Using Syntax to Learn Semantics: An Experiment in Language Acquisition with a Mobile Robot

Tim Oates, Zachary Eyler-Walker and Paul R. Cohen
Computer Science Department, LGRC
University of Massachusetts, Box 34610
Amherst, MA 01003-4610
{oates,zwalker,cohen}@cs.umass.edu

Abstract

Children learn natural languages by hearing utterances while interacting with their physical environment. We investigate one aspect of language acquisition by similarly situated, embodied artificial agents - using information about syntax to learn linguistically relevant semantic features. The agent is assumed to have no innate knowledge of syntax, and instead leverages the weak information about syntax available in word co-occurrences. Similarity of context (i.e. the surrounding words) is used to hierarchically cluster words, with clusters corresponding to sets of words that are similar syntactically and, often, semantically. The goal is to identify semantic features captured by the clusters. The leaves of the hierarchy are individual words, which are semantically very specific, and movement up the hierarchy leads to less specificity. The results of an experiment are discussed in which human subjects generated unrestricted natural language utterances to describe the activities of a Pioneer1 mobile robot. The combination of word clustering on this corpus and a common subsequence algorithm applied to the time series of sensor values recorded by the robot made it possible for the Pioneer1 to learn a variety of semantic features.

Introduction

Children learn a staggering number of semantic features that are relevant to their native language. For example, consider differences in the meanings of individual words, which can hinge on semantic features at various levels of detail. The meanings of *push* and *shove* are similar in that both involve *contact* between two entities and the application of *force* by one entity to the other. However, the meanings of these words differ in the magnitude and duration of the force that is applied. Depending on the level of detail considered, these two words have either the same semantic features or similar, but different, features.

This paper investigates how situated, embodied artificial agents can use the same inputs available to children, i.e. utterances and the physical context in which they occur, to learn the same kinds of linguistically relevant semantic features that children learn. We assume that the meanings of words are learned in an associationist manner as proposed by John Locke (Locke 1975), through repeated exposure to utterances of a word in the presence of its referent. Later sections describe such a learning mechanism that has been implemented and tested on a Pioneer1 mobile robot. This form of learning, driven by occurrences of individual words, yields highly specific semantic features, such as the precise magnitude and duration of the force that needs to be applied to *shove* something.

In contrast to semantic features specific to individual words, we are particularly interested in the acquisition of more abstract semantic features, such as the presence or absence of *contact* between two objects. Features of this kind are typically shared by many different words. The importance of this type of feature can be seen in the work of Talmy, who investigated the semantic content of verbs describing the motion and location of objects across a large number of languages (Talmy 1985). He found that each language selected a small set of semantic elements to express in the meanings of such verbs. For example, English encodes the fact that motion occurs and the manner or cause of motion (e.g. *roll*, *bounce* and *run*), while Spanish encodes the fact that motion occurs and the path that motion follows (e.g. *entró*, which means to move in, and *salió*, which means to move out). If the learner can determine which of these semantic features are expressed by verbs in the target language, the meanings of new verbs can be learned more quickly by focusing on only those aspects of the physical context that are relevant for associationist learning.

In addition to the role that semantic features play in providing the meaning of words, they play a role in syntax as well. Verbs with similar meanings appear in similar syntactic constructions (Zwicky 1971). Pinker goes so far as to suggest that a very small number of abstract semantic features, such as *motion* and *contact*, can be used to formulate rules for determining verbs' argument structures (Pinker 1989).

