Learning Visually Grounded Words and Syntax of Natural Spoken Language

Deb Roy
MIT Media Laboratory
20 Ames Street, Rm. E15-488, Cambridge, MA 02142, USA
dkroy@media.mit.edu
(617) 253-0596

Running head: Learning Visually Grounded Words and Syntax

January 15, 2002

Abstract

Properties of the physical world have shaped human evolutionary design and given rise to physically grounded mental representations. These grounded representations provide the foundation for higher level cognitive processes including language. Most natural language processing machines to date lack grounding. This paper advocates the creation of physically grounded language learning machines as a path toward scalable systems which can conceptualize and communicate about the world in human-like ways. As steps in this direction, two experimental language acquisition systems are presented.

The first system, CELL, is able to learn acoustic word forms and associated shape and color categories from fluent untranscribed speech paired with video camera images. In evaluations, CELL has successfully learned from spontaneous infant-directed speech. A version of CELL has been implemented in a robotic embodiment which can verbally interact with human partners.

The second system, DESCRIBER, acquires a visually-grounded model of natural language which it uses to generate spoken descriptions of objects in visual scenes. Input to DESCRIBER's learning algorithm consists of computer generated scenes paired with natural language descriptions produced by a human teacher. DESCRIBER learns a three-level language model which encodes syntactic and semantic properties of phrases, word classes, and words. The system learns from a simple 'show-and-tell' procedure, and once trained, is able to generate semantically appropriate, contextualized, and syntactically well-formed descriptions of objects in novel scenes.

Introduction

Humans have been designed by an evolutionary process which is driven by the structure of the physical world. Our physiology is ultimately a reflection of the effects of gravity, the characteristics of light and sound propagation, and a host of other properties of the natural world. In other words, evolutionary design is grounded in the physical world. Various layers of sensory, motor, and control adaptations are causally linked to how the world works. The resulting physiology provides the 'hardware' in which representational structures and processes (the 'software') are likewise shaped by physical world constraints. Aspects of the world which are sensed or acted upon by an agent must be represented either implicitly or explicitly within in the agent. The particular aspects of the world which are represented and the manner in which those aspects are represented is highly dependent on the evolved physiological mechanisms at the agent's disposal. Higher level cognitive abilities including language build on top of this grounded foundation. Thus mental representations are grounded. Representational grounding is essential for situated agents to perceive, plan, act, and communicate. The link between evolutionary grounding and grounding of mental representations is intimately intertwined. The former gives rise to the latter. An analogous relationship between evolutionary design and representational grounding can be applied to the creation of artificial intelligence (AI) systems.

Is the mind a symbol manipulator as posited in the classical view of AI articulated by (Newell & Simon, 1976)? Searle's well known Chinese room experiment (Searle, 1980) questions the viability of purely symbolic AI. Searle argues that even if a machine could be programmed to manipulate symbols intelligently, the machine would not intrinsically understand the meaning of those symbols. Symbol processors blindly push symbols around using preprogrammed rules. Even if the result is intelligent behavior, the machine does not actually understand what it is doing. For example, a web search engine finds relevant web sites based on natural language queries, yet it does not *really* understand what the queries mean. A human interpreter is needed to read meaning into the actions of the machine.

Over the past two decades, research in artificial intelligence has shifted emphasis from purely symbolic tasks such as chess and theorem proving, towards embodied and perceptually grounded problems (cf. Brooks, 1986; Agre, 1988). A key issue in establishing intrinsic meaning for a machine is for all symbols to be bound, either directly or indirectly, to the machine's physical environment (Harnad, 1990). For a machine to understand the concept of *hammer*, it must be able to recognize one when it perceives it and know what its function is. Thus both perception and action play a crucial role in representing the concept of a hammer in a way which is meaningful for the machine, as opposed to a token which has meaning only to an external human observer.

In addition to grounding symbols, the rules by which symbols are manipulated in cognitive processes must also be grounded. Returning to the notion of cognition as symbol manipulation, what we seem to need are symbols that capture the 'sensorimotor shape' of what they represent. But this is exactly what symbols are not. Symbols, such as words, are arbitrary patterns which do not resemble the form of what they signify. This situation motivates the development of an expanded notion of signs that include both symbols with arbitrary form and environmentally shaped 'non-symbols'. The philosopher Charles Peirce did exactly this in his analysis of signs (Peirce, 1932). I will briefly introduce Peirce's taxonomy of signs and show how it can be interpreted as a framework for cognitive processing of grounded language.

Signs are physical patterns which signify meaning. Peirce made the distinction between three types of signs based on how they are associated with their meanings: icons, indices, and symbols. Icons physically resemble what they stand for. Interpreted in cognitive terms, an icon is the pattern of sensory activation resulting from an external object[1]. Icons are the result of projections of objects through the perceptual system and are thus strongly shaped by a combination of the nature of the physical world and the nature of the perceptual system. Icons *are* the physical world as far as an agent is concerned. There is not a more direct way to know what is 'out there' than what the perceptual system reports.

Peirce defines indexical signs as a set of sensory properties which point to some other entity of interest to the agent. For example, footsteps can serve as indexical signs which indicate a human presence. Words can also serve as indices. For instance, 'this' and 'here' can be used to pick out objects and locations in the world. The entire meaning of an index is derived from its direct connection to its referent. Indexical reference is equivalent to the problem of achieving joint reference which is critical to success in language acquisition (cf. Bloom, 2000).

The third level of representation is symbolic. Words are archetypical examples of symbols. Symbols can be removed from their context. They can be used when their referents are not present. Symbols gain meaning in large part from their relationships with other symbols. Possible relationships between symbols include, 'part-of' (leaves and branches are parts of trees), 'kind-of' (elms and oaks are kinds of trees), and 'need-a' (people need food and sleep), to mention but a few. In this paper I will focus on the distinction between iconic and symbolic signs. Although indexical signs are needed to establish connections to the world, they are beyond the scope of this discussion.

Peirce's distinction of icons and symbols provides a framework for developing cognitive models of communication (for example, see Harnad, 1990; Deacon, 1997). Rather than thinking of the mind as a symbol manipulator, we can think of the mind as a *sign manipulator*. In other words, a manipulator of both icons and symbols. Icons are by their very nature grounded. They directly reflect experiential data and may form the basis of perception and action. The rules for combining and manipulating icons are constrained by the agent's physiology/hardware and are guided by experience. For example, the design of the agent's visual system might allow for the combination of icons representing *red* and *ball* (a ball with the property red), or *ball* and *bounce* (a ball which bounces), but not *red* and *bounce* (since the perceptual system cannot bind a color to an action without an object present). Thus the nature of how the perceptual system represents the external world influences the rules by which symbols may be meaningfully combined. Experience will teach the agent that things tend to fall down (not up), along with a large number of other rules about how the world works. These two types of knowledge result in grounded rules of icon manipulation. Symbols are associated with icons or categories of icons and thus directly inherit iconic groundings. Similarly, grounded icon manipulation rules are inherited for symbol manipulation.

