Semantic Based Learning of Syntax in an Autonomous Robot: Preliminary Results

Matthew McClain
University of Illinois at Urbana-Champaign
405 N. Mathews Ave.
Urbana, IL 61801
mrmcclai@uiuc.edu

Stephen Levinson
University of Illinois at Urbana-Champaign
405 N. Mathews Ave.
Urbana, IL 61801
sel@ifp.uiuc.edu

Abstract - **It is the goal of the Language Acquisition Group at the University of Illinois at Urbana-Champaign (LAR-UIUC) to build a robot that is able to learn language as well as humans through embodied sensorimotor interaction with the physical world. This paper proposes cognitive structures to enable an autonomous robot to learn the syntax of two-word sentences using its understanding of lexical semantics. Production rules of syntax in Chomsky Normal Form will be explicitly represented using a hidden Markov Model. Preliminary results, in the form of simulated experiments, show that these models can learn representations of syntax in this form.**

Index Terms – language acquisition, developmental robotics, syntax, semantics

I. Introduction

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 having any understanding of its meaning. This is the central reason why the pursuit has been unsuccessful. The Language Acquisition and Robotics Group at the University of Illinois at Urbana-Champaign (LAR-UIUC) is attempting to construct robots that are able to use language as well as humans by learning the semantics of language through embodied interaction with the world. We posit that the memory should be associative, as its primary function is to correlate the robot's various sensori-motor inputs into a meaningful model of the world. Reinforcement learning, both supervised and unsupervised, is required to train the associative memory. As well, the robot's sensori-motor capabilities should be complex so that robust models of the world can be developed using them.

LAR-UIUC is using an embodied robotic framework to study intelligence and language from a bottom-up perspective. The current work is another step toward our goal. In this sense, it builds upon a great deal of work done by past members and should easily lend itself to work by future members.

The title of this paper is meant to reflect our belief that semantics plays a central role in language acquisition. We reject the idea that language acquisition is a result of an innate language faculty in humans. The work in this paper will show that information about syntax can be inferred using semantics, and does not require innate knowledge.

The next section of this paper presents related work from various fields. The third section presents the robotic framework developed by LAR-UIUC, including recent experiments. The problem that this paper seeks to address is presented in the fourth section, and the fifth section describes the proposed solution. Preliminary experiments and their results are presented in the sixth section. Finally, the seventh section proposes an experiment in LAR-UIUC's framework.

II. Related Work

A. Syntax Acquisition and Representation

The acquisition of syntax by machines in an embodied framework has been studied by Sugita and Tani [1] and Roy [2, 3]. Sugita and Tani use a pair of recursive neural networks (RNNs) to learn the compositional semantics of two-word command sentences. Roy has implemented algorithms that learn adjective-noun phrases using visual information in both simulated visual scenes and a robotic implementation.

Brown [4] has proposed five stages of children's sentence production from the study of the development of children's speech. 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. Bloom's [5] study of children's sentence production in speech goes beyond the formal structure of children's sentences and takes into account the context in which the child utters the sentence.

Chang, Dell, and Bock [6] have developed a dual-path connectionist model that is able to learn rules of syntax and lexical categories, and uses this information to produce syntactically correct sentences.

Kamp and Reyle [7] give examples of how predicate argument structures have been used to describe the compositional semantics of syntactic constructions.

Formal grammars have been developed by Chomsky [8] to provide a mathematical representation of the syntax of natural language. This representation contains four elements: V_T, a set of terminal symbols (words, in natural language); V_N, a set of non-terminal symbols (lexical or phrase categories); S, a special non-terminal symbol that represents a well-formed sentence of the grammar; and R, a set of production rules that dictate how the symbols of the grammar replace (or re-write) each other. The Chomsky Hierarchy [9] defines different classes of grammars based on the forms of the production

rules. Of relevance to this paper are context-free grammars, in which the production rules can be written in Chomsky Normal Form (CNF). In CNF, the production rules can be one of two types: X→x or X→YZ, where the lower-case letters are terminal symbols and upper case letters are non-terminals. The first type of rule determines lexical categorization and the second type gives sentence production rules, which lead to a binary tree interpretation of a sentence. While context-free grammars cannot represent all of the complexity of natural language, they are able to capture a significant amount of the structure of natural language with relatively simple production rules.

