SEMANTIC BASED LEARNING OF SYNTAX IN AN AUTONOMOUS ROBOT

BY

MATTHEW R MCCLAIN

B.S., Rensselaer Polytechnic Institute, 1999
M.S., University of Illinois at Urbana-Champaign, 2003

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2006

Urbana, Illinois

ABSTRACT

The ability to program a digital computer to use a natural language has evaded humans for over half a century. I believe that this is because all previous attempts have ignored the role of the physical world in language acquisition. It is the goal of the Language Acquisition Group at the University of Illinois at Urbana-Champaign to build a robot that is able to learn language as well as humans through embodied sensorimotor interaction with the physical world. I propose cognitive structures for an autonomous robot to enable it to learn the syntax of noun-verb sentences using its semantic understanding of the sentences' words. Specifically, production rules of syntax in Chomsky normal form will be explicitly represented using a hidden Markov model. These structures will be integrated into the robot's long-term associative memory, so that the representations of syntax that the robot acquires are dependent upon the robot's sensorimotor experience. The robot will then use this representation of syntax to produce sentences, some of which are novel, that describe what the robot observes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1 Language Acquisition

1.1.1 History

The ability to create artifacts with "living" qualities has always been the interest of humans. Wiener [1] states, "At every age of technique since Daedalus or Hero of Alexandria, the ability of the artificer to produce a working simulacrum of a living organism has always intrigued people" (p. 39). Since human beings display the most advanced behavior of any living creature, creating a humanlike machine, or simply a machine that can think like a human, has been the loftiest of such interests. These desires first appeared in myths and science fiction, from the golems that were given life from clay by the Rabbi of Prague in 1580 to humanoid robots in the 1926 movie *Metropolis* and L. Frank Baum's books about the Land of Oz in the early 20th century.

Wiener [1] has noted that each era of technology has its own models for the human mind as metaphors of the high technologies of the day. Not until the advent of the digital computer in the 20th century has that technology made a thinking machine feasible. Motivated by this, Turing in 1950 connected the ability to think to a machine's ability to use language in his proposal of the Turing test [2]. Although the validity of this test has been argued, there is an intuition that if a machine could use a natural language as well as a human, then it must be that it can perform symbolic manipulation of this language, i.e., that it can think.

The following history of language acquisition is guided by Levinson [3].

1.1.1.1 Cybernetics

Wiener coined the term "cybernetics" to encompass the study of control and communications in several different fields (engineering, neurophysiology, sociology, mathematics, and so on), the commonality being the use of feedback to regulate a process. While studying neurophysiology with Rosenblueth of the Harvard Medical School, he noticed that the ability for a human to pick up a pencil is strikingly similar to the control problem in aiming an antiaircraft gun [1]. Remarkably, control engineering's phenomena of feedback and overshoot have direct correlates in neurophysiological disorders of *tabes dorsalis* (the lack of proprioceptive feedback) and purpose tremor (the

tendency to make exaggerated corrections in response to proprioceptive feedback), respectively. This led Wiener to rethink the traditional model of the human mind:

> The central nervous system no longer appears as a self-contained organ, receiving inputs from the senses and discharging into the muscles. On the contrary, some of its most characteristic processes are explicable only as circular processes, emerging from the nervous system into the muscles, and re-entering the nervous system through the sense organs, whether they be proprioceptors or organs of the special senses. [1, p. 8]

In 1948, there had already been many machines built using feedback as a means of control. Early devices like Drebbel's temperature regulators invented in the 1600s and Watt's governor invented in the late 1700s were succeeded by various electronic sensors and effectors. Weiner remarks, "[T]he many automata of the present age are coupled to the outside world both for the reception of impressions and for the performance of actions. They contain sense organs, effectors, and the equivalent of a nervous system to integrate the transfer of information from the one to the other" [1, p. 43].

In fact, this technology, along with the emerging digital computer, had been so developed that Wiener made his own claim for the possibility of a thinking machine in the near future:

> It has long been clear to me that the modern ultra-rapid computing machine was in principle an ideal central nervous system to an apparatus for automatic control; and that its input and output need not be in the from of numbers or diagrams but might very well be, respectively, the readings of artificial sense organs, such as photoelectric cells or thermometers, and the performance of motors or solenoids. With the aid of strain gauges or similar agencies to read the performance of these motor organs and to report, to "feed back," to the central control system as an artificial kinesthetic sense, we are already in a position to construct artificial machines of almost any degree of elaborateness of performance. [1, p. 26-27]

Of course, 45 years later, we have still yet to reach this goal. The wisdom of Wiener's claim, however, comes from his adherence to sense and motor organs (he gives a synopsis of how to build artificial limbs earlier). I believe that this embodied mind approach to building a thinking machine holds much promise, especially in contrast to the dominant approach to artificial intelligence.

1.1.1.2 Artificial intelligence

The field of artificial intelligence (AI) as it is largely practiced today can be said to have emerged from the Church-Turing thesis and the Turing test presented in Turing's landmark 1950 paper [2]. The Church-Turing thesis, which itself was a product of the failed attempt to axiomatize mathematics, proved the undecidability theorem, which says that the existence of unprovable theorems of mathematics cannot be predicted in a formal manner. This proof required a mechanism for performing any possible computation; Church did this by way of recursive functions while Turing used his Universal Turing Machine.

The appeal of Turing's approach was that it had a physical implementation in the emerging digital computer. Thus, if the human brain was a computational machine, no matter how complex, it could be emulated by a computer. Indeed, even Wiener [1] was so struck by its power, he wrote "it became clear to us that the ultra-rapid computing machine, depending as it does on consecutive switching devices, must represent almost an ideal model of the problems arising in the nervous system" [1, p. 14]. In Turing's 1950 paper [2], he details his own prediction for AI, saying that by the end of the past century, "one will be able to speak of machines thinking without expecting to be contradicted" (p. 442). Turing then goes on to prescribe a method for teaching a computer to use language for his imitation test, which amounts to teaching the computer language as you would a child. At this point, he makes two incorrect assumptions.

The first of these incorrect assumptions comes in Turing's [2] statement, "Our hope is that there is so little mechanism in the child brain that something like it can be easily programmed" (p. 456). The study of child cognitive development, which had not gained serious academic attention until the 1970s, has shown that children are born with a great deal of structure in their brain. The learning they perform is not a simple absorbing of information on a blank sheet, but an intricately structured and coordinated activity. In the 1950s, however, the scientific community considered children to be merely blank slates with no knowledge, which explains Turing's misconception.

Turing's second incorrect assumption concerns the character of the learning that children undergo that enables them to use language. He suggests that a computer could be taught language in the same manner in which a child is taught in school. It is in this

assumption that Wiener's and Turing's approaches to thinking machines differ most markedly. Wiener recognized the importance of equipping a computer with a myriad of senses and motors, whereas Turing suggests that the computer's sensorimotor abilities are not a central issue, stating, "We need not be too concerned about the legs, eyes, etc." [2, p. 456]. Turing's approach does not take into account the semantic meaning of language that children acquire by interacting with the world around them. He is misled by the case of Helen Keller, the deaf and blind woman who was linguistically functional despite her disabilities. It was the fact that she did not lose her sight and hearing until she was two years old that left her with the ability to use language. Turing redeems himself somewhat at the end of his paper, stating, "It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English" [2, p. 460]. However, he still does not account for motor function and the learning that children do outside of the classroom.

The study of artificial intelligence has since followed Turing's idea that sensori-motor function is not important. Even his suggested use of sense organs has largely been ignored. The power of the digital computer seduced the scientific community into believing that the computer alone could easily achieve artificial intelligence. Thus, for the last 50 years, the search for strong AI has been approached with the idea that computers can use language as well as humans without an embodied understanding of its meaning. This is the central reason why the pursuit has been unsuccessful.

1.2 Language Acquisition and Robotics Group at the University of Illinois at Urbana-Champaign

In the light of this failure of strong AI, the Language Acquisition and Robotics Group at the University of Illinois at Urbana-Champaign (LAR-UIUC) is attempting to build an autonomous robot with the ability to acquire language in the vein of Norbert Wiener's vision of an embodied mind.

1.2.1 Theory and assumptions

We posit that previous attempts at strong AI have failed because they separate language from the world, where its semantic meaning is embedded. Language in humans developed through evolution to enable us to better operate in the world in which we live. Thus, our approach is based on the idea that language use is grounded in a semantic

understanding of the world. Bloom [4] has noted that "Studies of children learning English and certain other languages have revealed that the semantics of early sentences have to do with ideas about objects that originate in the development of sensorimotor intelligence in the child's first 2 years" (p. 41). We also posit that this semantic understanding is best acquired through complex sensorimotor interaction with the world, associative long-term memory, and reinforcement learning. The memory should be associative, as its primary function is to correlate the robot's various sensorimotor inputs into a meaningful model of the world, and then use these models to guide its decisions. Reinforcement learning, both supervised (a reward or punishment is given from a benign teacher) and unsupervised (correlated occurrences are repeated in the environment), will be used to train the associative memory and tune the robot's decision making. Complex sensorimotor abilities are necessary to have an embodied mind, as in Wiener's vision of thinking machines. We believe that a once our robots have acquired a descriptive spatio-temporal understanding of the world as a basis for semantics, more abstract cognition can occur by way of analogy. Because much of our research is guided by human language acquisition, we would like our robot's senses and motor functionality to be as anthropomorphic as possible.

We must also be aware of some basic assumptions about how the world around us is structured and how we are able to perceive it. First, we must assume that things exist outside of our minds that are perceivable by our senses and that the physical laws that govern the world around us are constant. Second, we assume that the continuous signals that we receive from our senses represent useful information about the world around us that can be discretized and classified in meaningful ways. The third assumption that we make is that there are useful generalizations to be made about the things in the world that we can perceive. For example, there are very few things in the world that look like ducks, sound like ducks, and act like ducks, but are not actually ducks. Further, even if this is not the case (maybe it is a goose), other important information can be generalized from what we know about ducks (geese can also fly and swim, just like ducks). We posit, then, that language is a system of symbols that not only represents the things in the world that our minds have learned to classify, but is also used as a medium for thinking about the world.

1.2.2 Research method

The goal for LAR-UIUC is to develop computational cognitive models that enable robots to learn and use language as well as humans. It is important to make the point that our aim is not to directly mimic human language acquisition, although we do draw greatly upon the work of psychologists to guide our research. It is simply not feasible at this time to implement all of the known aspects of human language acquisition in a robot. By the nature of this endeavor, LAR-UIUC is approaching the phenomenon of language from a bottom-up perspective. Our methodology is develop enough lower-level functionality in our robots to be able to address higher levels of linguistic behavior, and not to wait until our robots can demonstrate all of the complexity of human language acquisition at one level before we attempt to investigate the next. Thus, it is necessary to remove some of the complexity from the natural environment in which humans learn language. Of course, such a license must be used with care – a balance must be struck between feasibility and relevance to the important questions about language that we seek to address. We believe that this bottom-up methodology has significant contributions to make the understanding of language. As seen in the second chapter of [5], many significant contributions to the understanding of the acoustics of human speech came from the development of the telephone.

One significant benefit of this research method is that there is access to the information regarding what the robot has learned about language. Specifically, at any point in the robot's development, the computer program that contains the language models can output the current states of the models. This information can then be interpreted by a human to find out exactly what the robot has or has not learned in order to give detailed explanations of the robot's linguistic behavior. This, of course, cannot be done when studying language in humans and is a significant barrier to understanding the linguistic behavior of children during language development.

1.2.3 Current implementation

Figure 1.1 shows the model of embodied cognition that has been developed for our project by Levinson et al. [6]. The somatic system incorporates all of the robots' functionalities that are necessary for cognition: senses for perceiving the world, a motor system for acting on the world, and a noetic system that is the brain of the robot. This

model is meant to be analogous to a human model of the mind. Our implementation, however, uses metal and silicon instead of flesh and bone.

The sensory system receives information from the world. In humans, this would include the five senses (sight, hearing, smell, touch, and taste) as well as proprioceptive feedback about the body (for example, tension in neck muscles that indicate the position of the head). In our robots, we are able to emulate vision through stereo cameras and hearing through stereo microphones. As well, various sensors on the robot are able to provide rudimentary touch sense and proprioceptive feedback.

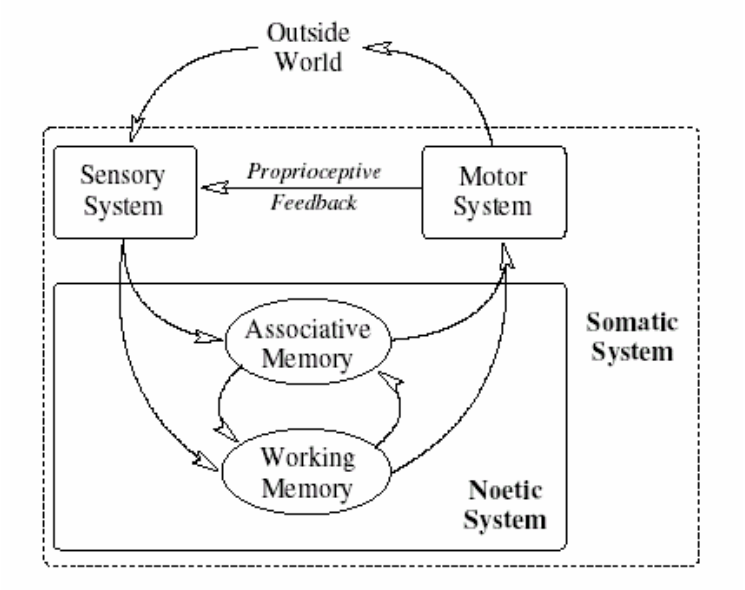

Figure 1.1: LAR-UIUC's model of embodied cognition

The motor system is crucial for interaction with the world. It is not enough to be able to simply perceive the world – our robots must be able to act on the world and perceive the effects of their actions because they must learn that they are part of the world. The motor functions of humans that we seek to replicate are the ability to change perceptual positions (locomotion, head motion), the ability to interact with objects (grabbing, pushing), and the ability to interact with humans through speech.

The noetic system is the brain of the robot. This system is responsible for determining the robot's actions as a function of its sensori-motor experience and instincts. Before the robot has acquired any sensori-motor experience, it will use a set of instincts to guide its behavior. Human infants exhibit strong instincts to explore the

world and acquire sensorimotor experience [7] as well as instincts for survival (fear of heights and loud noises, for example). Thus, our robots should have similar instincts to explore and survive.

The noetic system is divided into a working memory and an associative memory. The associative memory is the robot's long-term memory where the robot learns to classify the discretized information from the sensory system. Here, the robot builds its models of the world and develops connections with language. The working memory controls the robot's actions. Before the robot has learned anything, the working memory contains a set of instincts that governs how it will behave, guides the robot to acquire information for its associative memory, and performs survival tasks. The decisions that the working memory makes include commands to the motor system as well as commands to the associative memory that specify when to train the cognitive models using the information from the sensory system.

1.2.3.1 Current robotic implementation

1.2.3.1.1 Hardware

Currently, LAR-UIUC has three robots: Illy, Norbert, and Alan (shown in Figure 1.2). The robotic platform is from Arrick Robotics. This is a mobile robotic platform, with the following motor abilities:

- A front gripper that can pick up small objects, with a push switch on the inside of the gripper for tactical sensing
- A head with vertical tilt and horizontal panning
- Wheels driven by dual-differential drive with DC gear motor and encoders

To this robotic platform, we have added the following:

- On-board computer for receiving and distributing raw sensori-motor features
- Two cameras mounted on the head for stereo vision
- Two microphones for binaural hearing
- Wireless receiver/transmitter for communicating with desktop computers that handle most of the computational load of the robotic cognitive system

1.2.3.1.2 Software

The robot's base software includes programs written by Squire [8] to distribute the robot's raw sensorimotor features. Built on top of this are programs to process the raw sensorimotor features: a sound-source localization program developed by Li [9], speech processing and production programs that use warped linear prediction developed by Kleffner [10], and a visual processing program developed by Lin [11] that segments objects from the background and extracts shape and color features. As well, McClain [12] has implemented a distributed shared memory for higher-level features and a finite-state machine to act as the central controller and working memory for the robot during exploration.

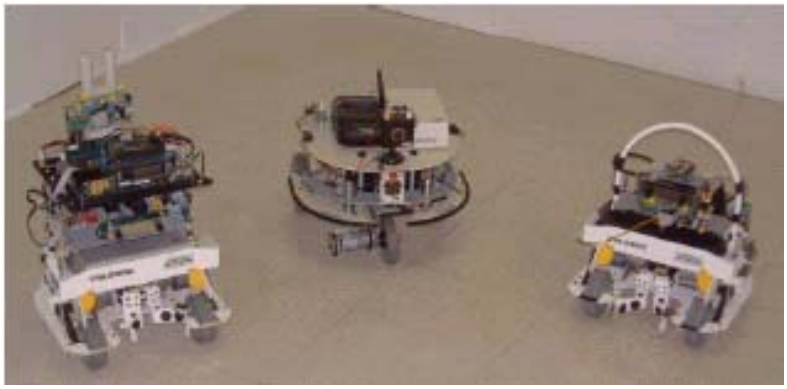

Figure 1.2: LAR-UIUC's robots (from left to right) Illy, Alan, and Norbert

1.3 Syntax Acquisition

The goal of this dissertation is to develop cognitive structures for a mobile, autonomous robot that enable it to learn and use basic syntax based on the robot's learned semantic understanding of the world. The rule of syntax to be learned is that a sentence can be composed of a noun followed by a verb.

The choice to make the syntax of two-word sentences the goal of this research is not arbitrary. In terms of language development, this is analogous to the two-word phase in young children. According to Brown [13], "At about eighteen months children are likely to begin constructing two-word utterances; such a one, for instance as *Push car*…A construction such as *Push car* is not just two single-word utterances spoken in a certain order" (p. 77). Brown [13] signifies this as Stage I of I-V, each of which correspond to different processes of sentence construction. Stage I corresponds the child learning the

semantic roles of the constituents of sentences. This is demonstrated by the fact that "for English, at least, Stage I children are able to use order appropriately in spontaneous speech and to make correct discriminating responses to contrasting orders" (p. 64). This is a significant step in language development, as it is the first use of syntax, which separates human language from the sort of communication used by other animals and which allows human language to be so expressive.

This stage follows one in which children produce only single words, naming individual objects ("doggie," "mama," etc.) or desires ("want," "mine," etc.). Bloom [14] even notes, "Many investigators have reported that the child's single-word utterances, before the development of syntax, seem to function as one-word sentences." (p. 10) This shows that children appear to want to communicate complex thoughts, but do not yet possess the linguistic competence to structure words into sentences.

The next of Brown's [13] stages, Stage II, is characterized by children beginning to demonstrate the use of "semantic modulations such as number, specificity, tense, aspect, mood, etc., expressed by inflections or free forms belonging to small closed classes" (p. 32). This represents a higher level of complexity of sentence formation that builds upon the children's knowledge from Stage I.

The basic information about sentence construction that children demonstrate in Stage I can also be used for increasing children's vocabulary of verbs. This process of syntactic bootstrapping, discussed in [15-17], is a method for determining what the verb in a sentence is referring to in the learning context. A child who has learned the rules of syntax that govern noun-verb sentences understands that the action described by the verb is happening to the object described by the noun. Then, if this child hears an unknown word following a noun, they can assume that the unknown word is a verb and that it refers to something happening to the object referred to by the noun.

1.3.1 Word learning

In order to enable robots to learn the proposed syntax, our robot must first know some words that can be used as nouns and as verbs. As mentioned in Section 1.2.2, previous work has resulted in the ability for robots to learn names of objects, which can be used as nouns. Thus, part of the current work must be to enable robots to learn words that can be used as verbs.

Psychologists who study childhood language acquisition have varying viewpoints on how children learn verbs. Most of the early research regarding word acquisition focuses on noun learning and the specific techniques that children use to assign word labels to objects. An overview of this research, along with reasons that psychologists have avoided verb word acquisition can be found in [18].