Symbols and icons provide complementary strengths. The retention of iconic represen-

[1] For brevity I will use the word 'object' to mean anything in the world which can be referred to including physical objects, but also properties, events, states, etc.

tations is essential even after symbolic representations have been acquired. Without iconic representations, the agent would be unable to recognize new percepts or perform motor control. In addition, some kinds of reasoning are most efficiently achieved using iconic representations. For example an image-like representation of a set of spatially arranged objects can be used to efficiently store all pair-wise spatial relations between objects. A non-iconic representation such as a list of logical relations ('Object A is directly above Object B') grows factorially with the number of objects and is vastly less efficient as the number of objects becomes large.

On the other hand, symbols are unencumbered by experiential baggage. They can be efficiently copied, stored and manipulated. New symbols can be created in terms of existing icons and symbols to represent abstract concepts. The degree of abstraction is limited only by the cognitive capacities of the agent.

The conceptualist case for embodied and grounded representations has been argued by many authors including (cf. Jackendoff, 1983; Johnson, 1987; Lakoff, 1987; Harnad, 1990; Barsalou, 1999). I will not elaborate further on these issues since they are well explicated in the cited works. I will, however, briefly present one perspective on why ungrounded representations are problematic from an engineering design perspective. This perspective returns to my earlier point that a link exists between grounding the design process of systems, and the resulting representations that emerge in those systems.

One way to look at the problem of ungrounded symbolic representations is that the representations in the machine are based on the intuitions of human designers. We, as builders of these systems, intuit the form and nature of our own mental representations and build them into the machines. The obvious danger in this approach is that our intuitions may be wrong. Symbolic representations designed on the basis of intuition can, however, be extremely powerful. Witness the success of chess playing machines and countless powerful number crunching systems. Both of these examples operate in worlds created by people for people. The world of chess and the world of numerical computation are domains in which intuition based representations undoubtedly excel. In most natural domains which are not constructed by people, we should be less comfortable in relying on our intuitions.

An alternative to symbolic representations based on the intuitions of human designers is to construct physically grounded machines (cf. Brooks, 1986; Agre, 1988). In doing so, representations in the machines are shaped by the same factors which have shaped human design – the physical world. Grounded mechanisms can serve as a foundation for higher level representations. The link between evolutionary grounding and representational grounding in humans provides a lesson in how to proceed in designing intelligent machines. If we build grounded machines that process raw sensor data and act in a physical environment, then whatever representations we design for these machines will be shaped by properties of the physical world. The effects of gravity, light and sound propagation, and so forth, will mold the representations which arise in the machines. Rather than rely solely on human intuitions, we submit our designs to the tests of physical reality. In so doing, I believe we will develop representations that are more likely to scale in human-like ways. I do not suggest that building grounded systems is sufficient to assure human-like representations, but I believe that it is *necessary.*

This paper presents two implemented learning systems which represent progress toward our goal of building machines that learn to converse about what they see and do. By choosing the problem of linking natural language to perception and action, we are forced to design systems which connect the symbolic representations inherent in linguistic processing to grounded sensorimotor representations. The first system, CELL, demonstrates perceptual grounding for learning the names of shapes and colors. The second system, DESCRIBER, learns hierarchical grammars for generating visually-grounded spatial expressions.

Our work is related to several other research efforts. One aspect of CELL is its ability to segment continuous speech and discover words without a lexicon. This problem has also been studied by a number of researchers including (Aslin, Woodward, LaMendola, & Bever, 1996; de Marcken, 1996; Brent, 1999). In contrast to these efforts which process only linguistic input (speech or speech transcriptions), CELL integrates visual context (what the speech is about) into the word discovery process. Both CELL and DESCRIBER make use of 'cross-situational learning', that is, integration of partial evidence from multiple learning examples. Along similar lines, Siskind (1992) developed a model of cross-situational learning which learned word semantics in terms of symbolic primitives. Although Siskind (2001) has also developed visual representations of actions suitable for grounding verb semantics, he has not to date integrated these visual representations into a model of language acquisition. Several researchers (cf. Steels & Vogt, 1997; Cangelosi & Harnad, 2002, this volume) are studying models of the evolution of communication, i.e., the origin of language as an adaptive mechanism. In contrast, our work does not model evolutionary processes. Evolutionary processes are effectively replaced by iterative design processes carried out by human engineers. Another important difference is that we focus on the challenges faced by a single agent which must learn natural language in order to communicate with human 'caregivers'. Thus the target language is static from the perspective of the learning agent rather and an evolving target. Our work is most closely related to several projects which also seek to design grounded natural language learning systems such as Sankar and Gorin (1993), Regier (1996), Bailey, Feldman, Narayanan, and Lakoff (1997), and in this volume, Steels and Kaplan (2002). Our long term focus is to construct interactive robotic and virtual agents which can verbally interact with people and converse about both concrete and, to a limited degree, abstract topics using natural spoken language. We look to infant and child language acquisition for hints on how to build successively more complex systems to achieve this long term goal.

CELL: Learning Shape and Color Words from Sights and Sounds

CELL is a model of cross-modal early lexical learning. The model has been embodied in a robot shown in Figure 1 which learns to generate and understand spoken descriptions of object shapes and colors from 'show-and-tell' style interactions with a human partner. To evaluate CELL as a cognitive model of early word acquisition, it has been tested with acoustic recordings of infant-directed speech. In this paper a summary of the model and evaluations are provided. For more detailed presentations of CELL as a cognitive model the reader is referred to (Roy & Pentland, 2002). For further details of cross-modal signal analysis and integration into a robotic platform see (Roy, In press).

