# Exploring the Role of Attention in Modeling Embodied Language Acquisition

**Chen Yu (yu@cs.rochester.edu)**
Department of Computer Science, University of Rochester
Rochester, NY 14627 USA

**Dana H. Ballard (dana@cs.rochester.edu)**
Department of Computer Science, University of Rochester
Rochester, NY 14627 USA

**Abstract**

Language is about symbols and those symbols must be learned during infant development. Most recently, there has been an increased awareness of the essential role of inferences of speakers' referential intentions in grounding those symbols. Experiments have shown that these inferences serve as an important driving force in language learning at a relatively early age. The challenge ahead is to develop formal models of language acquisition that can shed light on the leverage provided by embodiment and attention. This paper describes a computational model of embodied language acquisition that can simulate some of the formative steps in infant language acquisition. The novelty of our work is that the model shares multisensory information with a real agent in a first-person sense, and eye gaze is utilized as deictic reference to spot temporal correlations between different modalities. As a result, the system can build meaningful semantic representations that are grounded in the physical world. We test our model's ability to associate spoken names of objects with their visually grounded meanings and compare the results of our approach with the case that does not use referential intentions.

## Introduction

Humans develop based on their sensorimotor experiences with the physical environment. Different levels of abstraction are necessary to efficiently encode those experiences, and one vital role of human brain is to bridge the gap from embodied experience to its expression as abstract symbols. As emphasized by Harnad (1990), to simulate human intelligence, a challenge in computational modeling is to specify how to ground symbolic representations from non-symbolic sensorimotor information.

Human infants solve many complex learning problems during their first year of life. Among them, one of the most challenging tasks is to learn words (symbols) in their native language by discovering the sound patterns of words and associating them with their meanings embodied in the physical environment (non-symbolic information). Around 6 to 10 months is the stage of grasping the first words. A predominant proportion of most children's first 50 words consist of middle-size object names, such as animals (dog, cat), toys (ball, block), specific persons (mama, dad) and other artifacts (fork, chair). Object names are special in the early lexical development because they constitute a much larger proportion of children's early vocabularies compared with the vocabularies of adults. This is true for almost every language that has been studied (Caselli, Casadio, & Bates, 2000). Gillette et al. (Gillette, Gleitman, Gleitman, & Lederer, 1999) provided strong evidence that learnability of a word is primarily based upon its imageability or concreteness. Therefore, nouns are learned before most of verbs because they are more observable.

In light of the concreteness of infant language acquisition, we have proposed and implemented a computational model of embodied language learning that can build word-meaning pairs grounded in the physical environment around us. To learn a word, infants need to discover its sound pattern, recognize its meaning, and associate these two. Here we focus on the second and third problems in this work. Specifically, this paper addresses how to associate spoken words (object names, etc.) with their visually grounded meanings. The essential structure models the computational role of body movement as deictic reference in creating and simplifying brain representations (Ballard, Hayhoe, Pook, & Rao, 1997). The movements of different body parts, such as eye movements and head movements, can be utilized to infer the speaker's referential intentions, which can then be used to establish relevant word-object links from a multitude of co-occurrences between words and objects in the physical world. In this way, body movements provide a substrate for understanding human language development.

## Modeling the Role of Mind Reading in Language Acquisition

A central task of word learning is to figure out which entities specific words refer to. The speech children hear and the environments they are situated in provide a multitude of temporally co-occurring word-object pairs, and children must determine which co-occurrences are

relevant. Thus, they need to learn the meanings of words by associating verbal labels with the information from other modalities, such as visual perception.

A popular approach to this correspondence problem is *associationism* which assumes that language acquisition is based on statistical learning of co-occurring data from linguistic modality and non-linguistic modalities. The idea has been explored in both cognitive modeling and experimental psychology with success. The studies of natural language processing of child-directed English showed that distributional information is a valuable cue for speech segmentation (e.g. Brent & Cartwright, 1996). Roy and Pentland (2002) use the correlation of speech and vision to associate spoken utterances with a corresponding object's visual appearance. In their model, a lexicon is acquired by finding statistically consistent inter-modal structure. Meanwhile, recent experimental evidences in both children and adults show that the cognitive system is sensitive to features of co-occurrence statistics in language (e.g. Saffron, Aslin, & Newport, 1996). As a result, it seems a reasonable assumption that the cognitive system is likely to exploit those simple and useful sources of information in early language acquisition. 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. Nevertheless, the experimental results of parent-child interactions showed that about 30% to 50% of the time that young children heard a word, they are not attending to the object that adults refer to.