I propose that in the initial stages of language development, it is unnecessary to assume that a language user must have a priori knowledge of lexical categories, i.e., that words can be categorized as nouns or verbs. Tomasello [19] writes, "It is important to conceptualize the child's early cognition not solely in terms of objects and properties, as many theories do, but rather in terms of event structures, with objects being no more prominent in the child's conception of the world than the activities and events in which they are embedded" (p. 137). Children begin to produce language at a single-word level, and it is very difficult to categorize single words into lexical categories. Winograd [20] gives an account of this. Jurafsky and Martin [21] note, "Traditionally the definition of parts-of-speech has been based on morphological and syntactic function….While word classes do have tendencies toward semantic coherence, this is not necessarily the case, and in general we don't use semantic coherence as a definitional criterion for parts-of-speech" (p. 289). Lexical categories are more reliably assigned to words based on their position in a syntactic structure than by the semantics of the words themselves. Thus, unless we assume that children have innate knowledge about the morphological and syntactic rules of their language (which is not accepted by this thesis), there appears to be a paradox, which will be addressed in the next section.

1.3.2 Learning syntax using semantics

The proposed solution to the paradox presented in the previous section is that semantics is used to build the foundation for understanding lexical categories and syntax. Once this basic understanding is developed, adultlike competence can be acquired. According to Brown [13], "It was shown that the nouns used by young English-speaking children were more reliably the names of things and their verbs more reliably the names of actions than is the case for the nouns and verbs used by English-speaking adults" (p. 26). This proposed solution is related to the semantic bootstrapping hypothesis developed by Pinker [22], which will be discussed in Section 2.4.2.

From an early language learner's point of view, the semantics of words seem to fit together to form sentences without imposing classifications upon them a priori. Words that we call "nouns" generally describe physical objects, which are identifiable by features of their form or function, and "verbs" describe actions or temporal states. The fact that "nouns" and "verbs" encode different information about the world allows them to be combined together to create an exponential number of sentences.

In addition, a two-word sentence with a noun and a verb coveys more information than the two words individually. Interpreting the sentence "ball move" as "there is a ball and it is moving" explains more about what is happening than interpreting the words individually as "there is a ball" and "something is moving." The use of syntax allows for a more complete linguistic description of the world.

To show that a robot has learned syntax, it must be able to generalize the production rule to new situations. The cognitive structures that are developed in this thesis must be able to determine if a known word can be used in a certain syntactic context. In a sense, they must be able to classify words into lexical categories.

1.3.3 Requirements of this dissertation

The cognitive structures that are to be developed in this thesis should use general learning algorithms as much as possible. A great deal of debate has surrounded the question of whether humans' ability to learn syntax is due to an innate language faculty or generic statistical learning algorithms. The LAR-UIUC's project is based on the idea that language is primarily driven by semantics that are grounded in sensorimotor interaction with the world, and thus we deny the claim that human language learning is driven by an innate language faculty that allows humans to learn the syntax of language. In fact, questions such as these present a unique opportunity for this line of research to make a significant contribution to the understanding of language. While the work in this thesis cannot prove that humans do not have an innate language faculty, showing that robot can learn basic syntax using only generic learning algorithms leaves open the possibility that humans can learn syntax in a similar fashion.

1.4 Contributions and Layout of Dissertation

My own contributions to the LAR-UIUC project are as follows. First, I developed an autonomous exploration mode to enable the robots to explore their environment. This was implemented as a finite state machine (FSM), which has also served as working memory component for other experiments. Included in the autonomous exploration programs is a shared memory that allows programs to share information in an asynchronous fashion. Second, I developed two programs to assist with the robot's vision. One of these programs was simply to display the images captured by the robot's cameras onto a desktop monitor in a streaming fashion. The other program allows the robot's camera parameters (brightness, contrast, etc.) to be adjusted online, instead of having these parameters hard-coded. Third, I assisted in the design and building of the group's newest robot, Norbert. My final contribution is to add the robot's ability to learn basic syntax to the existing associative long-term memory:

1. Extending the current object-learning system developed by Squire [8] to learn the some basic action words and their perceptual representations
2. Implementing a small-vocabulary word recognizer that can use the robot's audio channel and can also perform speech production tasks.
3. Developing and simulating stochastic models for syntax acquisition
4. Implementing the above models for syntax learning into the robotic framework described in Section 1.2

The rest of this dissertation is organized as follows. Chapter 2 is a review of relevant research from various fields. Chapter 3 gives descriptions of the cognitive models and algorithms that will be used to learn syntax based on semantics, including results of simulations to test these algorithms. Chapter 4 will include the details of the robotic implementation of the cognitive models and the experiments to be performed with the robot. In Chapter 5, the results of the robotic experiments will be presented and discussed. Chapter 6 will contain ideas for further developing the research presented here.

CHAPTER 2 LITERATURE REVIEW

2.1 Automatic Speech Recognition and Natural Language Processing

Since its beginnings in the middle of the 20th century, there have been many significant developments in automatic speech recognition and natural language processing. The sound spectrograph of Koenig et al. [23] in 1946 allowed research to be performed using spectral characteristics to classify speech signals, which lead to the first machine speech recognizers at Bell Labs by Davis et al. [24] in 1952. Further advances were made with the advent of cepstral processing by Oppenheim et al. [25] in 1968 and speech coding with linear prediction coefficients by Atal and Hanauer [26] in 1971.

Text-based language processing was attempted with the ELIZA program by Weizenbaum [27] in 1976 and SHRDLU by Winograd [28] in 1972. ELIZA is most well known for its imitations of a psychologist in which it turns the user's responses into another question. SHRDLU acted on a simulated world of blocks in response to natural language commands. Both of these programs drew much attention to their success; however, their approaches to language processing are completely nonembodied and wholly dependent on the symbols for language that are programmed into them. Thus, their success is only in the limited domain of language for which they are programmed.

The hidden Markov model (HMM) marked a key innovation in statistical speech processing. An HMM is a stochastic model in which the underlying structure is composed of unobservable discrete states which generate observable outputs, and the next-state behavior is Markovian (the next-state is only dependent on the previous state). When the model is trained with a sequence of observations, it converges to a local maximum in the model's parameter space, which represents the most likely model to have generated the observation sequence. Baker [29] and Jelinek et al. [30] independently were the first to use HMMs for speech processing in 1975.

The significance of HMMs was demonstrated in two early experiments by Cave and Neuwirth [31] and Poritz [32]. These experiments showed that HMMs can be powerful tools in representing linguistic structure. Cave and Neuwirth [31] trained an HMM on a corpus of English text using the characters as the observations. In the training corpus, only the 26 letters of the alphabet and the space character were used. The result was that the HMM learned categories of letters and lexical structure of written English

without prior knowledge of these two phenomena. This information is represented by the resulting state observation probabilities and state-transition probabilities, respectively. For instance, depending on the number of states used, one or more states will have non-zero output probabilities for only vowels. This means that the HMM has discovered that these letters are in a special category and have a special purpose in the structure of language. The output probabilities for all of the states can be analyzed in the same way. Then, by viewing the state-transition matrix, one can observe the structural rules that the HMM has learned. For example, with a sufficient number of states, one state will only output the space character. Then, the possible next-states after this state contain the letters that begin words. Poritz [32] performed a similar experiment on speech in which the states of the HMM converged to represent broad phonetic categories. In this case, the state-transition matrix of the HMM captures rules of phonology.

Baker and Jelinek also made use of algorithms to transform strings of phonemes into words. Baker [29] used the Viterbi algorithm, which was first used for speech processing by Vintsyuk [33] in 1968. Jelinek et al. [30] utilized the stack decoder, or A* decoder, which was developed by Jelinek [34] in 1969. Two other notable speech recognition systems were developed at Carnegie-Mellon University: Harpy by Lowerre [35] and Hearsay-II by Lesser et al. [36].

For a more in-depth history of automatic speech recognition and natural language processing, see Jurafsky and Martin [21].

2.2 LAR-UIUC Past Work

The following is a selection of the work by members of LAR-UIUC to the present.

2.2.1 Learned voice control of robot movement

Liu [37] developed programs in which the robot learns to act according to voice commands. The robot is trained by an experimenter who speaks the movement command while pushing a touch sensor on the robot. The sensors are positioned on the robot so that pushing the button exerts a force on the robot in the direction that corresponds to the spoken movement command. For example, the button for moving the robot forward is at the back of the robot. Once trained in this fashion, the robot's movement can be

controlled by the experimenter using voice commands. This early experiment demonstrates the robot's ability to associate linguistic input with the robot's sensorimotor system.

2.2.2 Autonomous exploration

A program for autonomous exploration has been implemented by McClain [12]. In this experiment, the robot continuously searches its environment for objects and investigates them to obtain useful information about them. The robot also has simple survival mechanisms that enable it to explore its environment for long periods of time. This experiment demonstrates how the robot can exhibit higher-level exploratory behavior from a set of low-level instinctual behaviors that have been coded into the robot.

2.2.3 Cascade of hidden Markov models as a semantic associative memory

Squire [8] has developed a cognitive structure based on a cascade of hidden Markov models (CHMM) that enables the robot to learn the names of objects in its environment and their semantic meaning. This model incorporates three separate HMMs into one structure, with two lower-level models that classify visual and audio (linguistic) input, respectively. The third HMM is used to learn correlations between these two modalities. The robot uses these to associate spoken names for objects with their physical appearance in its visual inputs, thus developing a semantic understanding of the words. The algorithm developed uses a recursive maximum likelihood estimator (RMLE), also developed by Squire, to update the HMMs while the robot is exploring its environment.

2.2.4 Binaural sound source localization

Li [9] has developed algorithms for LAR-UIUC's robots that enable them to accurately determine the direction of sounds. These algorithms essentially use time differences from the robot's two microphones and compute the angle on the half-plane in front of the robot that indicates the sound's source. In experiments, the robot uses this to turn toward sound sources. This is used to help direct the robot as well as to indicate that a sound has grabbed the robot's attention.

2.2.5 Visual navigation using PQ-learning

An algorithm for learning visual navigation by reinforcement called PQ-Learning has been developed by Zhu and Levinson [38]. This algorithm has allowed LAR-UIUC's robots to learn optimal action policies for reaching objects much more efficiently than traditional Q-learning.

2.2.6 Visual maze navigation

Lin [39] has developed algorithms for LAR-UIUC's robots that enable them to learn how to navigate through a maze using visual information and reinforcement. In this experiment, the robot learns effective actions for navigating the maze as well as a cognitive representation of the maze, which allows it to find the goal more efficiently.

2.2.7 Mental map learning as manifold learning from time series

Algorithms for manifold learning from inputs in a time series have been developed by Lin [11]. This has been applied to LAR-UIUC's robotic framework to learn a mental map of the locations of the objects in the robot's environment. The robot is then able to use this mental map to estimate the robot's position and orientation in the environment when an object is present in a visual scene.

2.3 Embodied Cognition

In this section, related work in embodied cognition from various fields will be presented.

2.3.1 Embodied cognition in humans

The following is a brief review of the research concerning the role of embodiment in human cognition.

Piaget was perhaps one of the first psychologists to attempt to explain the ways in which children's minds work. In [40], Piaget explores the development of sensorimotor intelligence in very young children. He begins with innate reflexes (sucking, in particular) and continues by describing how children begin to coordinate amongst their different sensorimotor modalities. From this, Piaget concludes that intelligence in humans is grounded in our sensorimotor interaction with the world around us.

Lakoff and Johnson [41] share Piaget's conclusions and make connections to language by hypothesizing that all language is grounded in basic sensorimotor understanding of the world by way of metaphors. They give many examples of how

abstract notions such as love, economics, and morality are conceptualized in language as metaphors of sensorimotor experience.

Smith and Gasser [42] offer "six lessons" from developmental psychology for building intelligent embodied agents: use many different modalities that interact; learn incrementally, not in large leaps; interact with the physical world; use exploration that is not goal-oriented; use social interaction; and learn a language.

Experiments that demonstrate the role of embodiment in human language use have been performed by Kaschak et al [43]. These experiments probe the question of how sensorimotor experience is used in language comprehension in the human brain. The findings were that a visual stimulus that matches the content of a sentence interferes with the processing of the sentence when presented concurrently. The conclusion that is drawn from this is that the same part of the brain is used in both visual and language tasks, which demonstrates the role of embodiment in natural language.

2.3.2 Multiagent systems

Research has been done on how multiple agents acting in simulated environments can develop their own ways of conceptualizing and communicating about their world. Wang and Gasser [44] give a framework for mutual online concept learning (MOCL) using a connectionist algorithm that they call the mutual perceptron convergence algorithm. MOCL is distinguished from other multi-agent learning systems by the fact that there is no predefined concept for the agents to learn. Therefore, the concepts that the agents learn must be developed by the interactions of the agents. Komarova and Niyogi [45] have developed a measure of mutual intelligibility between two languages and used it to address questions about language learning between two agents. These questions include what language a learner should acquire to optimize the mutual intelligibility with a target language and what algorithms are useful for acquiring this optimal language, both under various conditions.

2.3.3 Developmental robotics

New approaches to artificial intelligence are being explored that deal with the issue of embodiment. The new field has been called "developmental robotics" or "epigenetic robots", and is characterized by using robots, embodied with sensorimotor

capabilities, to discover for themselves the structure in their environment and their own embodiment instead of having most of their functionality pre-programmed.

Rodney Brooks was one of the first researchers to articulate the failures of traditional AI as a paradigmatic failure [46]. He argues that the foundations of human intelligence are in how we evolved to survive in the world, and so attempting to create intelligent machines that do not interact with the world is doomed to failure. Brooks' first ideas involved implementing insect robots (with Flynn and Tavrow) [47]. At Massachusetts Institute of Technology Artificial Intelligence Laboratory (MIT AI Lab), Brooks worked with Cog, a robot designed to emulate a human torso with arms, neck, and head that was used for studying sensori-motor coordination [48, 49]. Also at MIT AI Lab, Breazeal used Kismet, a robotic head with complex gesturing capabilities, was used to study social interaction with humans [50].

Weng of Michigan State University focuses on what he calls "autonomous mental development" [51]. This approach is meant to parallel human mental development in that the robots learn by interacting with their environment, without task-specific goals. Weng has implemented his ideas in the robots SAIL-1, SAIL-2, and Dav, and has been able to demonstrate the ability to learn complex tasks such as obstacle avoidance in autonomous navigation (with Zeng) [52] and auditory development (with Zhang) [53].

Sporns and Alexander of Indiana University have conducted research in neurophysiological modeling using a system of rewards to modulate how their robots interact with their environment [54, 55].

Kuipers has developed methods for unsupervised robotic learning in which the robot initially has no knowledge of its own sensors, effectors, or the environment. In his work, Kuipers demonstrates that the robots are able to learn the relationships between their sensors, effectors and environment [56], position and orientation in their environment [57], and object representations [58] using only domain-independent statistical learning methods. These methods are based on learning the Spatial Semantic Hierarchy [59], which describes the different levels of a cognitive map from raw sensorimotor inputs to abstract representations.

Oudeyer et al. [60] have developed a method for autonomous learning from sensorimotor experience called Intelligent Adaptive Curiosity. In this method, the robot

seeks out high-potential learning situations by measuring the error in making predictions based on the robot's previous experiences.

2.3.3.1 Machine syntax acquisition

The acquisition of syntax by machines has been studied by Sugita and Tani [61] and Roy [62, 63]. Sugita and Tani use a pair of recursive neural networks (RNNs) to learn the compositional semantics of two-word command sentences with a robot. This is accomplished by learning the correspondences between sentences and behavioral patterns, and generalizing them to novel contexts. Roy has implemented algorithms that learn adjective-noun phrases using visual information in both simulated visual scenes with DESCRIBER [62] and a robotic implementation called CELL in the robot Toco [63]. The CELL system works by computing a mutual information measure between the robot's audio and visual modalities to discover the associations between words and their semantic representations.

2.4 Psychology and Psycholinguistics

Many researchers in the field of psychology have studied the development of language. Many of the ideas from their research have been used for this thesis. In the first section, the works cited concern children's stages of syntax development. The second section involves the "semantic bootstrapping hypothesis" of syntax acquisition. The third section includes research in computational modeling of syntax.

2.4.1 Childhood language and syntax development

From Brown's studies of the development of children's speech, he has determined five stages of children's language [13]. Brown designates these stages both by the mean length of utterance (MLU, measured in morphemes) and on the syntactic complexity that the children's utterances display. The first two stages have been described in Section 1.3. Brown's stage III is characterized by the use of a variety of sentence modalities, such as questions, negatives, and imperative. In stage IV, children begin to embed simple sentences within one another. Stage V is designated by the coordination of simple sentences.

Bloom's [14] study of children's speech goes beyond the formal structure of children's sentences and takes into account the context in which the child utters the

sentence. Her claim is that in order to understand the mental processes of children, it is crucially important to know what the child means to communicate about the world around it.

2.4.2 Syntax acquisition and semantic bootstrapping

Of relevance to this thesis is the semantic bootstrapping hypothesis. The basis of this theory is that, although lexical categories that are necessary for syntax are not universally defined by semantics, children must initially use semantics to learn them. Then, once a language learner creates these lexical categories, the learner can begin to acquire the rules of syntax and then use the rules of syntax to perform lexical categorization on words that do not fit the semantic definitions. According to this hypothesis, children would first learn words for which the semantic meaning is based directly upon sensori-motor experience: words like "dog," "toy," and "cup" correspond to things that have observable features of form and function, and words like "give", "run", and "eat" correspond to observable changes in state. Then, if the child understands a sentence like "He gave me the toy," the child can learn from the sentence "He gave me the idea" that the word "idea," which does not have semantic meaning in sensorimotor experience, is the same type of word as "dog," "toy," or "cup."

This idea was first articulated by Wexler and Culicover [64], Grimshaw [65], and Macnamara [66], and was further developed (and given the name "semantic bootstrapping hypothesis") by Pinker [22]. Pinker cites as evidence to this theory the fact that children's first words in syntactic contexts fit the semantic definitions of lexical categories and the fact that parents also follow this pattern in speech to their children.

2.4.3 Computational models and statistical learning

Saffran [67] has studied the extent to which humans can learn linguistic structures from statistical information. In experiments with adults and children using an artificial language that contains predictive dependencies in its grammatical structure, Saffran was able to conclude that humans use statistical learning mechanisms to learn the dependencies in a language.

Chang et al. [68] have developed a Dual-path connectionist model that is able to learn rules of syntax and lexical categories, and use this information to produce syntactically correct sentences. The model is trained with grammatical sentences, along with

information about the meaning of the sentence. Although they do not use and embodied approach (the information about the meaning of the sentence is symbolically encoded), Chang et al. take care in explaining how the information that they do provide to the model would be salient to an embodied language learner.

2.5 Linguistics

2.5.1 Formal grammars

Chomsky [69] has made enormous contributions to the understanding of language by formulating mathematical representations of syntax, called formal grammars. A formal grammar is a means of representing all of the allowable sequences of words (sentences) of a language using a generative model. A formal grammar is composed of four elements, represented symbolically by V_N, V_T, S, and R. V_N is a finite set of nonterminal symbols and V_T is a finite set of terminal symbols, disjoint from V_N. The terms "terminal" and "nonterminal" mean that the symbol can or cannot appear in a sentence of the language, respectively. In natural language, terminal symbols are words and nonterminal symbols represent lexical categories like N ("noun") or V ("verb") and phrase categories such as NP ("noun phrase", for example "the wet brown dog"). The symbol S is a special nonterminal symbol that signifies a valid sentence of the language. The valid sequences of terminal symbols (words) that comprise a language are generated using the set of production rules R. These rules tell how the symbols (both terminal and nonterminal) can be replaced by other symbols and are written as $X \rightarrow Y$, where both X and Y can be combinations of terminal and nonterminal symbols. For this reason, the production rules are also sometimes called rewrite rules. One example of a type of production rule are those that describe which words belong to which lexical categories in natural language. In the example above, X would be a single nonterminal that represents a lexical category and Y would be a nonterminal (word) that is of that type, for example N→"ball." This production rule means that the word "ball" can be used as a noun. The sentences of a language, then, are generated by starting with the nonterminal S and applying the production rules to get a sequence of terminal symbols.