CELL discovers words by searching for segments of speech which reliably predict the presence of visually co-occurring shapes. Input consists of spoken utterances paired with images of objects. This approximates the input that an infant receives when listening to a caregiver and visually attending to objects in the environment. The output of CELL's

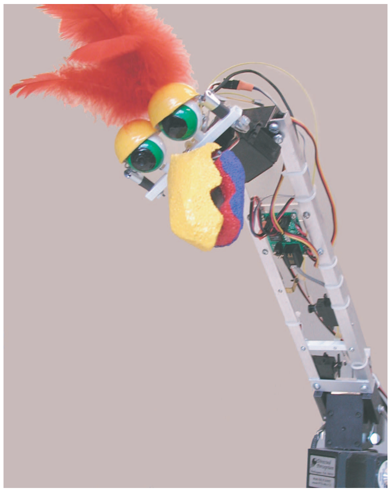


Figure 1. Toco is a robot which can be taught by 'show-and-tell'. Toco learns early language concepts (shapes, colors) by looking at objects and listening to natural spoken descriptions. After learning, Toco can engage in interactive visually-grounded speech understanding and generation.

learning algorithm consists of a lexicon of audio-visual items. Each lexical item includes a statistical model (based on hidden Markov models) of a spoken word and a visual model of a shape or color category.

To acquire lexical items, the system must (1) segment continuous speech at word boundaries, (2) form visual categories corresponding to word semantics, and (3) form appropriate correspondences between speech and visual models. These learning problems are difficult. Continuous speech does not contain reliable acoustic cues at word boundaries. Thus problem 1 may be regarded as a search problem in an extremely large and noisy space. In problem 2, visual categories are not built into CELL but must be learned by observation. The choice of categories must integrate visual clustering properties of physical objects with the conventions of the target language. Problem (3) is difficult because, in general, the linguistic descriptions provided to CELL contain multiple words. Of those words, only a subset refer directly to visual context. In many utterances, *none* of the words refer to the context. To address the inference problem inherent in (3), evidence from multiple observations must be integrated.

Input to CELL is grounded in sensors. Linguistic input consists of acoustic signals transduced by a microphone. Context for inferring visual referents of words is provided by a color video camera. A set of feature detectors have been designed which extract salient features from sensor data. These detectors determine the nature of iconic representations in CELL.

The acoustic front-end processor converts spoken utterances into sequences of phoneme probabilities. The ability to categorize speech into phonemic categories was built-in since similar abilities have been found in pre-linguistic infants after exposure to their native language (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992). At a rate of 100Hz, this processor computes the probability that the past 20 milliseconds of speech belong to each of 39 English phoneme categories or silence. The phoneme estimation was achieved by

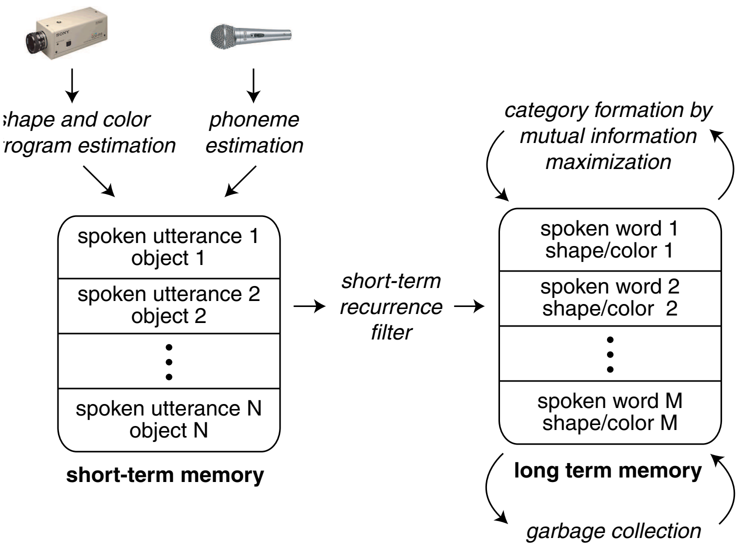


Figure 2. Architecture of the CELL model.

training an artificial recurrent neural network similar to (Robinson, 1994). The network was trained with a database of phonetically transcribed speech recordings of adult native English speakers (Garofolo, 1988). Utterance boundaries are automatically located by detecting stretches of speech separated by silence. Figure 3 shows the output of the acoustic analyzer for the short utterance “bye, ball” (extracted from our infant-directed spontaneous speech corpus described below). The display shows the strength of activation of each phoneme as a function of time. The brightness of each horizontal display line indicates how confident the acoustic processor is in the presence of each phoneme. The acoustic representation in Figure 3 is the icon formed in CELL in response to the spoken utterance. The icon may be thought of as the result of projecting the spoken utterance through the perceptual system of the machine. CELL’s perceptual system is biased to extract phonemic properties of the acoustic signal.

A visual processor has been developed to extract statistical representations of shapes and colors from images of objects. The visual processor generates shape and color histograms of local and global features. In a first step, edge pixels of the object are located using a background color model which segments the foreground region. For each pair of edge points of the object, the normalized distance between points and the relative angle of edges at the two points are computed. All distances and angles are accumulated in a two-dimensional histogram. This histogram captures a representation of the object’s silhouette which is invariant to scale and in-plane rotation. Three-dimensional shapes are represented with a collection of two-dimensional shape histograms, each derived from a different view of the object. Color is also represented using a histogram of all RGB values corresponding to the object. Figure 4 displays the shape histograms for two objects used in evaluations. These histograms depict the iconic representation of shapes in CELL. Although the relationship between the original shape of the object’s silhouette and the ‘shape’ of iconic representation in the histogram’s activation patterns is complex, a direct causal link nonetheless exists

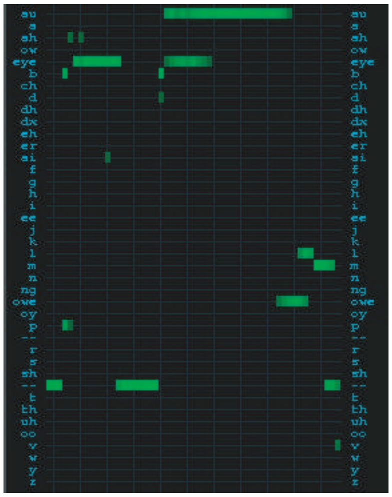


Figure 3. Iconic representation of the spoken utterance “bye, ball” in CELL.

between the former and the latter.

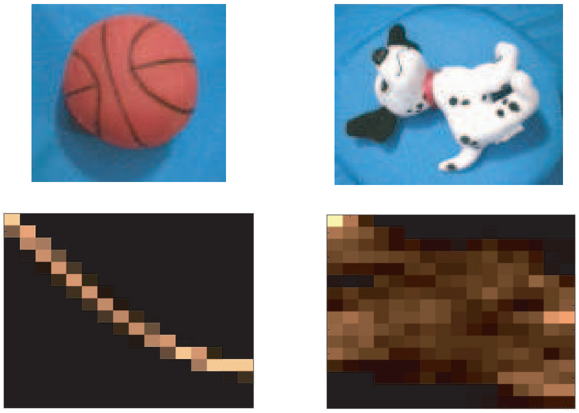


Figure 4. Iconic representation of two objects in CELL. Input images of objects are shown in the upper row. The resulting shape histograms are shown below.