The question that remains to be answered is how a situated, embodied language learner can identify linguistically relevant semantic features. As noted previously, associationist learning driven by the occurrence of individual words yields highly specific semantic features. To identify more abstract semantic features we take advantage of the relationship between meaning and syntax mentioned earlier. The learner is assumed to have no innate knowledge of syntax, and instead leverages the weak information about syntax available in word co-occurrences. Given a corpus of sentences in a language, similarity of context (i.e. the surrounding words) can be used to hierarchically cluster words. Clusters correspond to sets of words that are similar both syntactically and semantically by virtue of the relationship between syntax and semantics. The leaves of the hierarchy are individual words, and movement up the hierarchy leads to clusters containing increasingly many words whose shared

semantic features are necessarily more abstract. Locke essentially proposed using associationist mechanisms to learn semantic features at the leaves of this hierarchy, but they can also be used to learn the features shared by words further up as well. Later sections describe how we operationalize this idea to identify abstract semantic features that Talmy and Pinker argue are so vitally important in language acquisition.

The remainder of the paper is organized as follows. The next section describes two learning mechanisms, one for identifying clusters of semantically related words given a set of utterances and one for learning the semantic features shared by words in clusters in an associationist manner. The latter mechanism assumes that the language learner's knowledge of the context in which utterances occur is provided by a set of sensors that produce values over time. We then describe the results of an experiment in which human subjects generated unrestricted natural language utterances to describe the activities of a Pioneer1 mobile robot. The combination of word clustering on this corpus and associationist learning applied to the time series of sensor values recorded by the robot made it possible for the Pioneer1 to learn a variety of semantic features. The final section concludes and points to future work.

Learning Mechanisms

As described in the previous section, learning linguistically relevant semantic features at different levels of abstraction is a two stage process. First, the learner constructs a cluster hierarchy of semantically related words by grouping words that occur in similar syntactic contexts. Second, an associationist learning mechanism is used to identify the semantic features shared by the words in a cluster based on the current physical context when members of the cluster are uttered. The level of abstraction inherent in the semantic features identified in this manner depends on the size of the cluster, with small clusters containing few words leading to highly specific features, and large clusters containing many words leading to highly abstract features. This section describes solutions to both of these learning problems.

Word Clustering

Many different methods for clustering words based on similarity of syntactic context exist (Brown *et al.* 1992; Hindle 1990; Pereira, Tishby, & Lee 1993; Redington, Chater, & Finch 1993). Consider an example of the kind of output they produce based on a simple grammar that generates syntactically and semantically well-formed sentences using the nouns, verbs and adjectives shown below:

- nouns (people and things) – boy, girl, cat, dog, dish, chair, block, shoe
- nouns (places) – hallway, kitchen, doorway
- verbs – turn, start, stop, avoid, follow, bump, hit, push, move
- adjectives – small, large, red, blue

Sentences produced by the grammar include "The boy avoided the large dog" and "The girl turned", but do not include semantically ill-formed sentences such as "The chair hit the cat."

Figure 1 shows one part of the cluster hierarchy produced by the method described in (Brown *et al.* 1992) on a corpus containing 500 sentences from the grammar. All of the words at the leafs of this part of the hierarchy are verbs. The set of all verbs is divided into two subsets, those that take a direct object and those that do not. The verbs that take a direct object are further divided into those that involve contact and those that do not. Finally, the former set is divided into punctual and non-punctual verbs. Remarkably, all of this structure was identified given only a corpus of sentences in the language and no a priori knowledge of syntax. Although it is a trivial matter for an English speaker to label the interior nodes in Figure 1 as *punctual* or *contact* or *direct object*, out goal is to allow a situated, embodied agent to, in effect, learn these labels for itself given exposure to sentences and the context in which they are uttered.

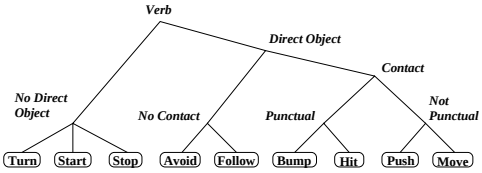


Figure 1: Part of a cluster hierarchy based on a simple grammar.