2.5.2 Chomsky hierarchy

The Chomsky hierarchy [70] categorizes grammars according to the forms that the production rules R can take. When describing these production rules, uppercase letters will be used to designate nonterminal symbols and lowercase letters will designate terminal symbols. In the Chomsky hierarchy, each category uses a subset of the rules of the category above it. At the highest level of the hierarchy is phrase-structure grammars in which the production rules are unrestricted. In the next lowest level are context-sensitive grammars, which may contain production rules of the form $xYz \rightarrow xyz$. Context-free grammars occupy the next lowest level, and are characterized by the fact that nonterminal symbols on the left-hand side of the production rules are rewritten as terminal symbols regardless of the symbols surrounding it. At the lowest level of the Chomsky hierarchy are regular grammars, in which the rules can have the forms $X \rightarrow x$ and $X \rightarrow xY$.

Of special interest to this thesis are context-free grammars. Although context-free grammars cannot fully represent the syntax of natural languages, they still provide enough representational power to explain many linguistic phenomena, without a great deal of computational complexity. Any context-free grammar can be written in what is called Chomsky normal form (CNF) in which all rules are of the form $X \rightarrow YZ$ or $X \rightarrow x$. This formulation of context grammars leads to an intuitive binary tree interpretation of sentence generation, with the S symbol at the root of the tree.

2.5.3 Predicate argument structures

Linguists have used predicate argument structures as a way of representing the combinational semantics of a sentence as formalized logical expressions. Works that have used this representation of semantics are Montague [71], Batali [72], Parsons [73], and Kamp and Reyle [74]. Predicate argument structures do not take into account semantic information about the world that underlies the language, and thus, will not be used in this thesis.

2.6 Computational Algorithms for Syntax

2.6.1 Parsers

This section reviews several algorithms that determine if a sentence is grammatical, and the set of production rules it was generated by. Deterministic algorithms have been developed for parsing regular grammars by Djikstra [75], for context-free grammars by Younger [76], and context-sensitive grammars by Tanaka and Fu [77]. Details of these algorithms, as well as algorithms that take into account uncertainty in the words (for example, in parsing sentences from speech input), can be found in Levinson [3].

2.6.2 Grammatical inference

Algorithms for inferring the rules of a grammar given sentences generated by that grammar are less common. Levinson [3] gives an algorithm for inferring rules of a regular grammar based on Baum's algorithm [78]. Baker's [79] inside-outside algorithm provides a way of inferring the rules of a CNF grammar from a large set of sentences generated by the grammar.

Klein [80] has developed algorithms for inferring grammatical structure in an unsupervised manner. Three different methods were used on corpora from different languages: one which induces bracketed tree structure, another which induces word-to-word dependency structures, and a third method that merges the previous two. Klein's algorithm to induce bracketed tree structure uses a generative constituent-context model (CCM) where the lexical items surrounding a bracket (the "context") are used in addition to the lexical items within the bracket (the "constituents"). The word dependency structure is inferred using a generative dependency model with valence (DMV). Improvements to unsupervised grammatical inference have also been made by Smith and Eisner [81], using structural annealing and contrastive estimation techniques.

2.7 Summary

This chapter presents a broad view of research related to this thesis. These works cited present many different ways in which researchers have attempted to further the understanding of human language. This thesis attempts to bring together ideas from

many of these different approaches. Those works that will be drawn upon directly by this thesis have been explained more thoroughly. The next chapter will describe in detail the problem that this thesis seeks to address, followed by the proposed solution.

CHAPTER 3 SEMANTIC BASED LEARNING OF SYNTAX

3.1 Scope and Definition

The problem addressed by this dissertation is the learning of syntax of noun-verb sentences by an autonomous robot, using its semantic understanding of the world that is grounded in its sensori-motor capabilities.

For this dissertation, learning the syntax of noun-verb sentences means learning the lexical categorization, the sequence of the lexical categories, and the compositional semantics associated with the CNF production rule S→NV. The lexical categorization is to be based on sensorimotor experience, where N is the nonterminal symbol that represents the category of nouns, and V is the nonterminal symbol that represents the category of verbs, and S is a meaningless delimiter. It is assumed that the robot will have previously learned all of the words that will be used in the learning of this rule of syntax. That is, the robot will have been sufficiently trained on the names of the objects in its environment and words that describe the objects' actions. The robot's name learning system has been developed by Squire [8] and was introduced in Section 1.3. The proposed solution for this thesis will build upon this work, and so a more thorough explanation follows in Section 3.1.1. To train the robot, a human observer will speak noun-verb sentences to the robot that describe what the robot observes. Once the robot has learned the S→NV rule of syntax, it should be used to describe novel occurrences in the world. The implementation should be robust so that it can be used in LAR-UIUC's robotic framework in future experiments.

The point that the robot's acquisition of syntax is equivalent to learning a syntax production rule in CNF is for the purpose of extending the work proposed in this thesis to learning an entire grammar. The topic of extending the current work to learning additional production rules in CNF is discussed in Chapter 6.

3.1.1 Current name learning system

This section describes the name learning system implemented in LAR-UIUC's robots by Squire [8]. The purpose of this system is to enable a robot to gain a semantic understanding of the names of objects, through its sensori-motor capabilities, in the framework of a semantic associative memory. The framework that was developed,

shown in Figure 3.1, allows the robots to associate visual representations of objects in the visual model with words recognized by the auditory model.

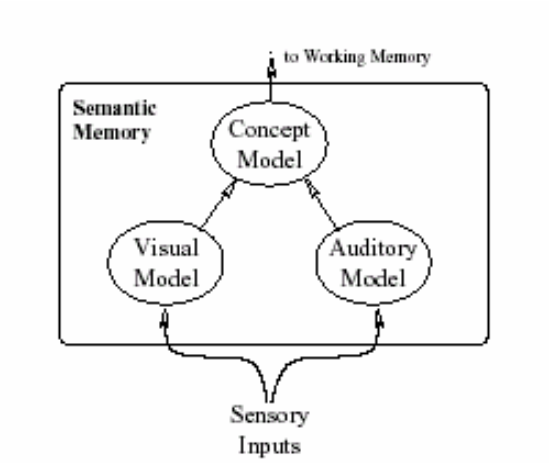

Figure 3.1 Model of semantic associative memory for object name learning

The auditory model in Figure 3.1 performs word recognition on the robot's audio input channel. This includes a speech detection component that determines the onset and completion of utterances, as well as algorithms to compute spectral information for classifying the speech segments. The identified word is chosen from a set of learned words, and outputted to the concept model.

The visual model recognizes co-occurrences from the robot's video input channel. This model receives visual features that are extracted from the images taken from the robot's cameras. The video processing that produces these features performs segmentation to find the objects in the images and then computes color and shape features for each of the objects. The visual model is trained on the features from one object at a time to associate the color and shape information for each of the objects in the robot's environment. These associations are used to classify observed objects, and this information is used as an input to the concept model.

The concept model is used to associate outputs from the visual model (classified objects) with outputs from the auditory model (classified words). To train this model, an experimenter speaks the name of the object that the robot is observing. The concept model learns to associate the acoustically identified word with the visually identified

object, which can be interpreted as learning the name of the object. Once these associations are learned, the concept model provides the robot with a semantic understanding of the names of the objects in its environment.

The semantic memory in Figure 3.1 was implemented as a CHMM, with each of the models in the figure implemented as an HMM. The HMMs used by the auditory and visual models can be viewed as classifiers, in which the states of the HMMs correspond to the classes to be identified. For example, the states of the visual model HMM correspond to the visual representations of each of the objects in the robot's environment. These visual representations are joint probability distributions over the visual inputs, and the parameters of these distributions are the observable outputs of the HMM's states. Similarly, the auditory model's HMM has one state per word that is learned, the observable outputs of which are the parameters of a joint probability distribution over the auditory features. The sequence information learned by these HMMs can be viewed as prior information when making classifications. The HMM used for the concept model is better viewed as learning associations between words (states of the auditory HMM) and visual representations (states of the visual HMM).

In order to implement the semantic memory in the robotic framework, Squire developed algorithms for the RMLE of the HMM parameters. Traditional algorithms for estimating the parameters of an HMM require a large data set a priori. To fit the semantic memory into the robotic framework, however, models that learn in an on-line fashion are required. See [8] for a derivation of the algorithm as well as a discussion of the algorithm's convergence properties.

3.2 Proposed Solution

This section presents the proposed solution for the robotic syntax acquisition.

3.2.1 Cognitive model

The proposed cognitive model for syntax acquisition is shown in Figure 3.2. This model builds upon the semantic memory in Figure 3.1, leaving the connections in that model in tact, and adding the necessary connections to the syntax model.

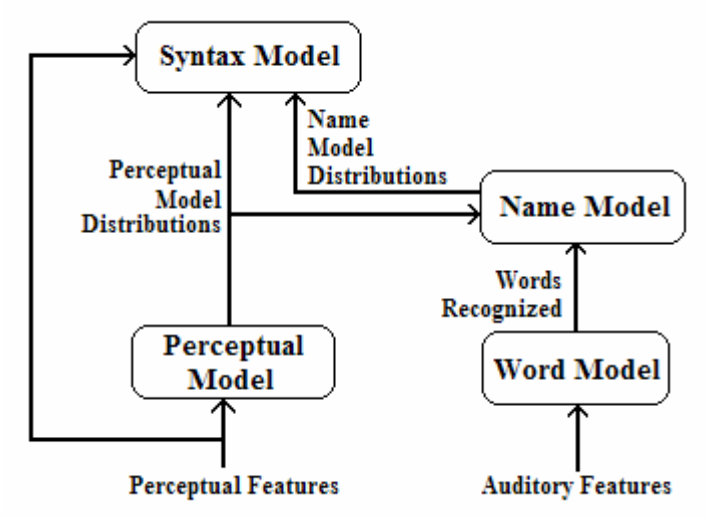

Figure 3.2: Cognitive model for syntax acquisition

The names of the models from Figure 3.1 have changed. The "visual model" has been changed to the more general "perceptual model," to include sensorimotor inputs from different modalities. This model will receive one feature vector for each object in the robot's visual field. The syntax model will receive information about the learned perceptual associations from this model. The "auditory model" has been renamed the "word model," since its role is to identify words in the audio input channel. The "concept model" has been changed to the "name model" to reflect that fact that it is used to name objects and actions identified by the perceptual model. In this cognitive model, the name model provides the syntax model with information regarding the perceptual representations that are associated with words recognized by the word model.

3.2.2 Syntax model

The purpose of the syntax model is to learn lexical categorization, the sequence of these categories in the production rule, and the compositional semantics. This information will be inferred from hearing noun-verb sentences that describe what the robot is observing.

3.2.2.1 Representation of syntactic information

The information to be learned in acquiring the production rule S→NV is the lexical categorization, the ordering of the lexical categories, and the compositional semantics. As stated in Section 1.3.2, the lexical categories will be defined in terms of the robot's sensori-motor experience. This means that words in the N (noun) category

will describe the features of the objects in the robot's environment, and words in the V (verb) category will describe features of the objects' actions. This information will be represented as the probability that each of the words in the two-word sentence describe each of the perceptual features. For this purpose, the perceptual features will be grouped into perceptual feature modalities. For example, in the robotic implementation, color is described by several color histogram bins, each of which is a single perceptual feature. These color features will be grouped into the color perceptual modality. Thus, the lexical categorization information will be the probability that each word in the two-word sentence describes each of the perceptual modalities. In this formulation, the following constraints will be imposed. First, each word must describe at least one perceptual modality, since the words uttered to describe the world cannot be meaningless. Second, both words will not be allowed to describe the same perceptual modality. For example, both words cannot describe the color of what the robot is observing. It is noted that while this constraint applies to noun-verb sentences, it may not apply in general to natural language. The ordering of the lexical categories will be represented by a probability matrix of each of the categories (including the delimiter) being followed by each of the categories.

The compositional semantics information to be learned for the S→NV production rule is that both words in a noun-verb sentence describe information about the same object. This will be represented in the syntax model by the probability that the features described by each word correspond to the same object that the robot is observing. In the robotic implementation, all of an object's perceptual features are stored in a single vector. Thus, the robot must learn whether the two words it hears describe perceptual features from the same feature vector or from different feature vectors. In the case where the robot is only observing one object, determining this will be trivial. When more than one object is present, this will have to be inferred. In this case, another constraint will be imposed that each word can only describe perceptual modalities from the same perceptual feature vector. This constraint does not apply to language in general (in the sentence "The cat plays with the ball," the word "plays" has something to do with both the cat and ball). However, for the S→NV production rule, this constraint is appropriate. It is noted here that there is more compositional semantics associated with this production rule than

is used here. For example, the robot will assume that the two-word sentences that it hears describe what it is observing, which would distinguish the noun-verb sentence from a verb-noun sentence which expresses a command. It is hoped that the proposed representation of compositional semantics in this thesis can be easily expanded for additional types of information. For these reasons, the main focus of this dissertation is the learning of lexical categories and their ordering.

3.2.2.2 Computation of syntactic parameters

During training of the syntax model, the cognitive model will work as follows. Each of the words uttered will be identified by the word model and outputted to the name model. Based on this information, the name model will provide the syntax model the representations in the perceptual model that correspond to the identified words as well as information about the word order. The syntax model will then infer which of the perceptual representations is most likely to describe each of the observed perceptual modalities. The compositional semantics will be inferred by determining whether the two words generated features corresponding to the same object or different objects. An illustration of this process is given in Figures 3.3 and 3.4. In this example, the robot has already learned perceptual representations of the words "ball," "cat," "stay," and "move." The arrows connecting the word, name, and perceptual models represent these learned associations. A ball is observed moving and a cat is observed remaining still, and the two-word sentence "ball move" is heard. As shown in the figures, the perceptual modalities that are best described by the "ball" perceptual representation are shape and color and the modality that is best described by "move" is the change in position. The compositional semantics information is then inferred from the fact that both words describe aspects of the same object in the perceptual scene.

3.2.2.3 Proposed implementation of the syntax model

The proposed implementation of the syntax model is an HMM. In this HMM, the states will represent the lexical categories and the state transitions will represent their sequence. As shown by Cave and Neuwirth [31], HMMs can learn both categorization and sequence information simultaneously, and so they are well suited to represent lexical categorization information syntactic information. The lexical categories will be determined by the semantics of the words, and so the states will be defined by the

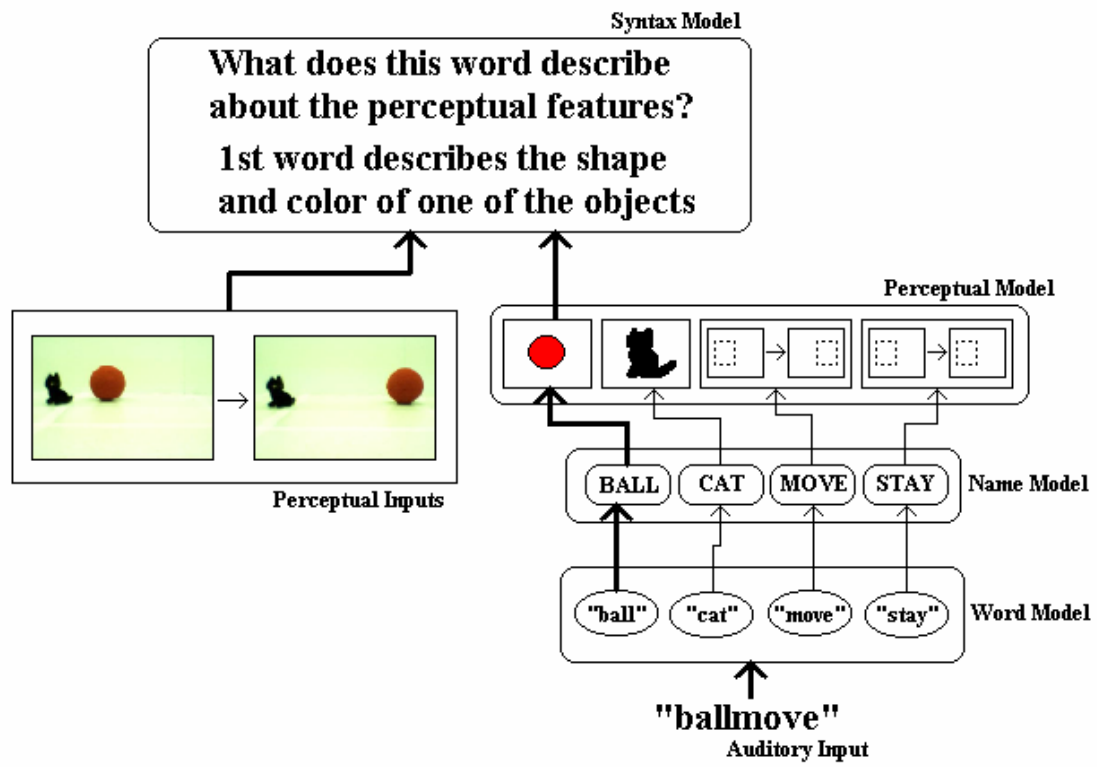

Figure 3.3: Estimation of lexical category information for first word

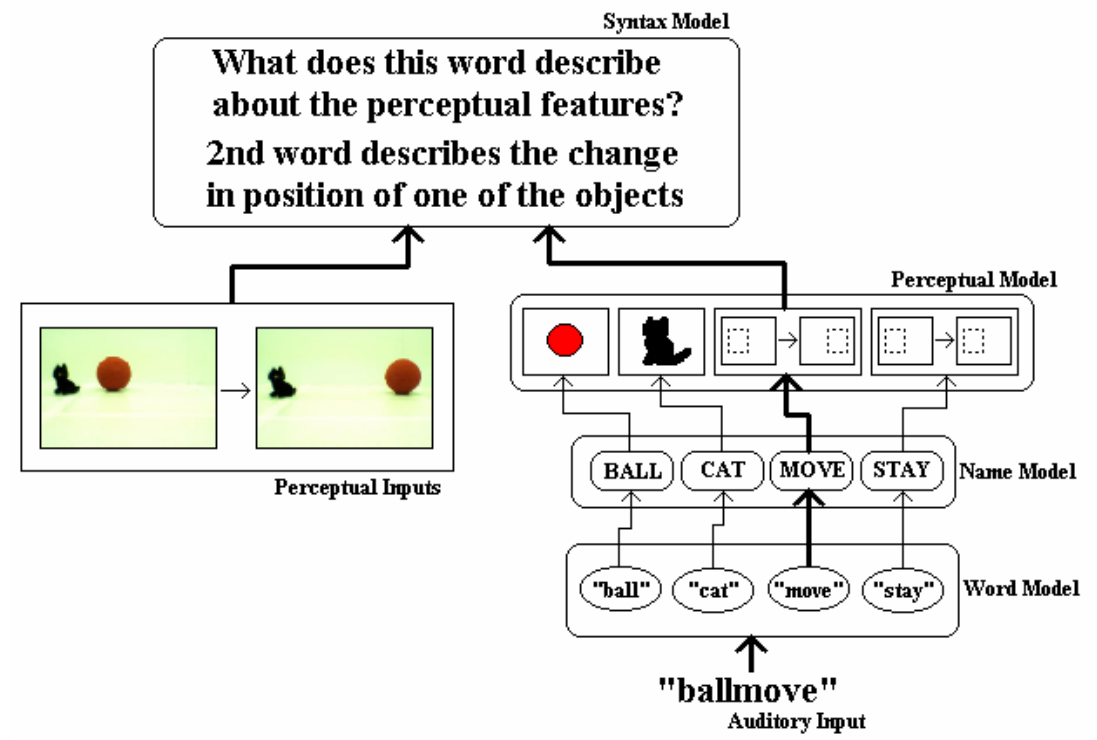

Figure 3.4: Estimation of lexical category information for second word

perceptual feature modalities that are described by them. Thus, the observable outputs of the states will be the probability that the state describes each of the perceptual modalities. As well, the HMM will learn the ordering of the lexical categories by learning the state transition probabilities.

