

The Role of Embodied Intention in Early Lexical Acquisition

Chen Yu (yu@cs.rochester.edu)

Dana H. Ballard (dana@cs.rochester.edu)

Department of Computer Science; University of Rochester
Rochester, NY 14627 USA

Richard N. Aslin (aslin@cvs.rochester.edu)

Department of Brain and Cognitive Sciences; University of Rochester
Rochester, NY 14627 USA

Abstract

We examine the influence of inferring interlocutors' referential intentions from their body movements at the early stage of lexical acquisition. By testing human subjects 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 paper presents a computational model that can identify the sound patterns of individual words from continuous speech using non-linguistic contextual information and employ body movements as deictic references to discover word-meaning associations. To our knowledge, this work is the first model of word learning which not only learns lexical items from raw multisensory signals to closely resemble natural environments of infant development, but also explores the computational role of social cognitive skills in lexical acquisition.

Introduction

To acquire a vocabulary item, a young language learner must discover the sound pattern of a word from continuous speech since spoken language lacks the acoustic analog of blank spaces of written text. Furthermore, learning a word involves mapping a form, such as the sound "cat", to a meaning, such as the concept of cat. The child senses a multitude of co-occurrences between words and things in the world, and he or she must determine which co-occurrences are relevant.

In the last ten years, there has been tremendous progress in understanding infants' ability to segment continuous speech, discover words and learn their meanings. Most research focuses on the role of linguistic information as the central constraint. 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. These include prosodic patterns (e.g., Cutler & Butterfield, 1992), phonotactic regularities (e.g., Mattys & Jusczyk, 2001), allophonic variations (e.g., Jusczyk, Hohne, & Bauman, 1999) and distributional probability (e.g., Aslin, Woodward, laMendola, & Bever, 1996; Brent & Cartwright, 1996). Recent computational approaches on child-directed corpora have also revealed that relatively simple statistical learning mechanisms could make an important contribution to certain aspects of language acquisition (for review see Brent, 1999).

Recently, a popular explanation of the word learning problem termed *associationism* assumes that language

acquisition is solely based on statistical learning of co-occurring data from the linguistic modality and non-linguistic context. 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. Roy and Pentland (2002) have developed a computational model of infant language learning, in which they used the temporal correlation of speech and vision to associate spoken utterances with a corresponding object's visual appearance. It seems quite reasonable to assume that the human cognitive system exploits this statistical information. However, 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, which has not been verified by experimental studies of infants (Bloom, 2000).

In addition to temporal co-occurrences of multisensory data, recent psycholinguistic studies (e.g., Baldwin et al., 1996; Bloom, 2000; Tomasello, 2001) have shown that other major sources of constraints in language acquisition are social cognitive skills, such as children's ability to infer the intentions of adults as adults act and speak to them. These kinds of social cognitions are called mind reading by Baron-Cohen (1995). Bloom (2000) argued that children's word learning actually draws extensively on their understanding of the thoughts of speakers. His claim has been supported by the experiments in which young children were able to figure out what adults were intending to refer to by speech. Baldwin et al. (1996) referential intent when determining the reference of a novel label. showed that infants established a stable link between the novel label and the target toy only when that label was uttered by a speaker who concurrently showed his attention toward the target, and such a stable mapping was not established when the label was uttered by a speaker who was out of view and hence showed no signs of attention to the target toy.

In a complementary study of embodied cognition, Ballard, Hayhoe, Pook, and Rao (1997) proposed a cognitive system of implicit reference termed deictic, in which the body's pointing movements are used to bind objects in the world to variables in cognitive programs of human brain. Also, in the studies of speech production, Cooper (1974) found speakers have a strong tendency to look toward objects referred to by speech.