The clustering method used in this paper is based on (Brown *et al.* 1992). The algorithm begins by creating one cluster for each unique word in the corpus. The mutual information between each pair of clusters, $\mathcal{C}_i$ and $\mathcal{C}_j$, is then computed as follows:

$$q(i,j) = p(\mathcal{C}_i\mathcal{C}_j) \log \frac{p(\mathcal{C}_i\mathcal{C}_j)}{p(\mathcal{C}_i)p(\mathcal{C}_j)}$$

$p(\mathcal{C}_i)$ is the probability of a word in cluster i occurring in the corpus and $p(\mathcal{C}_i\mathcal{C}_j)$ is the probability of a word in cluster i preceding a word in cluster j in the corpus. The total mutual information between pairs of clusters is then:

$$I = \sum_{i,j} q(i,j)$$

Merging two clusters reduces I, and the pair of clusters that results in the smallest loss of mutual information is merged. This merging process repeats until a single cluster containing all unique words in the corpus is created. For example, at some point during the clustering

process on the sample corpus mentioned earlier, the clusters $\{bump\}$ and $\{hit\}$ were merged to create the cluster $\{bump, hit\}$, as were the clusters $\{push\}$ and $\{move\}$ to create $\{push, move\}$. These larger clusters were later merged to create a single cluster containing all verbs involving contact – $\{bump, hit, push, move\}$.

From Sensors to Semantics

Given a word cluster, $\mathcal{C}$, how might a situated, embodied artificial agent learn the semantic features associated with that cluster? For the sake of concreteness, assume the learner is a mobile robot and that its knowledge of the physical environment is provided by a set of sensors. We assume that the presence of a word's referent in the physical environment induces a pattern, $\mathcal{P}$, in the time series produced by the robot's sensors. Let $p(\mathcal{P}|\mathcal{C})$ be the probability of the pattern occurring in the sensor data gathered when a member of $\mathcal{C}$ is uttered, and let $p(\mathcal{P}|\overline{\mathcal{C}})$ be the probability of the pattern when a member of $\mathcal{C}$ is not uttered. Under the reasonable assumption that words are uttered more frequently when their referent is present than when it is absent, it will be the case that $p(\mathcal{P}|\mathcal{C})$ is significantly different from $p(\mathcal{P}|\overline{\mathcal{C}})$.

Each time a word $w \in \mathcal{C}$ is uttered, the robot can record the values of its sensors over a window of time centered on the occurrence of the word. The result is a set of sensor time series that co-occurred with utterances of words in $\mathcal{C}$. The referent of w, and thus $\mathcal{P}$, may appear before, after or at the same time that w is uttered, and the relative timing of the two may change from one utterance to another. There is initially a total lack of knowledge concerning the location and the nature of $\mathcal{P}$ in the individual time series. The task facing the learner is to identify $\mathcal{P}$ given a set of time series gathered in this manner, and thereby to identify the semantic features associated with $\mathcal{C}$. We call this process finding *distinctive* subsequences because patterns identified in this manner serve to distinguish time series gathered in the presence of words in $\mathcal{C}$ from those gathered in their absence.[1]

The first step toward the discovery of variable-length distinctive subsequences is the identification of a set of fixed-length subsequences that capture patterns occurring in the robot's sensors. Let $\mathcal{S}$ denote the robot's sensor array. Fixed length patterns are identified by randomly sampling sequences of length L, called L-sequences, from $\mathcal{S}$. Given n L-sequences and a measure of similarity between multivariate, real-valued time series, we construct an n-by-n similarity matrix. The matrix is used to cluster the L-sequences and to select a prototype from each cluster by finding the sequence that minimizes the average distance to all other sequences in the cluster. The measure of similarity that we use is Dynamic Time Warping (DTW) (Sankoff & Kruskall 1983). DTW is a generalization of classical algorithms for comparing discrete sequences (e.g. minimum string edit distance (Cormen, Leiserson, & Rivest 1990)) to sequences of continuous values. It was used extensively in speech recognition, a domain in which the time series are notoriously complex and noisy, until the advent of Hidden Markov Models, which offered a unified probabilistic framework for the entire recognition process (Jelinek 1997).