To represent the S→NV production rule, three states will be used, with the intent that each state in the model will converge to one of S, N, and V. The S state should represent a delimiter that signifies that either no sentences have been heard yet or that the end of a sentence has been heard and will be meaningless otherwise. The N state should have a high probability to describe features that correspond to object identification, and the V state should have a high probability to describe the action modalities features. As well, the state transition matrix of the HMM will be initialized to be ergodic, with the intention that it will converge to a left-to-right model.

The compositional semantics will be represented by a parameter outside of the HMM. The justification for this is that the compositional semantics is information about the entire production rule and not about any individual lexical category. Thus, the compositional semantics parameter will be a single number that reflects the probability that the two words heard describe perceptual modalities from the same object. The cognitive model will receive the perceptual features from each object as a feature vector. Thus, the compositional semantics parameter will be estimated according to whether the two words that the robot hears describe features in the same perceptual feature vectors. When there is only one object present, this decision will be trivial, but when this is not the case, this will need to be inferred.

3.2.2.4 Syntactic parameter estimation

The HMM will be trained by hearing two-word sentences that fit the S→NV construction and describe something that the robot is observing. The parameters that represent the syntax will be determined by maximizing the likelihood of the syntax parameters, the perceptual representations associated with the two words spoken, and the robot's perceptual inputs. This likelihood function will be the sum of the likelihoods of all possible interpretations of how the two words spoken describe the perceptual information. These various interpretations, or hypotheses, determine the lexical categorization and compositional semantics information.

Each lexical categorization hypothesis will be represented by a binary vector whose length is the number of perceptual modalities. The value of an element in the vector gives the word (position 0 or 1) that is hypothesized to describe the perceptual modality that corresponds to the vector's index where that element exists. For example, hypothesis vector [1,0,1] means that the word in position 1 describes the first and third perceptual modalities and the word in position 0 describes the second perceptual modality. With the constraint that each word must describe at least one perceptual modality, the number of lexical categorization hypotheses is 2^n-2, where n is the number of perceptual modalities. In the likelihood function, the lexical categorization parameters will be represented by a matrix $\boldsymbol{\alpha}$, where α_{ij} is the probability that the perceptual representation of word j describes perceptual modality i. Thus, a lexical categorization hypothesis h will correspond to a product of α_{ij} values where $j = j_h(i)$ represents the i^{th} element of the lexical categorization hypothesis vector j_h.

The compositional semantics parameters in the likelihood function will be a matrix of values $\boldsymbol{\beta}$, where β_{xy} is the probability that the word 0 generates perceptual features in vector x and word 1 generates features in vector y. The compositional semantics parameter to be learned by the syntax model is β_{syn}, which represents the probability that the two words heard describe perceptual information about the same object.

The likelihood function is:

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{v}, \vec{w}) = \sum_{h=0}^{2^n-2} \sum_{x=1}^{m} \sum_{y=1}^{m} \beta_{xy} \prod_{i=1}^{n} \alpha_{ij_h(i)} [\prod_{j_h(i)=0} P(v_{xi} \mid w_0)][\prod_{j_h(i)=1} P(v_{yi} \mid w_1)] \qquad (3.1)$$

where m is the number of objects (perceptual feature vectors) present, n is the number of possible feature modalities, w_a is word a ($a \in [0,1]$) and corresponds to the word's position in the two-word sentence, v_{xi} is perceptual modality i in feature vector x, and $P(v_{xi} \mid w_k)$ is the probability that the perceptual representation of word k describes feature modality i in feature vector x.

The intuition behind this likelihood function is that it computes all of the possible ways that the two words spoken describe what the robot is observing. The probabilities in Equation (3.1) need to be computed from the distributions learned by the name and

perceptual models using Bayes' rule, as shown in Equation (3.2), where v_{xi} is perceptual modality i in feature vector x, w_k is word k recognized by the word model, n_i is state i of the name model, and p_j is state j of the perceptual model.

$$\mathrm{P}(v_{xi} \mid w_k) = \sum_i \sum_j \mathrm{P}(n_i \mid w_k)\mathrm{P}(p_j \mid n_i)\mathrm{P}(v_{xi} \mid p_j) \tag{3.2}$$

In general, a perceptual modality may represent more than one perceptual feature. Thus, $\mathrm{P}(v_{xi} \mid p_j)$ will be computed by:

$$\mathrm{P}(v_{xi} \mid p_j) = \prod_{k=0}^{d_i} \mathrm{P}(v_{xi}(k) \mid p_j) \tag{3.3}$$

where d_i is the number of perceptual features in modality i and $v_{xi}(k)$ is feature k of modality i in feature vector x.

To learn the syntax, the parameter matrices α and β will be estimated in an expectation-maximization (E-M) fashion, starting with the estimation of the α values. To estimate the values for α, Equation (3.1) will me maximized with respect to the lexical categorization hypotheses with the constraint that for each perceptual modality i:

$$\sum_j \alpha_{ij} = 1 \tag{3.4}$$

Equation (3.5) will be used to compute this maximization. Here, the previously learned value of β_{syn} will be used to create the matrix β by making the values on the diagonal of β equal to β_{syn} and the values off of the diagonal equal to 1-β_{syn}. The values of α that correspond to the hypothesis $j_{\hat{h}}$ are estimated to be 1, and the rest of the values in α are estimated to be 0.

$$\hat{h} = \arg\max_h \left(\sum_{x=1}^{m}\sum_{y=1}^{m} \beta_{xy}\left[\prod_{j_h(i)=0} \mathrm{P}(v_{xi} \mid w_0)\right]\left[\prod_{j_h(i)=1} \mathrm{P}(v_{yi} \mid w_1)\right]\right) \tag{3.5}$$

The syntax HMM is then trained with these estimates. This training takes three iterations, one for the each word, using the corresponding α_{ij} values (i = 0 for the first word and i = 1 for the second), and a third to mark the end of the utterance. The syntax feature vector used in the third iteration contains all zeros because there are no words to generate any of the perceptual features. Thus, the HMM will first be trained with the vector of estimates that correspond to the features described by the first word, [α_{00} α_{10} …

α_{n-10}], then the estimates that correspond to the second word, [α_{01} α_{11} … α_{n-11}], and finally with [0 0 … 0]. For example, let $i = 0$ represent the color modality in α_{ij}, $i = 1$ represent the shape modality, and $i = 2$ represent the change in position modality. In the "ball move" example shown in Figures 3.1 and 3.2, the correct hypothesis is j_h = [0 0 1]. Then, the first training iteration of the HMM will be [1 1 0], the second training iteration will be [0 0 1], and the third will be [0 0 0].

To estimate the values of β_{syn}, Equation (3.1) will be maximized with respect to β with the constraint:

$$\sum_x \sum_y \beta_{xy} = 1 \tag{3.6}$$

Equation (3.7) will be used to compute this, using the learned values of α from the syntax model.

$$\hat{x}, \hat{y} = \arg\max_{x,y} \left(\sum_{h=1}^{2^n-2} \prod_{i=1}^{n} \alpha_{ij_h(i)} \left[\prod_{j_h(i)=0} \mathrm{P}(v_{xi} \mid w_0) \right] \left[\prod_{j_h(i)=1} \mathrm{P}(v_{yi} \mid w_1) \right] \right) \tag{3.7}$$

Since the learned α values are associated with the states of the syntax HMM and the sequence of the states is still unknown, the values for α in Equation (3.7) must be estimated using the state probabilities from the training of HMM described above. Equation (3.8) shows how these values are computed.

$$\alpha_{ij} = \sum_{k=0}^{2} \alpha_{ik} \mathrm{P}(\mathrm{state}(t) = k \mid \mathrm{word}(t) = j) \tag{3.8}$$

The syntax model parameter β_{syn} will be trained by first computing β_{est} according to Equation (3.9):

$$\beta_{est} = \begin{Bmatrix} 1, \text{if } \hat{x} = \hat{y} \\ 0, \text{else} \end{Bmatrix} \tag{3.9}$$

This estimated value will then be used to update the value of β_{syn} using the gradient update in Equation (3.10), where ${\beta_{syn}}^k$ is the estimated value of β_{syn} after the k^{th} iteration and ε is the learning rate.

$${\beta_{syn}}^{k+1} = {\beta_{syn}}^k + \varepsilon(\beta_{est} - {\beta_{syn}}^k) \tag{3.10}$$

3.2.2.5 Sentence production using the learned syntax model

As mentioned in Section 3.1, the robot will be required to generate two-word sentences that describe what it observes, using what it has learned about syntax as well as the previously learned information in the perceptual, name, and word models. The words will be generated by initializing the state of the HMM to the most likely state after the delimiter state. Then, the word that best describes the information to be conveyed by the current HMM state will be determined using Equation (3.11), where v_{xi} is perceptual modality i in feature vector x, w_k is word k of the word model, $\alpha_s(i)$ is the probability that the current HMM state describes perceptual modality i. The probabilities in this equation are calculated according to Equations (3.2) and (3.3). In the case that the robot is observing more than one object, the robot will decide which feature vector will be described by the first word, reflected by the variable x in Equation (3.11).

$$\arg\max_k \sum_{i=1}^{n} \alpha_s(i) \mathrm{P}(v_{xi} \mid w_k) \tag{3.11}$$

After this word is chosen, the next state is determined by choosing the most likely next state in the syntax HMM, and the word to output is chosen by Equation (3.12). Here, the compositional semantics information is included, where x is same value as in Equation (3.11). Here, β is determined using β_{syn} using the same method used for Equation (3.4).

$$\arg\max_k \sum_{y=1}^{m} \beta_{xy} \sum_{i=1}^{n} \alpha_s(i) \mathrm{P}(v_{yi} \mid w_k) \tag{3.12}$$

This is repeated until the delimiter state is detected. A state is determined to be the delimiter state if each of the $\alpha_s(i)$ values for that state are less than 0.5.

3.3 Numerical Simulations

In order to test the proposed solution, numerical simulations were performed using Matlab. Matlab functions implemented by Squire [8] for HMMs, including the RMLE training algorithm, were used. The name, word, and perceptual models are all assumed to have been previously learned.

In the simulations, the number of perceptual modalities is set to three, each containing a single perceptual feature. In the first simulation, the first and third

perceptual modalities will be described by the first word (the "noun") and the second perceptual modality is described by the second word (the "verb"). In the second simulation, the first word describes the first modality, the second word describes the second modality, and the third modality is not described by either. The number of words is set to four, corresponding to two nouns and two verbs.

The likelihood function used in the simulations includes a bias factor for the compositional semantics. When the α parameters were estimated, this bias was uniform. When the β parameters were estimated, the likelihoods of the parameters on the matrix diagonal were biased by β_{syn} while the likelihoods of the parameters not on the matrix diagonal were biased by $1-\beta_{syn}$. This biasing was necessary for the simulated syntax model to consistently learn the compositional semantics parameter. However, this bias was not needed in the robotic experiments described in Chapter 4, and so was omitted from Equations (3.1), (3.4), and (3.6).

3.3.1 Simulation procedure

In each of the simulations, the syntax model is trained as follows. First, the number of feature vectors generated is chosen randomly to be one or two, representing one or two objects with corresponding actions. To generate the perceptual features, the object and its corresponding actions are chosen from the two available objects and actions. When two feature vectors are used, the choices are constrained so that the two objects chosen are not the same. From the chosen objects and actions, the feature vectors are generated from gaussian distributions. The pairs of objects and actions generate the features in the modalities they describe using means of 0 and 1 and a variance of 1. In the second simulation, where the third perceptual modality is generated randomly, a gaussian distribution with mean 0.5 and variance 10 is used. The words corresponding to the objects and actions are always generated correctly, and the word and name models are deterministic, so that the correct perceptual model is chosen to be associated with the words.

The perceptual model used in the simulation has the same distributions as above for the modalities that the objects and actions describe. The distributions used for the modalities that are not described by a perceptual representation (for example, the

distribution of a verb feature in the representation of a noun) have a mean of 0.5 and variance of 10.

3.3.2 Simulation results

The following results are obtained using the procedure described above with the following initial conditions. The syntax model's HMM state transition matrix is set to be ergodic. Two of the parameters of the HMM that represent the lexical categorization are set to have a slight bias to be generated by one of the states, using a different state for each parameter. These slight biases in the HMM's initial conditions are necessary for it to converge to a non-uniform model. The initial β parameter value in the syntax model is set to 0.5, to represent no prior information about the compositional semantics.

Both of the simulations described in Section 3.3.1 were successful. Figures 3.5 through 3.9 show the results of the simulation where the first and third perceptual modalities are generated by the first word (the "noun") and the second modality is generated by the second word (the "verb").

Figure 3.5 shows that the model learned is a left-to-right model, which is the appropriate form for representing the S→NV production rule. Figures 3.6 through 3.8 show that the first state in the model does not describe any of the perceptual modalities, the second state describes the first and third modalities, and the third state describes the second perceptual modality. This reflects a correct learning of the lexical categorization. The compositional semantics parameter correctly learned that the two words in the noun-verb sentence describe the same object, as seen in Figure 3.9.

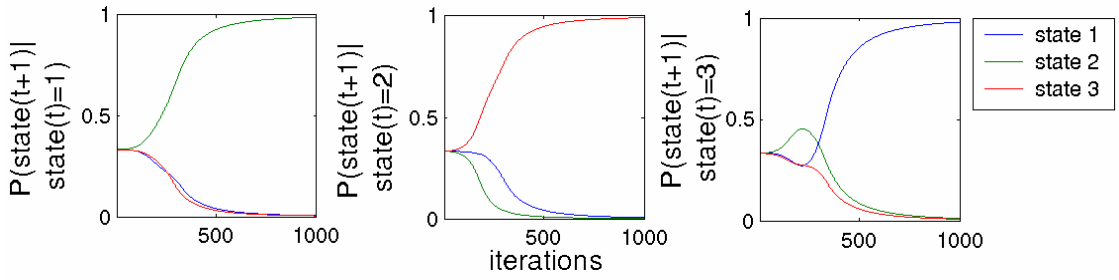

Figure 3.5: State transition parameters for the syntax HMM, first simulation

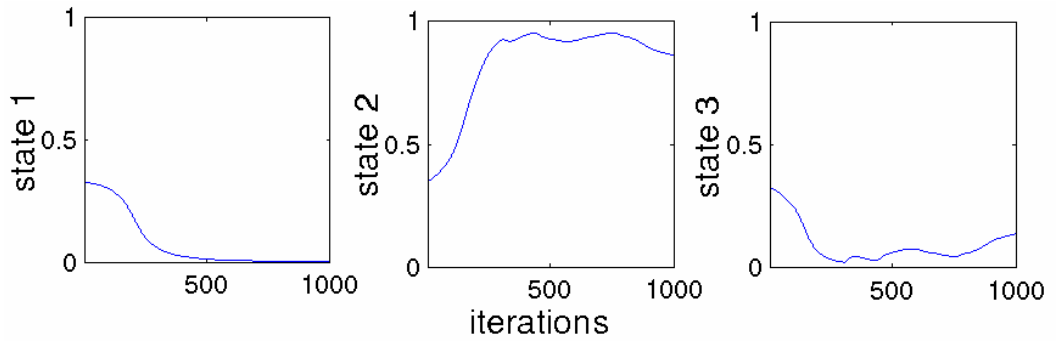

Figure 3.6: Probability that each state describes the first perceptual modality, first simulation

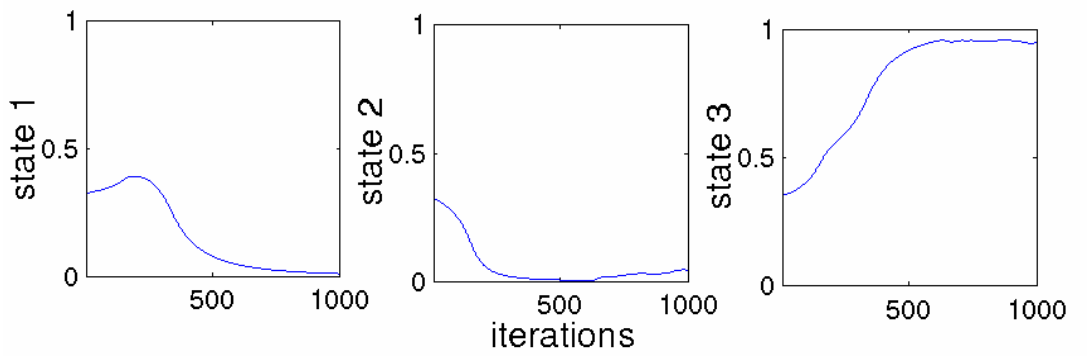

Figure 3.7: Probability that each state describes the second perceptual modality, first simulation

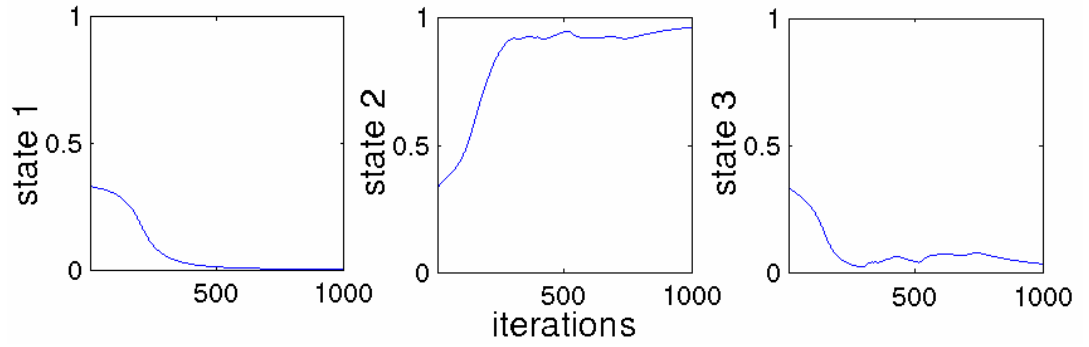

Figure 3.8: Probability that each state describes the third perceptual modality, firsr simulation

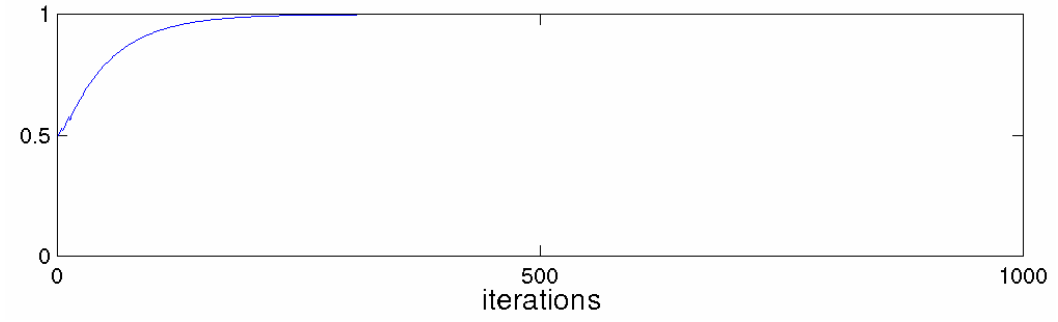

Figure 3.9: Probability that both words describe modalities in the same perceptual feature vector, first simulation

The results of the second simulation are shown in Figures 3.10 through 3.14. Again, the sequence of the lexical categories is correctly learned to be a left-to-right sequence. The first and second words generate the first and second perceptual modalities, respectively, and the third modality is generated randomly. This is how these modalities were generated in the experiment, and so the lexical categorization was correctly learned.