By putting together all those ideas on shared attention and intention, we propose that speakers' body movements, such as eye, head and hand movements, can reveal their referential intents in verbal utterances, which could possibly play a significant role in early language development. A plausible starting point of learning the meanings of words is the deployment of speakers' intentional body movements to infer their referential intentions which we term *embodied intention*. This work takes some first steps in that direction by examining the problem through both empirical research and computational modeling with the hope to obtain a more complete picture. The next section presents the experiments that use adult language learners exposed to a second language to study the role of embodied intention in infant language acquisition. In light of the human subject study, we then propose a computational model of word learning to simulate the early stage of infant vocabulary learning. The implemented model is able to build meaningful semantic representations grounded in multisensory inputs. The essential structure models the computational role of the inference of speakers' referential intentions by making use of body movements as deictic references (Ballard et al., 1997), and employs non-linguistic information as constraints on statistical learning of linguistic data.

Human Simulations

Previous language-learning studies have shown similar findings for adults exposed to an artificial language and children or even infants exposed to the same type of language (Saffran, Newport, & Aslin, 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 pre-linguistic concepts (internal representations of the world) in their brains prior to the development of 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 built up. As pointed out by Gillette et al. (Gillette, Gleitman, Gleitman, & Lederer, 1999), although the representations of concepts of adults may differ from those of young children, 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 this, our first experiment was conducted with monolingual adults exposed to a second language to shed light on the role of embodied intention in the early stage of infant language learning. The experiment consists of two phases. In the training phase, subjects were asked to watch a video and try to discover lexical items. In the testing phase, they were given the tests of both speech segmentation and lexical learning.

Methods

Participants. 18 monolingual English speaking students at the University of Rochester participated in this study, and were paid for their participation. Subjects were randomly assigned to two experimental conditions,

with 9 subjects in each condition.

Stimuli. Subjects were exposed to the language by video. In the video, a person was reading the picture book of "I went walking" (Williams & Vivas, 1989) in Mandarin. The book is for 1-3 year old children, and the story is about a young child that goes for a walk and encounters several familiar friendly animals. The speaker told the story in a way similar to a caregiver describing it to a child. For each page of the book, subjects saw a picture and heard verbal descriptions. The study included two video clips that were recorded simultaneously when the speaker was reading the book, and provided different learning conditions for subjects: audio-visual condition and intention-cued condition. In audio-visual condition, the video was recorded from a fixed camera behind the speaker to capture a static view. In the intention-cued condition, we recorded video from a head-mounted camera to get a dynamic view. Furthermore, an eye tracker was utilized to track the course of the speaker's eye movements and gaze positions were overlapped on the video to indicate what the speaker was attending to. Auditory information is same in both videos. Figure 1 shows the snapshots from two video clips.



Figure 1: The snapshots when the speaker uttered "the cow is looking at the little boy" in Mandarin. **Left:** a snapshot from the fixed camera. **Right:** a snapshot from a head-mounted camera with the current gaze position (the white cross).

Procedure. Subjects were shown video clips on a computer monitor and asked to try to identify both sound patterns of 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.

Test. Subjects were given two written multiple-choice tests: a speech segmentation test and a word learning test. There were 18 questions in each test. For every question in the first test, subjects heard two sounds and were asked to select one that they thought was a word but not a phrase or a syllable. They were given as much time as they wanted to answer each question. A second test was used to evaluate their knowledge of lexical items learned from the video. The images of 12 objects in the picture book were displayed on a computer monitor. Subjects heard one isolated spoken word for each question and were asked to select an answer from 13 choices (12 objects and also an option for none of the above).

Results

Figure 2 shows the average correct answers of two tests. In the speech segmentation test, subjects made significantly more errors in the audio-visual condition ($M = 12.1$, $SD = 1.2$) than in the intention-cued condition

($M = 14.6, SD = 1.5$). The further analysis revealed a significant main effect of conditions $F(1, 16) = 23.19, p < 0.001$. For word learning, a direct comparison of the intention-cued condition ($M = 12.2, SD = 2.3$) with the audio-visual condition ($M = 4, SD = 2.2$) also revealed a significant difference ($F(1, 16) = 54.67, p < 0.001$). This human subject study provides substantial evidence for the hypothesis that embodied intention plays an important role in language acquisition. This proposal suggests that a formal model that explores the computational role of embodied intention in lexical development, should show similar advantages to intentional cues.