Visual input is captured from a video camera mounted in the head of the robot. The robot has four degrees-of-freedom (two at the neck and two at the ’waist’) for controlling the orientation of the camera. The robot captures images of objects from various vantage points. The chi-squared divergence statistic is used as a visual distance metric for comparing histograms. This statistic has been shown to work well for visual histogram comparison (Schiele & Crowley, 1996).

CELL learns using an on-line procedure. Input consists of images of objects paired with spoken descriptions. Phonemic representations of the speech utterances and co-occurring color and shape histograms are temporarily stored in a short term memory (STM). The STM has a capacity of five utterances, corresponding to approximately 20 words of infant-directed speech. As input is fed into the model, each new [utterance,shape,color] entry replaces the oldest entry in the STM. A short-term *recurrence filter* searches the contents of the STM for recurrent speech segments which occur in matching visual contexts. The STM focuses initial attention to input events which occur in close temporal proximity. By limiting analysis to a small window of input, computational resources for search and memory for unanalyzed sensory input are minimized, as is required for cognitive plausibility. To determine matches, an acoustic distance metric has been developed (Roy, 2000) to compare each pair of potential speech segments drawn from the utterances stored in STM. This metric estimates the likelihood that the segment pair in question constitutes variations of similar underlying phoneme sequences, and thus represents the same word. The chi-squared divergence metric described earlier is used to compare the visual components associated with each STM utterance. If both the acoustic and visual distance are small, the speech segment and shape/color histogram are copied into the LTM. Each entry in the LTM is a pairing of a 'speech icon' with a 'shape/color icon' which I refer to collectively as a *cross-modal icon*.

In a final learning step, cross-modal icons from LTM are consolidated into categorical models of spoken words paired with shapes/colors. A category is defined by a prototype cross-modal icon in conjunction with an acoustic and a visual radius which specify allowable deviation from the prototype. To understand how the consolidation step works, let us assume that the LTM contains N cross-modal icons. We designate one of these as the prototype and assign a radius of allowable variation to the constituent speech and visual icons. Each of the remaining $N-1$ items can be compared to the prototype to determine whether they match the prototype acoustically and visually. If the distance from the test icon to the prototype is less than the corresponding radius, then we set an indicator variable to 1, otherwise to 0. Thus, for each comparison between a cross-modal icon and the prototype, we obtain the value of an acoustic and a visual indicator variable.

Each of the $N-1$ comparisons can be thought of as an experiment, the outcome of which is the value of two random variables which indicate acoustic and visual matches. The mutual information (MI) between the two variables is estimated in order to test whether the two variables are correlated (and thus indicative of a good speech-to-visual association). Mutual information is a measure of the amount of uncertainty removed regarding the value of one variable given the value of the other (Cover & Thomas, 1991). It is assumed that semantically linked speech-shape/color icons will lead to random variables with high MI. Cross-modal icons with high MI are selected to form categorical representations. The radii are determined by an optimization process which searches for the combination of speech and visual radii values which maximizes cross-modal MI (see Roy (In press) for details). During on-line learning, each new cross-modal icon placed in LTM by the recurrence filter is treated as a prototype. If the resulting MI is high, it leads to a permanent lexical item. A garbage collection process periodically removes cross-modal icons from LTM which are not selected by the consolidation process.

A human partner can interact with the robot implementation of CELL in one three modes. In the Learning Mode, the person may present objects to the robot and describe them using natural fluent verbal descriptions. Each object-utterance pair enters the CELL's STM and is processed. Once CELL has acquired a set of visually-grounded words, the robot may be switched to either the Understand Mode or the Generate Mode. In the Understanding Model, a person may speak one-word or two-word phrases (shape, color, or color-shape). In response, the robot will attempt to recognize the speech (using the acoustic models in its acquired lexicon) and then find an object in its environment which best fits the visual model associated with the recognized word or phrase. In the Generate Mode, a novel object can be presented to the robot for description. The system builds a three-dimensional model of the object by viewing it from multiple angles. The robot generates a spoken description using a phonetic speech synthesizer of the object by retrieving the best matching entries in its lexicon.

Evaluation with Infant-Directed Speech

CELL has been evaluated as a model of how infants might segment fluent spontaneous speech and discover word units. A corpus of audio-visual data was gathered from infant-directed interactions (Roy, 1999). Six caregivers and their pre-linguistic (7-11 months) infants were asked to play with objects while being recorded. Seven classes of objects commonly named by young infants were selected (Huttenlocher & Smiley, 1994): balls, shoes, keys, toy cars, trucks, dog, horses. A total of 42 objects, six objects for each class, were obtained. The objects of each class varied in color, size, texture, and shape.

Each caregiver-infant pair participated in 6 sessions over a course of two days. In each session they played with 7 objects, one at a time. All caregiver speech was recorded onto DAT using a wireless head-worn microphone. In total, approximately 7,600 utterances were collected comprising of 37,000 words across all six speakers. Most utterances contained multiple words with a mean utterance length of 4.6 words.

The objects were imaged by the robot. A total of 209 images from varying perspectives were collected for each of the 42 objects, resulting in a database of 8,778 images.

The speech recordings from the caregiver-infant play sessions were combined with the images taken by the robot to provide multimodal input to CELL. To prepare the corpus for processing, the following steps were performed: (1) The speech recordings were segmented at utterance boundaries. This was done automatically by finding contiguous frames of speech detected by the recurrent neural network. (2) For each utterance, a random set of 15 images of the object which was in play at the time the utterance was spoken was selected.

The lexicons extracted from the corpus were evaluated using three measures. The first measure, segmentation accuracy (M1), is the percentage of lexical items with boundaries set at English word boundaries. The second, word discovery accuracy (M2), is the percentage of lexical items which were complete English words with an optional attached article. The third measure, semantic accuracy (M3), is the percentage of lexical items which passed M2 and were paired with semantically correct visual models.

For comparison, the system was also run with only acoustic input. In this case, it was not meaningful to use the MI maximization. Instead, the system searched for globally recurrent speech patterns, i.e., speech segments, which were most often repeated in the entire set of recordings for each speaker. This acoustic-only model may be thought of as a rough approximation to a minimum description length approach to finding highly repeated speech patterns which are likely to be words of the language.