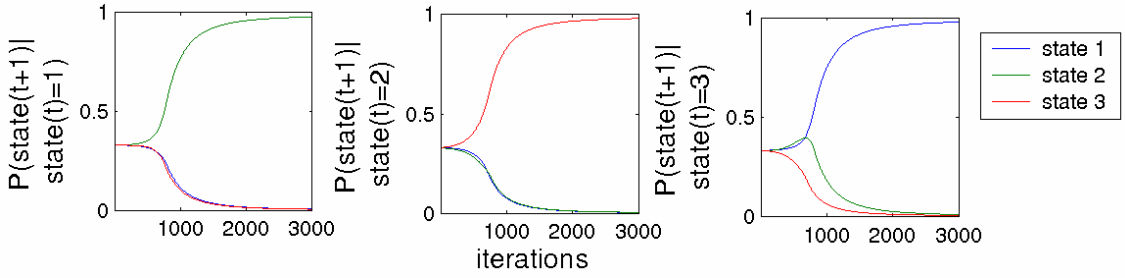

Figure 3.10: State transition parameters for syntax HMM, second simulation

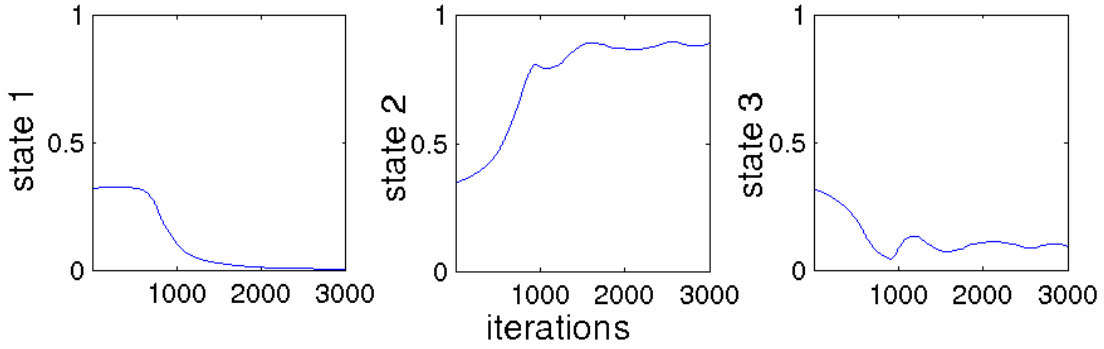

Figure 3.11: Probability that each state describes the first perceptual modality, second simulation

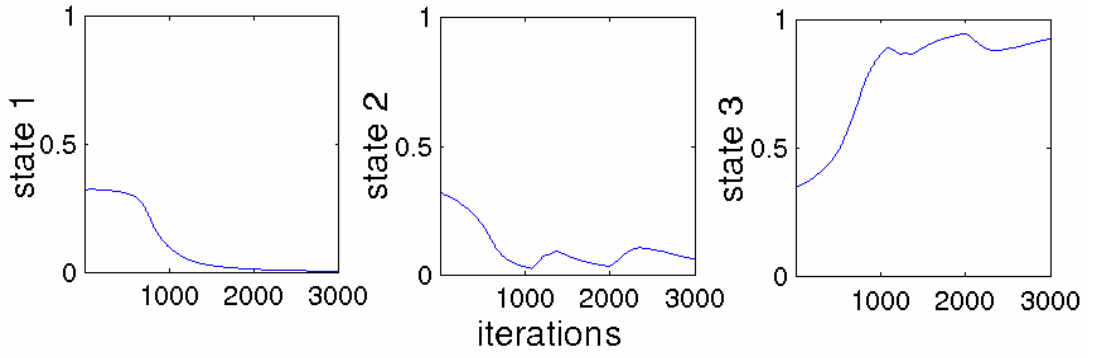

Figure 3.12: Probability that each state describes the second perceptual modality, second simulation

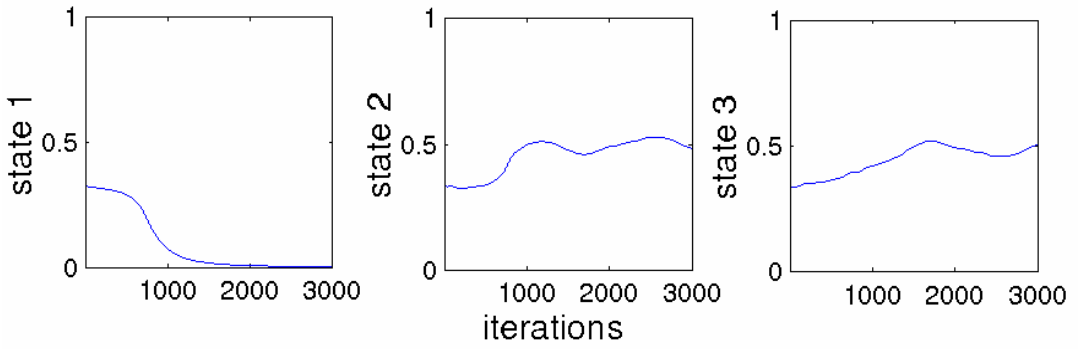

Figure 3.13: Probability that each state describes the third perceptual modality, second simulation

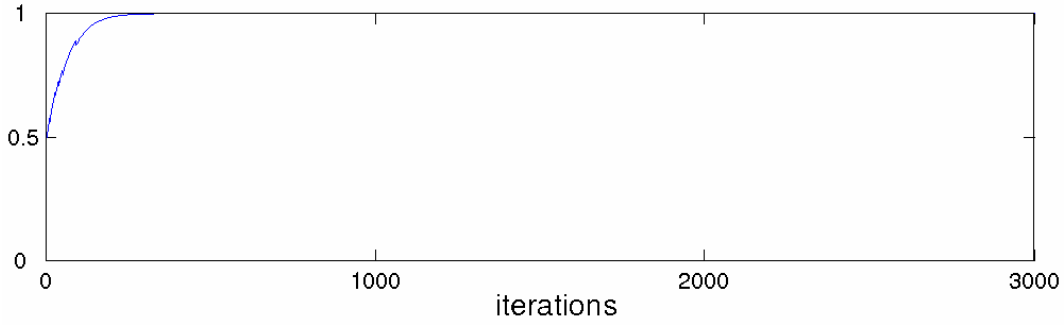

Figure 3.14: Probability that both words describe modalities in the same perceptual feature vector, second simulation

3.4 Summary

This chapter presents a novel solution to the problem of machine syntax acquisition. The experimental results described in Section 3.3.2 show that this solution produces the desired outcome in a numerical simulation, and so has the potential to succeed when implemented into LAR-UIUC's robotic framework. The next chapter describes the proposed robotic experiments for this thesis.

CHAPTER 4 ROBOTIC IMPLEMENTATION

The cognitive models and algorithms for syntax learning developed in Chapter 3 will be implemented in LAR-UIUC's robotic framework. As stated in Section 3.2.2, the robot will be trained by hearing example noun-verb sentences that describe what the robot observes. The following section gives a more detailed account of how this training will occur, as well as experiments in which the robot demonstrates its learned knowledge of syntax.

4.1 Robotic Environment

The robotic environment for this work is an approximately 6' by 5' pen, with white walls and a white floor. The robot used in all of the experiments is Illy, shown in Figure 1.2. The objects used in this experiment are a toy dog, a toy cat, and a soda can, which can be seen in Figure 4.1. To create the objects' actions that will be described by the verbs, the robot will navigate to an object and then choose one of three actions to perform: move forward 6 in, move forward 18 in, or raise the front gripper. After performing this action, the robot will perform a counter-action: move backward 6 in, move backward 18 in, or lower the front gripper, respectively, and observe the effect on the object. Each of the robot's cognitive models will be trained using these objects and actions.

Figure 4.1: The objects used in the experiment

It is noted here that the robotic environment described here constitutes a simplified version of the world. While ideally, language learning should occur in the real world, this is simply not yet feasible with LAR-UIUC's current robotic implementation. As stated in Section 1.2.2, these simplifications are necessary in order to study higher language function. Even in the robotic environment used in these experiments, the learning tasks proposed are not trivial.

4.1.1 Perceptual feature set

The perceptual modalities that the robot will use to represent the nouns and verbs are color, shape, the robot's action, and the object's change in position. The color and shape features are computed using the visual processing program developed by Lin [11]. This program uses loopy belief propagation to segment the foreground and background pixels in the images captured by the robot's camera. The foreground pixels are then grouped to find the objects in the visual scene. The color features are computed as the proportion of the object's pixels that are in a set of color histogram bins. The shape features are the ratio of the object's width to height and a shape moment, which measures the average squared distance of the object's pixels from the object's center.

Nine total perceptual features are used in this experiment, divided amongst the perceptual modalities as follows. The color modality used for the syntax experiment uses five features, which are color histogram bin populations, as described above. The shape modality contains two features which are the ratio and moment features computed by the visual processing program. It is noted here that the shape parameters do not constitute a three-dimensional representation of the objects, the repercussion of which will be discussed further in Chapter 5. The robot's action modality will be a single feature that is a ternary number, representing one of the three possible actions described in Section 4.1. The object's change in position will also be a single feature, computed as the difference in the object's position in the robot's visual field before the chosen action is taken and after the counter-action is taken. The object's bottom-left corner in the right camera is recorded as the object's position. The change in position is calculated as the L1 norm of the object's x- and y-position before and after the actions, with the y-dimension weighted by a factor of 2. The extra weight is given to the y-dimension because the object's change in position is induced by the robot pushing the object forward, which will be

reflected in change in the object's *y*-dimension position. In the case that the object leaves the robot's field of vision after the counter-action, a large number is chosen for the change in position measurement. This number was chosen to be larger than any possible measurement if the object remained in the robot's field of vision.

4.1.2 Word set

The set of words used for the experiment are "kitten," "puppy," "can," "stay," "move," and "gone." The word "kitten" is used to name the toy cat, "puppy" is used to name the toy dog, and "can" is used to name the soda can. The word "stay" is used to describe no change in position of the object, "move" describes the case when the object's position changes and remains in the robot's field of vision, and "gone" describes a large change in position or the case where the object leaves the robot's field of vision. These words were chosen to be acoustically different so that the word recognition would be robust. For example, if the word "cat" was chosen for the toy cat, it would be easily confused with the word "can."

4.2 Cognitive Model Implementation

Sections 4.2.1 through 4.2.4 describe the specific implementation of the cognitive models in Figure 3.2.

4.2.1 Word model

The word model used in this experiment uses log area ratios to represent the spectral information of each segment of the speech signal. The programs written by Kleffner [10] provide are used, which provide feature extraction and speech synthesis functions. The speech signal is recorded at 22 kHz, and computes eight log area ratios on 300 ms windows of speech every 100 ms. Single word recognition is performed by computing the distance between a test and example word using the method of dynamic time warping, originally used by Vintsyuk [33], where the distance between two segments of speech is the L2 norm of the LARs of the two segments. The word with the minimum distance is chosen as the recognized word. Two word sentences are recognized using a similar method to single word recognition, except that the test utterance is compared against all concatenated combinations of the single words.

4.2.2 Perceptual model

An HMM with six states is used for the robot's perceptual model. The states of the HMM are meant to represent each of the perceptual events (three objects and three actions) in the robot's environment. The observable outputs of these states are the nine perceptual features described in Section 4.1.1, each modeled with discrete distributions. For the shape and color features, which are noninteger values, seven histogram bins are used for the discrete distribution outputs. These features are quantized so that each bin has an equal (or as close to equal as possible) population amongst the training samples. The action and change in position features are modeled with discrete distributions containing three bins. For the action feature, there is simply one bin for each possible action. The change in position feature is quantized so that each bin has an equal (or as close to equal as possible) population amongst the training samples. The state transition matrix of this HMM is set to be uniform, since the perceptual model is only used as a perceptual classifier where the sequence information is irrelevant.

4.2.3 Name model

The name model is implemented by an HMM with six states (one for each object/action and the corresponding words that name them). This model uses seven observable outputs, which are modeled by discrete distributions with six possible outcomes. One of these outputs is the word recognized by the word model, in which each outcome is one of the six possible words. The remaining six outcomes are the probabilities of each of the states of the perceptual model. These are quantized into six bins such that each bin has an equal (or as close to equal as possible) population amongst the training samples. It is necessary to use the outputs of the perceptual model in this way, instead of using a single output which gives the perceptual state recognized because there is a bias toward recognizing the object states as opposed to the action states even though both are actually present in a single feature vector. This is due to the fact that the object states have more perceptual features that describe them (color and shape modalities, which constitute seven total features) than do the action states (robot action and change in position, which contain two total features). Thus, if the perceptual model is to classify a perceptual feature vector, it is much more likely to label it as the object that it is observing than the action. For example, the robot will learn the meaning of the

word "stay" in the context of each of the objects. If the name model only receives a decision from the perceptual model about which perceptual state was observed, it is likely that the state that describes the object will be chosen. Thus, the robot will learn that "stay" is likely to be a name for each of the objects. However, if the name model receives a probability of each of the perceptual model states, it can learn that the correct perceptual model state has a consistently high probability across all samples of the word "stay." As well, because the robot learns the word "stay" in the context of each of the objects, the average probability of each of the object states will be low.

The name model's learned distributions over the perceptual states will be used to compute the probability of each of the perceptual states given a name state. These probabilities are computed using Equation (4.1):

$$\mathrm{P}(p_x \mid n_y) = \sum_{i=0}^{5} b_y(x,i) * q(x,i) \tag{4.1}$$

where p_x is perceptual state x, n_y is name state y, $b_y(x,i)$ is the value in the i^{th} bin of name state y's distribution for perceptual state x, and $q(x,i)$ is the quantization lower bound of the i^{th} bin for perceptual state x's distributions in the name model.

4.2.4 Syntax model

The syntax model is implemented by an HMM, using three states and four observable features. The three states each represent a non-terminal symbol of formal grammar, as described in Section 3.2.2.3. The four observable features are the lexical categorization parameters, which represent the probabilities that the current state describes each of the perceptual feature modalities: shape, color, robot action, and change in position. Binary distributions will be used for these features, so that one symbol represents the probability that the current state does produce the perceptual feature modality and the other represents the probability that the current state does not.

Choosing three states for the HMM introduces some prior knowledge to the syntax learning task. While ideally this would not be necessary, some number of states needs to be chosen for the HMM. The syntax model assumes that it hears two-word sentences, and so the choice of three states is reasonable in order to represent each of the words as well as a delimiter.

4.3 Cognitive Model Training

The following section describes the training that was performed on the cognitive models used in the syntax experiments.

4.3.1 Perceptual and name model training

The perceptual and name models were trained by first initializing the HMM states to be close to the correct distributions and then training using the RMLE training algorithm developed by Squire [8]. The distributions for both of the models were estimated using 30 samples of each of the objects recorded by the robot. Within each of the 30 samples, the possible robot actions were performed 10 times each. Each of the objects' 30 samples was used to initialize a state of the perceptual model. Also, the 90 total samples were clustered into three groups using *k*-means on the change in position feature. These resulting groups were used to initialize the remaining three states of the perceptual model, as representations of "stay," "move," and "gone."

The name model was then initialized by computing the likelihood of each of the perceptual states for each of these 90 training samples, including a word model output for the seventh name model feature. The states of the name model were then initialized using the same method as for the perceptual model.

The training of the perceptual and name models was completed using the RMLE algorithm by randomly choosing from the set of 90 training samples. In this training, the perceptual model was first updated, and then the probabilities of each of the perceptual states for the current sample are used as inputs to the name model, along with a word model output. The word used was randomly chosen to describe either the object or the action. The model was trained in sets of 1000 randomly chosen samples. The criteria used to determine convergence was if the difference in values of the state transition matrices after the set of 1000 samples was below a threshold.

Figures 4.2 through 4.5 show the discrete distributions of the nine perceptual features for each of the states of the perceptual HMM. In the figures, although the states are labeled using their corresponding names, this is only for clarity and is not meant to indicate that the states are labeled a priori. Excepting the action feature, the bins have no relevant meaning, and so they are not explicitly labeled. It is important to note here how each of the perceptual features correlate with each of the perceptual states. As expected,

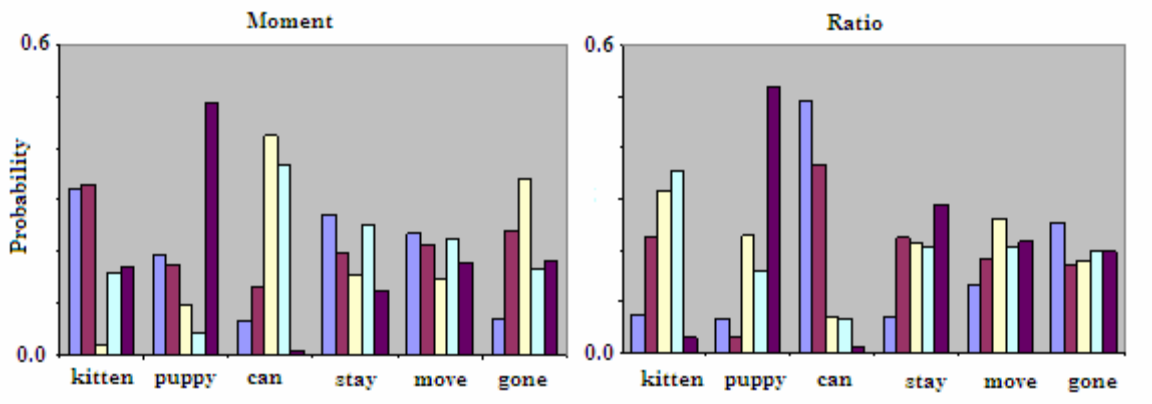

Figure 4.2: Distributions of the shape features in the perceptual model

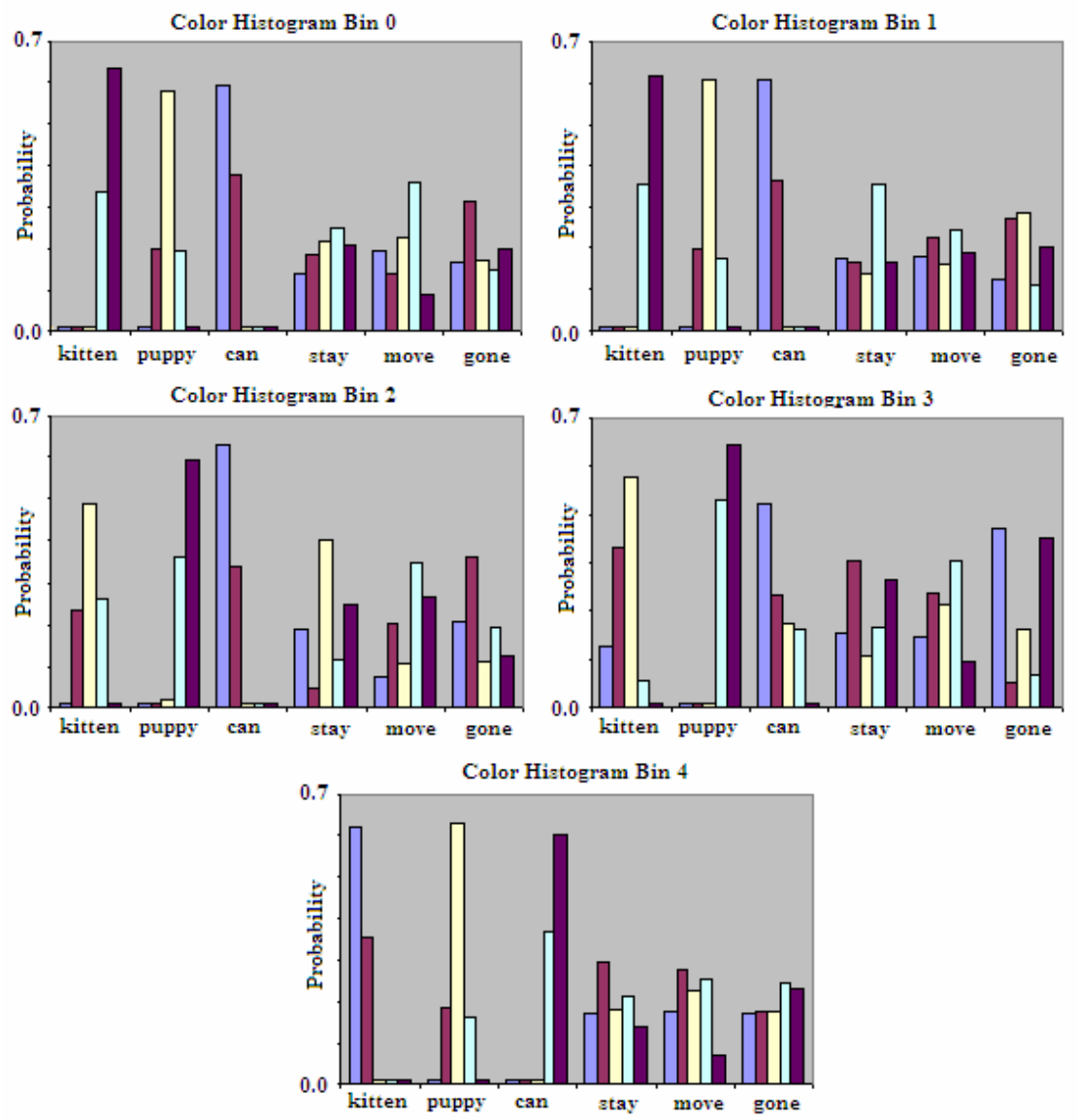

Figure 4.3: Distributions of the color features in the perceptual model

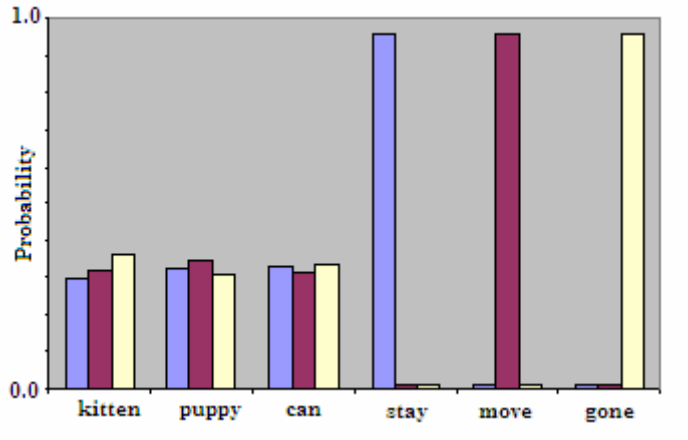

Figure 4.4: Distributions of the change in position feature in the perceptual model