B. Hidden Markov Models

Hidden Markov Models (HMMs) have been shown to be able to learn aspects of the structure of language without prior knowledge. The Cave and Neuwirth [10] experiment gives a significant example of this. 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 an HMM is trained with a sequence of observations, the parameters converge to a local maximum in the model's parameter space, which represents the most likely model to have generated the observation sequence.

C. Embodied Cognition

Many researchers are studying embodied cognition in different contexts. Yu [11] has researched the effects of multi-modal learning in language acquisition. Weng [12] is studying embodied cognition using a developmental robotics approach.

III. ROBOTIC FRAMEWORK

A. Cognitive Cycle

The cognitive cycle in Figure 1 is used by LAR-UIUC to guide our implementation. Central to this cycle are the associative and working memories, which comprise the noetic system. This is where the robots learn their models of the world based upon their sensori-motor experience, and make decisions based on these models. The cognitive cycle is completed by the outside world, in which the robot is able to perceive the affects of its own actions. As well, proprioceptive feedback is available to help make robust measurements of the outside world.

B. Robotic Implementation

1) Hardware: Humans are our only examples of natural language users, and so we would like the physical implementation of our robots to be as anthropomorphic as possible. Our robots' motor abilities must allow them to move about their environment and interact with the environment through articulated movement (for example, grasping at objects and outputting speech). The sensory system must provide the robot with enough information to build

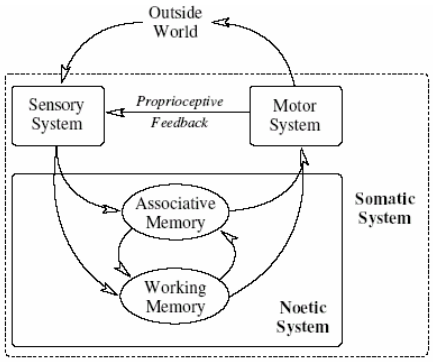

Figure 1: Cognitive Cycle

useful models of its environment. We have chosen to focus on visual and audio inputs, as these are the main sources of sensory information in humans and (possibly not coincidentally) have the most readily available physical implementations as cameras and microphones. As well, our robots have some touch sense, implemented by various sensors.

As the base of our robotic implementation we have chosen Arrick Robotic's Trilobot. This provides our robots with the ability to move about their environment, a front arm and gripper for grasping objects, and a head with pan and tilt ranges of motion. As well, sensors around the perimeter of the robot and in the gripper provide some basic tactile sense. To this platform we have added cameras for stereo vision and microphones for binaural hearing. An on-board computer has been added to handle some of the computational load of the robots' cognition. The robots have also been equipped with wireless internet to distribute the rest of the computing to desktop computers.

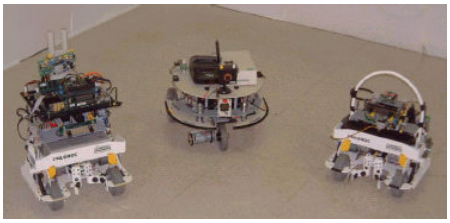
Figure 2: LAR-UIUC's robots Illy, Alan, and Norbert

2) Software: Various software programs have been developed for the robots to implement their cognitive abilities. A visual processing program has been implemented by Lin [13] to perform object segmentation and extract shape and color features. Kleffner [14] has implemented algorithms to process speech using warped linear prediction. Programs for distributing raw sensory information amongst computers have been developed by Squire [15]. A working memory has been implemented by McClain [16] that provides the robot with the ability to explore its environment. Zhu [17] has developed PQ-learning algorithms for visual navigation using reinforcement. Liu [18] has contributed programs that allow the learning of voice commands for controlling the robot's movement.

C. Experiments: Semantic Associative Memory

Squire [15] has implemented an associative long-term memory that allows the robots to learn the semantics of words based on sensory experience. The model used for the semantic associative memory has three components: a visual model, an auditory model, and a concept model.

Each of the components of the associative memory was implemented by Squire using an HMM. In the visual model, each state corresponds to the visual representation of one object as a joint probability distribution over the visual features. As well, the auditory model uses the each state of the HMM to represent one word using a joint probability distribution over the auditory features. The states of the concept model HMM have outputs that are distributions over the states of the visual and auditory models, which can be interpreted as a naming of objects.

