segmentation have begun to consider the learning problem from the point of view of the statistical properties of language and how this information might be stored and computed in the human brain. Saffran, Newport, and Aslin (1996) suggested that infants might compute the transitional probabilities between sounds in a language and use the relative strengths of these probabilities to hypothesize word boundaries. The method they developed treats syllables rather than phonemes as the fundamental units of input and calculates the probability of each syllable in the language conditioned on its predecessor. They argued that infants might segment utterances at low points of the transitional probabilities between adjacent syllables. Aslin,Woodward, LaMendola, and Bever (1996) proposed that infants learn the metrical and phonotactic properties of word boundaries by generalizing from utterance boundaries, which are then used to segment words within utterances. They introduced a connectionist model that successfully implemented this segmentation strategy. Christiansen, Allen, and Seidenberg (1998) employed a similar connectionist model and showed that using multiple cues (i.e., statistics and prosody) results in superior word-segmentation performance than one cue alone. Brent and Cartwright (1996) encoded information about the distributional regularity and phonotactic constraints in their computational model. Distributional regularity means that sound sequences occurring frequently and in a variety of contexts are better candidates for the lexicon than those that occur rarely or in few contexts. The phonotactic constraints include both the requirement that every word must have a vowel and the observation that languages impose constraints on word-initial and word-final consonant clusters. More recently, Brent (1997, 1999a) proposed a model called incremental distributional regularity optimization (INCDROP). INCDROP asserts that the process of segmenting utterances and inferring new wordlike units is driven by the recognition of familiar units within an utterance. It posits a single mechanism that can discover new units by recognizing familiar units in an utterance, extracting those units, and treating the remaining contiguous stretches of the utterance as novel units. When an utterance contains no familiar units, the whole utterance is treated as a single novel unit, so there is no need to assume a special bootstrapping device that discovers the first units. A good survey of the related computational studies of speech segmentation can be found in (Brent, 1999b), in which several methods are explained, their performances in computer simulations are summarized, and behavioral evidence bearing on them is discussed. Most of these studies, however, use phoneme transcriptions of text as input and do not deal with raw speech. Transcriptions do not show the acoustic variability of spoken words in different contexts and by various talkers and, thus, do not capture all the difficulties that young children face with natural speech input.

*2.2.2. Lexical learning*

Compared with the studies on speech segmentation, relatively few computational learning methods of lexical acquisition have been proposed and implemented. Among them, 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 nonlinear 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 presegmented 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. Li, Farkas, and MacWhinney (2004) proposed a developmental model based on self-organized networks, which learns topographically organized representations for linguistic categories over time.

Different from the symbolic models of vocabulary acquisition described previously, Steels (1997) reported experiments in which autonomous visually grounded agents bootstrap meanings through adaptive language games. He argued that language is an autonomous evolving adaptive system maintained by a group of distributed agents without central control, thereby enabling the lexicon to cope with new meanings as they arise. Roy and Pentland (2002) implemented a model of early language learning that learns words and their semantics from raw sensory input. They used the temporal correlation of speech and vision to associate spoken utterances with a corresponding object's visual appearance. However, the audiovisual corpora were collected separately in Roy's system. Specifically, audio data were gathered from infant–caregiver interactions, whereas visual data were captured by a charge-coupled device (CCD) camera on a robot. Thus, to simplify the problem, audio and visual inputs were manually correlated based on two assumptions: temporal co-occurrences of words and their meanings, and the uniqueness of the semantic representation of each utterance. Whereas this work was groundbreaking, the simplifying assumptions do not represent the natural case. The second assumption is obviously not true in most cases including infant-directed speech. We will show in the next section that the first assumption is also not reliable for modeling language acquisition.

To summarize, recent computational models suggest an associative basis of word learning and use general-purpose learning mechanisms, such as rational inference and associative learning, to tackle the inductive problem in word learning with success. However, two questions are left open and seem well worth pursuing. The first question is about social cognitive skills in language acquisition. Empirical findings (Baldwin, 1993; Tomasello, 2001) have demonstrated that young language learners seem to rely on their interpretations of the gaze and pointing behaviors of others to infer others' mental states and then guide their word learning. There is no corresponding computational model to provide a mechanism for how social cues are used in word learning and to answer the question of whether social cues and associative learning can be integrated together. The second question concerns the role of embodiment in language learning. As pointed out earlier, most models simplify the learning problem by using synthesized or artificial data instead of raw multisensory data collected in natural contexts. First of all, these artificial data do not capture all the difficulties that young children face with natural speech input. Moreover, we argue that many useful constraints are encoded at the sensorimotor level and can be inferred from the interactions between brain, body, and environment. Symbolic modeling is not able to extract and use those constraints.

## 3. The role of embodied intention

A common conjecture in models of lexical learning is that children map sounds to meanings by seeing an object while hearing an auditory word form. The most popular computational mechanism of this word-learning process is *associationism,* which assumes that language acquisition is solely based on statistical learning of co-occurring data from the linguistic modality and nonlinguistic context (see a review by Plunkett, 1997). 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. Studies in intermodal perception (e.g., Bahrick, Lickliter, & Flom, 2004; Gogate,Walker-Andrews, & Bahrick, 2001; Slater, Quinn, Brown, & Hayes, 1999) have also shown that infants are able to learn intermodal relations, suggesting that intermodal temporal synchrony is an important cue to pair objects and sounds. Despite the merit of this idea, associationism is unlikely to be the whole story because it is based on the assumption that words are always uttered when their referents are perceived. Bloom (2000) argued that around 30% to 50% of the time, when young language learners hear a word, they are not attending to the object referred to by the speech. Therefore, if children hear a word and associate it with whatever is perceived at the time that the word is used, they will make lots of mismappings. But in fact, children rarely make mistakes of this type in word learning.

In addition to spatiotemporal contiguity of visual context and auditory input, recent studies (e.g., Baldwin, 1993; Baldwin et al., 1996; Bloom, 2000; Tomasello, 2000, 2001; Woodward & Guajardo, 2002) have shown that another major source of constraints in language acquisition is in the area of social cognitive skills, such as children's ability to infer the intentions of adults during face-to-face discourse. This kind of social cognition was called *mind reading* by Baron-Cohen (1995) or more generally, theory of mind (Wellman & Liu, 2004). 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 2nd year of life actually draws extensively on their understanding of the thoughts of speakers. His claim has been supported by experiments in which young children were able to figure out what adults were intending to refer to by speech based on social cues. 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. 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.

The problem with the hypothesis that infants and young children use social cognitive cues in word learning is that the empirical evidence is based on macrolevel behaviors (e.g., head orientation or pointing) in constrained contexts (e.g., Baldwin, 1993), rather than on microlevel behaviors (e.g., gaze and body position) that unfold in real time during natural contexts. The studies at the macrolevel demonstrated many intelligent behaviors in infant word learning, but they cannot provide a formal account of the underlying mechanisms. Thus, one wants to know not only *what* learners can do using social cues but also *how* they make use of those cues. To answer the second question, one needs to tackle the problem at the microlevel and study real-time sensitivities to body cues in natural contexts. Recent studies of adults performing visual–motor tasks in natural contexts have suggested 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 a timescale 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. Also, in studies of language production, recent work (e.g., Griffin & Bock, 2000; Meyer, Sleiderink, & Levelt, 1998) has shown that speakers have a strong tendency to look toward objects referred to by speech. Meyer et al. 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. Note that these body movements operate on a timescale that is much more rapid than the typical head and hand movements used in studies of infant sensitivity to an adult speaker's intentions.

The goal of this study is to combine the foregoing perspectives on language development, embodied cognition, and speech production to create a computational model of the three key tasks of early lexical development reviewed in Section 1.We propose that speakers' body movements, such as eye movements, head movements, and hand movements, can reveal their referential intents in verbal utterances, and in turn may play a significant role in early language development (Yu & Ballard, 2003; Yu et al., 2003). A plausible starting point for learning the meanings of words is the deployment of speakers' intentional body movements to infer their referential intentions. To support this idea, we provide a first formal account of how the intentions derived from body movements, which we term *embodied intention,* facilitate the early stage of vocabulary acquisition. We argue that infants learn words through their sensitivity to others' intentional body movements. They use temporal synchrony between speech and referential body movements to infer referents in speech. Our work takes some first steps in that direction by examining the problem through both empirical research and computational modeling. In the next section, we present the methods and results of experiments with adult language learners who are exposed to a second language, to study the role of embodied intention in a context that mimics some of the features of infant language acquisition. In the following section, we then propose a computational model of word learning to simulate the early stage of infant vocabulary learning. Our implemented model is able to build semantic representations grounded in multisensory input, using the principles of embodied intention. The essential structure of the model is that it assigns an important computational role for making inferences

about the speakers' referential intentions, by using body movements as deictic references (Ballard et al., 1997), thereby employing nonlinguistic information as constraints on statistical learning of linguistic data.

## 4. Simulations using human adults

To study the role of extralinguistic factors in lexical development, such as sensitivity to attention cued by eye gaze, we conducted a first experiment using adults. The purpose of this experiment was to establish a "proof of concept" that our hypotheses about the role of embodied intention in the early stage of language learning had merit. Using adults, of course, 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 these adults may not have the same level of plasticity as infant learners. Nonetheless, previous language learning studies have shown similar findings for adults exposed to an artificial language (Saffran, Newport, et al., 1996), and for children (Saffran, Newport, Aslin, Tunick, & Barrueco, 1997) or even infants exposed to the same types of language materials (Saffran, Aslin, et al., 1996). This suggests that certain mechanisms involved in language learning are available to humans regardless of age. Lakoff and Johnson (1999) argued that children have already built up prelinguistic concepts (internal representations of the world) in their brains prior to the development of the lexicon. Thus, if we assume that those concepts are already established, the lexical learning problem would mainly deal with how to find a sound pattern from continuous speech and associate this linguistic label with a concept previously established nonlinguistically. Furthermore, Gillette et al. (1999), Snedeker and Gleitman (2004), and Gleitman et al. (2005) designed the Human Simulation Paradigm in which they used adults to model the target population (infants). The argument is that although adults have different perceptual, cognitive, and memory systems, and the representations of concepts held by adults may differ from those of young children, a considerable part of vocabulary learning is not limited by immaturities in early conceptual development, but rather by solving the word-to-world mapping problem. Therefore, there should be little difference between adults and children with regard to acquiring simple words as long as they are provided with the same information. In light of these considerations, our first experiment was conducted with monolingual adults exposed to a second language to shed light on the role of embodied intention at the earliest stage of infant language learning. The experiment consisted of two phases. In the training phase, participants were asked to watch a video and try to identify which sound sequences were words in the language and associate them with their meanings. In the testing phase, they were given tests to assess both speech segmentation and lexical learning.

### *4.1. Methods*

#### *4.1.1. Participants*

Twenty-seven monolingual English-speaking students at the University of Rochester participated in this study and were paid $10 for their participation. Participants were randomly assigned to three experimental conditions, with 9 participants in each condition.