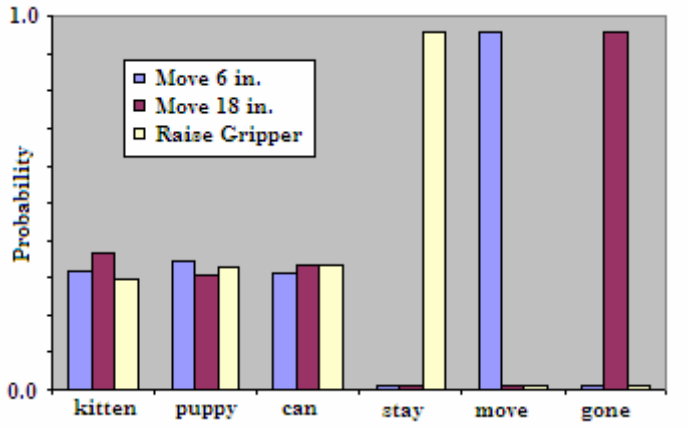

Figure 4.5: Distributions of the robot action feature in the perceptual model

the features that are not meant to be described by a state (for example, the shape features for the verb states and the action feature for the noun states) have high variance in their distributions. As well, the features that are meant to be described by a state (the shape feature the noun states and the action feature for the verb states) have low variance in their distributions. As a result, it is expected that the syntax model will be able to correctly infer which words describe which perceptual modalities. It is important to note, however, that the variances are not equal in all cases for the modalities that are meant to be described by the perceptual states. The variance of the change in position and action modality distributions for the verb perceptual states is much lower than the variance of the color and shape modality distributions for the noun perceptual states. Thus, it is expected that more errors will be made when inferring the lexical categorization parameters of the first word than with the second word in the noun-verb sentences.

Specifically, if low probability shape features for a noun state are encountered in which the verb states have higher probabilities, the syntax model may infer that the second word uttered in the two word sentence is meant to describe the shape features. However, since the models proposed here are meant to be robust, the syntax model should be able to learn the two-word syntax as long as the number of these errors is small.

Figures 4.6 and 4.7 show the distributions in the name model. Again, the states are labeled for clarity but these labels not meant to indicate that the states are labeled a priori. Figure 4.6 shows the probabilities of each of the perceptual HMM states for each of the name states. This figure shows the probability of each name state given a perceptual HMM state. These probabilities were computed from the name HMM's discrete distributions using equation 4.1. These values were then normalized so that the probability of each of the name states given a perceptual state input sums to one. Figure 4.6 shows the probability of each name state given the word recognized by the word model. These figures show that the name model correctly correlates the perceptual HMM states and the words that name them.

4.3.2 Word model training

The word model was trained with a single example of each word. The audio was recorded using the robot's on-board microphones instead of by a close-talk microphone that has traditionally been used by LAR-UIUC. This change as made in order to improve the aesthetics of the experiments, despite the degraded quality of the speech recorded from the robot's microphone. Used in this way, the accuracy of the single-word recognition is about 98%, and the accuracy of recognizing words in two-word sentences is approximately 90%. Most of the errors in recognition were due to difficulties in segmenting the speech in the presence of the background noise.

4.3.3 Syntax model training

The syntax model was trained only using the RMLE algorithm. The model's parameters were initialized so that the state transition matrix was ergodic. The lexical categorization parameters were initialized with a slight bias to prevent the model from converging to a non-uniform solution. The compositional semantics parameter was initialized to have no bias ($\beta_{syn}^{0} = 0.5$). For training, 60 samples were recorded of the robot performing an action on an object and hearing a noun-verb description by the

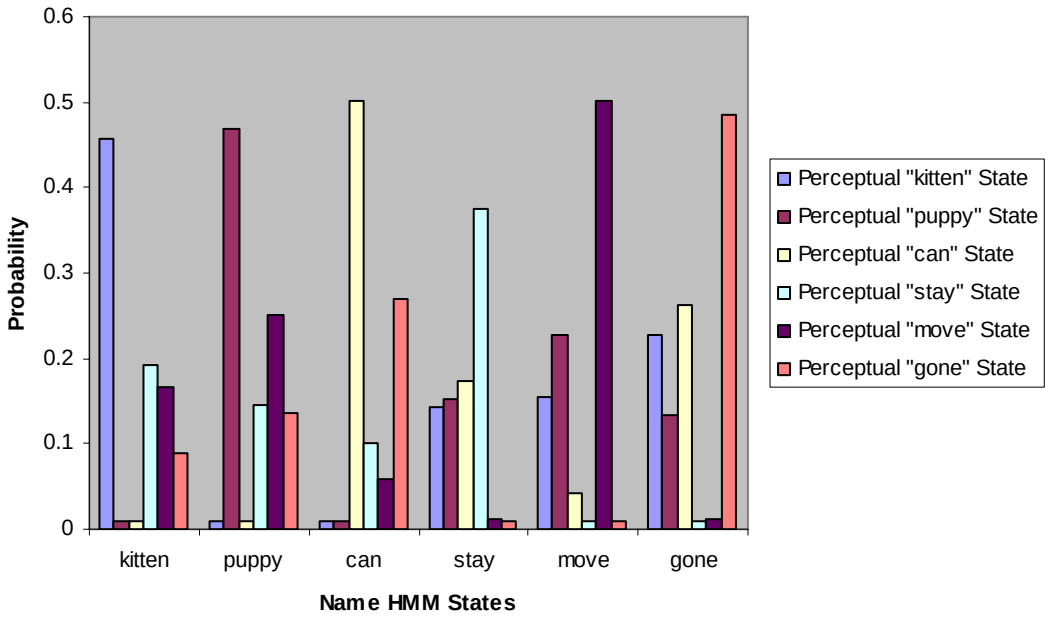

Figure 4.6: Distributions of the perceptual HMM state features in the name model

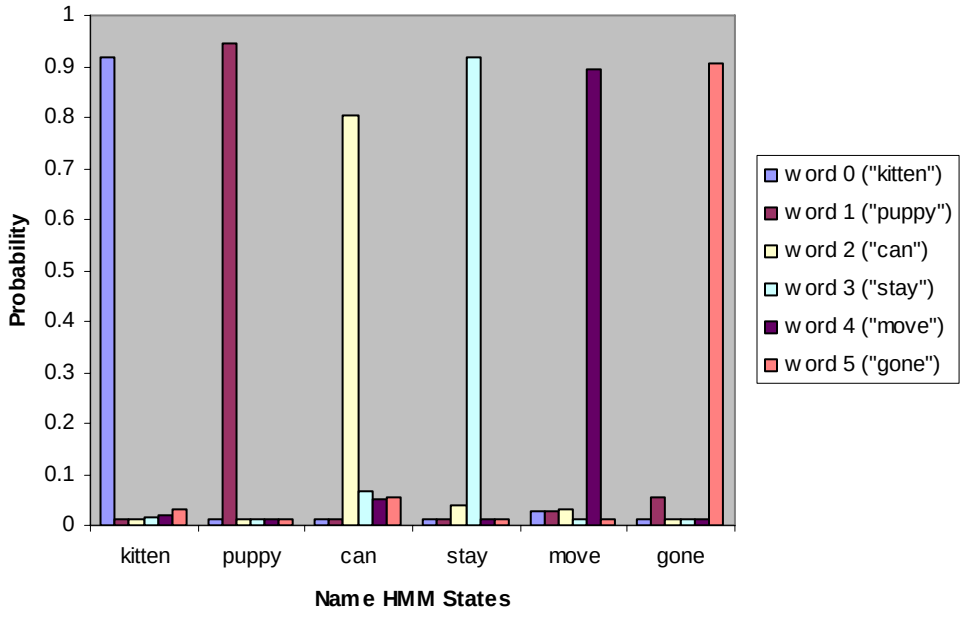

Figure 4.7: Distributions of the word feature in the name model

experimenter. These 60 samples comprised 30 trials with the toy kitten and 30 with the can. The toy dog was intentionally left out of the syntax training so that it could be used during testing to demonstrate the ability to produce novel sentences. Within each of the

groups of 30 trials, each of the actions was performed 10 times. The model was then trained with blocks of 1000 iterations, each using a randomly chosen sample from the group of 60. Thus, each training iteration corresponds to hearing one noun-verb sentence, which results in three training iterations of the syntax HMM. In each training iteration, a second feature vector had a 50% chance of being present, representing a second object that the robot is observing. If a second feature vector is present, it is randomly chosen from the set of training samples involving the object not present in first feature vector. For the RMLE algorithm, an exponentially decreasing learning rate was used where the initial value was 0.005, and the exponent was 0.25.

To determine convergence, the difference in all of the parameter values (state transition parameters, state output parameters, and compositional semantics parameter) from before and after the block of 1000 samples is computed. If the sum of these differences was below a threshold, the model was determined to have converged. The results of the syntax model training are shown in Chapter 5.

4.4 Syntax Model Testing

The following section describes the tests that were performed on the syntax model to determine its success.

4.4.1 Initial test

The syntax model learned by the robot was tested by using the model to produce two-word sentences that describe what the robot observes. Here, Equations (3.9) and (3.10) were used to produce the sentences. In this test, the robot chose an action to perform on an object (placed in front of it by the experimenter) and outputted a two-word sentence using the method described in Section 3.2.2.5. The results of this experiment are discussed in the next chapter. A video showing the syntax test can be found at *http://www.ifp.uiuc.edu/speech/acquisition/acquisition.html*, in the "Demonstration Videos" section, under "Semantic Based Learning of Syntax."

4.4.2 Demonstration

In order for the ideas developed in this thesis to be useful to LAR-UIUC, the syntax model was incorporated into the robot's autonomous exploration program. In this demonstration, the robot finds the objects through exploration using its perceptual model

or can be prompted by the experimenter to find a specific object. Figure 4.8 shows a diagram of the FSM used as the robot's controller in this experiment.

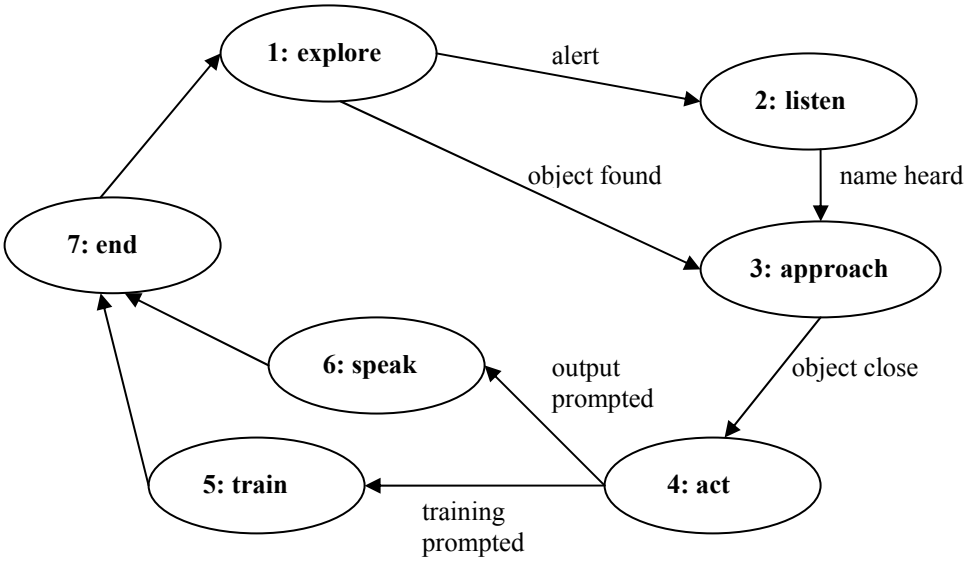

Figure 4.8: The FSM controller for the syntax demonstration

In the first state, the robot looks for objects in its environment based on the visual (color and shape) modalities. Once the robot finds an object of interest, the robot navigates toward the object in state 3. In state 1, the experimenter can also direct the robot to a specific object by first alerting the robot. When the robot hears this alert, the FSM will transition to state 2, and the robot will turn toward the experimenter using its sound source localization ability. In state 2, the robot listens for the name of an object. When a name is recognized, the FSM will transition to state 3, and the robot will look for an object that fits the corresponding perceptual state.

In state 3, the robot will navigate toward the object of interest, given that it is in the robot's visual field. The robot uses the object's *x*-position in the visual field as well as its stereo vision to guide the navigation. If the robot determines that the object is close to it (within about 6 in), the FSM will transition to state 4. If the object of interest is not seen in the visual field, the robot will search its environment for it. In the case that the robot cannot find the object of interest in a certain number of searching iterations, the FSM will return to state 1 (not shown in Figure 4.8).

In state 4, the robot will choose one of its predefined actions to perform on the object (move forward 6 in, move forward 18 in, or raise the front gripper), perform the action, and then perform the appropriate counteraction. After this, the experimenter prompts the robot to either output the two-word sentence that describes what the robot observed (state 6 of the FSM) or to listen for a two-word sentence for training (state 5 of the FSM).

In state 5, the robot listens for a two-word sentence, estimates the syntax parameters, and performs the parameter updates. In state 6, the robot produces the two-word sentence that describes what it observed. This sentence is produced using Equations (3.9) and (3.10). After each of these states, the FSM transitions to state 7, in which the robot backs up to prepare to search for another object and remembers the object that it just found so that the same object is not used as the next object of interest. Last, the FSM transitions back to state 1, the start state.

A video showing the syntax demonstration can be found at *http://www.ifp.uiuc.edu/speech/acquisition/acquisition.html*, in the "Demonstration Videos" section, under "Semantic Based Learning of Syntax."

4.5 Summary

This chapter describes the robotic implementation used in the syntax learning experiments in this thesis. The training of the name, word, and perceptual models was presented. The learned perceptual distributions are of significant importance, due to their influence on the syntax acquisition described in Chapter 5.

CHAPTER 5 ROBOTIC EXPERIMENT RESULTS AND DISCUSSION

5.1 Syntax Model Training Results

This section presents the results of the syntax HMM training as described in Section 4.3.3.

5.1.1 Lexical categorization parameters

Figures 5.1 through 5.4 show the training of the lexical categorization parameters. Although 5000 training iterations were used, the parameters reached their final values after fewer than 1000 training iterations. The final values of these parameters show that the syntax model has correctly learned a semantic definition of the two lexical categories used in this experiment: one state of the syntax model (state 0) is most likely to describe the color and shape features, another (state 1) is most likely to describe the change in position and action features. These two states can be interpreted as noun and verb lexical categories. As well, the remaining state does not generate any features, and so can be interpreted as a delimiter state.

These figures also show the expected errors that occurred in training. As noted in Section 4.3.1, the color and shape feature distributions in the noun states have higher variance than the change in position and action feature distributions in the verb states. As expected, this resulted in errors in training that are reflected in the noise and final values of state 0's parameters in Figures 5.1 and 5.2. In contrast, Figures 5.3 and 5.4 show that the parameters for state 1 reach the maximum value with less noise.