IV. SEMANTIC BASED LEARNING OF SYNTAX

A. Background

The associative memory developed by Squire [15] can be viewed as the single-word learning stage in childhood language development. According to Brown [19], "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" (p77). Thus, using childhood language development as a guide, the next step for LAR-UIUC is the task of enabling our robots to learn to use two-word sentences.

B. Proposed Research

We propose to enable our robots learn the syntax of two-word sentences, using production rules in CNF as a basis of representing the syntactic information. Once a production rule has been learned, the robot should be able to use this information in new contexts. The specific production rule that is to be learned is S→NV: a sentence can be composed of a noun followed by a verb. The information that is to be learned is the ordering of the lexical categories and the compositional semantics of the production rule.

C. Lexical Categories

We propose that in the initial stages of language development, it is unnecessary for a language user to have knowledge of lexical categories. Tomasello [20] notes, "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." (p137). Words are more reliably assigned to lexical categories based on their position in a syntactic structure than by the semantics of the words themselves. Jurafsky and Martin [21] note that "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." (p289). Thus there appears to be a paradox: knowledge of syntax is required to learn lexical categories and vice-versa. The solution to this paradox is that lexical categories are initially learned using semantics. Once this basic understanding is developed, adult-like competence can be acquired. According to Brown [19], "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." (p26) This proposed solution is related to the semantic bootstrapping hypothesis developed by Pinker [22].

The semantics used by LAR-UIUC's robots are grounded in the robots' sensori-motor experience, which are encoded as perceptual features. The semantic understanding of lexical categories will be based on these same perceptual features. Specifically, a lexical category will be defined by the subset of the perceptual features that members of that category describe.

D. Compositional Semantics

The compositional semantics of a sentence is the information that is not described by the lexical semantics of the words of the sentence. For example, the sentence "ball move" does not just mean "there is a ball' and "something is moving". The compositional semantics of the production rule S→NV relevant to this paper is that the noun and the verb describe the same object. In LAR-UIUC's robotic implementation, perceptual features from each object that the robot is observing are associated into a single vector. Thus, the robot will need to learn that the two words in the production rule S→NV describe features from the same feature vector.

V. IMPLEMENTATION

A. Cognitive Framework

The proposed cognitive framework can be seen in Figure 3. This framework extends Squire's semantic associative memory described in section III.C by adding a syntax model as and renaming the concept, visual, and auditory models as the name, perceptual, and word models, respectively, to better reflect their role in the syntax acquisition framework.

B. Inferring Syntactic Information

The syntax model in Figure 3 receives the necessary information to infer the syntactic parameters: perceptual features, the learned distributions over the perceptual features in the perceptual model, and distributions over the perceptual states that correspond to the words uttered.

As stated in section III.C, the lexical categorization is defined by the subset of perceptual features that members of that category describe. This will be represented by an n by 2

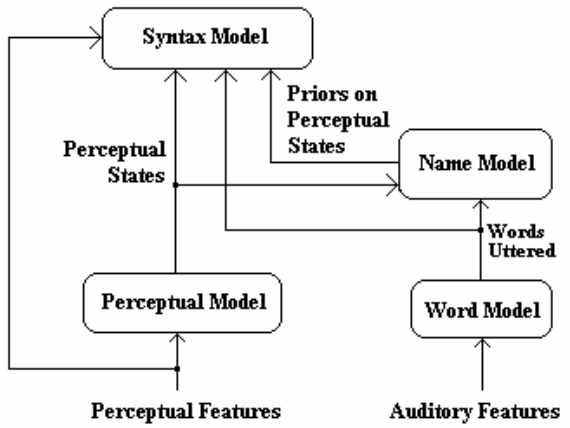

Figure 3: Cognitive Model for Syntax Acquisition

parameter matrix α, where n is number of perceptual features. Element α_{ij} represents the probability that feature number i is described by word j.

The compositional semantics information will be represented by the parameter β, which signifies the probability that the two words spoken describe the same object.