*4.1.2. Stimuli*

Participants were exposed to non-native language materials by watching a videotape with a sound track in Mandarin Chinese. In the video, a native speaker of Mandarin described in his own words the story shown in a picture book entitled, *I Went Walking* (Williams & Vivas, 1989). The book is for 1- to 3-year-old children, and the story is about a young child who goes for a walk and encounters several familiar friendly animals. The speaker was instructed to narrate the story in the same way that a caregiver would speak to a child. For each page of the book, the speaker saw a picture and uttered verbal descriptions. The study included two video clips and one audio clip that were recorded simultaneously when the speaker was narrating the story. These materials provided three different learning conditions for the adult participants, all of whom were native speakers of English who had no familiarity with Mandarin: audio-only, audiovisual, and intention-cued conditions. In the audio-only condition, the only information participants received was the auditory signal. In the audiovisual 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 intention-cued condition, the video was recorded from a head-mounted camera to provide a dynamic first-person view. Furthermore, an eye tracker was used 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. Specifically, the speaker's monocular eye position was monitored with an Applied Science Laboratories (ASL; Bedford, Massachusetts) eye tracker. The eye position signal was sampled at 60 Hz and had a time delay of 50 msec. The accuracy of the eye-in-head signal was approximately 1° over a central 40° field. Both pupil and first Purkinje image centroids were recorded, and horizontal and vertical eye-in-head positions were calculated, based on the vector difference between the two centroids. This technique reduces artifacts due to any movement of the headband with respect to the head. The ASL headband held a miniature "scene camera" to the left of the speaker's head, aimed at the scene (the picture book). The tracker creates a cursor, indicating eye-in-head position, which is merged with the video from the scene camera, thereby providing a video record of the scene from the speaker's perspective, with the cursor indicating the intersection of the participant's gaze with the picture book. Because the scene camera moves with the head, the eye-in-head signal indicates the gaze position with respect to the world. Fig. 1 shows snap-

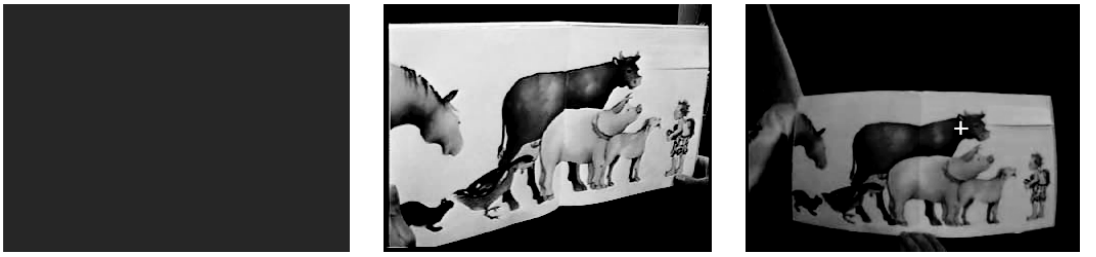


Fig. 1. The snapshots when the speaker uttered "The cow is looking at the little boy" in Mandarin. *Left:* No nonspeech information in audio-only condition. *Center:* A snapshot from the fixed camera. *Right:* A snapshot from a head-mounted camera with the current gaze position (the white cross).

Table 1
The translation of verbal descriptions in English

| |
|---|
| – A little boy gets up and goes to walk |
| – so first of all he finds a black cat |
| – the boy is giving the cat a hug on his neck |
| – and the little boy is taking the black cat |
| – then the little boy and the cat walk along |
| – the cat is following the little boy |
| – then he sees something that looks like a horse's tail |

shots from two video clips. Auditory information was the same in all three conditions, and the total length of the story was 216 sec. Some samples of verbal descriptions are translated into English, as shown in Table 1.

*4.1.3. Procedure*

Participants were divided into three groups: audiovisual, intention-cued, and audio-only. The first two groups were shown video clips on a computer monitor and asked to try to identify both the sound patterns that corresponded to individual words and their meanings. They watched the same video five times before being tested and were given the opportunity to take a break in the middle of each session, but few did. The audio-only group was provided with the audio recording and then tested after listening to it five times. Thus, all three groups received the same audio information but different levels of visual information.

*4.1.4. Test*

The participants in the audiovisual and intention-cued conditions were given two written multiple-choice tests: a speech-segmentation test and a word-learning test. The participants in the audio-only condition were just given the first test. There were 18 questions in each test. For every question in the first test, participants heard two sounds and were asked to select one that they thought was a word but not a multiword phrase or some subset of a word. They were given as much time as they wanted to answer each question. There are two types of distractors. One type just randomly removed a syllable at either the beginning or the end of a word. For example, the positive instance "ya zi" (duck) was paired with a distractor "ya." The other type was built by extracting some phrases that consist of more than one word. For instance, the positive instance "nan hai" (boy) was paired with "na hai zou" (boy + walk). The second test was used to evaluate their 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. Participants 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 previously mentioned).

*4.2. Results*

Fig. 2 shows the average percentage correct on the two tests. In the speech-segmentation test, a single-factor analysis of variance revealed a significant main effect of the three condi-

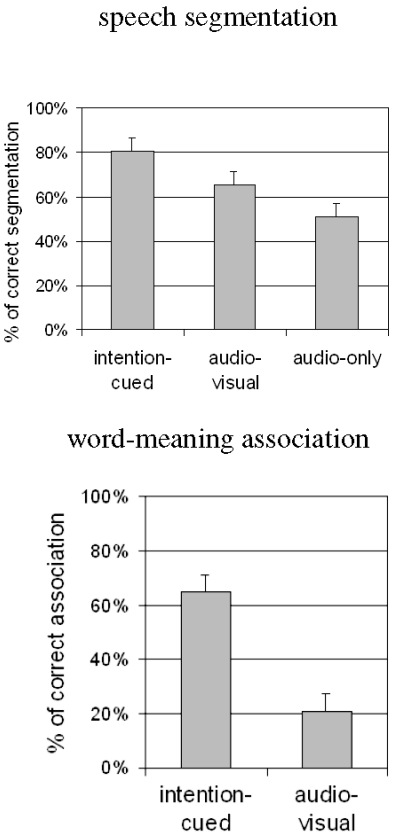



Fig. 2. The mean percentages of correct answers in tests.

tions, $F(2, 24) = 23.52$, $p < .001$. Post hoc tests showed that participants gave significantly more correct answers in the intention-cued condition, $M = 80.6\%$, $SD = 8.3\%$, than in the audiovisual condition, $M = 65.4\%$, $SD = 6.6\%$, $t(16) = 4.89$, $p < .001$. Performance in the audio-only condition did not differ from chance, $M = 51.1\%$, $SD = 11.7\%$. Participants 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 min without any visual context. Therefore, they were not asked to do the second test. For the word-learning test, performance in the intention-cued condition was much better than in the audiovisual condition, $t(16) = 8.11$, $p < .0001$. Note also that performance in the audiovisual condition was above chance, $t(8) = 3.49$, $p < .005$, one-sample $t$ tests.

### *4.3. Discussion*

The results of this study of word learning in adults exposed to a second language provide substantial evidence in support of the hypothesis that embodied intention plays an important role in language acquisition by suggesting that language learners are sensitive to gaze cues

in real-time interaction. This finding goes beyond the claims by Baldwin (1993) and Tomasello (2001) that referential intent as evidenced in gaze affects word learning. Our results suggest that social cues not only play a role in high-level learning and cognition but also influence the learning and the computation at the sensory level. However, the precise linkage between the visual cues available in the intention-cued and audiovisual conditions has not been specified.

To establish a formal model that explores the computational role of embodied intention in lexical development, a more fine-grained analysis of the information available to the learners in each condition is needed. To quantitatively evaluate the difference between the information available in the audiovisual and intention-cued conditions, the intention-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 later. 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 1,019 spoken words, and 116 of them were object names of pictures in the book. A straightforward hypothesis about the difference in information between the intention-cued and audiovisual conditions is that participants 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 intention-cued condition (but not the audiovisual condition), it could explain the superior performance on both tests in the intention-cued condition. For instance, if the onset of spoken words was always 300 msec after a saccade, then participants could simply find the words based on this delay interval. To analyze this possible correlation, we examined the time relation of eye fixation and speech production. We first spotted the keywords (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 Fig. 3, 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. From this analysis, we conclude that in this kind of natural task, eye movements and speech production are not perfectly time locked. However, the vast majority of eye movements to objects are made within 900 msec prior to the spoken word that refers to that object. 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 participants in the intention-cued condition on tests of both speech segmentation and word-meaning association. In the next section, we describe a computational model of embodied intention that is also able to use the information encoded by this dynamic synchrony to learn words.

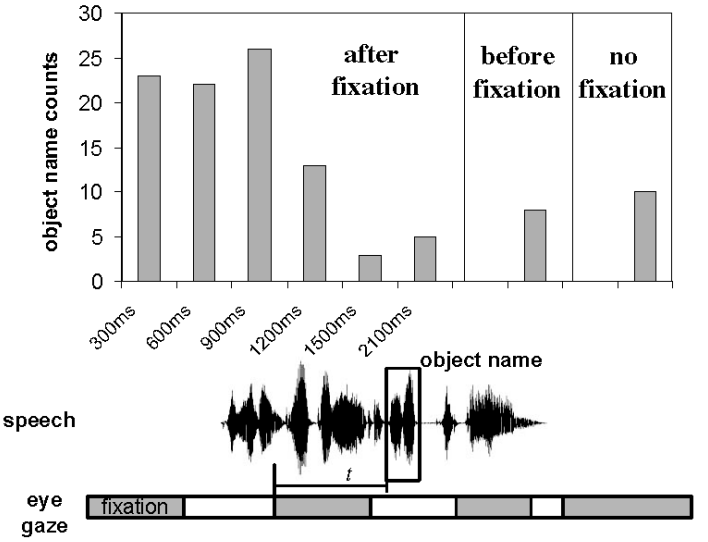



Fig. 3. The level of synchrony between eye movement and speech production. Most spoken object names were produced after eye fixations, and some of them were uttered before eye fixations. Occasionally, the speaker did not look at the objects at all when he referred to them in speech. Thus, there is no perfect synchrony between eye movement and speech production.

## 5. The computational model

The foregoing lexical-learning experiment in Mandarin Chinese suggests that online information about eye gaze facilitates the acquisition of new vocabulary items in a novel second-language context. However, the precise mechanism by which the adults in our study segmented words from fluent speech and mapped these sounds onto meanings remains unclear. Moreover, as pointed out earlier, studies of adults learning a second language may not reflect the same underlying mechanisms used by infants and young children as they learn their first language. Thus, in this section, which represents the bulk of this research program, we build a computational model that learns lexical items from raw multisensory signals to more closely resemble the difficulties infants face during the early phase of language acquisition. 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.

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 participants who were asked to act as caregivers and perform some everyday activities, one of


Fig. 4. The computational model shares multisensory information like a human language learner. This allows the association of coincident signals in different modalities.

which was narrating the picture book (used in the preceding experiment) in English for a young child, thereby simulating natural infant–caregiver interactions. As shown in Fig. 4, 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 young 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 the caregiver's referential intentions. In sum, the model we are building is essentially an ideal observer; if the simulated infant learner can acquire from realistic multisensory input, under plausible conditions, the kinds of information we know infants actually acquire in the natural environment, then we can use this model to further explore language acquisition as a simulation, rather than relying solely on empirical findings from the literature on infants and young children. Such a model is useful not only as a description of underlying mechanisms of language learning, but also as a tool for making predictions about aspects of language development that have not been studied empirically or that would be difficult to study in the natural environment.

To learn words from caregivers' spoken descriptions (shown in Fig. 5), three fundamental problems need to be addressed: (a) object categorization to identify grounded meanings of words from nonlinguistic contextual information, (b) speech segmentation and word spotting to extract the sound patterns of the individual words that might have grounded meanings, and (c) association between spoken words and their meanings. To address those problems, our model consists of the following components as shown in Fig. 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 representations of the objects that the speaker is attending to at attentional points in time and categorizes them into groups.