Results of the evaluation shown in Table 1 indicate that the cross-modal algorithm was able to extract a large proportion of English words from this very difficult corpus (M2), many associated with semantically correct visual models (M3). Typical speech segments in the lexicons included names of all six objects in the study, as well as onomatopoetic sounds such as “ruf-ruf” for dogs, and “vroooom” for cars. Comparison with the audio-only system

Table 1: Performance results on word segmentation, word discovery, and semantic accuracy averaged across six speakers. Results shown for cross-modal learning using CELL, and acoustic-only learning.

	Segmentation Accuracy (M1)	Word Discover (M2)	Semantic Accuracy (M3)
audio only	7±5%	31±8%	13±4%
audio-visual	28±6%	72±8%	57±10%

clearly demonstrates improved performance when visual context is combined with acoustic evidence in the clustering process. For the difficult test of word boundary detection (M1), multimodal input led to a four-fold improvement over acoustic-only processing. Cross-modal structure enabled the system to find and extract useful knowledge without the aid of manual annotations or transcriptions.

DESCRIBER: Learning Word and Grammar Semantics in a Spatial Description Task

The word learning experiments based on CELL focused on learning surface forms of words, and the acquisition of associated visual categories. A second system, DESCRIBER, investigates the acquisition of a limited visually grounded grammar in a spatial description task. The description task is based on images of the sort shown in Figure 5. The computer generated image contains a set of ten non-overlapping rectangles. The height, width, x-y position, and red-green-blue (RGB) color of each rectangle is continuously varying and chosen from a uniform random distribution. DESCRIBER addresses the following problem: Given a set of images, each with a *target object* and a natural language description of the target, learn to generate *syntactically correct, semantically accurate, and contextually appropriate* referring expressions of objects embedded in novel multi-object scenes.

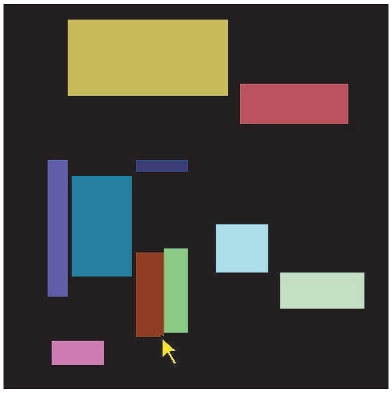


Figure 5. A typical scene processed by DESCRIBER. The arrow indicates the target object.

DESCRIBER acquires a visually-grounded language model which enables it to generate referring expressions. In an assessment of semantic ambiguity made by human judges, the visual descriptions have been found to be comparable to human generated descriptions. This paper provides a sketch of the principal mechanisms underlying DESCRIBER. A more detailed account of the system will be available in an upcoming paper.

The ‘perceptual system’ of DESCRIBER consists of a set of feature extractors which operate on synthetic images. In comparison to CELL, visual processing in DESCRIBER is trivially available since we have direct access to the source of the images (i.e., access to the program which generated the images). A set of visual attributes including shape, size, location, color, and brightness, is extracted from each rectangle in a scene. The features for the set of objects constitute the iconic representation of a scene. Learning in DESCRIBER consists of six stages:

Stage 1: Word Class Formation

In order to generate syntactically correct phrases such as ‘large red square’ as opposed to ‘red large square’ or ‘square red’, word classes that integrate syntactic and semantic structure must be learned. Two methods of clustering words into syntactically equivalent classes were investigated. The first relies on distributional analysis of word co-occurrence patterns. The basic idea is that words which co-occur in a description are unlikely to belong to the same word class since they are probably labeling different aspects of the scene. The second method clusters words which co-occur in similar visual contexts. This method uses shared visual grounding as a basis for word classification. We have found that a hybrid method which combines both methods leads to an optimal clustering of words.

Stage 2: Feature Selection for Words and Word Classes

A subset of visual features is automatically selected and associated with each word. A search algorithm finds the subset of visual features for which the distribution of feature values conditioned on the presence of the word is maximally divergent from the unconditioned feature distribution. Features are assumed to be normally distributed. The Kullback-Leibler divergence is used as a divergence metric between word-conditioned and unconditioned distributions. This method has been found to reliably select word features in an eight dimensional feature space. Word classes inherit the conjunction of all features assigned to all words in that class.

Stage 3: Grounding Adjective/Noun Semantics

For each word (token type), a multidimensional Gaussian model of feature distributions is computed using all observations which co-occur with that word. The Gaussian distribution for each word is only specified over the subset of features assigned to that word in Stage 2.

Stage 4: Learning Noun Phrase Word Order

A class-based bigram statistical language model is learned and models the syntax of noun phrases. The visually grounded word classes acquired in Stage 1 form the basis for this Markovian model of word order.

Stage 5: Grounding the Semantics of Spatial Terms

A probabilistic parser uses the noun phrase bigram language model from Stage 4 to identify noun phrases in the training corpus. Utterances which are found to contain two noun phrases are used as input for this stage and Stage 6. Multi-noun-phrase utterances are usually of the form ‘TARGET_NP [spatial relation] LANDMARK_NP’, that is, a noun phrase describing the target object, followed by a spatial relation, followed by a *landmark* noun phrase. A typical utterance of this type is, ‘The large square slightly to the left of the vertical pink rectangle.’. An automatic process based on bigram word pair probabilities is used to tokenize commonly occurring phrases (e.g., ‘to the left of’ is converted to the token ‘to_the_left_of’). Any words in the training utterance which are not tagged as noun phrases by the parser are treated as candidate spatial terms. A set of visual spatial primitives based on Regier (1996) is introduced in this stage. The procedures in Stages 2 and 3 are re-used to ground spatial words in terms of these spatial features.

Stage 6: Learning Multi-Phrase Syntax

Multi-noun-phrase training utterances are used as a basis for estimating a phrase-based bigram language model. The class-based, noun phrase language models acquired in Stage 4 are embedded in nodes of the language model learned in this stage.

To train DESCRIBER, a human participant was asked to verbally describe approximately 500 images of the kind shown in Figure 5. Each spoken description was manually transcribed, resulting in a training corpus of images paired with utterance transcriptions[2]. Figures 6 and 7 illustrate the results of the learning algorithm on this training corpus. The language model has a three-layer structure. At the highest level of abstraction (left side of Figure 6), phrase order is modeled as a Markov model which specifies possible sequences of noun phrases and connector words, most of which are spatial terms. Transition probabilities have been omitted from the figure for clarity. Two of the nodes in the phrase grammar designate noun phrases (labeled TARGET_OBJECT and LANDMARK_OBJECT) and are diagrammatically linked by dashed lines to the next level of the model. Note that at the phrase level, the semantics of relative noun phrase order are encoded by the distinction of target and landmark phrases. In other words, the system knows that the first noun phrase describes the target and the second describes the landmark. This distinction is learned in Stage 6 (details of how this is learned can be found in Roy (In review).