A major additional source of constraints in language acquisition is the importance of inferences of the speaker's referential intentions in language acquisition. These kind of inferences is 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 fact that an infant at age around nine months will naturally follow his mother's gaze and her pointing gesture. This finding gives rise to the argument that based on inference of the intention of adults, young children may figure out what adults are intending to refer to when words are heard. Several experimental results verified the important role of intentional cues in language learning (see Baldwin, Markman, Bill, Desjardins, Irwin, & Tidball, 1996; Baron-Cohen, Baldwin, & Crowson, 1997; Tomasello & Barton, 1994). Specifically, Baldwin et al. (Baldwin et al., 1996) showed that children associate object names with objects only if they believe that the acts of adults are naming. In their experiments, 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 attention toward the target, and such stable mapping was not established when labels were uttered by a speaker who was out of view and hence showed no signs of attention to the target toy. Based on this fact, they concluded that enriching cues to referential intent might be one of the very importance factors to enhance vocabulary acquisition.

A complementary picture emerged from studying computational models of the brain. Ballard and colleagues (Ballard, Hayhoe, Pook, & Rao, 1997) proposed that at time scales of approximately one-third of a second, orienting movements of the body play a crucial role in cognition and form a useful computational level, termed the embodiment level. At this level, the constraints of the body determine the nature of cognitive operations. From a computational perspective, this level provides a map between lower-level actions and higher-level symbols. The way it is done is through a system of implicit reference termed deictic, whereby the body's pointing movements are used to bind objects in the world to cognitive programs. This gives rise to the idea that different parts of body movements, especially eye and head movements, can be useful cues to guide the associations of data from different modalities. Furthermore, in the context of studying language understanding, Cooper (1974) found that people have a strong tendency to look toward objects referred to in conversations. He showed that the response system of eye movements in the presence of an on-going conversation is always characterized by a high degree of linguistic sensitivity.

Based on the above analyses, a challenge ahead is to build a formal model to explore the computational role of mind reading in learning words. In the model we developed, the associations between spoken words and their grounded meanings are established by employing gaze position and head direction to estimate a speaker's focus of attention.

## An Embodied Language Learning System

In order to ground language, we attach different kinds of sensors to a real person, as shown in Figure 1. Those sensors include a head-mounted CCD camera to capture a first-person point of view, a microphone to sense acoustic signals uttered by the agent, an eye tracker to track gaze positions that indicate the agent's attention, and position sensors attached to the head and hands of the agent to simulate proprioception in sense of motion. The functions of those sensors are similar to human sensory system and they allow our model to collect user-centric multisensory data to simulate the development of language acquisition capabilities.

The advantages of this approach are as follows:

- The system is embodied by sharing sensory experiences about the physical world with a real agent. Thus, the system sees as the agent sees, hears as the agent hears, and experiences the life and the environment of the agent in a first-person sense.
- The computational role of attention in language

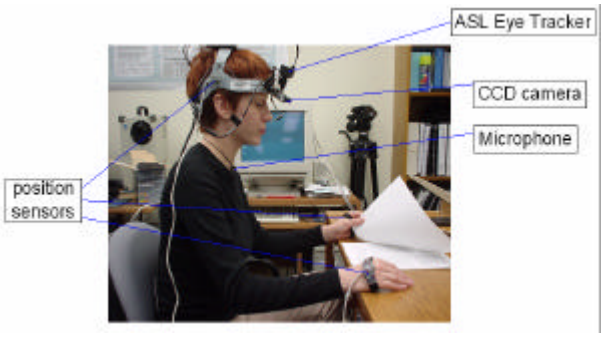


Figure 1: The system shares multisensory information with a real agent in a first-person sense. This allows the association of coincident signals from different modalities.

acquisition is modeled as deictic reference by extracting temporal information of eye movements. This is based on the unique property of eye movement: it implicitly carries information of the speaker's attention.