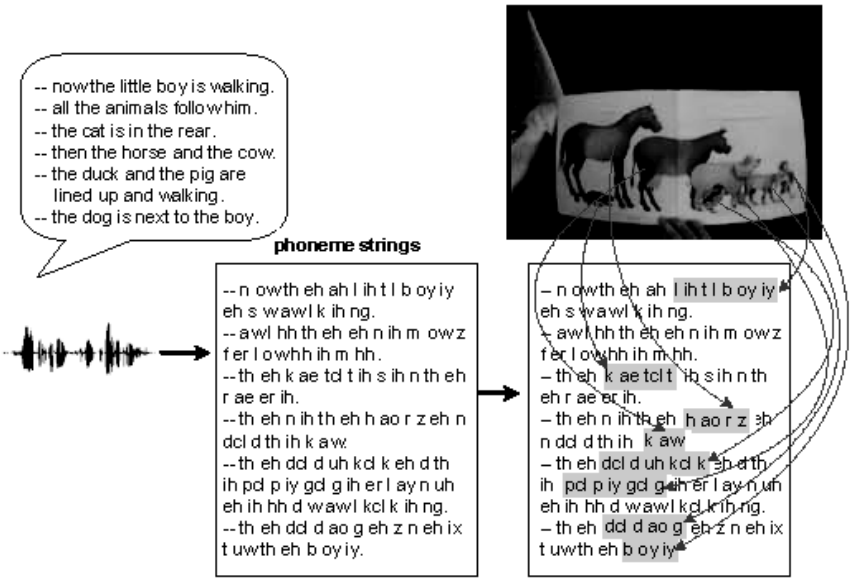



Fig. 5. The problems in word learning. The raw speech is first converted into phoneme sequences. The goal of our method is to discover phoneme substrings that correspond to the sound patterns of words and then infer the meanings of those words from nonlinguistic modalities.

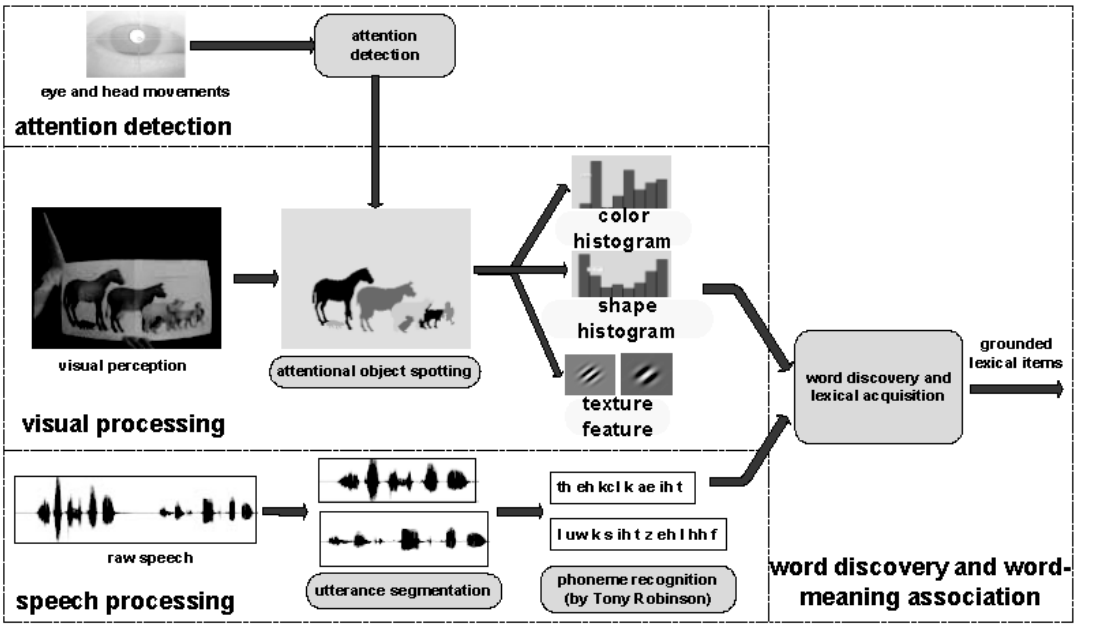



Fig. 6. The overview of the system. The system first estimates participants' focus of attention, then uses spatiotemporal correlations of multisensory input at attentional points in time to associate spoken words with their perceptually grounded meanings.

- *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 to cluster those subsequences.
- *Word learning* is the crucial step in which information from different modalities is integrated to discover words from fluent speech and map them to their grounded meanings extracted from visual perception.

### 5.1. 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 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 hidden Markov model consists of a set of $N$ states $S = \{s_1, s_2, s_3, \ldots, s_N\}$, the transition probability matrix $A = a_{ij}$, where $a_{ij}$ is the transition probability of taking the transition from state $s_i$ to state $s_j$, prior probabilities for the initial state $\pi_i$, and output probabilities of each state $b_i(o(t)) = P\{o(t)|s(t) = s_i\}$. Salvucci and Anderson (1998) first proposed a HMM-based fixation identification method that uses probabilistic analysis to determine the most likely identifications for a given protocol. Our approach is different from theirs in two ways. First, we use training data to estimate the transition probabilities instead of setting predetermined values. Second, we noticed that head movements provide valuable cues to model the focus of attention. This is because participants almost always orient their heads toward the object of interest, thereby keeping their eye-position with respect to the head in the center of their visual field. Therefore, head position was integrated with eye position as the input to the HMM.

A two-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 two-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. Specifically, we need to compute the means ($\mu_{j1}$, $\mu_{j2}$) and variances ($\sigma_{j1}$, $\sigma_{j2}$) of two-dimensional Gaussians (four parameters) for each state and the transition probabilities (two parameters) between two states. 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. Using the actual evidence from the training data, a new estimate for the respective output probability can be assigned:

$$\bar{\mu}_j = \frac{\sum_{t=1}^{T} \gamma_t(j) O_t}{\sum_{t=1}^{T} \gamma_t(j)} \tag{1}$$

and

$$\bar{\sigma}_j = \frac{\sum_{t=1}^{T} \gamma_t(j)(O_t - \bar{\mu}_j)(O_t - \bar{\mu}_j)^T}{\sum_{t=1}^{T} \gamma_t(j)} \tag{2}$$

where $\gamma_t(j)$ is defined as the posterior probability of being in state $s_j$ at time $t$, given the observation sequence and the model.

As a result, 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. An example of the results from this eye-data analysis is shown in Fig. 7.

### 5.2. Visual processing

The attention-detection information, based on gaze information from eye and head movements, must be linked to the objects contained in the visual scene. Thus, there must be a mechanism to define what an object is and where it is located in the scene on a moment-to-moment

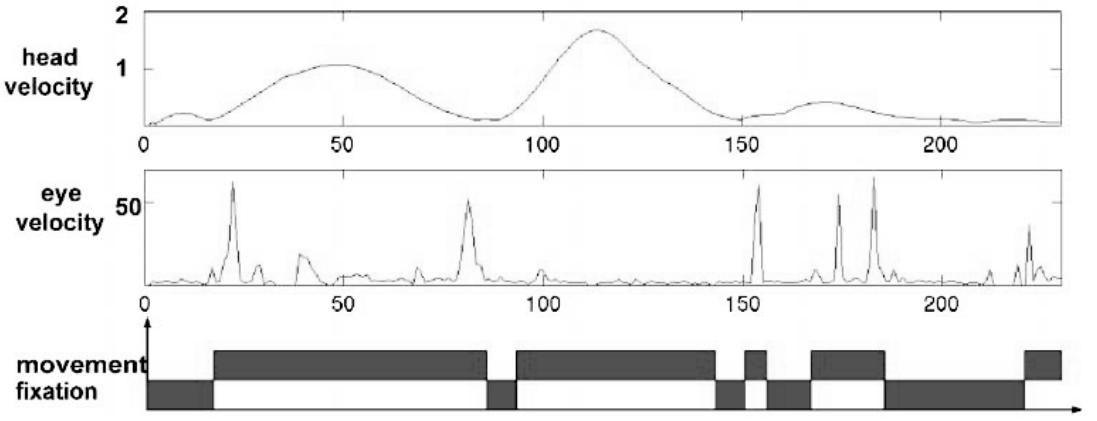



Fig. 7. Eye-fixation finding. *The top plot:* the velocity profile of the head. *The middle plot:* point-to-point velocities of eye movements. *The bottom plot:* a temporal state sequence of HMM (the *fixation* label indicates the fixation state and the *movement* label represents the saccade state).

basis as the participant is shifting visual attention. The visual data provided by the head-mounted camera (the scene) and by gaze information form the contexts in which spoken utterances are produced. Thus, the possible referents of spoken words are encoded in those contexts, and we need to extract those perceptually grounded meanings (i.e., the objects that serve as the referents of the words) from raw sensory inputs. As a result, we will obtain a temporal sequence of possible referents depicted by the box labeled *intentional contexts* in Fig. 9 on p. 983. Our method first uses eye and head movements as cues to estimate the participant's focus of attention as described in the previous section. Attention, as represented by eye fixation, is then used for spotting the objects of the participant'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, and then we extract a perceptual representation based on the visual appearance of the object. In this way, the referential intentions can be directly inferred from attentional objects. This approach consists of two steps: attentional object spotting and object categorization.

*5.2.1. Attentional object spotting*

Object spotting detects the visual objects fixated by speakers at critical points as speech output unfolds in real time, which provides perceptually grounded meanings in word learning. Knowing attentional states allows for automatic object spotting by integrating visual information with eye gaze data, which consists of two steps. First, the snapshots of the scene are segmented into blobs using low-level image features (Wang & Siskind, 2003). The result of image segmentation is illustrated in Fig. 8(b), and only blobs larger than a threshold are used. Next, we group those blobs into semantic objects. Our approach starts with a segmented image, uses gaze positions as seeds, and repeatedly merges the most similar regions to form new groups until all the blobs are labeled. Eye gaze in each attentional time is then used as a cue to extract the object of speaker interest from all the detected objects.

*5.2.1.1. Similarity measurement*

We use color as the similarity feature for merging regions. $L * a * b$ color space is adopted to overcome undesirable effects caused by varied lighting conditions and achieve more robust illumination–invariant segmentation. $L * a * b$ color consists of a luminance or lightness component ($L^*$) and two chromatic components: the $a^*$ component (from green to red) and the $b^*$ component (from blue to yellow). To this effect, we compute the similarity distance between two blobs in the $L^* a * b$ color space by employing the histogram intersection method proposed by Swain and Ballard (1991). If $C_A$ and $C_B$ denote the color histograms of two regions $A$ and $B$, their histogram intersection is defined as:

$$h(A,B) = \frac{\sum_{i=1}^{n} min(C_A^i, C_B^i)}{\sum_{i=1}^{n} (C_A^i + C_B^i)} \tag{3}$$

where $n$ is the number of the bin in the color histogram, and $0 < h(A, B) < 0.5$. Two neighboring regions are merged into a new region if the histogram intersection $h(A, B)$ is between a thresh-

old $t_c (0 < t_c < 0.5)$ and 0.5. Although this similarity measure is fairly simple, it is remarkably effective in determining color similarity between regions of multicolored objects.

*5.2.2. Merging process*

The approach of merging blobs is based on a set of regions selected by speakers' gaze fixations, termed *seed regions.* We start with a number of seed regions $S_1, S_2, \ldots, S_n$, in which $n$ is the number of regions that participants were attending to. Given those seed regions, the merging process then finds a grouping of the blobs into semantic objects with the constraint that the regions of visual objects are chosen to be as homogeneous as possible. The process evolves iteratively from the seed regions. Each step of the algorithm involves the addition of one blob to one of the seed regions and the merging of neighboring regions based on their similarities.

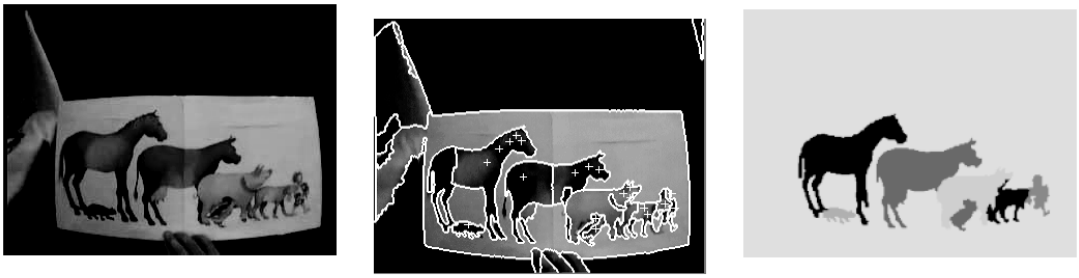
Our method is implemented using a sequentially sorted list (Adams & Bischof, 1994), which is a linked list of blobs ordered according to some attribute. In each step, we consider the blob at the beginning of the list. When adding a new blob to the list, we place it according to its value of the ordering attribute so that the list is always sorted, based on the attribute. Let $N(A)$ be the set of immediate neighbors of the blob A, which are seed regions. For all the regions $N(A)_1, N(A)_2, \ldots, N(A)_n$, the seed region that is closest to $A$ is defined as:

$$B = \arg\max_i h(A, N(A)_i); 1 \leq i \leq n \tag{4}$$

where $h[A, N(A)_i]$ is the histogram function measuring the similarity distance between region $A$ and $N(A)_i$, based on the selected similarity feature. The ordering attribute of region $A$ is then defined as $h(A, B)$. The merging procedure is illustrated in Appendix A. Fig. 8 shows how these steps are combined.