Prototypical L-sequences are obtained in this manner by sampling from $\mathcal{S}$ without regard to whether words in $\mathcal{C}$ occur. Because distinctive patterns, by definition, occur more or less frequently in the presence of words in $\mathcal{C}$ than in their absence, sampling in this manner ensures that clustering has access to the full range of patterns that can occur within a window of width L. Given k prototypical L-sequences, $\mathcal{P}_1$ through $\mathcal{P}_k$, we now want to determine which of them are distinctive. That is, we want to identify those prototypes for which $p(\mathcal{P}_i|\mathcal{C})$ is significantly different from $p(\mathcal{P}_i|\overline{\mathcal{C}})$.

Estimation of $p(\mathcal{P}_i|\mathcal{C})$ and $p(\mathcal{P}_i|\overline{\mathcal{C}})$ requires a set of sequences obtained from $\mathcal{S}$. This set must contain some sequences that co-occurred with $\mathcal{C}$ and some that did not. A window of width L is passed over each sequence, and DTW is used to determine which of the k prototypes is most similar to each of the resulting L-sequences. The L-sequences obtained in this manner are drawn from larger sequences that either did or did not co-occur with $\mathcal{C}$. If an L-sequence is most similar to prototype i and the former case holds, the counter $n_{i,\mathcal{C}}$ is incremented. If the latter case holds the counter $n_{i,\overline{\mathcal{C}}}$ is incremented. It is then a simple matter to estimate the probabilities of interest:

$$p(\mathcal{P}_i|\mathcal{C}) = \frac{n_{i,\mathcal{C}}}{\sum_{j=1}^{k} n_{j,\mathcal{C}}} \qquad p(\mathcal{P}_i|\overline{\mathcal{C}}) = \frac{n_{i,\overline{\mathcal{C}}}}{\sum_{j=1}^{k} n_{j,\overline{\mathcal{C}}}}$$

To determine whether these probabilities are significantly different we use a two-tailed t-test as follows. Consider a random variable whose value is either 1 or 0 depending on whether an L-sequence matches or does not match the i^{th} prototype. Given two such random variables, one associated with only those L-sequences that co-occurred with $\mathcal{C}$ and one associated with the L-sequences that did not, it is easy to compute their means and variances. A standard t-test is then used to determine the probability of making an error in rejecting the null hypothesis that the means are the same, i.e. that prototype i is not a distinctive L-sequence. If that probability is below a given significance level, then the L-sequence is said to be distinctive.

Given fixed-length prototypical L-sequences, variable-length distinctive subsequences that span more than L time steps are identified as follows. Note that prototype i can be distinctive for one of two reasons. Either $p(\mathcal{P}_i|\mathcal{C}) \gg p(\mathcal{P}_i|\overline{\mathcal{C}})$ or $p(\mathcal{P}_i|\mathcal{C}) \ll p(\mathcal{P}_i|\overline{\mathcal{C}})$. If the former

[1] For more information on the algorithm for finding distinctive subsequences see (Oates 1999).

condition holds we say that all L-sequences that are more similar to $\mathcal{P}_i$ than any other prototype are *frequent* L-sequences. Such L-sequences occur more frequently in the presence of $\mathcal{C}$ than in its absence. If the latter condition holds we say that all of the L-sequences that are more similar to $\mathcal{P}_i$ than any other prototype are *infrequent* L-sequences. Finally, L-sequences matching prototypes that are not distinctive are said to be *neutral*. A subsequence of length greater than L is frequent if all of the L-sequences that it contains are either frequent or neutral. The subsequence is infrequent if those L-sequences are either infrequent or neutral. In both cases, the subsequence is distinctive. It is possible to locate all of the frequent and infrequent variable-length subsequences in a larger time series in time that is linear in the length of the time series.

Experiments

An experiment was designed and executed to gather data for the clustering algorithms in which a Pioneer1 mobile robot was filmed completing a series of simple actions. Several volunteers watched the film and wrote sentences describing the robot's behavior. The goal of this experiment was to evaluate the utility of the algorithms presented in the previous section given inputs very similar to those available to human language learners.