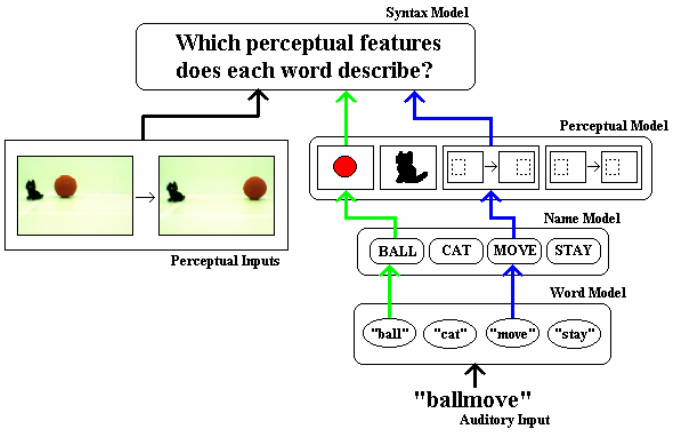

Figure 4: Example of Inferring Syntactic Information

As an example, let perceptual feature number 1 be shape, 2 be color, and 3 be change in position. Let it also be assumed that the associative memory has learned the names of two objects, "cat" and "ball", and two actions, "stay" and "move". This is represented in Figure 4 by the connections between the word, name, and perceptual models, shown for the two words being recognized.

Using this information, the syntactic information can be inferred by observing example two word sentences in context. This is done by computing which of the perceptual features being observed are meant to be described by which word, using the perceptual models associated with each word through the name model. Figure 4 gives an illustration of this.

In this example, the syntax model would infer that that the first word describes the shape and color information of one of the objects, and that the second word describes the change in position feature. This would be represented by α_{11}=1, α_{21}=1, α_{31}= 0, α_{12}= 0, α_{22}=0, and α_{32}=1. As well, both words describe features from the same object, and so β will be 1.

C. Syntax Model Implementation

In the syntax model, an HMM will be used to represent the production rule to be learned. The experiment by Cave and Neuwirth [10] shows that HMMs can learn categorization and sequence information, which is required for the lexical categorization component of the syntax. Therefore, an HMM will be used to explicitly represent a production rule of the form X→YZ , where each non-terminal symbol will be represented by a state, and the perceptual features described by the lexical category will be the observable outputs of each state.

The compositional semantics parameter will be represented by a parameter outside of the HMM because the information that it carries is associated with the whole production rule, not any single lexical category.

D. Likelihood Function

The estimates of the syntactic parameters will be computed by maximizing the joint likelihood of the syntactic parameters, perceptual inputs, and perceptual models associated with the words spoken. This likelihood computes all possible interpretations of how the two-word sentence describes the perceptual input. Each of these hypothesized interpretations is represented by a unique binary vector j, whose length is equal to the number of perceptual features. Each element of j is either 1or 2, indicating that the perceptual feature associated with the element's index is described by the first or second word, respectively, of the two-word sentence. This likelihood function is shown in (1).

$$L(\alpha,\beta,w,v)=\sum_{j}\sum_{k=1}^{m}\sum_{l=1}^{m}\beta_{kl}\,g(\beta)\,*$$
$$\prod_{i=1}^{n}\alpha_{ij(i)}[\prod_{j(i)=1}P(v_{ki}\mid w_1)][\prod_{j(i)=2}P(v_{li}\mid w_2)] \quad (1)$$

where:
m = number of feature vectors (objects present)
β_{kl} = probability that the features described by the 1st word come from feature vector k and the features described by the 2nd word come from vector l
$g(\beta)$ = prior distribution on β
v_{ki} = feature i from feature vector k
w_i = word number i of the two word sentence

The number of possible hypotheses in the first sum term are restricted by the assumptions that each word must describe at least one feature and each feature is described by exactly one word. It is also assumed that if a word describes more than one feature, those features are present in the same perceptual feature vector, which is included in the likelihood function.

1) Estimating Syntax Parameters: The syntactic parameters (α and β) will be estimated in an EM (expectation – maximization) fashion. That is, the α parameters will be estimated using the most recent estimate of β. Then, β will be estimated using the new estimate of α.

Each hypothesis in j corresponds to a set of α_{ij}'s with a coefficient of the form:

$$\sum_{k=1}^{m}\sum_{l=1}^{m}\beta_{kl}\,g(\beta)\,*$$
$$\prod_{i=1}^{n}\alpha_{ij(i)}[\prod_{j(i)=1}P(v_{ki}\mid w_1)][\prod_{j(i)=2}P(v_{li}\mid w_2)] \quad (2)$$

Here, the prior distribution $g(\beta)$ will be uniform. The α_{ij}'s with the maximum coefficient of the likelihood function will be estimated to be 1, and the rest will be 0.