*5.2.3. Clustering visually grounded meanings*

The extracted objects are represented by a model that contains color, shape, and texture features (Yu & Ballard, 2004). Based on the work of Mel (1997), we construct the visual features of objects that are large in number, invariant to different viewpoints, and driven by multiple visual cues. Specifically, 64-dimensional color features are extracted by a color indexing method



(a) Original image (b) Image segmentation with gaze positions (c) Object spotting

Fig. 8 The procedure of attentional object spotting.

(Swain & Ballard, 1991), and 48-dimensional shape features are represented by calculating histograms of local shape properties (Schiele & Crowley, 2000). Gabor filters with three scales and five orientations are applied to the segmented image. It is assumed that the local texture regions are spatially homogeneous, and the mean and the standard deviation of the magnitude of the transform coefficients are used to represent an object in a 48-dimensional texture feature vector. Thus, feature representations consisting of a total of 160 dimensions are formed by combining color, shape, and texture features, which provide fundamental advantages for fast, inexpensive recognition.

Most classification algorithms, however, do not work efficiently in high dimensional spaces because of the inherent sparsity of the data. This problem has been traditionally referred to as the curse of dimensionality. In our system, we reduced the 160-dimensional feature vectors into 15 vectors by using principle component analysis (Aggarwal & Yu, 2000), which represents the data in a lower dimensional subspace by pruning away those dimensions that result in the least loss of information. Next, because 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 (Hartigan, 1975). Finally, we select a centroid (a feature vector in the visual space) of each cluster as the perceptually grounded representation of word meanings.

### *5.3. Speech processing*

Infants begin to organize their phoneme categories in an adultlike manner by the age of 6 months (Jusczyk, 1997). Therefore, our computational model first converts acoustic signals into phoneme strings to simulate this capability. Then we need a method to compare phoneme strings and identify words embedded in continuous speech. The methods of phoneme recognition and phoneme string comparison, which provide a basis for building word-meaning associations, are described in this subsection (Ballard & Yu, 2003).

#### *5.3.1. Phoneme recognition*

We implemented an endpoint detection algorithm to segment the speech stream into spoken utterances. Each spoken utterance contains one or more spoken words. Then the speaker-independent phoneme-recognition system developed by Robinson (1994) is employed to convert spoken utterances into phoneme sequences. The method is based on recurrent neural networks (RNN) that perform the mapping from a sequence of the acoustic features extracted from raw speech to a sequence of phonemes. The training data of RNN are from the TIMIT database—phonetically transcribed American English speech—which consists of read sentences spoken by 630 speakers from eight dialect regions of the United States. To train the networks, each sentence is presented to the recurrent back-propagation procedure. The target outputs are set using the phoneme transcriptions provided in the TIMIT database. Once trained, a dynamic programming match (Kruskal, 1999) is made to find the most probable phoneme sequence of a spoken utterance—for example, the boxes labeled *phoneme strings* in Fig. 9. The output consists of 61 phonemic and phonetic symbols used in the TIMIT lexicon and in the phonetic transcriptions.

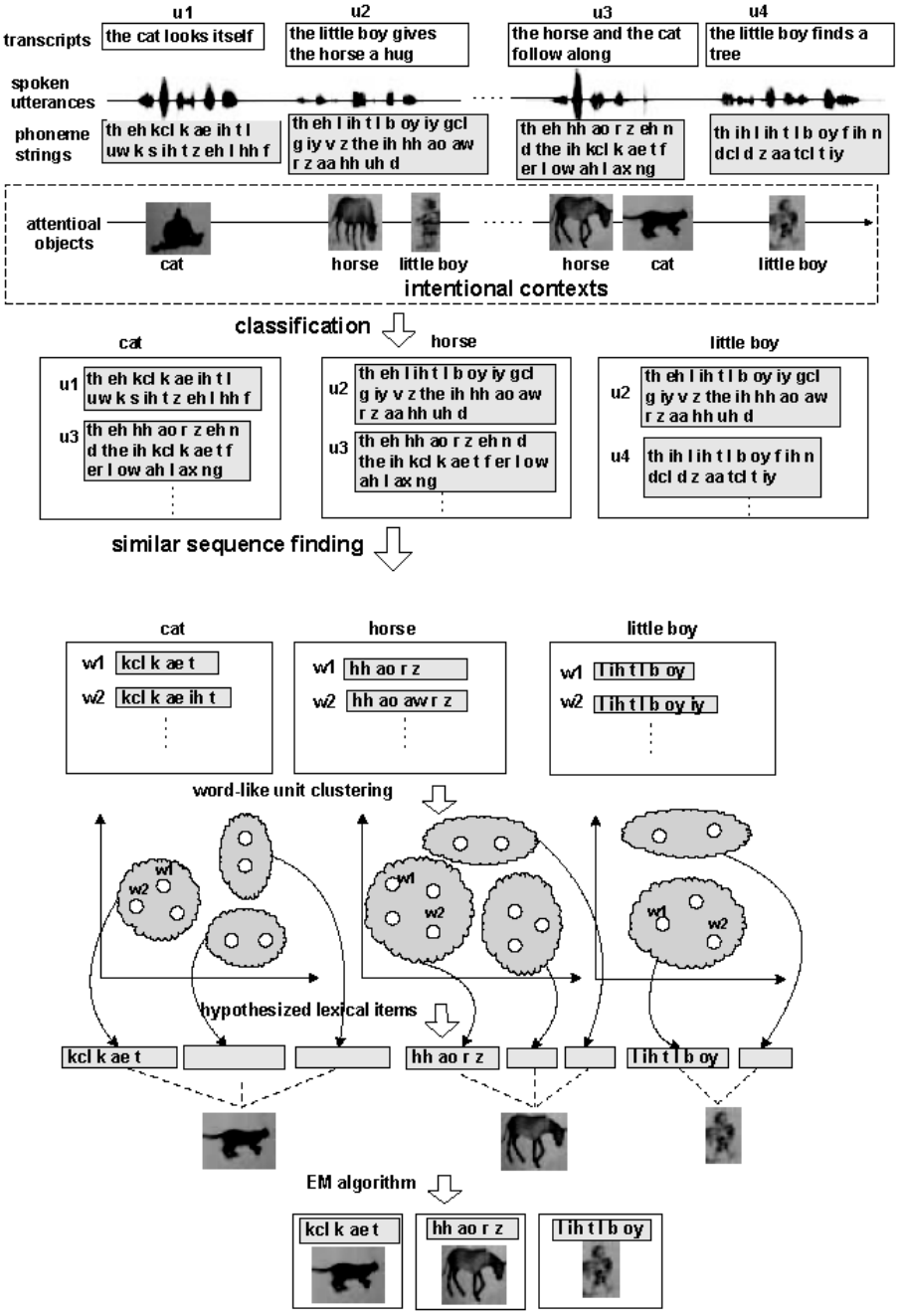



Fig. 9. Overview of the method for word learning. Spoken utterances are categorized into several bins that correspond to temporally co-occurring attentional objects. Then we compare any pair of spoken utterances in each bin to find the similar subsequences that are treated as wordlike units. Next, those wordlike units in each bin are clustered, based on the similarities of their phoneme strings. The expectation maximization (EM) algorithm is applied to find lexical items from hypothesized word-meaning pairs.

*5.3.2. Comparing phoneme sequences*

In our model, the comparison of phoneme sequences has two purposes: One is to find the longest similar substrings of two phoneme sequences (wordlike unit spotting described in Subsection 5.4.1), and the other is to cluster segmented wordlike units represented by phoneme sequences into groups (wordlike unit clustering presented in Subsection 5.4.2). In both cases, an algorithm for the alignment of phoneme sequences is a necessary step. Given raw speech input, the specific requirement here is to cope with the acoustic variability of spoken words in different contexts and by various speakers. Due to this variation, the outputs of the phoneme recognizer previously described are noisy phoneme strings that are different from phoneme transcriptions of text. In this context, the goal of phoneme string matching is to identify sequences that might be different actual strings, but have similar pronunciations.

*5.3.2.1. Similarity between individual phonemes.* To align phoneme sequences, we first need a metric for measuring distances between phonemes. We represent a phoneme by a 12-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. Based on Ladefoged (1993), the features we selected are consonantal, vocalic, continuant, nasal, anterior, coronal, high, low, back, voicing, strident, and sonorant. Each feature vector is binary, that is, the number 1 represents the presence of a feature in a phoneme and zero represents the absence of that feature. When two phonemes differ by only one distinctive feature, they are known as being minimally distinct from each other. For instance, phonemes /p/ and /b/ are minimally distinct because the only feature that distinguishes them is "voicing." We compute the distance $d(i, j)$ between two individual phonemes $i$ and $j$ as the Hamming distance, which sums up all value differences for each of the 12 features in two vectors. The underlying assumption of this metric is that the number of binary features in which two given sounds differ is a good indication of their proximity. Moreover, phonological rules can often be expressed as a modification of a limited number of feature values. Therefore, sounds that differ in a small number of features are more likely to be related. We compute the similarity matrix, which consists of $n \times n$ elements, where $n$ is the number of phonemes. Each element is assigned a score that represents the similarity of two phonemes. The diagonal elements are set to be zeros, and the other elements in the matrix are assigned negative values $[-d(i, j)]$ that correspond to the Hamming distance of distinctive features between two phonemes. In addition, a positive value is set as the reward of two matching phonemes in two strings.

*5.3.2.2. Alignment of two phoneme sequences.* The outputs of the phoneme recognizer are phoneme strings with the time stamps of the beginning and the end of each phoneme. We subsample the phoneme strings so that symbols in the resulting strings contain the same duration. We then apply the concept of similarity to compare phoneme strings. A similarity scoring scheme assigns positive scores to pairs of matching segments and negative scores to pairs of dissimilar segments. The optimal alignment is the one that maximizes the overall score. Fundamental to the algorithm is the notion of string-changing operations of dynamic programming (Kruskal, 1999). To determine the extent to which two phoneme strings differ from each

other, we define a set of primitive string operations. By applying several string operations, one phoneme string can be aligned with the other. The measurement of the similarity of two phoneme strings then corresponds to the sum of both the cost of individual string operations in alignment and the reward of matching symbols. To identify the phoneme strings that may be of similar pronunciation, the method needs to consider not only the similarity of phonemes but also their durations.

Thus, each phoneme string is subject to variations in speed (the duration of the phoneme being uttered). Such variations can be considered as compression and expansion of the phoneme with respect to the time axis. In addition, additive random error may also be introduced by interpolating or deleting original sounds. One step toward dealing with such additional difficulties is to perform the comparison in a way that allows for deletion–insertion as well as compression–expansion operations. In the case of an extraneous sound that does not delay the normal speech but merely conceals a bit of it, deletion–insertion operations permit the concealed bit to be deleted and the extraneous sound to be inserted, which is a more realistic and perhaps more desirable explanation than that permitted by additive random error. The detailed technical descriptions of our phoneme comparison method can be found in Appendix B.