The Pioneer1 is a small, wheeled robot, filling a space approximately two feet long, one-and-a-half feet wide, and one foot tall. It has two independently driven wheels near its front and a swivelling castor in the rear. It is able to act on objects in its environment through a single, two-fingered parallel gripper in its front. This gripper can open approximately eight inches, and is able to lift objects several inches off the floor. As set up for this project, the robot was not able to independently open or close and raise or lower its gripper; rather, when the gripper opened it was simultaneously lowered to its low position, and when it closed it was raised to its high position.

The robot is also fitted with a variety of sensors for observing its environment. The most prominent are an array of seven Polaroid-type sonar distance sensors, five pointed in a shallow arc approximately ten degrees on either side of straight ahead, and one each pointing directly towards the right and left. A video camera on the top front of the robot faces forwards. It provides a wide-angled view and can be aimed either straight ahead or at an angle down towards its gripper. For the purposes of this experiment, a "blob-vision" system was in place which picks out the nearest object colored either red or blue, and excludes everything else. The final set of sensors on the robot are at its grippers. There are two bump switches at the ends of the fingers, activated whenever they are driven against a suitably immobile object. Lastly, there is a light beam switch between the two gripper fingers, which is registers whenever an object enters the gripper.

Before the robot could be filmed engaging in various behaviors, it was necessary to develop a list of the actions it is able to make in a simple environment with a small number of props. Having done this a number of short sentence-scripts was drafted, spanning a range of complexities. This was pared down after taking into consideration the fact that the robot must have enough sensory information to observe the salient features of the actions. Among other things, this led to an exclusion of scripts with objects other than the robot as actor. Another aim of the script generation process was to include the complements of as many actions as possible, which it was thought might lead to richer word clusters later. By complementary we mean all the possible actions of a particular category or on a particular object; for example, the robot went into, around, through, and came out of the box. Below are a few of the final scripts:

- The robot spun.
- The robot went into the box.
- The robot touched the blue cup.
- The robot gave the red ball to the car.
- The robot moved the red cup from the red mat to the blue mat.

Ultimately, 41 scenes were filmed, with the order of the scripts randomized beforehand. Each scene involved the robot acting out one sentence-script, and was framed by several seconds of empty screen to delineate scene changes. Due to a certain degree of imprecision in the manual control of the robot, it was not always possible to exactly perform the actions scripted– a touch might become a nudge, for example.

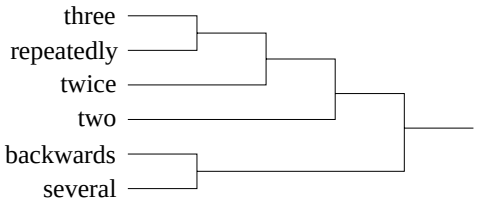
The next phase was to collect some volunteers to view the video and write sentences describing the robot's behavior. The volunteers were instructed to begin all sentences with "The robot ..." and to keep the sentences as simple as reasonably possible, perhaps as though they were speaking to a young child or playing a text-adventure. Otherwise the sentences they produced were unrestricted. Eight volunteers produced 328 sentences describing the 41 scenes. Below are some of the sentences produced by different subjects after viewing one of the scenes scripted above:

- The robot picked the red ball up and put it down in front of the red car.
- The robot put the ball next to the car.
- The robot lifted the red ball and placed it next to the car.
- The robot carried the ball to the car.

Before building a cluster hierarchy from the collection of sentences, a few minor transformations were made on them. Particularly, "negative" phrases were removed.

For example, "The robot went around the box without touching the red mat" became "The robot went around the box." Object names were also standardized, though all other terms were left untouched. Thus, "bin," "trash bin," "wastebasket" and "trash can" all became "trash-can."