- This work focuses on learning word in purely unsupervised mode. In contrast to several other implemented models, our algorithm makes no assumption about the presence of facilitative prior knowledge. Furthermore, because we do not use any language-specific knowledge in building and implementing our model, it could be applied to any language and shed light on understanding general properties of language acquisition.
- Besides detecting the agent's attention by using eye gaze as a cue, attention switches can be calculated and utilized to segment the agent's action sequences into action primitives, which is the first step toward understanding human behaviors and translating body movements into action verbs (Yu & Ballard, 2002).

The current implementation of the model is able to associate object names with visually grounded meanings. The overview of our approach is illustrated in Figure 2. We collected user's perspective video, eye and head movements in concert with acoustic data as system inputs. The Hidden Markov Model (HMM) is used to estimate the speaker's focus of attention. Then for each attentional point in time, we spot both the object fixated by eye gaze and the co-occurring spoken utterance. Based on clustering visual and audio data in their feature spaces separately, we select reliable visual-audio pairs that are grounded lexical items of objects. We describe the implementation of our system in the following subsections.

## Estimating the Focus of Attention

In our model, the primary objective of data analysis of eye movements is to determine where and when a speaker looks at the objects in the visual scene. Although there are several different modes of eye

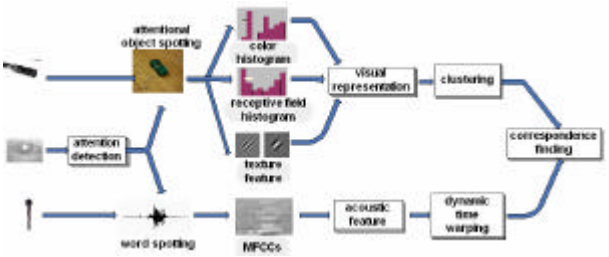


Figure 2: The overview of the approach to visually grounded word learning.

movement, the two most important modes for directing cognition are saccades and fixations. Saccades are rapid eye movements that allow the fovea to view a different portion of the visual scene. Each saccade is followed by a fixation when objects in the scene are viewed. Our goal is to find the fixations from continuous data stream of eye movements.

A 2-state HMM is used in our system for eye fixation finding. One state corresponds to saccade and the other represents fixation. The observations of HMM are 2-dimensional vectors consisting of the magnitudes of the velocities of head rotations in three dimensions and the magnitudes of velocities of eye movements. We model the probability densities of the observations using a two-dimensional Gaussian. The parameters of the HMM needed to be estimated comprise the observation and transition probabilities. The estimation problem concerns how to adjust the model $l$ to maximize $P(O \mid l)$ given an observation sequence $O$. We initialize the model with flat probabilities, and then the forward-backward algorithm allows us to train the model parameters using training data (see Rabiner & Juang, 1989).

As a result of training, the saccade state contains an observation distribution centered around high velocities and the fixation state represents the data whose distribution is centered around low velocities. The transition probabilities for each state represent the likelihood of remaining in that state or making a transition to another state. An example of the results of eye data analysis is shown in Figure 3.

## Visual Processing

The method of automatic object spotting by integrating visual information with eye gaze data consists of three steps. First, we apply a seeded region growing (SRG) algorithm (Admas & Bischof, 1994) to segment objects from the background, and eye gaze is utilized as a cue to extract the attentional object from all the objects in the scene. Next, the extracted object is represented by a feature vector that contains color, shape and texture features. Based on the work of Mel (1999), we constructed the visual features of objects that are large in number, invariant to different viewpoint, and are driven by multiple visual cues. Specifically, 64-

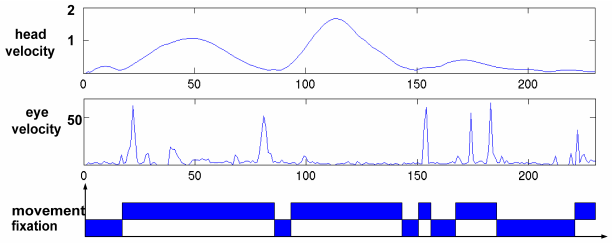


Figure 3: Eye fixation finding. **The top plot**: point-to-point velocities of head movement. **The middle plot**: the velocity profile of eye movement. **The bottom plot**: the results of fixation finding.