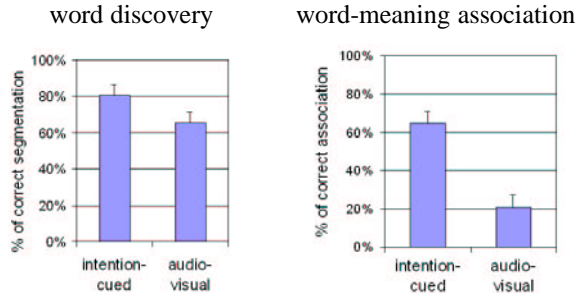


Figure 2: The mean percentages of correct answers in tests.

Computational Simulations

We supplement our empirical studies with a computational account. By implementing the descriptions of the theories or claims explicitly in computer simulations, we can not only test the plausibility of the theories but also gain the insights of both the nature of the problems and the possible solutions.

To simulate how infants ground semantics, our model needs to be embodied in the physical environment and sense the environment as a young child. To do so, we attached multiple sensors to an adult subject who was asked to act as a caregiver and perform some everyday activities, one of which was reading a picture book for a young child. Those sensors include a head-mounted CCD camera to capture visual information of the physical environment, a microphone to sense acoustic signals, an eye tracker to track the course of 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) can acquire multisensory data so that it shares the visual environment with the caregiver, hears infant-directed speech uttered by the caregiver and observes his or her body movements, such as gaze and head movements, which are deployed to infer the caregiver’s referential intentions.

The Model

To learn words from a caregiver’s spoken descriptions (shown in Figure 3), three fundamental problems needed to be addressed are: (1) object recognition to identify grounded meanings of words from visual perception, (2) speech segmentation and word spotting to extract the sound patterns of the individual words which might have grounded meanings, (3) association between spoken words and their meanings.

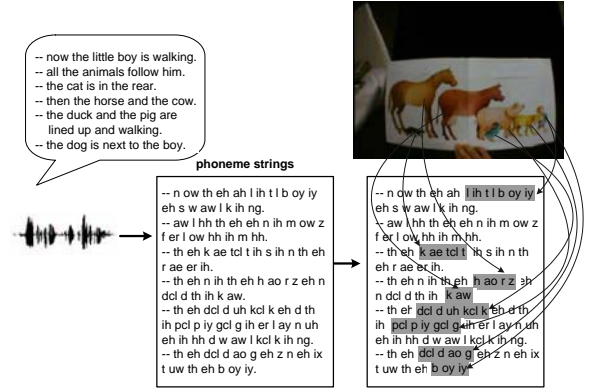


Figure 3: **The problems in word learning.** The raw speech is firstly converted into phoneme sequences. The goal of our method is to discover phoneme substrings that correspond to the sound patterns of words and then infer the meanings of those words from non-linguistic modalities.

Clustering Visually Grounded Meanings The non-linguistic inputs of the system consist of visual data from a head-mounted camera, head positions and gaze-in-head data. Those data provide the contexts in which spoken utterances are produced. Thus, the possible referents of spoken words that subjects utter are encoded in those contexts, and we need to extract those word meanings from raw sensory inputs. As a result, we will obtain a temporal sequence of possible referents depicted by the box labeled “intentional context” in Figure 4. Our method firstly utilizes eye and head movements as cues to estimate the subject’s focus of attention. Attention, as represented by eye fixation, is then used for spotting the target object of subject’s interest. Specifically, at every attentional point in time, we make use of eye gaze as a seed to find the attentional object from all the objects in a scene. The referential intentions are then directly inferred from attentional objects. We represent the objects by feature vectors consisting of color, shape and texture features. For further information see Yu, Ballard, and Zhu (2002). Next, since the feature vectors extracted from visual appearances of attentional objects do not occupy a discrete space, we vector quantize them into clusters by applying a hierarchical agglomerative clustering algorithm. Finally, for each cluster we select a prototype to represent perceptual features of this cluster.