Each word in the noun phrase language model is linked to an associated visual model. The grounding models for one word class are shown as an example in Figure 7. The words 'dark', 'light' and 'white' were clustered into a word class in Stage 1. The blue and green color components were selected as most salient for this class in Stage 2. The ellipses in the figure display isoprobability contours of the word-conditional Gaussian models in the blue-green feature space learned for each word in Stage 3. The model for 'dark' specifies low values of both blue and green components, whereas 'light' and 'white' specify high values. 'White' is mapped to a subset of 'light' for which the green color component is especially saturated. In summary, the phrase level language model is grounded through two levels of indirection in terms of sensory features of the system.

A planning system uses the grounded grammar to generate semantically unambiguous, syntactically well formed, contextualized text descriptions of objects in novel scenes. A concatenative speech synthesis procedure is used to automatically convert the text string to speech using the input training corpus. The final output of the system are spoken descriptions of target objects in the voice of the human teacher. In outline form, the planner works as follows:

Stage 1: Generate Noun Phrases

Using the noun phrase model as a stochastic generator, the most likely word sequence is generated to describe the target object, and each non-target object in the scene.

Stage 2: Compute Ambiguity of Target Object Noun Phrase

An ambiguity score is computed based on how well the phrase generated in Stage 1 describes non-target objects in the scene. The Viterbi algorithm (Rabiner, 1989) is used

[2] A natural extension of this work is to integrate the acoustic word learning methods from CELL to replace this manual transcription step.

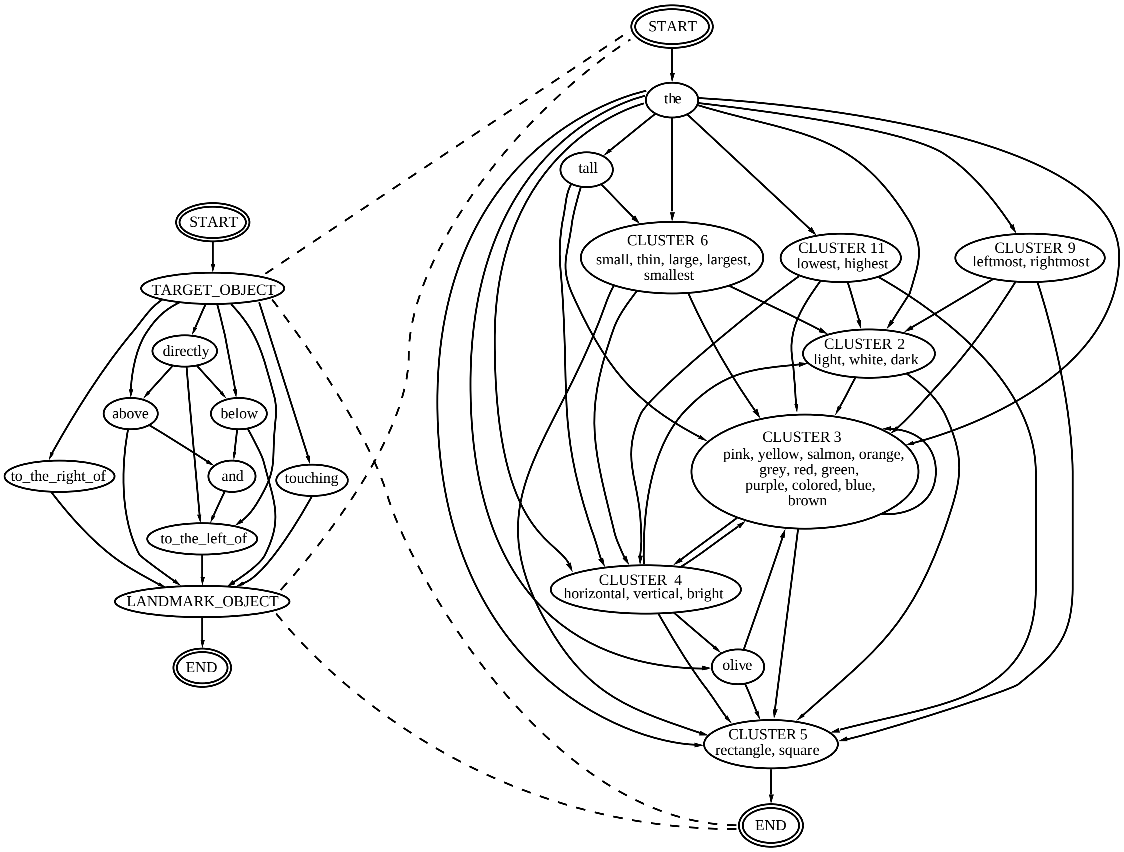


Figure 6. Noun phrase and word class structures acquired by DESCRIBER.

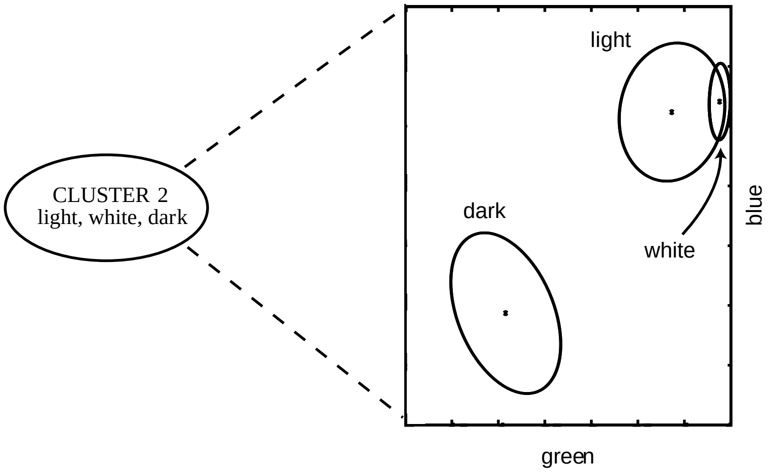


Figure 7. Visual grounding of words for a sample word class.

to compute the probability that each object in the scene matches the target phrase. If the closest competing object is not well described by the noun phrase, then the planner terminates, otherwise it proceeds to Stage 3.