dimensional color features are extracted by color indexing method, and 48-dimensional shape features are represented by calculating histograms of local shape properties. The 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 the object in a 48-dimensional texture feature vector. The 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. We then reduce the 160-dimensional feature vectors into the vectors with the dimensionality of 30 by principle component analysis (For further information, see Yu, Ballard, & Zhu 2002). Thirdly, given visual feature vectors of objects, we compare the similarity of two feature vectors by calculating the Euclidean distance. Based on the calculation results, those feature vectors are then clustered into several groups by applying a hierarchical agglomerative clustering algorithm. In this way, the visual representation of the attentional object at a point in time is quantilized into a symbol that corresponds to a specific cluster.

## Speech Processing

We have developed an utterance endpoint detection algorithm that segments a speech stream into several spoken utterances delimited by silence. In the implementation, short bursts of speech are ignored since they are likely due to environmental noise, and short silences within an utterance are merged into a single utterance. Then, each utterance is segmented into a sequence of overlapped frames. Mel-frequency cepstral coefficients (MFCC) are extracted from each frame that will be used as feature vector series for clustering those spoken utterances. In the feature extraction algorithm, Discrete Fourier Transform (DFT) is applied to obtain the speech spectrum of each frame, which then passes through a filter bank of mel-spaced triangular filters spaced on a linear-logarithm scale.

Next, let S denote a sequence of n MFCC such that $S = \{s_t \mid 1 \le t \le n\}$, where $s_t$ is a vector of values corresponding to the MFCC of the frame at time $t$. Given a set of $m$ sequences of MFCC, we want to cluster the time series into groups so that each group corresponds to a qualitatively different regime. To achieve this goal, we first compute the similarity of two time series of MFCC by applying the Dynamic Time Warping (DTW) algorithm. Given two time series $s_1$ and $s_2$, DTW finds the warping of the time dimension in $s_1$, which minimizes the difference between those two series. Given $m$ time series, we can construct a complete pairwise distance matrix by invoking DTW $m(m-1)/2$ times. We then employ a hierarchical agglomerative clustering algorithm that starts with one cluster for each time series and merges the pair of clusters with the minimum average intercluster distance. Based on the results of clustering, we can quantilize MFCC time series of spoken utterances into symbols that correspond to the clusters.

## Multimodal Learning

In the physical world, a powerful constraint in sensory data is spatio-temporal and cross-modal coherence. We utilize body movements as deictic references to label speech signals with approximately coincident information from other modalities. Thus, from the perspective of machine learning, it is an unsupervised (or self-supervised) learning procedure since training data are labeled by data from other modalities and no manually generated transcription or semantic label is involved. Figure 4 illustrates our method of associating word-object pairs from multimodal data.

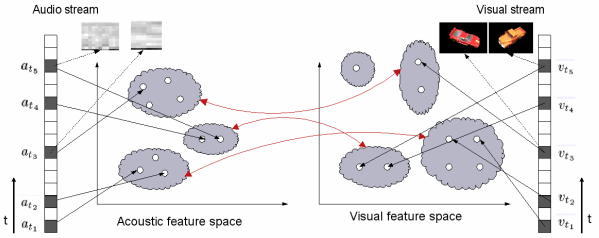


Figure 4: Grounded lexical items are built based on spatio-temporal and cross-modal constraints between audio-visual pairs. Based on clustering speech and visual signals separately, we associate spoken utterances with their possible meanings (e.g., the speaker's attentional objects) when those utterances are produced. As a result, we get multiple pairs of co-occurring audio and visual data that can be represented by a set $Q = \{(a_{t_1}, v_{t_1}), (a_{t_2}, v_{t_2}),..., (a_{t_n}, v_{t_n})\}$. From those hypothesized lexical items, we want to find audio-visual pairs that are correct lexical items of objects.