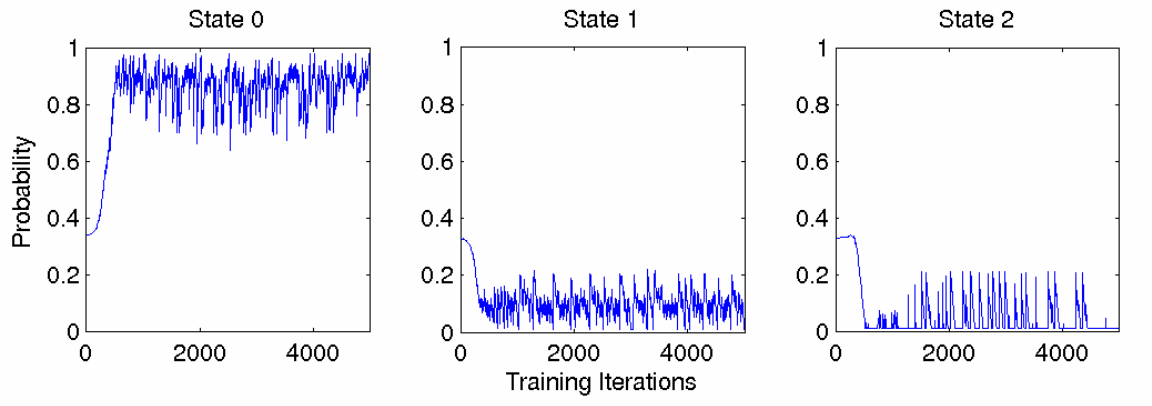

Figure 5.1: Probability that each syntax HMM state describes the color modality

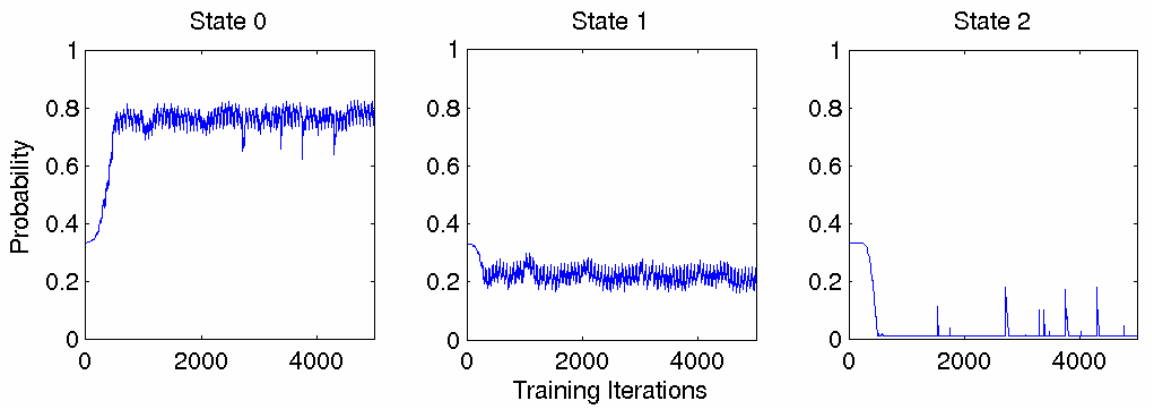

Figure 5.2: Probability that each syntax HMM state describes the shape modality

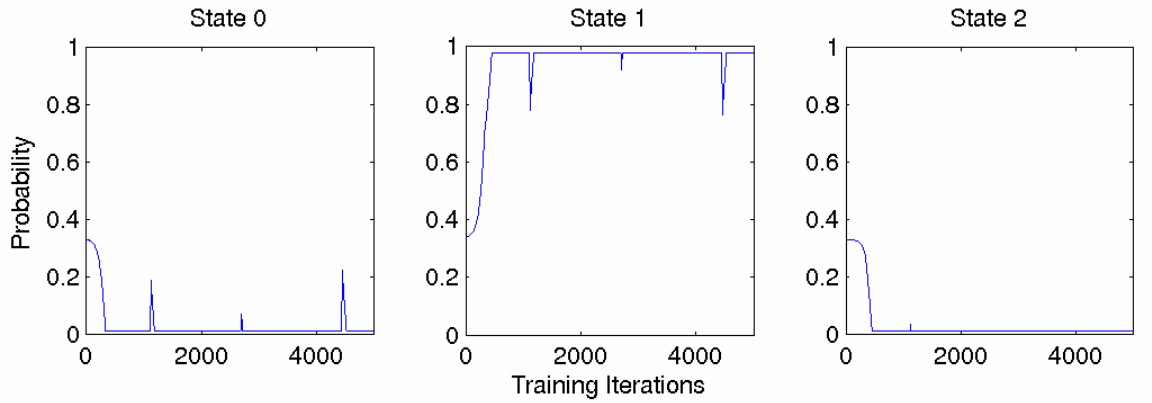

Figure 5.3: Probability that each syntax HMM state describes the change in position modality

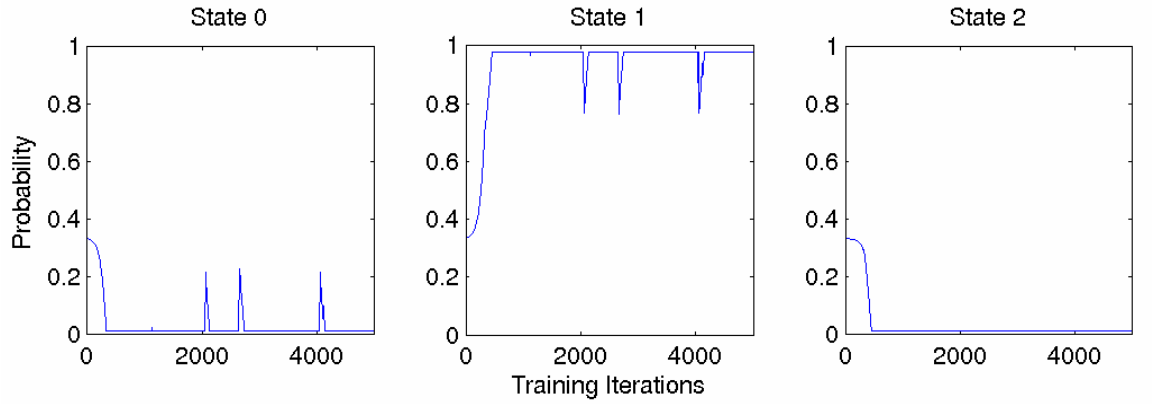

Figure 5.4: Probability that each syntax HMM state describes the robot action modality

During training, the estimates of the lexical categorization parameters for each training iteration were recorded. The counts of each of the hypotheses are shown in Table 5.1. Only hypotheses that were ever chosen are shown in this table. It can be seen that the correct hypothesis, [0 0 1 1], was chosen 71% of the time. Thus, despite the fact

that many errors were made, the model proposed here is robust enough to correctly learn the lexical categorization. This table also shows that most of the errors made were with inferring the lexical categorization of the shape and color parameters, as was expected (see Section 4.3.1).

Table 5.1: Lexical categorization hypothesis estimates

Hypothesized Word that Describes the Features				
Color	Shape	Change in Position	Action	Number of times chosen
1	0	0	0	60
1	1	0	0	20
1	1	1	0	4
1	1	0	1	2
0	0	1	1	3548
1	0	1	1	318
0	1	1	1	1048

5.1.2 Lexical category ordering parameters

The state transition parameters, along with the lexical categorization information, give the ordering of the lexical categories. Figure 5.5 shows the trends of the state transition parameters. These parameters reach their final values in fewer than 2000 training iterations. From this, it can be seen that the syntax HMM converges to a left-to-right model in which the state sequence is [2, 0, 1, 2, …]. With the learned lexical categorization information described in Section 5.1, the syntax HMM can be interpreted as representing the rule of syntax S→NV.

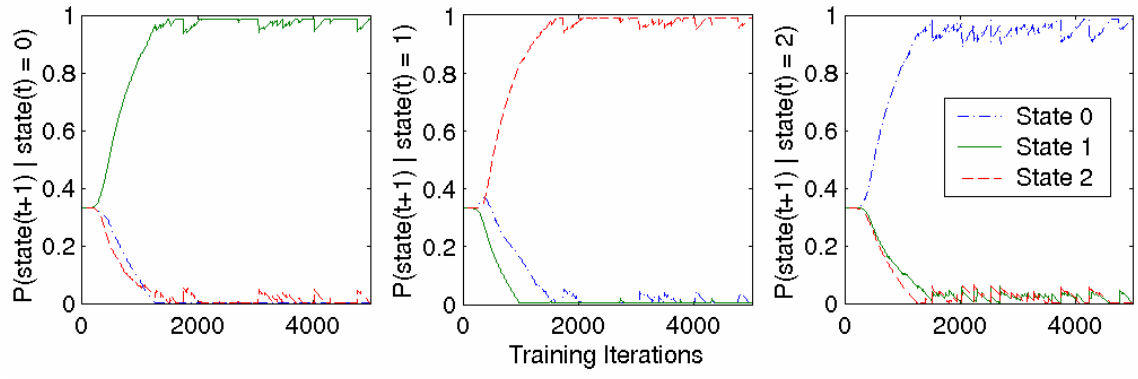

Figure 5.5: State transition parameters

5.1.3 Compositional semantics parameter

The compositional semantics parameter was defined for this rule of syntax to be the probability that the each word in the two-word sentence describes perceptual features regarding the same object. The training of this parameter is shown in Figure 5.6. Since the final value of this parameter is close to 1, this parameter was learned correctly.

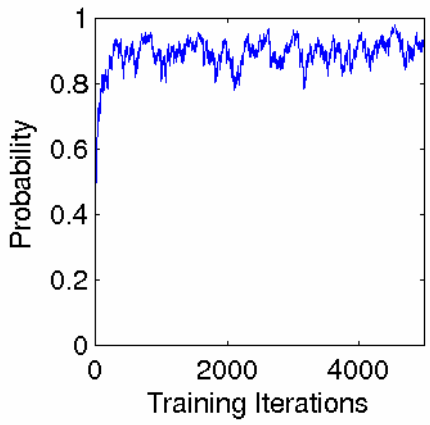

Figure 5.6: Probability that each word describes modalities from the same perceptual feature vector

As described in Section 4.3.3, the training algorithm was presented with two feature vectors (representing two objects present in the robot's field of view) half of the time. In these instances, the compositional semantics parameter may be estimated incorrectly. The results from the training were that the compositional semantics parameter was estimated correctly 79% of the time (1965 times out of 2480 possible).

5.2 Syntax Model Testing Results

The syntax model was tested on the robot by using it to produce two-word sentences that describe what the robot observes. The testing involved 45 trials: 15 trials with each of the objects and 5 each of the robot actions with each object. The robot was able to produce the correct sentence 41 out of the 45 trials (91% accuracy). In each case that the robot produced an incorrect sentence, the error occurred with the first word only. Thus, the accuracy of producing correct words was 96% (86 out of 90). As was required

by the model, the robot was able to produce two-word sentences describing the toy dog, which means that it was able to produce sentences that it never heard before.

5.2.1 Analysis of testing errors

In each of the errors in sentence production, the robot outputted an action word when the correct word was "puppy." An inspection of the perceptual features shows that in each of these cases, a low-probability shape or color feature was encountered for the perceptual state corresponding to the word "puppy." Thus, these errors occurred because of an inaccurate perceptual model of the "puppy" object and not because the syntax training did not include the toy dog object. The probability that a perceptual state describes a perceptual modality is computed as the product of the probabilities that the perceptual state describes the features in that modality (see Equation (3.4)). Thus, one low-probability feature can make the probability of its corresponding modality low as well. In addition, the perceptual states that correspond to the verbs have a high variance in the shape and color modalities, and so do not have very low probabilities for any measurement in these modalities (see Figures 4.1 and 4.2). Table 5.2 shows the low probability features that were encountered. This table shows the incorrect word that was outputted, the perceptual feature that had the low probability output for the "puppy" perceptual state ("Color 2" means color histogram bin number 2), and the probabilities of that feature for the "puppy" perceptual state and the perceptual state that corresponds to the incorrect word outputted.

Table 5.2: Low probability features encountered during testing

Incorrect word outputted	Perceptual Feature	P(feature | "puppy")	P(feature | incorrect word)
"gone"	Color 2	0.011	0.36
"move"	Color 2	0.011	0.20
"stay"	Ratio	0.067	0.26
	Color 2	0.020	0.40
"gone"	Ratio	0.067	0.13
	Color 2	0.020	0.11

Table 5.2 shows that most the errors were caused by low probability color features. This is most likely due to noise in the color measured by the cameras used on the robots. The low probability shape feature (ratio of width to height) is due to the fact that the observed width to height of the toy dog varies greatly according to the perspective in which the object is viewed. It is likely that the particular width to height ratio measured in the instances shown in Table 5.2 were rarely encountered in the perceptual model training.

5.3 Discussion of Results

The syntax model presented in this thesis is shown to correctly acquire and use the sentence production rule S→NV based on semantics. That is, the model learned the lexical categories N (noun) and V (verb) as defined by the perceptual feature modalities that the categories describe, the correct ordering of the lexical categories in the production rule, and the compositional semantic information as defined for this thesis. The syntax model's knowledge of lexical categories can be viewed as learning the production rules of the form $X \rightarrow x$, where X is either N or V and x is a word is the robot's lexicon. Thus, the syntax model has demonstrated the ability to learn each of the types of rules in CNF. If the model presented here can be extended to learn additional sentence production rules, the robot will have the ability to create and understand sentences of English generated by context-free grammars, which are able to capture a significant amount of the structure of natural language.

The robot used in the experiments was able to use the learned syntax to produce two-word sentences that describe what it observes, with high accuracy. The errors that were made in sentence production were due to deficiencies in the learned perceptual model and not errors in the learned syntax model. The robot was able to produce two-word sentences that it did not hear before, as was required.

The implementation of the syntax model presented in this thesis is shown to be robust. The model was able to converge to the correct values despite errors made during the training.

5.3.1 Issues with the proposed syntax model

This section will present some issues with the syntax model proposed in this thesis.

5.3.1.1 Tractability of the likelihood function

The computation of the likelihood function in Equation (3.2) becomes intractable if the number of perceptual modalities is large. This is because the number of compositional semantics hypotheses grows exponentially with the number of perceptual modalities. It is not known exactly what perceptual modalities humans use to build models of the world around them, let alone how many. Perhaps at the stage when children are beginning to learn syntax, only a small number is used. It could be argued that during semantic bootstrapping, a child would only use their five senses plus proprioception and information about their own mental states. This could constitute a small enough number of modalities for Equation (3.2) to be tractable. As well, there is much more information available to children in the language learning context than is used here, especially social cues, which may be able to be used to simplify the process of estimating lexical categories based on semantic descriptions. As well, the mental models that children develop are most certainly more descriptive than the models developed by the robots in these experiments, and may also able to simplify syntax as defined by this thesis.

5.3.1.2 Learning from two-word sentences

The syntax model used in this thesis requires that only two-word sentences of the form S→NV are heard by the robot to describe its world. This is obviously inconsistent with the natural speech that children are exposed to while learning language. As stated, in Section 1.2.1, directly mimicking human language acquisition is not the aim of this work. That being said, the problem of learning language from natural speech is much more difficult than the problem solved by this thesis. Not only must children learn syntax from long sentences, they must also be able to deal with many different forms of sentences ("look at the ball move," "move the ball over there," "do you see the moving ball"). Thus, the formulation of the robotic syntax acquisition in this thesis is meant to introduce a possible method for inferring syntactic information using semantics and

possibly lend plausibility to the idea that humans do the same, instead of being a finalized solution to the problem.

5.3.2 Effects of the perceptual model on syntax learning

The results in Section 5.1 show that the variance of the learned perceptual models directly affects the inference of the syntactic parameters. The shape and color modalities have high variance in the perceptual model for the object states (see Figures 4.1 and 4.2). Thus, when inferring the syntactic parameters it was not extremely unlikely to see feature values in one of these modalities that had a higher probability in one of the verb states than in one of the object states. This is in contrast to the perceptual models for the verb states, in which there is very little variance (see Figures 4.3 and 4.4). This variance in the perceptual model for the shape and color modalities resulted in the errors in syntactic parameter estimation seen in Table 5.1 and the convergence of the lexical categorization parameters seen in Figures 5.1 and 5.2 to values lower than those seen in Figures 5.3 and 5.4. Thus, performance of the syntax learning algorithm is affected by the variance in the perceptual model.

5.3.3 Lexical categorization and mental models

The fact that lexical categorization was required to be learned by the syntax model in this thesis leads to the attractive feature that it is consistent with the view that language is a method of describing mental models of the world. If the perceptual model used in this thesis is interpreted as the robot's mental model of the world, then the name, word, and syntax models would correspond to a language model. The significance of this is that this allows descriptive meanings of words to be learned without a priori restrictions on the syntactic roles of the words. This is important firstly because many words in natural language are used in many different syntactic roles. For example, the word "throw" can mean a verb that describes an action or it can be a noun that that describes the act of throwing. If these words were learned as verbs in the mental model of the world, then the mental model would need a separate representation for the same words used as nouns. Since the two different uses of the words are semantically similar, separate representations should not be necessary – it is the language model that determines how to use the word in a sentence to describe the world. In the syntax model used in this thesis, the lexical categorization information is learned by the syntax model,

not by the perceptual model. Secondly, this is significant because there is more meaning to words than what makes them appropriate for a syntactic role. For example, in order to know that the sentence "The truck rained pink resistance" is nonsense (except in a very metaphorical sense), one needs to know more about "truck" than what makes it a noun. It is noted here that this example is similar to the classic sentence from Chomsky [69], "Colorless green ideas sleep furiously." The point to be made is that although each of these examples have metaphorical semantic interpretations (see [3] for a valid semantic interpretation of "Colorless green ideas sleep furiously"), information beyond what makes the words appropriate for their lexical categories is necessary to know that only a metaphorical interpretation is possible.

5.3.4 Implications for human language development research

Although the research of LAR-UIUC is still in its early stages, it is hoped that the bottom-up approach to studying language will eventually lead to new answers to questions about human language development. One such possible answer that has emerged from this work concerns the phenomenon of children of different languages having a preference for nouns or verbs in their early speech production. Gopnik et al. [7] use the example of English children, who use more nouns in their early speech, and Korean children, who use verbs more frequently. It was found by Gopnik and Choi, a Korean colleague, that English-speaking mothers used more nouns when speaking to their children, while Korean-speaking mothers used more verbs. This was used to explain the children's preference for nouns or verbs. In addition, however, "[T]he Korean-speaking children learned how to solve problems like using the rake to get the out-of-reach toy well before the English-speaking children. English speakers, though, started categorizing objects earlier than the Korean speakers" [7, p. 89-90]. The conclusion, then, is that children's mental development (including language development) is influenced by the speech that they are exposed to.

While this is clearly interesting phenomena and while the conclusion makes sense intuitively, no account is given for how and why this occurs. It is here that findings from this thesis may be able to shed light on a possible explanation, and it comes from the cognitive model's similar bias for classifying a novel perceptual feature vector as a noun rather than as a verb. As noted in Section 4.2.3, this bias arises from the fact that from

the set of nine perceptual features, seven are used to discriminate the nouns and two are used to discriminate the verbs. Thus, it is hypothesized that children's preference for nouns or verbs comes from a similar bias in their mental models of the world. The way in which hearing more of nouns or verbs during language development could lead to this bias is explained as follows. In learning the perceptual representations for nouns or verbs, richer representations will need to be learned for the class with higher membership. That is, if a child learns the names of lots of different objects (nouns) but the names of few actions (verbs), the child will need to have more descriptive representations of the objects to be able to discriminate them. These more descriptive representations would be analogous to using more perceptual features. Then, during speech production, if children choose the words that best describe the world around then, they will choose the types of words for which they needed to create the most descriptive representations, i.e., the types of words that they heard most often while learning language. This explanation is consistent with the observation that Korean-speaking children learn to solve action related problems earlier and English-speaking children perform object categorization earlier. These phenomena could also be explained from the fact that Korean-speaking children are building more descriptive representations of the dynamics of the world around them while English-speaking children are finding interesting things to do with their rich representations of the objects in their environment.

5.4 Conclusions

The work in this chapter gives a method for acquiring information about syntax using semantics. The models and algorithms developed here are meant to be intuitive and straightforward. It is hoped that the results here give weight to theory that the human use of language is driven by semantics and that a wealth of innate knowledge of language is not necessary.

CHAPTER 6 FUTURE WORK

6.1 Extending the Current Work

This section presents ideas for extending the robotic syntax model presented in this thesis.

6.1.1 Compositional semantics representation

In order to create a syntax model capable of learning additional rules of sentence production, better representations of compositional semantics information must be developed. A shortcoming of the syntax model presented in this dissertation is that the information about the compositional semantics to be learned was defined by the designer of the model. In general, any such knowledge about syntax should be inferred from interpretations of how the sentences heard describe the world. A problem that must eventually be solved is how design a syntax model that can determine what information is relevant to lexical semantic and what is relevant to the compositional semantics. Such a solution will require more general representations of the semantic interpretation of a sentence than provided by Equation (3.1). As noted in Section 5.3.1.1, this equation already has issues with tractability and so a more general representation may require the use of other information such as social cues to be computationally feasible.

6.1.2 Learning additional production rules

The current thesis provides a means of representing one sentence production rule in CNF with an HMM. The extension of this work to multiple production rules is not trivial. The two types of production rules that I will discuss here, are other production rules that are of the form (in CNF) S→*XY* (rules that start sentences) and *Z*→*VW* (rules that do not start sentences).

In order to extend the current work to represent additional production rules of the form S→*XY*, it is reasonable to assume that simply two more states per additional production rule need to be added to the syntax HMM. The S state will serve as the start state for each of the production rules. After learning the rules of syntax, the state that follows the S state will no longer be deterministic, because the first state of each production rule will be possible. Thus, the state following the S state will need to be determined using compositional semantics information. For example, if the syntax model

has learned the two production rules S→NV and S→VN, the robot would need to understand the difference in the compositional semantics of each. Then, when interpreting a new sentence or generating a sentence, these two rules would be differentiated by the compositional semantics of the two rules, which have to do with the intentions of the speaker. The S→NV rule signifies that the speaker intends to describe something and the S→VN rule signifies a command. As noted in Section 6.1.1, the HMM implementation of the syntax model does not capture compositional semantics information of production rules. In the experiments in this thesis, only one production rule was to be learned and so there was an easy solution to representing a small amount of compositional semantics. However, if more than one production rule is to be learned simultaneously, representing this for production rule will be difficult unless it was assumed a priori which states of the HMM were to be used in which production rule.