### *5.4. Word learning*

We now describe our approach to integrating multimodal data for word acquisition (Yu & Ballard, 2004). Fig. 9 illustrates our method for spotting words and establishing word-meaning associations.

#### *5.4.1. Wordlike unit spotting*

The central idea of spotting wordlike units is to use nonspeech contextual information to identify a "chunk" of speech as a candidate "word." The reason for using wordlike units is that some objects are verbally described by noun phrases (e.g., "little boy") but not by single object names. The inputs shown in Fig. 9 are phoneme sequences ($u_1$, $u_2$, $u_3$, $u_4$) and possible meanings of words (attentional objects) extracted from nonspeech perceptual input. Those phoneme utterances are categorized into several bins, based on their possible associated meanings. For each meaning, we find the corresponding phoneme sequences uttered in temporal proximity and then categorize them into the same bin labeled by that meaning. For instance, $u_1$ and $u_3$ are temporally correlated with the object "cat," so they are grouped in the same bin labeled by the object "cat." Note that, because one utterance could be temporally correlated with multiple meanings in a perceptual context, it is possible that an utterance is selected and classified in multiple bins. For example, the utterance $u_2$ "the little boy gives the horse a hug" is produced while a participant is looking at both the object "little boy" and the object "horse." In this case, the utterance is put into two bins: one corresponding to the object "little boy" and the other labeled by the object "horse." Next, based on the method described in Subsection 5.3.2, we compute the similar substrings between any two phoneme sequences in each bin to obtain wordlike units. Fig. 10 shows an example of extracting wordlike units from the utterances $u_2$ and $u_3$ that are in the bin of the object "horse."

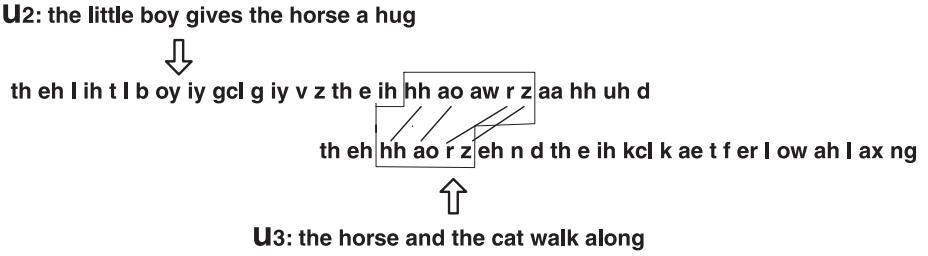



Fig. 10. An example of wordlike unit spotting. The similar substrings of two sequences are /hh ao aw r z/ (horse) and /hh ao r z/ (horse).

### 5.4.2. *Wordlike unit clustering*

As shown in Fig. 9, the extracted phoneme substrings of wordlike units are clustered by a hierarchical agglomerative clustering algorithm that is implemented based on the comparison method described in Subsection 5.3.2. The centroid of each cluster is then found and adopted as a prototype to represent this cluster. Those prototype strings are associated with their possible grounded meanings to build hypothesized lexical items. Among them, some are correct, such as /kcl k ae tcl t/ (cat)[1] associated with the object of "cat," and some are not relevant to the attentional objects. Now that we have hypothesized word-meaning pairs, the next step is to select reliable and correct lexical items.

### 5.4.3. *Multimodal integration*

In the final step, the co-occurrence of multimodal data selects meaningful semantics that associate visual representations of objects with spoken words, which are illustrated in Fig. 11. We take a novel view of this problem as being analogous to the word alignment problem in machine translation. For that problem, given texts in two languages (e.g., English and French), computational linguistic techniques can estimate the probability that an English word will be translated into any particular French word and then align the words in an English sentence with the words in its French translation. Similarly, for our problem, if different meanings can be looked at as elements of a "meaning language," associating meanings with object names can be

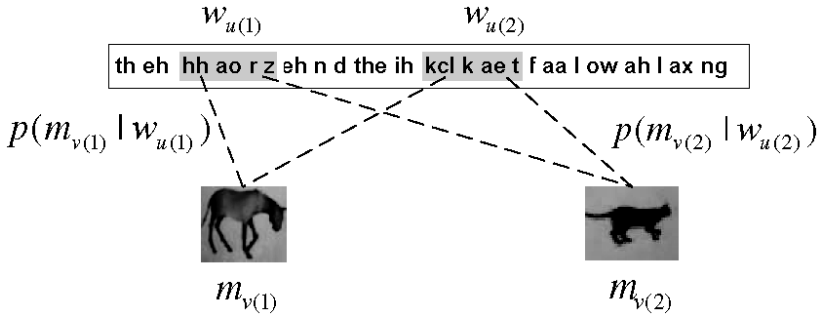



Fig. 11. Word meaning association. The wordlike units in each spoken utterance and co-occurring meanings are temporally associated to build possible lexical items.

viewed as the problem of identifying word correspondences between English and "meaning language." Thus, a technique from machine translation can address this problem. The probability of each word is modeled as a combination of the association probabilities of each word with its possible meanings. In this way, an EM algorithm can find the reliable associations of spoken words and their grounded meanings. Informally, we first derive the likelihood function of observing the data set. The EM algorithm starts with randomly assigned values of association probabilities and then iteratively alternates two steps: the expectation step (E-step) and the maximization step (M-step). In the E-step, it computes the expected likelihood of generating observation data, given these estimates of association probabilities. In the M-step, it reestimates those probabilities by maximizing the likelihood function. Once we have a new set of estimates, we can repeat the E-step and M-step. This process continues until the likelihood function converges.

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 wordlike units and $M$ is the number of perceptually grounded meanings. Let $S$ be the number of spoken utterances. All data are in a set $\chi = \{(S_w^{(s)}, S_m^{(s)}), 1 \le s \le S\}$, 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 $S_m^{(s)}$ includes $l$ possible meanings $m_{v(1)}, m_{v(2)}, \ldots, m_{v(l)}$, and the value of $v(j)$ is from 1 to $M$. We assume that every word $w_n$ can be associated with a meaning $m_m$. Given a data set $X$, we want to maximize the likelihood of generating the "meaning" corpus, given English descriptions:

$$p(S_m^{(1)}, S_m^{(2)}, \ldots, S_m^{(S)} \mid S_w^{(1)}, S_w^{(2)}, \ldots, S_w^{(S)}) = \prod_{s=1}^{S} p(S_m^{(s)} \mid S_w^{(s)}) \tag{5}$$

The independence assumption is satisfied reasonably well because it is less likely that a word in a spoken utterance refers to a visual object not occurring in this context but in other contexts. We use the method similar to that of Brown, Pietra, Pietra, and Mercer (1993). The joint likelihood of a meaning string given a spoken utterance:

$$\begin{aligned} p(S_m^{(s)} \mid S_w^{(s)}) &= \sum_a p(S_m^{(s)}, a \mid S_w^{(s)}) \\ &= \frac{\varepsilon}{(r+1)^l} \sum_{a_1=0}^{r} \sum_{a_2=0}^{r} \cdots \sum_{a_l=0}^{r} \prod_{j=1}^{l} p(m_{v(j)} \mid w_{a_{v(j)}}) \\ &= \frac{\varepsilon}{(r+1)^l} \prod_{j=1}^{l} \sum_{i=0}^{r} p(m_{v(j)} \mid w_{u(i)}) \end{aligned} \tag{6}$$

where the alignment $a_{v(j)}$, $1 \le j \le l$ can take any value from 0 to $r$ that indicates which word is aligned with the $j$th meaning. The term $p(m_{v(j)}|w_{u(i)})$ is the association probability for a word-meaning pair, and $\varepsilon$ is a small constant. In this way, the joint likelihood function of meaning strings given paired word strings can be expressed in terms of those word-meaning association probabilities.

We wish to find the association probabilities so as to maximize $p(S_m^{(s)} \mid S_w^{(s)})$ subject to the constraints that for each word $w_n$:

$$\sum_{m=1}^{M} p(m_m \mid w_n) = 1 \tag{7}$$

The assumption here is that a word refers to one meaning in a specific context. Note that multiple words could potentially refer to the same meaning. For instance, both the word *dog* and the word *doggie* refer to the object "dog." On the other hand, the model does not accommodate homonyms. Thus, multiple meanings compete with each other to obtain one linguistic label, and they could not be assigned to the same word. We believe that learning homonyms involves many exposures to a word in different contexts and through multiple sessions. Because the data in this study are collected from reading a single picture book, the issue of learning homonyms is beyond the goals of this model.

Next, we introduce Lagrange multipliers $\lambda_n$ and seek an unconstrained maximization:

$$h(t,\lambda) = \sum_{s=1}^{S} p(S_m^{(s)} \mid S_w^{(s)}) - \sum_{n=1}^{N} \lambda_n \left( \sum_{m=1}^{M} p(m_m \mid w_n) - 1 \right) \tag{8}$$

We then compute derivatives of the previously described objective function with respect to the multipliers $\lambda_n$ and the unknown parameters $p(m_m|w_n)$ and set them to be zeros. As a result, we obtain that:

$$\lambda_n = \sum_{m=1}^{M} \sum_{s=1}^{S} c(m_m \mid w_n, S_m^{(s)}, S_w^{(s)}) \tag{9}$$

$$p(m_m \mid w_n) = \lambda_n^{-1} \sum_{s=1}^{S} c(m_m \mid w_n, S_m^{(s)}, S_w^{(s)}) \tag{10}$$

where

$$c(m_m \mid w_n, S_m^{(s)}, S_w^{(s)}) = \frac{p(m_m \mid w_n)}{p(m_m \mid w_{u(1)}) + \ldots + p(m_m \mid w_{u(r)})} \times \sum_{j=1}^{l} \delta(m_m, v(j)) \sum_{i=1}^{r} \delta(w_n, u(i)) \tag{11}$$

The intuition behind the numerator in Equation 11 is that the more often a word and a meaning co-occur, the more likely that they are to be mutual translations. The denominator indicates that the co-occurrence count of a word and a meaning should be discounted to the degree that the meaning also correlates with other words in the same pair. The EM-based algorithm sets an initial $p(m_m|w_n)$ to be a flat distribution and performs the E-step and the M-step successively until convergence. In the E-step, we compute $c(m_m \mid w_n, S_m^{(s)}, S_w^{(s)})$ by Equation 11. In the M-step, we reestimate both the Lagrange multipliers and the association probabilities using Equations 9 and 10. When the association probabilities converge, we obtain a set of $p(m_m|w_n)$ and need to select the correct lexical items from many possible word-meaning associations.

Compared with the training corpus in machine translation, our experimental data are sparse and consequently cause some words to have inappropriately high probabilities to associate with the meanings. This is because those words occur very infrequently and are in a few specific contexts. We therefore use two constraints for selection. First, only words that occur more than a predefined number of times are considered. Moreover, for each meaning $m_m$, the system selects all the words with the probability $p(m_m|w_n)$ greater than a predefined threshold. In this way, one meaning can be associated with multiple words. This is because people may use different names to refer to the same object, and the spoken form of an object name can be expressed differently. For instance, the phoneme strings of both "dog" and "doggie" correspond to the object "dog." Therefore, the system is constrained to learn all the spoken words that have high probabilities in association with a particular meaning.

### 5.5. Experimental setup

A Polhemus 3D tracker (Polhemus, Colchester, Vermont) was used to acquire 6-DOF (Degree of Freedom) head positions at 40 Hz. A participant wore a head-mounted video-based, infrared reflection eye tracker from ASL. The headband of the ASL held a miniature scene camera to the left of the participant's head, which provided the video of the scene. The video signals were sampled at the resolution of 320 columns × 240 rows of pixels at the frequency of 15 Hz. The gaze positions on the image plane were provided at a frequency of 60 Hz and had a real-time delay of 50 msec. The acoustic signals were recorded using a headset microphone at a rate of 16 kHz with 16-bit resolution. Six participants, all native speakers of English, took part 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. The model was designed to learn spoken words of object names that were referred to by speech, such as dog, duck, horse, and pig. For evaluation purposes, we manually annotated the speech data and calculated the frequencies of words. We collected approximately 660 spoken utterances, and on average, a spoken utterance contained six words, which illustrates the necessity of word segmentation from connected speech. Among all these words, only approximately 15% of them are object names that we want to spot and associate with their grounded meanings. This statistic further demonstrates the difficulty of learning lexical items from naturally co-occurring data. It is important to note that these hand-coded annotations were only used for evaluation purposes; our model did not use these data as extra information to constrain the learning process.