Stage 3: Generate Relative Spatial Clause

A landmark object is automatically selected which can be used to unambiguously identify the target. Stage 1 is used to generate a noun phrase for the landmark. The phrase-based language model is used to combine the target and landmark noun phrases.

Sample output from DESCRIBER is shown in Figure 8 for four novel scenes which were not part of the training corpus. In each scene, the target object is indicated with an arrow. Note that the descriptions take into account the relative context of each target object. In the lower two scenes, Stage 1 failed to produce an unambiguous noun phrase, so DESCRIBER generated a complex utterance containing a relative landmark. These descriptions represent DESCRIBER's attempt to strike a balance between syntactic, semantic, and contextual constraints[3].

An evaluation was conducted to assess how accurately human listeners can select target objects from scenes, given DESCRIBER's spoken utterances, in comparison to the original human trainer's descriptions. Four judges selected the correct object 81.3% of the time based on DESCRIBER's output versus 89.8% when given human generated descriptions. These results suggest that DESCRIBER acquires the grounded language capacity necessary for this task.

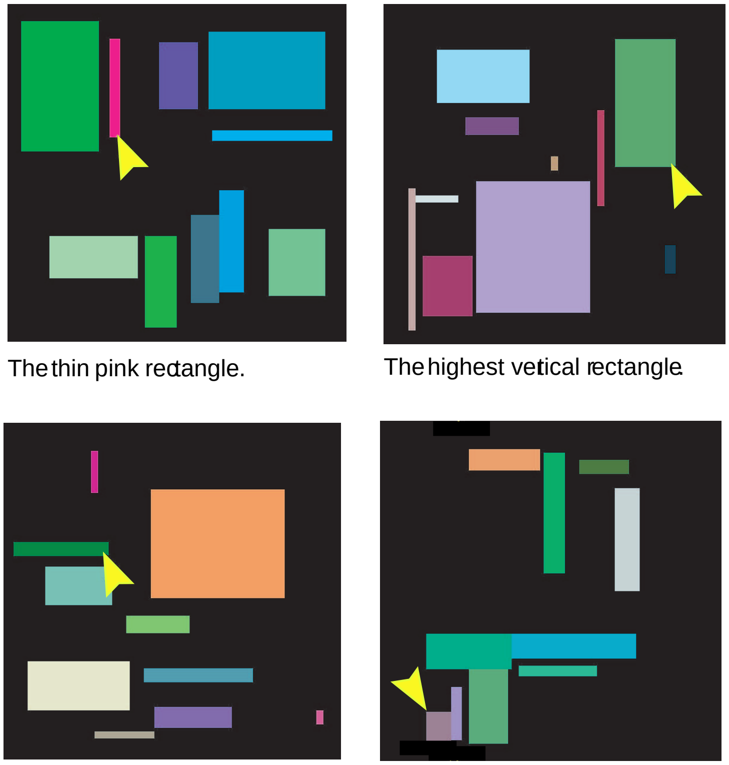
We have recently 'transplanted' a real time vision system in place of the synthetic vision system. In initial tests with controlled visual environments, we have been able to successfully run DESCRIBER on real world objects. Similar to CELL, the grounded language model underlying DESCRIBER can be used for language understanding, however, we have not yet implemented this capability.

Discussion

A goal in implementing CELL and DESCRIBER was to design representations which were at least partially shaped by the machine's environment, rather than the intuitions of human developers. To varying degrees this was accomplished. Representations in CELL and DESCRIBER are causally linked to sensory experience. Sensory input gives rise to internal representations which are used by each system to generate, and in the case of CELL, understand, spoken language. The iconic representations of word semantics enables each system to relate natural language to the visual environment.

Does either system achieve symbolic representation? DESCRIBER comes closer than CELL. In CELL, lexical items are audio-visual associations. The associations are formed on the basis of cross-modal mutual information. Mutual information is a symmetric metric (i.e., $MI(A;B) = MI(B;A)$). In other words, as far as CELL is concerned, visual categories could just as well serve as labels for speech or vice versa. Likewise, it is possible that in very early language learning, infants are also unaware of the referential capacity of language, and instead use it in purely associationist ways. In any case, the lexical items in CELL are clearly not symbolic. They are, however, grounded and thus could serve as the foundation for grounded symbols.

[3] An utterance length penalty was introduced as a fourth constraint to reward concise descriptions.



The thin pink rectangle.

The highest vertical rectangle.

The dark green rectangle above the light green rectangle.

The dark purple rectangle touching the light purple rectangle.

Figure 8. Sample output generated by DESCRIBER for target objects indicated by arrows in the images. Relative spatial clauses are automatically generated to reduce ambiguity when needed (bottom two scenes).

Words acquired by DESCRIBER come closer to being symbolic. DESCRIBER's acquired hierarchical language structures encode relations between words independent of their grounding. These relations enable DESCRIBER to generate verbal descriptions for novel scenes. This generative capacity is a result of the formation and use of word classes. Statistical rules of word order acquired from observation of some words are mapped to other words on the basis of shared class membership. For example, if the sequence 'large blue square' is observed, the sequence 'small red rectangle' can be generated if the appropriate word classes have been acquired. Since word classes are formed partially on the basis of semantic similarity, bottom-up visual grounding directly influences the application of syntactic rules to words. Thus, the rules of symbol manipulation in DESCRIBER are influenced by subsymbolic (iconic), visually grounded structure.

Conclusions and Future Directions

The systems presented in this paper represent steps forward toward machines which can converse about what they see and do. In contrast to conventional natural language processing systems, the semantics of words and utterances in CELL and DESCRIBER are grounded in visual representations. The non-symbolic links to perceptually derived representations constitute a step toward grounding semantics which will enable machines to autonomously acquire and use symbolic information for their own purposes. However, a limitation in this work is that neither CELL nor DESCRIBER can act on their environments. All they can do is talk and listen. Future work will address this important issue.

We are investigating new grounded architectures with two important features: the ability to act on their environment, and the ability to autonomously seek goals. Figure 9 shows a newly constructed robot which has been designed as a platform for supporting our current investigations. This robot has a much richer set of sensors than the robotic platform used with CELL, and also has the ability to manipulate objects in its environment. Aspects of CELL and DESCRIBER will be integrated and extended to support experiments with this new robot.