Assume that we could represent co-occurring audio-visual pairs in the form of cluster pairs expressed by

symbols, the problem we need to address is to identify symbol correspondence. We take a novel view of this problem as analogous to the word correspondence problem in machine translation. Thus, we have a representation of one form and want to translate it into another form. Our method is based on computing average mutual information and *t*-scores to find cluster correspondences (Gale & Church 1991). Given any pair of symbols (*u*,*v*) in which *u* represents a visual cluster and *v* represents an audio cluster, we can measure the association between *u* and *v* by making use of mutual information. $f^2$, a $c^2$-like statistic, seems to be a good measurement of correlation:

$$f^2 = \frac{(p(u,v)\overline{p}(u,v) - p(u)p(v))^2}{(p(u,v)+p(u))(p(u,v)+p(v))(p(u)+\overline{p}(u,v))(p(v)+\overline{p}(u,v))}$$

where *p(u,v)* is the probability of *u* and *v* co-occurrence, and *p(u)* is the probability of *u* occurrence but *v* is not in close temporal proximity. *p(v)* is the probability of *v* occurrence but *u* is not in close temporal proximity. $\overline{p}$ *(u,v)* is the probability that neither *u* nor *v* occurs. Using a series expansion, an expression may also be obtained for *var(* $f^2$ *)*.

To determine whether *(u,v)* is correlated, we apply the following rules based on the measurement of *t*-scores:

- If there is only one hypothesized pair that *u* and *v* co-occur, we measure the confidence by $t = f^2 \sqrt{\text{var}(f^2)}$ and compare it with a certain threshold.
- If there are multiple tokens $(v_1, v_2, ..., v_N)$ that could be associated with *u*, we can select the best pair $(u, v_i)$ as follows:

$$\max_i \left( \frac{2 * f^2(u,v_i) - \sum_{j=1}^{N} f^2(u,v_j)}{\sqrt{\sum_{j=1}^{N} \text{var}^2(f^2(u,v_j))}} \right)$$

We apply this method to build grounded lexicons in the experiment described in the following section.

## Experiment: Learning Object Names

In our experiments, the subjects were asked to sit at a table on which there were six objects: car, truck, motorcycle, jet, cow and elephant. They 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 from a first-person perspective. Six subjects participated in experiment and they were asked to randomly speak the names of six objects without prior instructions to look toward the corresponding objects when they utter the words. We recorded speech data as well as video from the first-person perspective, eye gaze and head position data. The eye position signals were reported at 60Hz and had a real time delay of 50 msec. The visual signals were sampled at the resolution of 320 columns by 240 rows of pixels at the frequency of 15Hz. The acoustic signals were recorded using a headset microphone at a rate of 16kHz with 16-bit resolution. In total, 198 spoken words were collected.

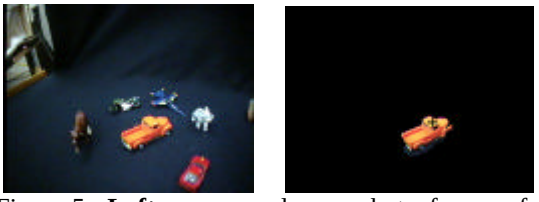

Figure 5: **Left:** an example snapshot of scene from the first-person view. **Right:** the object extracted from the left image using gaze position (black cross) as a cue.

We follow the method described in the previous section to process multisensory data. A perfect score would classify 198 audio-visual lexical items into six object groups. Based on eye movements as a temporal cue to associate audio and visual data, the correct rate of spotting audio-visual pairs is 92.9%. The errors are mainly caused by the fact that occasionally the subjects do not look toward the objects when they say the object names. 86.9% of lexical items are paired with semantically correct visual models. Figure 6 depicts the audio-visual pairs with their *t*-scores. With those large t values, we can select those audio-visual associations with confidence.

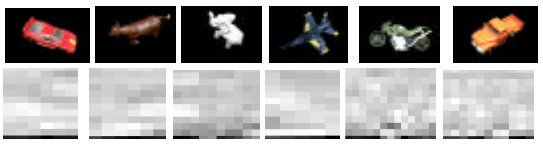


Figure 6: Visually grounded word learning. **Top**: visual images of six objects. The visual features of objects map to points in the visual feature space of Figure 4. **Middle**: Mel-frequency cepstral coefficients of spoken names of objects. Those acoustic features map to points in the acoustic feature space of Figure 4. **Bottom**: the *t*-scores of audio-visual pairs provide assurance for the correspondences. Only the scores for correct classifications are shown. The scores for incorrect associations are all less than 5.0.