Below is a fragment of the final hierarchy showing the order in which clusters are merged: Note that the root



of this fragment is a single cluster containing words that indicate repetition. Several clusters drawn from other interior nodes of the hierarchy are shown below:

{*middle, center, top, left*}
{*dropped, gave, carried, touched, set, placed, put*}
{*into, toward, through, over, by, inbetween, between, under, near, against*}

These clusters all consist of words that are not synonyms but have more abstract semantic relationships, such as position, verbs about moving objects, and prepositions. In the first cluster there is also an erroneous clustering of "backwards" with the words specifying repetition. Both the nature of good clusters and the appearance of bad clusters can be understood in terms of the clustering method. Because it relies on extremely local information, i.e. bigrams, it tends to pick out fairly superficial syntactic features. It is fortunate that semantically similar words tend to be used in similar syntactic contexts; nevertheless, some otherwise entirely dissimilar words can be found in identical contexts. Both these cases will lead to the clustering of terms, correctly or incorrectly, as above. Unfavorable clusterings are more likely to be produced from smaller input sets, such as that available here. Given larger corpora in which terms appear in more varied contexts, as is the case with human language learners, this problem will be minimized.

Given the cluster hierarchy constructed from the eight human subjects' natural language sentences, the next task was to identify the semantic features of word clusters. Because words are typically uttered more frequently in the presence of their referent than in its absence, patterns in the robot's sensors that occur significantly more frequently in the presence of words in a cluster than in their absence are deemed to be semantic features of the cluster. That is, the problem of identifying linguistically relevant semantic features is cast in terms of identifying distinctive subsequences in the robot's sensor data.

During the filming of each scene the values produced by the robot's sensors were recorded at a rate of 10Hz, and these time series were used to identify prototypical patterns in the sensor data. Given a word cluster, the time series were separated into two sets based on whether any of the human subjects used a member of the cluster when describing the associated scenes. These two sets of time series were used to determine which of the prototypes were distinctive for that cluster, and to identity variable length frequent subsequences in the time series that co-occurred with members of the word cluster.

This procedure was applied at every node in the fragment of the cluster hierarchy shown in Figure 2. Rather than using the values of all of the robot's sensors, many of which are irrelevant to the words in question, we used the time series produced by a single sensor that encodes the state of the robot's gripper.[2] As noted previously, the gripper can be up and closed (gripper state = 0.2), down and open (gripper state = 1.0), or moving between these two positions (gripper state = 0.4).

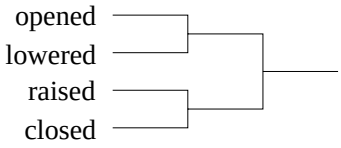


Figure 2: A fragment of the cluster hierarchy formed from the 328 natural language sentences produced by the human subjects. This part of the hierarchy contains words describing changes in the state of the robot's gripper.

The frequent sequences found in the gripper state sensor are shown in Figure 3. The leftmost plot in that figure shows the frequent subsequences associated with the clusters {*closed*}, {*raised*} and {*closed, raised*}. In all three cases, a single frequent subsequence was found that involved the gripper state transitioning from a value of 1.0 to a value of 0.2. This subsequence occurs exactly when the robot closes and raises its gripper. Because of the way the Pioneer's gripper operates, saying the the robot "closed" its gripper and that it "raised" its gripper denote the same transition, a fact that was identified by the distinctive subsequence algorithm. Likewise, the clusters {*opened*}, {*lowered*} and {*opened, lowered*} all yielded the same frequent subsequence, one in which the gripper state transitions from 0.2 (up and closed) to 1.0 (down and open). Finally, the only frequent subsequence identified for the cluster {*closed, raised, opened, lowered*} is a gripper state of 0.4, which occurs only when the gripper is in motion between one of its two resting states. That is, the semantic feature shared by all of

[2]Developing automated methods that will allow the robot to determine which sensors are relevant for a given word cluster is a non-trivial problem, and is the focus of ongoing research.