### 5.6. Results and discussions

To evaluate the results of the experiments, we defined the following four measures for segmenting wordlike units and building grounded lexical items.

- *Semantic accuracy* measures the categorization accuracy of clustering visual feature vectors of attentional objects into semantic groups. Each category corresponds to one perceptually grounded representation of word meaning.

- *Speech-segmentation accuracy* measures whether the beginning and the end of phoneme strings of wordlike units are word boundaries. For example, the string /kcl k ae tcl t/ is a positive instance corresponding to the word *cat,* whereas the string /kcl k ae tcl t i/ is negative. The phrases with correct boundaries are also treated as position instances for two reasons. One is that those phrases are consistent with some word boundaries but combine some words together. The other reason is that some phrases correspond to concrete grounded meanings, which are exactly the spoken units we want to extract (e.g., "little boy").
- *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 that are spotted by the model. This measure provides a quantitative indication about how much lexical knowledge can be acquired, based on a certain amount of exposure.

Table 2 shows the results for the four measures. 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 pretrained 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.

Table 2
Results of word acquisition

| Subjects (%) | Semantics (%) | Speech Segmentation (%) | Word-Meaning Association (%) | Lexical Spotting (%) |
|---|---|---|---|---|
| 1 | 80.3 | 72.6 | 91.3 | 70.3 |
| 2 | 83.6 | 73.3 | 92.6 | 73.2 |
| 3 | 79.2 | 71.9 | 86.9 | 76.5 |
| 4 | 81.6 | 69.8 | 89.2 | 72.9 |
| 5 | 82.9 | 69.6 | 86.2 | 72.6 |
| 6 | 76.6 | 66.2 | 83.1 | 72.8 |
| Average | 80.6 | 70.6 | 88.2 | 73.1 |

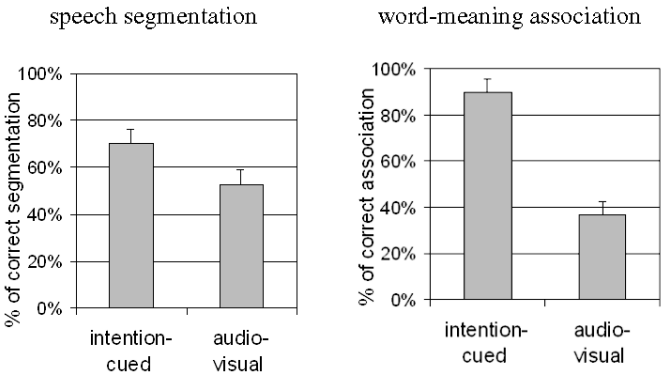



Fig. 12. A comparison of performance of the intention-cued method and the audiovisual approach.

To more directly demonstrate the role of embodied intention in language learning, we processed the data by another method in which the inputs of eye gaze and head movements were removed, and only audiovisual data were used for learning (e.g., Roy & Pentland, 2002). 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 intention-cued approach. Fig. 12 shows the comparison of these two methods.

The intention-cued approach outperforms the audiovisual approach in both speech segmentation, $t(5) = 6.94$, $p < .0001$, and word-meaning association, $t(5) = 23.2$, $p < .0001$. 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. In addition, the results obtained from this comparative study are very much in line with the results obtained from human participants, suggesting that not only is our model cognitively plausible, but the role of embodied intention can be appreciated by both human learners and by the computational model.

## 6. General discussion

### *6.1. The embodiment of word learning*

Computational models of development and cognition have changed radically in recent years. Many cognitive scientists have recognized that models that 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 that 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 this 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 (Bailey, Chang, Feldman, & Narayanan, 1998; Brent & Cartwright, 1996; Cohen, Oates, Adams, & Beal, 2001; Siskind, 1996). 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 (Arbib, 2005; 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.

### *6.2. The role of social 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-month-old infants exposed to Mandarin Chinese under audio-videotape or au-

ditory-only conditions did not show phoneme learning. Both studies indicate that learning is influenced by the presence of a live person generating social cues to attract infant attention and motivate learning. As reviewed in Section 3, 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, 2001; Tomasello & Farrar, 1986).

In this article, we focused on the ability of the young language learner to infer interlocutors' 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 social 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 intention-cued condition outperformed participants in the audiovisual condition in both word discovery (segmentation) and word-meaning tests, indicating that human participants can use 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 social cues.

Social cues are referential in nature. In the computational model described in the previous section, a speaker's referential intentions are estimated and used 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 using 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 embodied intentions 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 embodied intention serves to facilitate, and may in fact be a necessary feature of, learning the vocabulary in a new language.

### 6.3. The integration of social cues and statistical learning

Most computational models of word learning are based on associative mechanisms (see Regier, 2003, for a review). Probabilistic techniques, such as connectionist models, Bayesian inference, and latent semantic analysis, have been applied to model word learning. Many of these approaches use spatiotemporal contiguity to determine the referent of a word. However, parents do not carefully and explicitly name objects for their children in many cultures. Thus, words are not typically used at the moment their referents are perceived. For instance, Gleitman (1990) showed that most of the time the child does not observe something being opened when the verb "open" is used. Nevertheless, children have no difficulty in learning such words. Associative learning, without some further constraints or additional information, cannot explain this observation.

Our computational model demonstrates that intentional cues can be directly used to discover what objects in the world should get mapped to words. Thus, social cues are useful to address the spatial ambiguity in word learning by selecting correct lexical items from multiple co-occurring word-meaning pairs. The success of this model suggests that social cues could also be one of the driving forces to deal with the problem of temporal contiguity described previously. We propose that social cues could filter out irrelevant information and make the model focus on specific moments in time selected by social-object correlations. If the model just "zooms in" on those critical moments, it may find that words and meanings are most often temporally co-occurring. Therefore, the model does not need to process a large amount of irrelevant data, but concentrates on multisensory input at those critical moments. This mechanism suggests that in early word learning, although infants perceive multisensory input, the brain might refuse to process those data until the referential intentions of a speaker have been detected from social cues, such as the speaker's gaze. Then the brain only needs to process sensory data captured at those moments to learn words. In this way, social cues provide a gating mechanism that determines whether co-occurring data are relevant or not. This strategy seems to be more efficient compared with a purely statistical learning mechanism that needs to deal with a large amount of irrelevant data in calculating the statistical properties of the co-occurrences that are relevant from multimodal data.

The implementation of our computational model shows how the constraints of social cues and statistical learning can be integrated naturally. We compared the performance of our approach with the one based solely on associative learning, and the results demonstrated that our model outperformed the associative one in both speech segmentation and word-meaning association. However, we cannot claim that this social spotlight is an indispensable part of early word learning. Even though the results of associative learning are not as good as the intention-cued approach in both human experiments and computational modeling, associative mechanisms alone can nevertheless learn some correct word-meaning pairs. Previous work (e.g., MacWhinney, 1989; Plunkett et al., 1992; Regier et al., 2001; Roy & Pentland, 2002; Siskind, 1996; Tenenbaum & Xu, 2000) also showed that purely statistical or associative learning models can accomplish the word-learning task, based on multiple- or even one-trial exposures.

To sum up this discussion, our work shows that social cues could be seamlessly integrated into the framework of statistical learning in modeling early word learning. We also provide quantitative results on the effect of considering social cues. Based on our results and

previous work on infant word learning by Baldwin (1993) and Tomasello (2001), it seems clear that social cues are helpful in early language acquisition. From the perspective of machine learning, it is possible that one can build an associative model based on statistical learning and obtain very good results. Although that model can provide the machinery to learn language for machine intelligence, it might not be sufficient to simulate natural intelligence and infant development without including social cues. One extension of this work is to ask whether statistical learning is sufficient for word learning or whether social cues are necessary. To answer this type of question, we need to set up a word-learning experiment similar to the one used in studying speech segmentation (Aslin et al., 1996) in which all the other cues are removed, and only distributional information remains. Then we can gradually add different kinds of social cues to identify to what degree infants use social cues and under what situations.

### *6.4. Fast mapping*

One striking fact about early language acquisition is *fast mapping,* the rapid learning of a new word based on only a few exposures (Carey & Bartlett, 1978; Markson & Bloom, 1997). Our experiments do not try to explain the mechanism of fast mapping. However, the general principles of our computational model show great potential for solving this problem. Computationally, a key issue in modeling fast mapping is to find a referent of a word from the sea of ambiguity. A common scenario is a language learner who is faced with multiple words on the one hand and multiple referents on the other hand. The learner must filter out irrelevant information and discover the one-to-one mapping from many-to-many possible word-to-world associations. For example, an adult may say, "This is a car," while there is a toy car in the environment. On the language side, all the spoken words (*this, is, a,* and *car,* etc.) could refer to a toy car. Infants have to determine which word is the object's name. The EM algorithm in our computational model can address this issue. Besides providing a relatively limited number of the most probable lexical items, the EM algorithm also generates a large number of word-meaning pairs with uncertainty (low probabilities). This indicates that infants may potentially accumulate valuable information about many word–semantic associations long before these associations are unique. For example, they hear "this" numerous times in quite different contexts. Then the model already has multiple possible meanings linked to the word *this,* with very low probabilities due to the constraint in Equation 7 or syntactic knowledge. On the other hand, the new word *car* may just be perceived once or a very few times, but always in the context of the visual object "car." Again, based on Equation 7, the only possible meaning that could be assigned to the word *car* is the visual object "car." By simultaneously reducing the hypotheses in word and meaning spaces, a few exposures might be enough for the model to find the correct word from multiple candidates and the relevant meaning from the context.

On the semantics side, infants need to select a correct meaning from multiple referents in a natural environment. Markman (1995) proposed a set of particular constraints infants might use to map words onto meanings. These constraints include the whole-object assumption, mutual exclusivity, and the taxonomic assumption. In addition to those constraints, our computational model is able to observe a speaker's intentional body movements to figure out the

speaker's referential intentions in speech. Compared with models based on associative learning, this is a big step in addressing and partially solving the problem of referential indeterminacy (Quine, 1960). This model shows that body cues can be directly used to find the referent of the spoken name. However, it does not address the problem of what meaning is derived from that referent. In the previous example, the sound could refer to the object name, its position, its color, or any other possible meaning. This leaves an interesting question for future research: How deep can infants read the intent of the adult and use it to disambiguate the possible meanings of words?

### *6.5. Word discovery using visual context and joint attention*

Most studies of speech segmentation and word discovery focus on the role of linguistic information as a central constraint. Previous experiments involving human participants have found several linguistic cues that infants may use to detect words from connected speech, such as transitional probabilities (Saffran, Aslin, et al., 1996), utterance boundaries (Aslin et al., 1996), stress patterns of syllables (Jusczyk, Houston, & Newsome, 1999), and allophonic and phonotactic cues (Jusczyk, Hohne, et al., 1999). Inspired by empirical studies, computational modelers have proposed and developed several algorithms to simulate lexical segmentation in infancy (see a brief review in Section 3). However, both empirical and computational studies do not consider the role of nonlinguistic information in speech segmentation. A child does not learn language by closing his or her eyes and just hearing acoustic signals. Rather, the child is situated in a natural environment and learns language in this rich context. In our work, we propose that in addition to linguistic cues, nonlinguistic contextual cues could also play an important role in speech segmentation. This idea has been confirmed in both experimental and computational studies reported in this article. In our human simulation studies, we provided three learning environments for language learners: intention-cued, audiovisual, and audio-only conditions. The results from the audio-only condition, in which participants just listened to the same 216-sec audio recording five times, were close to chance. Participants told us that after listening to fluent speech in Mandarin for about 15 min, they could not get any information about what constitutes a word or what those words might mean (in the absence of a referential context). The superior performance in the intention-cued over the audiovisual condition showed that the more specific the context is, the more effective language learners can use it for speech segmentation. We propose that visual attention to an object that may be referred to by ongoing speech helps language learners to segment words from fluent speech. To support this idea, our computational model provides a formal account of the role of nonlinguistic cues in word discovery. We showed that the sound patterns frequently appearing in the same context are likely to have grounded meanings related to this context. Thus, by finding the frequently uttered sound patterns in a specific context (e.g., an object that speakers intentionally attend to), the model discovers wordlike sound units as candidates for building lexical items.