Comparing Phoneme Sequences We describe our methods of phoneme string comparison in this subsection. Detailed descriptions of algorithms can be obtained from Ballard and Yu (2003). First, the speaker independent phoneme recognition system is employed to convert spoken utterances into phoneme sequences. To fully simulate lexical learning, the phoneme recognizer does not encode any language model or word model. Therefore, the outputs are noisy phoneme strings that are different from phonetic transcriptions of text. Thus, the goal of phonetic string matching is to identify sequences that might be different actual strings, but have similar pronunciations. In our method, a phoneme is represented by a 15-dimensional binary vector in which every entry stands for a single articulatory feature called a distinc-

tive feature. Those distinctive features are indispensable attributes of a phoneme that are required to differentiate one phoneme from another in English. We compute the distance between two individual phonemes as the Hamming distance. Based on this metric, a modified dynamic programming algorithm is developed to compare two phoneme strings by measuring their similarity.

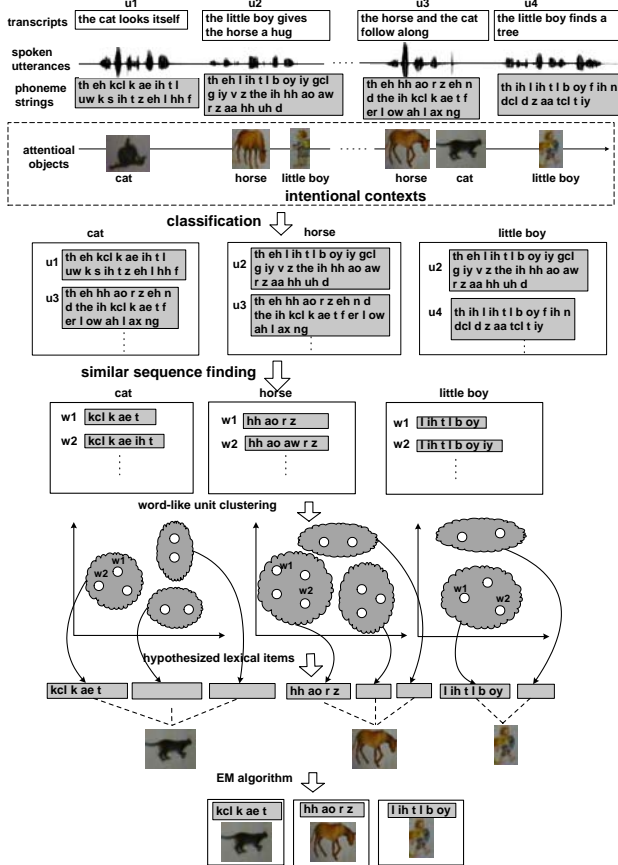


Figure 4: **Overview of the method.** Spoken utterances are categorized into several bins that correspond to temporally co-occurring attentional objects. Then we compare any pair of spoken utterances in each bin to find the similar subsequences that are treated as word-like units. Next, those word-like units in each bin are clustered based on the similarities of their phoneme strings. The EM-algorithm is applied to find lexical items from hypothesized word-meaning pairs.

Word Learning Figure 4 illustrates our approach to spotting words and establishing word-meaning associations, which consists of the following steps (See Ballard & Yu, 2003 for detailed descriptions):

- Phoneme utterances are categorized into several bins based on their possibly associated meanings. For each meaning (an attentional object), we find the corresponding phoneme sequences uttered in temporal proximity, and then categorize them into the same bin labeled by that meaning.
- The similar substrings between any two phoneme sequences in each bin are found and treated as word-like units.
- The extracted phoneme substrings of word-like units

are clustered by a hierarchical agglomerative clustering algorithm. The centroids of clusters are associated with their possible grounded meanings to build hypothesized word-meaning pairs.