The compositional semantics parameter β will be computed by first computing the coefficients of each of the β_{kl} parameters in the likelihood function. In this computation, the computed value of β from the last training iteration will be used as the prior $g(\beta)$. Then, if $k=l$ for the β_{kl} with the maximum coefficient, β is estimated to be 1. Otherwise, it is estimated to be 0.

E. Training the Syntax Model

To train the syntax model, the estimated α values associated with each word are incorporated into a vector, and used to train the syntax HMM. Thus, the HMM will first be trained with $[\alpha_{11},\alpha_{21},\alpha_{31}]$, and then $[\alpha_{12},\alpha_{22},\alpha_{32}]$. A third feature vector is then used to train the HMM to signify the end of the sentence. This feature vector should signify no information about the perceptual features, and so will be [0,0,0].

The values of α used in the computation of β must be obtained from the HMM in the syntax model. Since the state sequence is unknown, these values for α must be estimated from the state probabilities while training the HMM. Thus, the estimates for α are computed according to (3).

$$\alpha_{ij} = \sum_{k}\alpha_{ik}\,P(\text{state} = k \text{ after word } j) \quad (3)$$

The β parameter is trained by using the linear update equation in (4), where ε is the learning rate.

$$\beta^{(k+1)} = \beta^{(k)} + \varepsilon(\beta - \beta^{(k)}) \quad (4)$$

VI. EXPERIMENTS AND RESULTS

Numerical simulations were performed using Matlab as a basic sanity check of the ideas presented (not as a conclusive working model). The HMM in the syntax model was initialized to have a uniform state-transition matrix. Small biases were used in the initial output distributions of two of the three states to avoid the convergence to a degenerate solution. The compositional semantics parameter was initialized to have no bias.

A. Training

The syntax model was trained using purely symbolic data. In the training scenario, three perceptual features were used. The number of objects and actions were set to two each.

In each training iteration, the number of perceptual feature vectors used (representing the number of objects observed) is randomly chosen to be 1 or 2. Then, the objects and their actions are chosen randomly. If two feature vectors are used, the objects in each are different. The perceptual features associated with each lexical category are generated with gaussian distributions. The two members of each category use means of 0 and 1, and variance of 1.

The perceptual model is assumed to have been previously learned, but is similar to the distributions learned by the semantic associative memory in the robot. In this model, the distributions in each perceptual state are the same as the generating distributions described above for the relevant features. For the irrelevant features, distributions with mean ½ and variance of 10 are used.

The word and name models are deterministic in the experiments. Each iteration of the training represents one two-word sentence spoken in context. The models were trained until the parameters reached their final values.

B. Experiments

In the first experiment, the objects generate the first and third perceptual features and the actions generate the second feature. The second experiment uses the same setup, except that the third perceptual feature is generated randomly.

C. Results

The parameters of the syntax model converged to the desired values. In the first experiment, the state output parameters reflect that state one is a delimiter, state 2 describes the first and third perceptual features, and state 3 describes the second feature. These can be seen in Figures 5-7. Figure 8 shows that the HMM converges to a left-to-right model. This, along with the state output parameters, can be interpreted as the production rule "S→NV". The compositional semantic parameter reflects that the words in the production rule describe the same object, as shown in Figure 9.

The parameters in the second experiment also converged to their desired values. With the exception of Figure 8, the convergence of the parameters in this experiment was similar to those of the first experiment, using approximately 3000 training iterations to converge. Figure 10 shows the convergence of the third perceptual features, which was

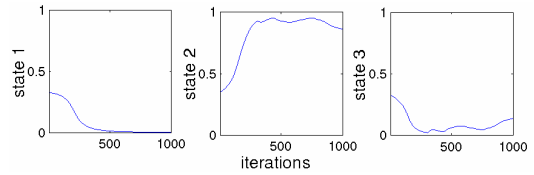

Figure 5: Probability that each state describes the first perceptual feature, 1st experiment