Because all objects in the scene are potential referents for any word that may appear, and the segmentation method is based on matching sound sequences in one context with those in another, without social spotlights the model needs to collect many more pairs of sound sequences for consideration. However, most of these pairs are not relevant and lead to false string

matches. This results in the performance difference between the intention-cued approach and the audiovisual approach.

### 6.6. Perceptually grounded word meanings

There is evidence that from an early age infants are able to form perceptually based category representations (Quinn, Eimas, & Rosenkrantz, 1993). Those categories are highlighted by the use of common words to refer to them. Thus, the meaning of the word *dog* corresponds to the category of dogs, which is initially a nonlinguistic mental representation in the brain. Furthermore, Schyns and Rodet (1997) argued that the representations of object categories emerge from the features that are perceptually learned from visual input during the developmental course of object recognition and categorization. In this way, object naming by young children is essentially about mapping words to selected perceptual properties.

Most researchers agree that infants generalize names to new instances on the basis of some similarity, but there are many debates about the nature of what defines similarity (see a review in Landau, Smith, & Jones, 1998). It has been shown that shape is generally attended to for solid rigid objects, and children attend to other specific properties, such as texture, size, or color for objects that have eyes or are not rigid (Smith, Jones, & Landau, 1996). In light of the perceptual nature of infant categorization, our model represents object meanings as perceptual features consisting of shape, color, and texture extracted from the visual appearances of objects. The categories of objects are formed by clustering those perceptual features into groups. Our model then chooses the centroid of each category in the perceptual feature space as a representation of the meaning of this category and associates this feature representation with linguistic labels. As a result, when the model, as a young language learner, perceives a novel exemplar, it will be able to find the corresponding category of this instance by comparing the similarities between the perceptual feature of the exemplar and the perceptual representations of categories that it has previously learned.

One interesting question in object naming is how does word learning during infancy influence the formation of categories to which those words refer? Waxman and Hall (1993) showed that linguistic labels may facilitate the formation of a category by infants at 16 to 21 months of age. Because adults create features to subserve the representations and categorizations of objects (Schyns, Goldstone, & Thibaut, 1998), the formation of categories in the presence of auditory input is possibly based on the similarities of perceptual features and linguistic labels. Our current model provides a framework to further investigate this question and offers a formal account of how linguistic, perceptual, and conceptual advances are linked (Landau, 2004; Roberts & Jacob, 1991; Waxman, 2004).

### 6.7. The interaction between word discovery and word-meaning mapping

Instead of performing a complete segmentation of any given utterance (or corpus), our model focuses on spotting specific kinds of sound patterns that have visually grounded meanings and therefore are necessary to build an early vocabulary. Here we do not claim that language learners just spot key words instead of the complete segmentation of speech. Actually, many studies (e.g., Saffran, Aslin, et al., 1996) have shown that infants do have the ability to

solve much of the speech-segmentation problem. However, we argue that it is unlikely that infants only begin to build their vocabulary once their ability to segment speech is fully developed.

Our model provides another cognitively plausible explanation that nonlinguistic context can also play an important role in word discovery. This is done by categorizing spoken utterances based on their contexts and extracting frequently uttered sound patterns from a specific context to form hypothesized wordlike units. Thus, our model suggests that there exists an interaction between speech segmentation and the mapping of words to their meanings. Specifically, the model addresses word discovery and word-meaning mapping problems simultaneously and integrates perceptual information from different modalities at the early stages of sensory processing. This approach is quite different from typical methods of multimodal integration, which first extract symbolic representations from sensory data and then merge those symbolic representations from different modalities to find correlations between symbols. Our work suggests that infants' initial perceptual, linguistic, cognitive, and social capabilities may be based on a general system in which those subsystems interact with one another and provide useful information to facilitate the development of other components.

### *6.8. Assumptions and limitations of the model*

The range of problems we need to address in modeling lexical acquisition in a purely unsupervised manner and from raw multimodal data is substantial. In our effort to make concrete progress, we made some assumptions to simplify the modeling task and to allow us to focus on the most fundamental problems in lexical acquisition. First, the model mainly deals with how to associate visual representations of objects with their spoken object names. This is based on the finding that a predominant proportion of infants' early vocabulary consists of nouns, which has been confirmed in various languages and under varying child-rearing conditions (Caselli et al., 2000). Also, the model is able to learn only object names that are grounded in visual perception but not other nouns that represent other unseen meanings or abstract notions, such as the word "yesterday." We believe that those initial and imageable words directly grounded in the physical environment serve as a foundation for the acquisition of abstract words that become indirectly grounded through their relations to those grounded words. Thus, we believe that the development of word hierarchies and category hierarchies, which happens at later stages of lexical development, is based on this learning process of grounding object names in visual perception and, therefore, is beyond the scope of this article.

Second, the model is not designed to simulate the development of infants' initial capabilities to recognize phonemes from acoustic input. By the age of 6 months, infants start to distinguish phonetic differences, but ignore other noncontrastive differences, and begin to organize their phoneme categories in an adult-like manner (see Jusczyk, 1997, for a review). Therefore, we assume that a language learner has knowledge of the phonemic structure of the language prior to lexical development. However, it is still an open question as to whether children have the innate capability to segment a speech stream into discrete alphabetic sound units like phonemes. However, to closely resemble natural environments that infants live in, our model does deal

with sound variation in spoken words as a function of phonetic context, speaking rate, and talker differences (see details in Subsection 6.1).

Third, in natural conditions, a language learner observes the body movements of an interlocutor and infers referential objects by means of monitoring the speaker's gaze direction. Due to the difficult logistical problem of tracking the speaker's gaze direction and head movement from the perspective of the listener (viewer) who is attempting to pick up on cues to learn the meanings of words, both our empirical study of adults and our computational model used information from an eye tracker and position sensors as input to the learner. This presumes that the actual infant learner has access to similar information in a natural context, and that question must await further experimental studies. Specifically, we would like to answer the following questions in future work: (a) If listeners receive less spatially precise information about gaze than we provided in this experiment, where a fixation cross was visible continuously, to what degree do they use gaze cues? (b) If listeners are provided with temporally dense gaze information, as in this experiment, but as a result they are not always looking at the speaker, to what degree do they make use of these gaze cues?

## 7. Conclusion

This work demonstrates a significant role of embodied intention in infant word learning through both simulation studies of adults learning a second language and a computational model. In both cases, regardless of whether the language learner is a human participant or a computational model, the intention-cued approach outperformed the audiovisual approach. We conclude that a purely statistical learning approach to language acquisition will be less efficient and may, in fact, fail because of the inherent ambiguity in natural learning environments. The inference of embodied intention, as one of infants' social cognitive skills, provides constraints to avoid the large number of irrelevant computations and can be directly applied as deictic references to associate words with visually grounded referents. In future work, we plan to apply the computational mechanism outlined in this model to actual infant data to further evaluate our proposal for the role of embodied intention in language acquisition.

## Notes

1. We used the TIMIT phoneme set that consists of 61 symbols.

## Acknowledgments

The authors wish to express their thanks to Mary Hayhoe, Elissa Newport and Mike Tanenhaus for fruitful discussions. Brian Sullivan was a great help in building the experimental system. We would like to thank Terry Regier, Raymond W. Gibbs, and Benjamin Bergen for insightful and detailed comments on an earlier version of this article.

## References


Adams, R., & Bischof, L. (1994). Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 16,* 641–647.

Aggarwal, C. C., & Yu, P. S. (2000). Finding generalized projected clusters in high dimensional spaces. In *Proceedings of the ACM Sigmod.* Dallas, TX: ACM.

Arbib, M. A. (2005). From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. *Behavioral and Brain Sciences., 28,* 105–167.

Aslin, R. N., Woodward, J. C., LaMendola, N., & Bever, T. (1996). Models of word segmentation in fluent maternal speech to infants. In J. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bahrick, L., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides infants's selective attention, perceptual and cognitive development. *Current Directions in Psychological Science, 13,* 99–102.

Bailey, D. (1997). *When push comes to shove: A computational model of the role of motor control in the acquisition of action verbs.* Unpublished doctoral dissertation, University of California, Berkeley.

Bailey, D., Chang, N., Feldman, J., & Narayanan, S. (1998, August). Extending embodied lexical development. Paper presented at *Twentieth Annual Meeting of the Cognitive Science Society* (*COGSCI-98*), Madison, WI. Retrieved September 11, 2005, from http://www.icsi.berkeley.edu/NTL/papers/cogsci98.pdf

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

Baldwin, D. A., Markman, E. M., Bill, B., Desjardins, R. N., Irwin, J. M., &Tidball, G. (1996). Infant's reliance on a social criterion for establishing word–object relations. *Child Development, 67,* 3135–3153.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20,* 1311–1328.

Ballard, D. H., & Yu, C. (2003). A multimodal learning interface for word acquisition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (Vol. 5, pp. 784–787).

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind.* Cambridge: MIT Press.

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

Booth, A. E., & Waxman, S. R. (2002). Object names and object functions serve as cues to categories for infants. *Developmental Psychology, 38,* 948–957.

Brent, M. R. (1997). Toward a unified model of lexical acquisition and lexical access. *Journal of Psycholinguistic Research, 26,* 363–375.

Brent, M. R. (1999a). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Journal of Machine Learning, 34,* 71–106.

Brent, M. R. (1999b). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Science, 3*(8), 294–301.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61,* 93–125.

Brown, P. F., Pietra, S., Pietra, V., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, 19*(2), 263–311.

Butterworth, G. (1991). The ontogeny and phylogeny of joint visual attention. In A. Whiten (Ed.), *Natural theories of mind: Evolution, development, and simulation of everyday mindreading* (pp. 223–232). Oxford, England: Blackwell.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. In *Papers and reports on child language development* (pp. 17–29). Stanford University.

Caselli, M. C., Casadio, P., & Bates, E. (2000). Lexical development in English and Italian. In M. Tomasello & E. Bates (Eds.), *Language development: The essential reading* (pp. 76–110). Oxford, England: Blackwell.

Christiansen, M., Allen, J., & Seidenberg, M. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13,* 221–268.

Church, K. W. (1987). Phonological parsing and lexical retrieval. *Cognition, 25,* 53–69.

Clark, A. (1997). *Being there: Putting brain, body and world together again.* Cambridge, MA: MIT Press.

Cohen, P. R., Oates, T., Adams, N., & Beal, C. R. (2001). Robot baby 2001. *Lecture Notes in Artificial Intelligence, 2225,* 32–56.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language, 31,* 218–236.

Ervin-Tripp, S. (1973). Some strategies for the first two years. In T. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 261–286). New York: Academic.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kucaj, II (Ed.), *Language development: Vol 2. Language, thought and culture* (pp. 310–334). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition, 73,* 135–176.

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

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

Gleitman, L., & Newport, E. (1995). The invention of language by children: Environmental and biological influences on the acquisition of language. In L. R. Gleitman & M. Liberman (Eds.), *An invitation to cognitive science* (2nd ed.,Vol. 1, pp. 1–24). Cambridge, MA: MIT Press.