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

Experimental Setup

A Polhemus 3D tracker was utilized to acquire 6-DOF head positions at 40Hz. A subject wore a head-mounted eye tracker from Applied Science Laboratories(ASL). The headband of the ASL held a miniature “scene-camera” to the left of the subject’s head, which provided the video of the scene. The video signals were sampled at the resolution of 320 columns by 240 rows of pixels at the frequency of 15Hz. The gaze positions on the image plane were reported at the frequency of 60Hz. The acoustic signals were recorded using a headset microphone at a rate of 16 kHz with 16-bit resolution. Six subjects participated in the experiment. They were asked to read the picture book (used in the previous experiment) in English. They were also instructed to pretend that they told this story for 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.

Results and Discussion

To evaluate experimental results, we define the following three measures: (1) **Semantic accuracy** is to measure the accuracy of clustering visual objects (e.g., animals) in the picture book. (2) **Word discovery accuracy** is to measure whether the beginning and the end of phoneme strings of word-like units are correct word boundaries. (3) **Word learning accuracy** is to measure the percentage of successfully segmented words that are correctly associated with their meanings.

Table 1: Results of word acquisition

Subjects	Semantics	Word discovery	Word learning
1	80.3%	72.6%	91.3%
2	83.6%	73.3%	92.6%
3	79.2%	71.9%	86.9%
4	81.6%	69.8%	89.2%
5	82.9%	69.6%	86.2%
6	76.6%	66.2%	83.1%
Average	80.6%	70.6%	88.2%

Table 1 shows the results of three measures. The recognition rate of the phoneme recognizer we used is 75% because it does not encode any language model or

word model. Based on this result, the overall accuracy of speech segmentation is 70.6%. Naturally, an improved phoneme recognizer based on a language model would improve the overall results, but the intent here is to study the learning procedure without pre-trained models. The error in word learning is mainly caused by a few words (e.g., “happy” and “look”) that frequently occur in some contexts but do not have visually grounded meanings. Considering that the system processes raw sensory data, and our learning method works in unsupervised mode without manually encoding any linguistic information, the accuracies for both speech segmentation and word learning are impressive.

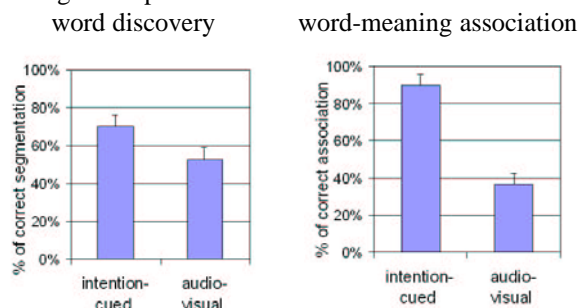


Figure 5: A comparison of performance of the intention-cued method and the audio-visual approach.

To demonstrate the role of embodied intention in language learning, we process data by another method in which eye gaze and head movements are ignored, and only audio-visual data are used for learning. In this approach, we have to classify spoken utterances into the bins of all the objects in the scene instead of just the bins of attentional objects. Except for this point, the method shares other implemented components with the intention-cued approach. Figure 5 shows the comparison of two methods. The intention-cued approach outperforms the other one in both speech segmentation and word-meaning association. The significant difference lies in the fact that there exists a multitude of co-occurring word-object pairs in natural environments that infants are situated in, and the inference of referential intents through body movements plays a key role in discovering which co-occurrences are relevant.

General Discussion

The Role of Embodied Intention

We propose that the ability of a young language learner to infer interlocutors’ referential intentions through the observations of their body movements may significantly facilitate lexical learning. This proposal has been verified by the empirical studies in which adult language learners exposed to a second language in the intention-cued condition outperformed the ones under audio-visual condition in both word discovery and word-meaning learning tests. Furthermore, in the computational model described in the previous section, a speaker’s referential intents are estimated and utilized to facilitate lexical learning in two ways. Firstly, possible referential objects in time provide cues for word spotting from a continuous speech stream. Speech segmentation without prior

language knowledge is a challenging problem and has been addressed by solely using linguistic information. In contrast, our method appreciates the importance of non-linguistic context in which spoken words are uttered. We propose that the sound patterns frequently appearing in the same context are likely to have grounded meanings related to this context. Thus, by finding frequently uttered sound patterns in a specific context (e.g., an object that subjects intentionally attend to), the model discovers word-like sound units as candidates for building lexicons. Secondly, a difficult task of word learning is to figure out which entities specific words refer to from a multitude of co-occurrences between words and things in the world. This is accomplished in our model by utilizing speakers’ intentional body movements as deictic references to establish associations between words and their visually grounded meanings.