Extending the syntax model used in this dissertation to represent production rules of the general form $Z \rightarrow VW$ also seems to be feasible. In this thesis, the learning of a rule of syntax is treated as the learning of lexical categorization and compositional semantics. Ideally (as stated in Section 6.1.1), this should occur by learning to attribute the description of the robot's perceptual inputs to either the first nonterminal symbol, the second nonterminal symbol, or the compositional semantics. Then, if other production rules are applied, the perceptual inputs associated with each of these nonterminal symbols can be divided again into two other nonterminal symbols and some compositional semantics. Take, for example, the sentence "red ball move" (extended from the example in Figures 3.1 and 3.2). In the framework proposed by this dissertation, the sentence would assign the shape and color information to the noun phrase "red ball" and the change in position information to the verb "move." Then, the noun phrase would be subdivided using the production rule NP→AdjN to assign the color information to the adjective "red" and the shape information to the noun "ball." Viewed in this way, the nonterminals on the left-hand side of the production rule and the compositional semantics divide up the information encapsulated by the nonterminal symbol on the left-hand side of the production rule.

In order to acquire a new sentence production rule of the type $Z \rightarrow VW$, the syntax model would need to have already learned a rule of the form S→XY, where Z in the first

rule is the same symbol as either *Y* or *Z* in the second. Then, when a three-word sentence is heard, the lexical categorization information will be inferred about each word, and a rule of the form $Z \rightarrow VW$ will need to be inferred that combines the lexical categorization information about the two possible pairs of the words into one of the lexical categories in the left hand side of the production rule $S \rightarrow XY$. Using the "red ball move" example from above, the rule of syntax already learned would be $S \rightarrow XY$, where the lexical category *X* describes shape and color information and *Y* describes change in position. Then, when the sentence "red ball move" is heard, a new lexical categorization will need to be inferred for each word: A→"red", B→"ball", and C→"move", where A describes color information, B describes shape, and C describes change in position. Then, a new rule of syntax will need to be inferred to transform the string ABC into *XY*. The intuition here is that the new rule $Z \rightarrow VW$ should mean that the lexical category *Z* describes all of the information described by *V* and *W*, with some additional compositional information as (note here that this is the same intuition behind the $S \rightarrow XY$ rule). So, the inference algorithm should find that $X \rightarrow AB$ is the relevant rule. As well, it will need to be inferred that *Y* and C are the equivalent lexical categories.

6.1.3 Learning abstract definitions of lexical categories

In order for the syntax model developed in this thesis to be able to represent adult-like syntactic competence, the path of semantic bootstrapping will need to be followed from semantically defined lexical categories to abstractly defined lexical categories. Although the solution to implementing this in a robotic framework is unknown and somewhat daunting, some broad ideas will be offered here. From the perspective that language is a means of describing a mental model of the world, the process of learning abstract lexical categories should start with the development of abstract concepts in the mental model that help to understand how the world works. Examining adult language, it can be inferred that a useful mental model of the world could include abstract concepts of things and their functions in the world, actions and changes in state that can be performed by or on these things, and a variety of properties that these things can have. These, of course, would correspond to abstract definitions of nouns, verbs, and adjectives, respectively. Developing computational methods to enable a robot to create such a model of the world is a very difficult problem to be solved.

6.2 Syntactic Bootstrapping

According to Fisher [16], syntactic information can be used by children to assist in the learning of new words. As described in Section 1.3, if a child had learned the rule of syntax S→NV, then if a sentence of this form is heard with a novel verb, the child can use the compositional semantics of the production rule to infer that the verb describes something about the object referred to by the noun. Production rules learned by the syntax model developed it this thesis can be used in the same way. When a sentence with a novel word is heard, the syntax model can identify the perceptual feature vector that the known word is meant to describe. Then, the syntax model can inform the perceptual models to train a new (or unused) state with the identified feature vector. As well, the name model would be informed to train a new (or unused) state to associate the new perceptual state with the new word. It should be noted that using syntactic bootstrapping would not be useful for the LAR-UIUC's current robotic name learning system because it assumes that the robot knows the correct referent of a new word. However, when LAR-UIUC's robots eventually leave their controlled environment, the problem of identifying what a new word refers to will have to be dealt with. In this case, syntactic bootstrapping will become quite useful.

6.3 Perceptual Model Improvements

The results in Chapter 5 show that the variance of the perceptual model in the syntax acquisition experiment had a direct affect on the accuracy of the syntactic parameter estimation. In this section, some ideas will be presented for improving the perceptual model in order to improve syntax acquisition.

First, improvements to the perceptual feature set are possible. As mentioned in Section 4.1.1, the shape features used in this thesis are not rotation invariant, and so cannot capture a three-dimensional representation of the objects in the robotic environment. Recent work in LAR-UIUC by Rivera [82] involves the learning of rotation invariant feature points for object recognition. Incorporating this into the syntax learning experiment should greatly improve the performance of the syntax model. Another improvement would involve using linear discriminant analysis (LDA) on the

color features. LDA is a computational method that finds a linear transformation on the feature space of labeled examples that maximizes the separation between the classes. Since labeled samples are required for LDA, it would be applied after the perceptual classes are learned and stored training samples are labeled.

Using different probability distributions in the perceptual model could also lead to better representations. Independent discrete distributions were used in this thesis because they are easy to use. Most of the distributions in Figures 4.1 through 4.4 could be modeled using Gaussian distributions, which can also capture covariance information that could be useful.

6.4 Summary and Final Words

This chapter has presented some ideas for extending and improving the syntax model described in this thesis. Although the current HMM implementation may not be appropriate for a full representation of a grammar, the formulation of syntax acquisition used in this thesis I believe has some merit and can be extended to the acquisition of an entire grammar.

REFERENCES

[1] N. Wiener, *Cybernetics: or Control and Communication in the Animal and the Machine*. Cambridge: MIT Press, 1961.

[2] A. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 443-460, 1950.

[3] S. E. Levinson, *Methods of Mathematical Linguistics: Applications to Speech Technology*. London, UK: John Wiley, 2005.

[4] L. Bloom, *Language Development From Two to Three*, New York: Cambridge University Press, 1991.

[5] *A History of Engineering and Science in the Bell System: Communication Sciences (1925-1980)*, Prepared by Members of the Technical Staff, AT&T Bell Laboratories, ed. by S. Millman, 1984, AT&T Bell Laboratories.

[6] S.E. Levinson, K. Squire, R.S. Lin, and M. McClain, "Automatic language acquisition by an autonomous robot." AAAI Spring Symposium on Developmental Robotics, March 21-23, 2005.

[7] A. Gopnik, A. Meltzoff, and P. K. Kuhl, *The Scientist in the Crib: What Early Learning Tells Us About the Mind*. New York: Harper Collins, 1999.

[8] K. Squire, "HMM-based semantic learning for a mobile robot," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2004.

[9] D. Li, "Computational models for binaural sound source localization and sound understanding," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2003.

[10] M. Kleffner, "A method of automatic speech imitation via warped linear prediction," M.S. thesis, University of Illinois at Urbana-Champaign, 2003.

[11] R. Lin, "Unsupervised learning of nonlinear manifolds for map building on an autonomous robot" Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2005.

[12] M. McClain, "The role of exploration in language acquisition for an autonomous robot," M.S. thesis, University of Illinois at Urbana-Champaign, 2003.

[13] R. Brown, *A First Language: The Early Stages*. Cambridge,MA: Harvard University Press, 1973.

[14] L. Bloom, *Language Development: Form and Function in Emerging Grammars*, Cambridge, MA: The MIT Press, 1970.

[15] S. Pinker, "The bootstrapping problem in language acquisition," in *Mechanisms of Language Acquisition*, Brian MacWhinney, Ed. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1987, pp. 399-442.

[16] C. Fisher, "The role of abstract syntactic knowledge in language acquisition: a reply to Tomasello (2000)," *Cognition* vol. 82, pp. 259-278, 2002.

[17] L.G. Naigles, A. Fowler, A. Helm, "Syntactic bootstrapping from start to finish with special reference to down syndrome," in *Beyond Names for Things*, M. Tomasello and W.E.Merriman, Eds. Hillsdale, NJ: Lawrence Erlbaum Associates, 1995, pp. 299-330.

[18] M. Tomasello and W.E. Merriman, "Introduction: verbs are words, too," in *Beyond Names for Things*, M. Tomasello and W.E. Merriman, Eds. Hillsdale, NJ: Lawrence Erlbaum Associates, 1995, pp. 1-18.

[19] M. Tomasello, "Pragmatic contexts for early verb learning," in *Beyond Names for Things*, M. Tomasello and W.E. Merriman, Eds. Hillsdale, NJ: Lawrence Erlbaum Associates, 1995, pp. 115-146.

[20] T. Winograd, *Language as a Cognitive Process, Volume 1: Syntax*. Reading, MA: Addison-Wesley, 1983.

[21] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Englewood Cliffs, NJ: Prentice-Hall, 2000.

[22] S. Pinker, *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press, 1984.

[23] W. Koenig, H. K. Dunn, and L. Y. Lacy, "The sound spectrograph," *Journal of the Acoustical Society of America*, vol. 18, pp. 19-49, 1946.

[24] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 637-642, 1952.

[25] A. Oppenheim, R. Schafer, and T. J. Stockham, "Nonlinear filtering of multiplied and convolved signals," *Proceedings of the IEEE*, vol. 56, no. 8, pp. 1264-1291, 1968.

[26] B. S. Atal and S. Hanauer, "Speech analysis and synthesis by prediction of the speech wave," *Journal of the Acoustical Society of America*, vol. 50, pp. 637-655, 1971.

[27] J. Weizenbaum, "ELIZA – A computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36-45, 1976.

[28] T. Winograd, *Understanding Natural Language*. New York: Academic Press, 1972.

[29] J. K. Baker, "The DRAGON system – An overview," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. ASSP-23, no.1, pp. 24-29, 1975.

[30] F. Jelinek, R. L. Mercer, and L. R. Bahl, "Design of a linguistic statistical decoder for the recognition of continuous speech," *IEEE Transaction on Information Theory*, vol. IT-21, no. 3, pp. 250-256, 1975.

[31] R. L. Cave and L. P. Neuwirth, "Hidden Markov models for English," in *Hidden Markov Models for Speech*, J. Ferguson, Ed. Princeton, NJ: IDA-CRD, 1980.

[32] A. B. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Proceedings of Int. Conf. on Speech and Signal Processing*, 1982, pp. 1291-1294.

[33] T. K. Vintsyuk, "Speech discrimination by dynamic programming," *Kibernetica*, vol. 4, no. 1, pp. 81-88, 1968.

[34] F. Jelinek, "Fast sequential decoding algorithm using a stack," *IBM Journal of Research and Development*, vol. 13, pp. 675-685, 1969.

[35] B. T. Lowerre, "The Harpy speech recognition system," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, 1968.

[36] V. R. Lesser, R. D. Fennell, L. D. Erman, and D. R. Reddy, "Organization of the Hearsay-II speech understanding system," *IEEE Transactions on Acoustic Speech and Signal Processing*, vol. ASSP-23, pp. 11-24, 1975.

[37] Q. Liu, "Interactive and incremental learning via a multisensory mobile robot," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2001.

[38] W. Zhu and S. E. Levinson, "PQ-learning: An efficient robot learning method for intelligent behavior acquisition," in *Proc. 7th Int. Conf. on Intell. Autonomous Systems*, vol. 1, Marina del Rey, CA, Mar. 2002, pp. 404-411.

[39] R. Lin, "Learning vision-based robot navigation," M.S. thesis, University of Illinois at Urbana-Champaign, 2002.

[40] J. Piaget, *The Origins of Intelligence in Children*. New York: International Universities Press, Inc., 1952.

[41] G. Lakoff and M. Johnson, *Philosophy in the Flesh: The Embodied Mind and its Challenges to Western Thought*. New York: Basic Books, 1999.

[42] L. Smith and M. Gasser, "The development of embodied cognition: Six lessons from babies," *Artificial Life*, vol. 11, no. 1-2, pp. 13-30, 2005.

[43] M. Kaschak, C. Madden, D. Therriault, R. Yaxley, M. Aveyard, A. Blanchard, and R. Zwaan, "Perception of motion affects language processing," in *Cognition*, 94, pp. B79-B89, 2005.

[44] J. Wang and L. Gasser, "Mutual online concept learning for multiple agents," in *Proceedings of AAMAS-2002*, Bologna, Italy. July 2002, pp. 362-369.

[45] N. Komarova and P. Niyogi, "Optimizing the mutual intelligibility of linguistic agents in a shared world," *Journal of Artificial Intelligence*, vol. 154 no. 1-2, pp. 1-42, 2004

[46] R. A. Brooks, "Achieving artificial intelligence through building robots," Massachusetts Institute of Technology, Artificial Intelligence Laboratory, Tech. Rep. 899, 1986.

[47] A. Flynn, R. A. Brooks, and L. Tavrow, "Twilight zones and cornerstones, a gnat robot double feature," Massachusetts Institute of Technology Artificial Intelligence Laboratory, A.I. Memo 1126, July 1989.

[48] R. A. Brooks, "Prospects for human level intelligence for humanoid robots," in*Proceedings of the First International Symposium on Humanoid Robots (HURO-96)*, Tokyo, Japan, October, 1996.

[49] R. A. Brooks, C. Brezeal, M. Marjanovic, B. Scassellati, M. M. Williamson, "The Cog project: Building a humanoid robot," in *Computation for Metaphors, Analogy, and Agents*, C. Nehaniv, Ed. New York: Springer, 1999, pp. 52-87.

[50] C. Breazeal, "Sociable machines: Expressive social exchange between humans and robots." Sc.D. dissertation, Massachusetts Institute of Technology, 2000.

[51] J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur and E. Thelen, "Autonomous mental development by robots and animals," *Science*, vol. 291, no. 5504, pp. 599-600, Jan. 26, 2000.

[52] S. Zeng and J. Weng, "Online-learning and attention-based approach to obstacle avoidance using a range finder," *Journal of Intelligent and Robotic Systems*, vol. 43, no. 2, June 2005.

[53] Y. Zhang and J. Weng, "Grounded auditory development by a developmental robot," in *Proc. INNS/IEEE International Joint Conference of Neural Networks 2001 (IJCNN 2001)*, Washington DC, July 14-19, 2001, pp. 1059-1064.

[54] O. Sporns, and W. H. Alexander, "Neuromodulation and plasticity in an autonomous robot," *Neural Networks*, vol. 15, pp. 761-774, 2002.

[55] O. Sporns and W. H. Alexander, "Neuromodulation in a learning robot: Interactions between neural plasticity and behavior," in *Proceedings IJCNN*, 2003, pp. 2789-2794.

[56] D. M. Pierce and B. J. Kuipers, "Map learning with uninterpreted sensors and effectors," *Artificial Intelligence*, vol. 92, pp. 169-227, 1997.

[57] B. Kuipers and P. Beeson, "Bootstrap learning for place recognition," in *Proc. 18th national Conf. On Artificial Intelligence*, pp. 174-180.

[58] J. Modayil and B. Kuipers, "Bootstrap learning for object discovery," presented at IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-04), 2004.

[59] B. Kuipers, "The spatial semantic hierarchy," *Artificial Intelligence*, vol.119, pp. 191-233, 2000.

[60] P. Y. Oudeyer, F. Kaplan, V. Hafner, and A. Whyte, "The playground experiment: task-independent development of a curious robot," in *Proc. of the AAAI Spring Symposium on Developmental Robotics*, 2005.

[61] Y. Sugita and J. Tani, "A holistic approach to compositional semantics: a connectionist model and robot experiments," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. K. Saul, and B. Scholkopf, Eds. Cambridge, MA: The MIT Press, 2004, pp. 969-976.

[62] D. Roy, "Learning visually-grounded words and syntax for a scene description task," *Computer Speech and Language*, vol. 16, no. 3, pp. 353-385, 2002.

[63] D. Roy, "Learning visually grounded words and syntax of natural spoken language," *Evolution of Communication*, vol. 4, no. 1, pp. 33-56, 2001.

[64] K. Wexler and P. Culicover, *Formal Principles of Language Acquisition*. Cambridge, MA: MIT Press, 1980.

[65] J. Grimshaw, "Form, function, and the language acquisition device," in *The Logical Problem of Language Acquisition*, C.L. Baker and J.J. McCarthy, Eds. Cambridge, MA: The MIT Press, 1981, pp.165-182.

[66] J. Macnamara, *Names for Things: a Study of Human Learning*. Cambridge, MA: Bradford Books/MIT Press, 1982.

[67] J. Saffran, "The use of predictive dependencies in language learning," *Journal of Memory and Language*, vo. 44, pp. 493-515, 2001.

[68] F. Chang, G. Dell, and K. Bock, "Becoming syntactic," *Psychological Review*, submitted for publication.

[69] N. Chomsky, *Syntactic Structures*. The Hague, Netherlands: Mouton, 1957.

[70] N. Chomsky, "On certain formal properties of grammars," *Information and Control*, vol. 2, pp. 137-167, 1959.

[71] R. Montague, "The proper treatment of quantification in ordinary English," in *Approaches to Natural Language*, by J. Hintikka, J. Moravcsik, and P. Suppes (eds.). Dordrecht, Holland: Reidel. 1973.

[72] J. Batali, "The negotiation and acquisition of recursive grammars as a result of competition among exemplars," in *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, T. Briscoe, Ed. Cambridge: Cambridge University Press, 2001.

[73] T. Parsons, *Events in the Semantics of English: a Study of Subatomic Semantics*. Cambridge, MA: MIT Press, 1990.

[74] H. Kamp and U. Reyle, *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht, The Netherlands: Kluwer Academic, 1993.

[75] E.W. Djikstra, "A note on two problems in connection with graphs," *Numerical Mathematics*, vol. 1, pp. 269-271, 1969.

[76] D.M. Younger, "Recognition and parsing of context free languages in time n^3," *Information and Control*, vol. 10, pp. 198-208, 1967.

[77] E. Tanaka and K.S. Fu, "Error correcting parsers for formal languages," *IEEE Transactions on Computers*, vol. C-27, pp. 605-615, 1978.

[78] L.E. Baum, "An inequality and associated maximum likelihood technique in statistical estimation for probabilistic functions of a Markov process," *Inequalities*, vol. 3, pp. 1-8, 1972.

[79] J.K. Baker, "Trainable grammars for speech recognition," in *Speech Communications Papers for the 97th Meeting of the Acoustical Society of America*, J.J. Wolf and D.H. Klatt, Eds., 1979, pp. 547-550.

[80] D. Klein, “The unsupervised learning of natural language structure,” Ph.D. dissertation, Stanford University, 2005.

[81] N. Smith and J. Eisner, “Annealing structural bias in multilingual weighted grammar induction,” presented at COLING-ACL, Syndey, 2006.

[82] R. Rivera, “360 degree object recognition using sift features with autonomous model building,” M.S. thesis, University of Illinois at Urbana-Champaign, 2006.

APPENDIX A: CD OF SOURCE CODE

AUTHOR'S BIOGRAPHY

Matthew R McClain received his Bachelor of Science in Electrical Engineering from Rensselaer Polytechnic Institute in 1999. After traveling and working abroad for a year, he attended graduate school at the University of Illinois at Urbana-Champaign where he received his Master of Science in Electrical Engineering with the Language Acquisition and Robotics Group at the Beckman Institute. His research interests are in computational modeling of language and cognition.