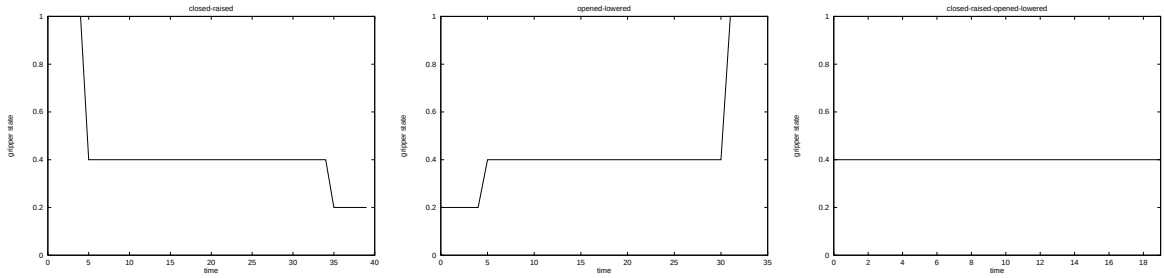


Figure 3: Distinctive time series identified for the clusters {*closed, raised*} (leftmost plot), {*opened, lowered*} (middle plot), and {*closed, raised, opened, lowered*} (rightmost plot).

the words in this cluster is that the gripper is moving between resting states.

Finally, the semantics of the word *pushed* were identified by looking for frequent subsequences in a subset of the robot's sensors containing the gripper state, the status of the break beam between the gripper paddles, and the status of the bump sensors on the tips of the gripper paddles. Three different frequent subsequences were identified in this case:

- gripper down, break beam on, bump switch off
- gripper down, break beam off, bump switch on
- gripper up, break beam off, bump switch on

These three configurations of the robot's sensors correspond to the following situations:

- pushing a small object between the robot's grippers on the floor
- pushing a large object that will not fit between the gripper paddles with the gripper down
- pushing an object with the gripper up and closed

There are three distinct ways that the robot can push objects, each of which was described with the word *pushed* by at least one subject, and the sensor time series that result from these situations were all identified as frequent subsequences.

Discussion

This paper presented a method that allows situated, embodied artificial agents to learn linguistically relevant semantic distinctions given the same input available to children, utterances and the physical context in which they occur. The utility of the method was evaluated in an experiment in which human subjects produced natural language utterances to describe the activities of a Pioneer1 mobile robot. Future work will involve attempting to scale the experiment reported herein to include vastly richer physical and linguistic interactions with the robot.

References

Brown, P. F.; deSouza, P. V.; Mercer, R. L.; Pietra, V. J. D.; and Lai, J. C. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18(4):467–479.

Cormen, T. H.; Leiserson, C. E.; and Rivest, R. L. 1990. *Introduction to Algorithms*. The MIT Press.

Hindle, D. 1990. Noun classification from predicate-argument structure. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 268–275.

Jelinek, F. 1997. *Statistical Methods for Speech Recognition*. MIT Press.

Locke, J. 1975. *An Essay Concerning Human Understanding*. Oxford : Clarendon Press. Original work published in 1690.

Oates, T. 1999. Identifying distinctive subsequences in multivariate time series by clustering. Submitted to ICML-99.

Pereira, F. C. N.; Tishby, N. Z.; and Lee, L. 1993. Distributional clustering of english words. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, 183–190.

Pinker, S. 1989. *Learnability and Cognition: the Acquisition of Argument Structure*. MIT Press.

Redington, F. M.; Chater, N.; and Finch, S. 1993. Distributional information and the acquisition of linguistic categories: A statistical approach. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*.

Sankoff, D., and Kruskall, J. B. 1983. *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley.

Talmy, L. 1985. Lexicalization patterns: Semantic structure in lexical forms. In Shopen, T., ed., *Language Typology and Syntactic Descriptions III*. Cambridge University Press. 57–149.

Zwicky, A. 1971. In a manner of speaking. *Linguistic Inquiry* 11:223–233.