To demonstrate the important role of attention in language learning, we process data using another method in which eye gaze and head movements are ignored, and only video and audio data are used for learning. Except for this point, this method shares most of the implemented components with the attention-based method. In this approach, since the first-person view captured from the head-mounted camera always contains several objects in the scene, a spoken word might be associated with any one of these co-occurring objects when it is uttered. Figure 7 shows the

comparison of two approaches. The result of the attention-based method is much better than the other one. The significant difference lies in the fact that there exist many possible correlations between temporally co-occurring audio-visual pairs. Inferences of the talker's referential intentions play a key role in discovering correct audio-visual correlations and then building grounded lexical items.

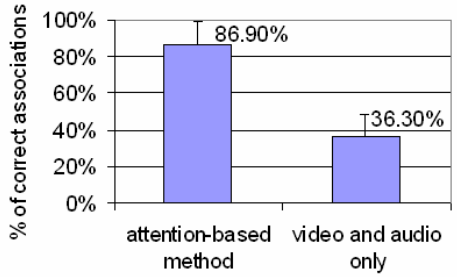


Figure 7: The comparison of the results of two methods provides a computational account of the importance of attention in language acquisition.

## Conclusion and On-going Work

We proposed a computational model of embodied language acquisition that can simulate the bootstrapping from speech to lexicons in early language acquisition. We implemented the proposed model and evaluated it by comparing our model with the similar one that does not utilize eye gaze and head direction as deictic reference. The results of our method are impressive, which demonstrate the importance of attention in language learning.

We constrained acoustic input to be isolated words in the current implementation. This is based on the assumption that exposure to isolated words may significantly facilitate vocabulary development at its earliest stage (Brent & Siskind, 2001). However, there is some evidence that infant-directed speech does not reliably provide isolated words and laboratory results show that infants have substantial segmentation capabilities by the onset of lexical acquisition. In light of this, we are working on developing a sub-system for speech segmentation that can simulate infants' ability to segment continuous speech. By integrating speech segmentation into our current system of word learning, we will have a more biologically plausible model of language acquisition, which will shed light on understanding human language development.

## References


Admas, R., & Bischof, L. (1994). Seeded region growing. *IEEE Transaction on Pattern Analysis and Machine Intelligence, 16(6),* 641-647.

Baldwin, D. A., Markman E. M., Bill, B., Desjardins, R. N., Irwin, J. M., & Tidball G. (1996). Infants' 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. *Behaviroal and Brain Sciences,20,*1311-1328.

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

Baron-Cohen, S., Baldwin, D. A., & Crowson M. (1997). Do Children with autism use the speaker's direction of gaze strategy to crack the code of language? *Child Development, 68(1),* 48-57.

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

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

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition, 81(B),* 33-B44.

Caselli, M. C., Casadio, P., & Bates, E. (2000). Lexical development in English and Italian. In Tomasello, M. & Bates, E. (Eds.) Language Development: The essential reading (pp. 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.

Gale, W., & Church, K. W. (1991). Identifying word correspondences in parallel texts. *Proceedings of the DARPA SNL Workshop*.

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

Harnad, S. (1990). The symbol grounding problem. *Physica D, 42,*335-346.

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.

Rabiner, L. R., & Juang, B. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77(2),* 257-286.

Roy, D., & Pentland, A. (2002). Learning words from sights and sounds: a computational model. *Cognitive Science, 26(1),* 113-146.

Saffron, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by eight-month old infants. *Science, 274,* 1926-1928.

Tomasello, M., & Barton, M. (1994). Learning words in non-ostensive contexts. *Developmental Psychology, 30,* 639-650.

Yu, C., Ballard, D. H., & Zhu, S. (2002). Attentional object spotting by integrating multimodal input. *Proceeding of the Fourth IEEE International Conference on Multimodal Interfaces* (pp 287-292).

Yu, C., & Ballard, D. H. (2002). Learning to recognize human action sequences. *Proceeding of the Second IEEE International Conference on Development and Learning* (pp 28-34).