Gogate, L., Walker-Andrews, A., & Bahrick, L. (2001). Intersensory origins of word comprehension: An ecological-dynamic systems view. *Developmental Science, 4,* 1–37.

Gopnik, A., & Choi, S. (1995). Names, relational words, and cognitive development in English and Korean speakers: Nouns are not always learned before verbs. In M. Tomasello & W. Merriman (Eds.), *Beyond names for things: Young children's acquisition of verbs* (pp. 63–80). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11,* 274–279.

Hartigan, J. (1975). *Clustering algorithms.* New York: Wiley.

Johnson, E., & Jusczyk, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44,* 548–567.

Jusczyk, P. W. (1997). *The discovery of spoken language.* Cambridge, MA: MIT Press.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants detection of the sound patterns of words in fluent speech. *Cognitive Psychology., 29,* 1–23.

Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics, 61,* 1465–1476.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39,* 159–207.

Kruskal, J. (1999). An overview of sequence comparison. In D. Sankoff & J. Kruskal (Eds.), *Time warps, string edits and macromolecules: The theory and practice of sequence comparison* (pp. 1–44). Boston: Addison-Wesley.

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences, 100,* 9096–9101.

Ladefoged, P. (1993). *A course in phonetics.* Orlando, FL: Harcourt Brace.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by.* Chicago: University of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought.* New York: Basic Books.

Landau, B. (2004). Perceptual units and their mapping with language: How children can (or can't?) use perception to learn words. In G. Hall & S. R.Waxman (Eds.), *Weaving a lexcon* (pp. 110–148). Cambridge, MA: MIT Press.

Landau, B., Smith, L., & Jones, S. (1998). Object perception and object naming in early development. *Trends in Cognitive Science 2*(1), 19–24.

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks, 17,* 1345–1362.

Macnamara, J. (1982). *Names for things: A study of child language.* Cambridge, MA: MIT Press.
MacWhinney, B. (1989). Linguistic categorization. In F. Eckman & M. Noonan (Eds.), *Linguistic categorization* (pp. 195–242). Philadelphia: Benjamins.
Mandel, D., Jusczyk, P., & Pisoni, D. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science, 6,* 314–317.
Markman, E. (1995). Constraints on word meaning in early language acquisition. In L. Gleitman & Barbara Landau (Eds.), *The acquisition of the lexicon* (pp. 199–228). Cambridge, MA: MIT Press.
Markson, L., & Bloom, P. (1997, February 27). Evidence against a dedicated system for word learning in children. *Nature, 385,* 813–815.
Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition, 78,* 91–121.
Mel, B. W. (1997). Seemore: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation, 9,* 777–804.
Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition, 66,* B25–B33.
Pinker, S. (1989). *Learnability and cognition.* Cambridge, MA: MIT Press.
Plunkett, K. (1997). Theories of early language acquisition. *Trends in Cognitive Sciences, 1,* 146–153.
Plunkett, K., Sinha, C., Miller, M., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. *Connection Science, 4,* 293–312.
Quine, W. (1960). *Word and object.* Cambridge, MA: MIT Press.
Quinn, P., Eimas, P., & Rosenkrantz, S. (1993). Evidence for representations of perceptually similar natural categories by 3-month old and 4-month old infants. *Perception, 22,* 463–475.
Rabiner, L. R., & Juang, B. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77,* 257–286.
Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism.* Cambridge, MA: MIT Press.
Regier, T. (2003). Emergent constraints on word-learning: A computational review. *Trends in Cognitive Sciences, 7,* 263–268.
Regier, T., Corrigan, B., Cabasan, R., Woodward, A., Gasser, M., & Smith, L. (2001). The emergence of words. In *Proceedings of the 23rd Annual Meeting of Cognitive Science Society* (pp. 815–820). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Richards, D., & Goldfarb, J. (1986). The episodic memory model of conceptual development: An integrative viewpoint. *Cognitive Development, 1,* 183–219.
Roberts, K., & Jacob, M. (1991). Linguistic versus attentional influences on nonlinguistic categorization in 15-month-old infants. *Cognitive Development, 6,* 355–375.
Robinson, T. (1994). An application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks, 5,* 298–305.
Roy, D., & Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science, 26,* 113–146.
Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month old infants. *Science, 274,* 1926–1928.
Saffran, J., Newport, E., Aslin, R., Tunick, R., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8,* 101–105.
Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35,* 606–621.
Salvucci, D. D., & Anderson, J. (1998). Tracking eye movement protocols with cognitive process models. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 923–928). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of Eye Tracking Research and Applications Symposium* (pp. 71–78). Palm Beach Garden, FL: ACM Press.
Schiele, B., & Crowley, J. L. (2000). Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision, 36*(1), 31–50.

Schyns, P., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 681–696.

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences, 21,* 1–54.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition, 61,* 39–61.

Siskind, J. M. (1999). *Visual event perception* (Tech. Rep. No. 99–033). Princeton, NJ: NEC Research Institute.

Slater, A., Quinn, P., Brown, E., & Hayes, R. (1999). Intermodal perception at birth: Intersensory redundancy guides newborn infants' learning of arbitrary auditory-visual pairings. *Developmental Science, 3,* 333–338.

Slobin, D. (1985). The cross-linguistic study of the language-making capacity. In D. Slobin (Ed.), *The cross-linguistic study of language acquisition* (1157–1256). Hillsdale NJ: Lawrence Erlbaum Associates, Inc.

Sloutsky, V. M., & Lo, Y.-F. (1999). How much does a shared name make things similar? Linguistic labels and the development of similarity judgment. *Developmental Psychology, 35,* 1478–1492.

Smith, L. (2000). How to learn words: An associative crane. In R. Golinkoff & K. Hirsh-Pasek (Eds.), *Breaking the word learning barrier* (pp. 51–80). Oxford, England: Oxford University Press.

Smith, L., Jones, S., & Landau, B. (1996). Naming in young children: A dumb attentional mechanism? *Cognition, 60,* 143–171.

Snedeker, J., & Gleitman, L. (2004). In Hall & S. Waxman (Eds.), *Weaving a lexicon* (pp. XX–XX). Cambridge, MA: MIT Press.

Steels, L. (1997). Synthesizing the origins of language and meanings from coevolution. In J. Hurford (Ed.), *Evolution of human language* (pp. 384–404). Edinburgh, Scotland: Edinburgh University Press.

Swain, M. J., & Ballard, D. (1991). Color indexing. *International Journal of Computer Vision, 7*(1), 11–32.

Tardif, T. (1996). Nouns are not always learned before verbs: Evidence from Mandarin speakers' early vocabularies. *Developmental Psychology, 32,* 492–504.

Tenenbaum, J., & Xu, F. (2000). Word learning as Bayesian inference. In L. Gleitman & A. Joshi (Eds.), *Proceeding 22nd Annual Conference of Cognitive Science Society* (pp. 517–522). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Thiessen, E., & Saffran, J. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39,* 706–716.

Tomasello, M. (2000). Perceiving intentions and learning words in the second year of life. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 111–128). New York: Cambridge University Press.

Tomasello, M. (2001). Perceiving intentions and learning words in the second year of life. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 132–158). New York: Cambridge University Press.

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

Wang, S., & Siskind, J. M. (2003). Image segmentation with ratio cut. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25,* 675–690.

Waxman, S., & Hall, D. (1993). The development of a linkage between count nouns and object categories: Evidence from fifteen- to twenty-one-month-old infants. *Child Development, 29,* 257–302.

Waxman, S. R. (2004). Everything had a name, and each name gave birth to a new thought: Links between early word learning and conceptual organization. In G. Hall & S. R. Waxman (Eds.), *Weaving a lexcon* (pp. 110–148). Cambridge, MA: MIT Press.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75,* 502–517.

Williams, S., & Vivas, J. (1989). *I went walking.* Chicago: Harcourt Brace.

Woodward, A., & Guajardo, J. (2002). Infants' understanding of the point gesture as an object-directed action. *Cognitive Development, 17,* 1061–1084.

Yu, C., & Ballard, D. H. (2003). Exploring the role of attention in modeling embodied language acquisition. In *Proceedings of the Fifth International Conference on Cognitive Modeling* (pp. 219–224). Bamburg, Germany: Universitats-Verlag Bamberg.

Yu, C., & Ballard, D. H. (2004). A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception, 1*(1), 57–80.

Yu, C., Ballard, D. H., & Aslin, R. N. (2003). The role of embodied intention in early lexical acquisition. In *Proceedings the Twenty Fifth Cognitive Science Society Annual Meetings* (pp. 1293–1298). Boston, MA.

## Appendix A

Fig. 13 describes the approach of semantic object spotting.

## Appendix B

Given two phoneme sequences $a_1, a_2, \ldots, a_m$ and $b_1, b_2, \ldots, b_n$, of length $m$ and $n$, respectively, to find the optimal alignment of two sequences using dynamic programming, we construct an $m$-by-$n$ matrix where the $(i^{th}, j^{th})$ element of the matrix contains the similarity score $S(a_i, b_j)$ that corresponds to the shortest possible time warping between the initial subsequences of $a$ and $b$ containing $i$ and $j$ elements, respectively. $S(a_i, b_j)$ can be recurrently calculated in an ascending order with respect to coordinates $i$ and $j$, starting from the initial condition at (1, 1) up to $(m, n)$. One additional restriction is applied on the warping process:

$$j - r \leq i \leq j + r$$

where $r$ is an appropriate positive integer called *window length.* This adjustment window condition avoids undesirable alignment caused by a too-excessive timing difference.

Let $w$ be the metric of the similarity score and $w_{del}[a_i] = min(w[a_i, a_i - 1]; w[a_i, a_i + 1])$ and $wins[b_j] = min(w[b_j, b_j - 1], w[b_j, b_j + 1])$. Fig. 14 contains the modified dynamic programming algorithm to compute the similarity score of two phoneme strings.

**Algorithm**: object segmentation based on gaze fixations
**Initialization:**
Compute the color histogram of each region.
Label seed regions according to the positions of gaze fixations.
Merge seed regions that are neighbors to each other and are close with respect to the similarity measurement.
Put neighboring regions of seed regions in the SSL.
**Merging:**
*While* the SSL is not empty
Remove the top region $A$ from SSL.
Compare the similarity between $A$ and all the regions in $N(A)$ and find the closest seed region $B$.
Merge the regions $A$ and $B$ and compute the color histogram of new region $I = A \cup B$.
Test each neighboring region $A_i$ of $A$:
If $A_i$ is labeled as a seed region
Merge the region with $I$ if they are similar.
Otherwise
Add the region to the SSL according to its color similarity with $I$, $h(A_i, I)$.

Fig. 13. The algorithm for merging blobs.

**Algorithm**: phoneme string comparison
**for** $i = 0$ to $m$ **do**
$s_{i0} = 0$
**end for**
**for** $j = 0$ to $n$ **do**
$s_{0j} = 0$
**end for**
**for** $i = 0$ to $m$ **do**
**for** $j = i - r$ to $i + r$ **do**

$$
\begin{aligned}
S_{ij} = max(&S_{i-1,j-3} + w[a_i, a_{j-2}] + \tfrac{1}{2}w[a_i, b_{j-1}] + \tfrac{1}{2}w[a_i, b_j],\\
&S_{i-1,j-2} + w[a_i, b_{j-1}] + \tfrac{1}{2}w[a_i, b_j],\\
&S_{i-1,j} + w_{ins}[a_i],\\
&S_{i-1,j-1} + w[a_i, b_j],\\
&S_{i,j-1} + w_{del}[b_j],\\
&S_{i-2,j-1} + w[a_{i-1}, a_j] + \tfrac{1}{2}w[a_i, b_j],\\
&S_{i-3,j-1} + w[a_{i-2}, a_j] + \tfrac{1}{2}w[a_{i-1}, b_j] + \tfrac{1}{2}w[a_i, b_j])
\end{aligned}
$$

**end for**
**end for**

Fig. 14. The algorithm for computing the similarity of two phoneme strings.