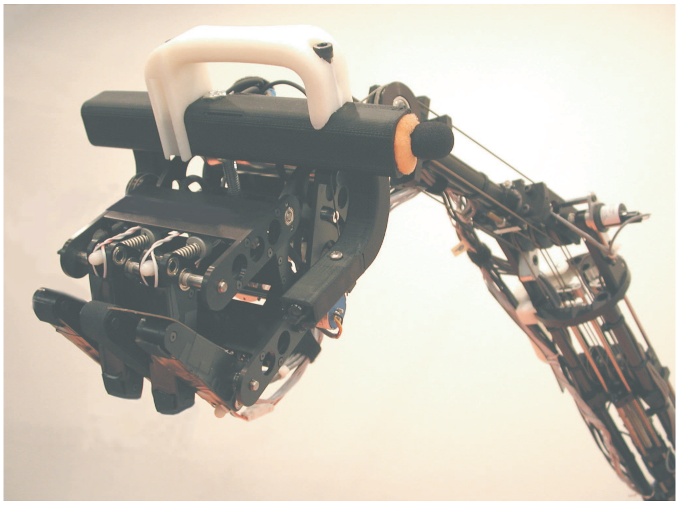


Figure 9. A new robot for experiments in sensorimotor representation grounding. Pictured here is the robot's 'head' with mouth closed. The sensor system includes stereo color cameras, stereo microphones, an artificial vestibular system, course grain touch sensors, joint position detectors, and artificial proprioceptive joint detectors. All degrees of freedom are compliant; a human may physically interact with the robot while it is in operation, providing a natural way to teach the robot how to act.

With continuing progress in the fields of perceptual computing, robotics, pattern recognition, and machine learning, the components required to create sophisticated, physically-grounded, language learning machines have become a reality. Simulated environments such as that used in DESCRIBER vastly simplify the development process and provide scaffolding for developing data structures and algorithms. However, for reasons discussed at the beginning of this paper, real physical environments must ultimately be addressed. A great challenge in building physically grounded machines such as CELL is the complexity of designing and debugging large heterogeneous hardware and software systems. To overcome this challenge, powerful data visualization and debugging tools must be developed in the future that are appropriate for these kinds of systems.

These projects help lead towards a new generation of communication machines which are grounded in the same physical reality as people. As one example of a practical outcome of this kind of work, I expect perceptually situated verbally controlled robots to become a reality by following this path of research. Both CELL and DESCRIBER have resulted in representations and learning algorithms which will serve as building blocks towards this goal. Ultimately, I believe that the lessons learned from designing grounded representational structures and processes can help lead to insights into the nature of the human conceptual and communication systems by letting us 'step out of ourselves' and perceive the world through the bodies and senses of our machines.

Acknowledgements

This paper benefited from comments from Peter Gorniak, Luc Steels, Rupal Patel, and Kai-Yuh Hsiao.

References

Agre, P. (1988). *The dynamic structure of everyday life* (Tech. Rep. No. 1085). MIT Artificial Intelligence Laboratory.

Aslin, R., Woodward, J., LaMendola, N., & Bever, T. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax* (p. 117-134). Mahwah, NJ: Erlbaum.

Bailey, D., Feldman, J., Narayanan, S., & Lakoff, G. (1997). Embodied lexical development. In *Proceedings of the nineteenth annual meeting of the cognitive science society.* Mahwah, NJ: Erlbaum.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioural and Brain Sciences*, *22*, 577-609.

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

Brent, M. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*, 71-106.

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, *2(1)*, 14-23.

Cangelosi, A., & Harnad, S. (2002). The adaptive advantage of symbolictheft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication.*

Cover, T., & Thomas, J. A. (1991). *Elements of information theory.* New York, NY: Wiley-Interscience.

de Marcken, C. (1996). *Unsupervised language acquisition.* Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Deacon, T. (1997). *The symbolic species : The co-evolution of language and the brain.* Norton.

Garofolo, J. (1988). *Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database.* Gaithersburgh, MD: National Institute of Standards and Technology (NIST).

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

Huttenlocher, J., & Smiley, P. (1994). Early word meanings: the case of object names. In P. Bloom (Ed.), *Language acquisition: core readings* (p. 222-247). Cambridge, MA: MIT Press.

Jackendoff, R. (1983). *Semantics and cognition.* Cambridge, MA: MIT Press.

Johnson, M. (1987). *The body in the mind.* Chicago: Univeristy of Chicago Press.

Kuhl, P., Williams, K., Lacerda, F., Stevens, K., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*, 606-608.

Lakoff, G. (1987). *Women, fire, and dangerous things.* Chicago, IL: The University of Chicago Press.

Newell, A., & Simon, H. (1976). Computer science as emperical inquiry: Symbols and search. *Communications of the ACM*, *19*, 113-126.

Peirce, C. (1932). Division of signs. In C. Hartshorne & P. Weiss (Eds.), *Collected papers of charles sanders peirce* (Vol. II). Cambridge, MA: Harvard Univeristy Press.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-285.

Regier, T. (1996). *The human semantic potential.* Cambridge, MA: MIT Press.

Robinson, T. (1994). An application of recurrent nets to phone probability estimation. *IEEE Trans. Neural Networks*, *5*(3).

Roy, D. (1999). *Learning words from sights and sounds: A computational model.* Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Roy, D. (2000). Integration of speech and vision using mutual information. In *Proc. of ICASSP.* Istanbul, Turkey.

Roy, D. (In press). Grounded spoken language acquisition: Experiments in word learning. *IEEE Transactions on Multimedia.*

Roy, D. (In review). Learning to generate visually grounded spoken language. *Computer Speech and Language.*

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

Sankar, A., & Gorin, A. (1993). Adaptive language acquisition in a multi-sensory device. In *Artificial neural networks for speech and vision* (p. 324-356). London: Chapman and Hall.

Schiele, B., & Crowley, J. (1996). Probabilistic object recognition using multidimensional receptive field histograms. In *ICPR'96 proceedings of the 13th international conference on pattern recognition, volume b* (pp. 50–54).

Searle, J. (1980). Minds, brains, and programs. *The Behavioural and Brain Sciences*, *3.*

Siskind, J. (1992). *Naive physics, event perception, lexical semantics, and language acquisition.* Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Siskind, J. (2001). Grounding the Lexical Semantics of Verbs in Visual Perception using Force Dynamics and Event Logic. *Artificial Intelligence Review*, *15*, 31-90.

Steels, L., & Kaplan, F. (2002). Aibo's first words. the social learning of language and meaning. *Evolution of Communication.*

Steels, L., & Vogt, P. (1997). Grounding adaptive language games in robotic agents. In C. Husbands & I. Harvey (Eds.), *Proceedings of the 4th european conference on artificial life.* Cambridge, MA: MIT Press.