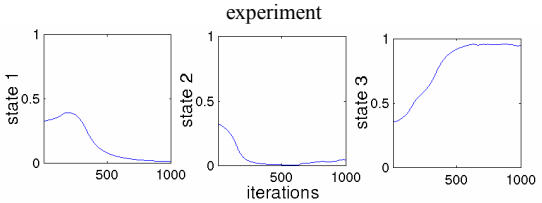

Figure 6: Probability that each state describes the second perceptual feature, 1st experiment

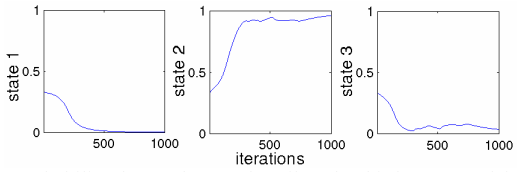

Figure 7: Probability that each state describes the third perceptual feature, 1st experiment

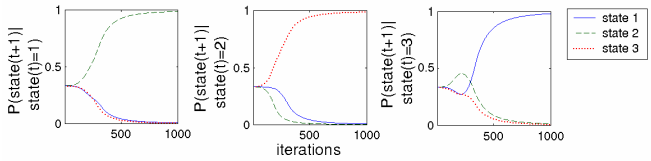

Figure 8: State transition parameters, 1st experiment

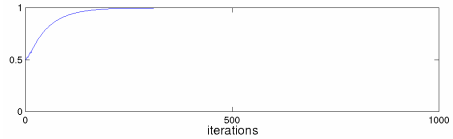

Figure 9: Compositional semantics parameter, 1st experiment

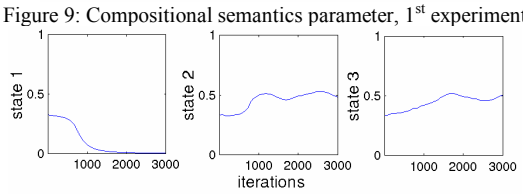

Figure 10: Probability that each state describes the third perceptual feature, 2nd experiment

not described by either word. Here, because the restriction was made that each feature must be described by one word, the parameter converged to mean that it is undecided which of the words describes this feature.

VII. FUTURE WORK

The cognitive models proposed in this paper will be implemented in LAR-UIUC's robotic framework. The robotic experiments will first involve the robot learning the names of three objects ("cat", "ball", and "dog") and three actions ("move", "stay", and "gone"). The robot will explore its environment, and perform actions on the objects (specifically, bumping into them) and a human experimenter will speak a two-word sentence with the production rule S→NV that describes what has happened to the object.

REFERENCES

[1] Y. Sugita and J. Tani, "A holistic approach to compositional semantics: a connectionist model and robot experiments," Advances in Neural Information Processing Systems 16 (NIPS2003), Vancouver and Whistler, Canada, (Eds) S. Thrun, L. K. Saul and B. Scholkopf, The MIT Press, pp.969-976, 2004.

[2] D. Roy, "Learning visually-grounded words and syntax for a scene description task," Computer Speech and Language, 16(3), pp353-385, 2002.

[3] D. Roy, "Learning visually grounded words and syntax of natural spoken language," Evolution of Communication, 4(1), pp33-56, 2001.

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

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

[6] F. Chang, G. Dell, and K. Bock, "Becoming Syntactic," Psychological Review, under review.

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

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

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

[10]R. L. Cave and L. P. Neuwirth, "Hidden Markov models for English," *Hidden Markov Models for Speech*, vol. IDA-CRD, Princeton, NJ, 1980.

[11]C. Yu, "The Emergence of Links between Lexical Acquisition and Object Categorization: A Computational Study", Connection Science, 17(3-4), 381-392,2005.

[12]J. Weng, "Developmental Robotics: Theory and Experiments", International Journal of Humanoid Robotics, vol. 1, no. 2, 2004.

[13]R. Lin, "Manifold learning from time series," Ph.D. Thesis, University of Illinois at Urbana-Champaign, 2005.

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

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

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

[17]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, pp404-411.

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

[20]R. Brown, *Psycholinguistics*, New York: The Free Press, 1970.

[21]M. Tomasello, "Pragmatic Contexts for Early Verb Learning," in *Beyond Names for Things*, ed. by M. Tomasello and W.E. Merriman, pp 115-146, 1995.

[22]D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Englewood, New Jersey: Prentice-Hall, 2000.

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