Sensory Level Modeling

The successful implementation of the model suggests that with advances in machine learning, speech processing and computer vision, modeling lexical learning at the sensory level is not impossible and it has some advantages over symbolic simulations by closely resembling natural environments in which infants develop. 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 to associate the body and the environment with words in language (Lakoff & Johnson, 1980). Second, infants learn words by sensing the environments from their perceptual systems and coping with several practical problems, such as the variability of spoken words in different contexts and by different talkers. To closely simulate infant vocabulary development, the computational model should also 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 phoneme transcripts).

Assumptions in 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, so to make concrete progress, some natural assumptions were made to simplify the modeling task and allow us to focus on the key 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 infant early vocabulary are nouns, which has been confirmed in various languages and under varying child-rearing conditions (Caselli, Casadio, & Bates, 2000). Also, the model is able to learn only object names that are grounded in visual perception but not other nouns that represent other meanings or abstract notions. We believe that those initial and imageable words directly grounded in the physical environment serve as a foundation for the acquisi-

tion of abstract words that become indirectly grounded through their relations to those grounded words. Second, the model does not intend to simulate the development of initial capabilities to recognize phonemes from acoustic input. We assume that a language learner has knowledge of the phonetic structure of the language prior to lexical development. Third, in natural conditions, a language learner observes the body movements of an interlocutor and infers referential objects by means of monitoring his/her gaze direction. Due to the difficulties to track the speaker's gaze directions and head movements, and then search for a target object in that direction from the learner's perspective, in both empirical studies and the computational simulation, an eye tracker and position sensors are utilized so that the language learner (i.e. a human subject or the computational model) can directly obtain the interlocutor's gaze and head movements, and also share the visual scene.

Conclusion

This work demonstrates a significant role of embodied intention in infant word learning through both human subject study and computational modeling. In both cases, no matter a language learner is a human subject or a computer program, the intention-cued approach outperformed the audio-visual approach. We conclude that the solely statistical learning of co-occurrences in data is less likely to explain the whole story of language acquisition. The inference of embodied intention, as one of infants' social cognitive skills, provides constraints to avoid the large amount of irrelevant computations and can be directly applied as deictic reference to associate words with visually grounded referents in the environment. Here we do not claim that young children employ the exact method presented in this paper. However, as a computational model, this work provides an existence proof for a machine learning technique that solves the lexical acquisition task. It leaves the open question of what techniques young children actually use to solve the problem by further empirical study. 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 to future research.

References

- 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*. Hillsdale, NJ: Erlbaum.
- 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*. Hong Kong.
- 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: The MIT Press.
- Brent, M. R. (1999). 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.
- 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* (p. 76-110). Blackwell publishers.
- Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6, 84-107.
- Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, 31, 218-236.
- Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, 73, 135-176.
- Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics*, 61, 1465-1476.
- 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.
- Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78, 91-121.
- Richards, D., & Goldfarb, J. (1986). The episodic memory model of conceptual development: An integrative viewpoint. *Cognitive Development*, 1, 183-219.
- Roy, D., & Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1), 113-146.
- 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.
- Tomasello, M. (2001). In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development*. Cambridge University Press.
- Williams, S., & Vivas, J. (1989). *I went walking*. Harcourt Brace and Company.
- Yu, C., Ballard, D. H., & Zhu, S. (2002). Attentional object spotting by integrating multisensory input. In *Proceedings of the 4th international conference on multimodal interface*. Pittsburg, PA.