Machine Learning Comprehension Grammars for Ten Languages

Patrick Suppes*
Stanford University

Michael Böttner*
Stanford University

Lin Liang*
Stanford University

Comprehension grammars for a sample of ten languages (English, Dutch, German, French, Spanish, Catalan, Russian, Chinese, Korean, and Japanese) were derived by machine learning from corpora of about 400 sentences. Key concepts in our learning theory are: probabilistic association of words and meanings, grammatical and semantical form generalization, grammar computations, congruence of meaning, and dynamical assignment of denotational value to a word.

1. Introduction

Our approach to machine learning of language combines psychological, linguistic, and logical concepts. We believe that the five central features of our approach—probabilistic association of words and meanings, grammatical and semantical form generalization, grammar computations, congruence of meaning, and dynamical assignment of denotational value to a word—are either new, or are new in their present combination. An overview of these concepts and related ones is given in Section 2.1. Two prior papers describing this approach, first presented at two conferences in 1991, are Suppes, Liang, and Böttner (1992) and Suppes, Böttner, and Liang (1995).

Using the theory embodying the concepts just listed, we report on our machine learning program of corpora from ten natural languages. Following our earlier work, we use a robotic framework. The computer program based on the theory learns a natural language from examples, which are commands occurring in mechanical assembly tasks (e.g., *Go to the screw, Pick up a nut, Put the black screw into the round hole*). A major improvement here, in comparison to Suppes, Böttner, and Liang (1995), is that the association relation is generalized from a unique correspondence between words and the program's internal representation of their meaning to a many-to-one relation, which permits different words to be associated to the same internal representation. This change is particularly important for the purpose of capturing case variation in word forms in inflecting languages such as Russian or German.

The robotic framework and the associated corpora we test our program on are certainly restricted, although we have implemented our learning program on Robotworld, a standard robot used in academic settings for development purposes. In the present paper, however, we have deliberately formulated the general learning axioms of our theory so they do not depend on the robotic framework. The axioms are meant to apply to many kinds of systematic language use, more or less in the sense of sublanguages, (see Kittredge and Lehrberger [1982]). We are already deep into our next

* CSLI, Ventura Hall, Stanford CA 94305-4115

area of application, machine understanding of physics word problems, and we have not needed to change the general formulation of our theory to accommodate this quite different problem of language learning.

In this paper, we first describe our theory of machine learning of natural language (Section 2), and then describe the corpora in ten languages that we used for experimental purposes (Section 3). The languages are: English, Dutch, German, French, Spanish, Catalan, Russian, Chinese, Korean, Japanese. In Section 4 we describe some empirical results, especially the comprehension grammars generated from learning the languages. Finally, in Section 5 we discuss related work and the most pressing unsolved problems.

2. Theory

The theory that underlies our learning program is given in terms of a system of axioms. We begin with a general formulation, which is then made more special and technical for the robotic framework of this paper.

Any learning program that is given the counterpart of the command *Get the black screw* in French *Prends une vis noire,* Russian *Voz'mi chjornuj vint,* or Japanese *Kuroi nejikugi o tore* finds itself confronted with at least the following problems: (i) to learn the meanings of words, (ii) to learn a grammar from a training set for each utterance to be comprehended, and (iii) to learn the semantic structure of each utterance. Our theoretical solution to these problems will become clear in what follows.

2.1 Central Concepts of the Theory

We give here an informal characterization of the key concepts used in our theory.

Association. This is the key learning concept. We use the classical concept of association to establish the connection between unknown words in a language and their meaning. The principle of association goes back at least to Aristotle, and certainly was used extensively by eighteenth-century philosophers like Hume long before psychology had become an experimental science. The fundamental role of association as a basis for conditioning is thoroughly recognized in modern neuroscience and is essential to the experimental study of the neuronal activity of a variety of animals. For similar reasons, its role is just as central to the learning theory of neural networks, now rapidly developing in many different directions.

We have not made explicit use of neural networks, but have worked out our theory of language learning at a higher level of abstraction. In our judgment, the difficulties we face need to be solved before a more detailed theory is developed. The choice of some level of abstraction is inevitable in all work of the present kind—a fact not always appreciated by interested bystander. Whatever the level of abstraction, there is one general issue about association that must be faced: is association probabilistic or deterministic? For us, the forming of associations is probabilistic rather than deterministic, which is a reflection of the complexity of the underlying process.

Here, formally, association is a binary relation between commands, words, and grammatical forms, on the one hand, and their corresponding representations in the internal language of the program, on the other hand. All words in each of the ten languages of the experiment are completely unknown to the program at the beginning of learning. The internal language, defined in Section 2.3, is not learned, but given from the beginning.

Working memory. The learning program has a short-term working memory for processing the command it is presented. The memory holds its content for the time period of a single trial. The first group of learning axioms describes the association computations that take place in working memory during the course of a trial.

Long-term memory. This memory also can change from trial to trial, but it stores associations and grammatical forms that remain unchanged when they are correct for the application being considered. The way the state of long-term memory changes from trial to trial is described in Section 2.4.1, in the second set of learning axioms.

Generalization. A distinct principle of generalization generates grammatical forms and their associated semantic forms. The methods used combine those of context-free grammars and model-theoretic semantics for formal languages.

The concept of generalization is widely used in psychological theories of learning. The kind of generalization used here is restricted to the generation of grammatical forms and grammatical rules from concrete utterances. For example, the phrase *the nut* generalizes to the grammatical form *the O*, where *O* is the category of objects. (More elaborate examples are considered later.)

Memory trace. When a generalization of any of the several kinds we use is made, the particular word association on which it is based is stored with it in long-term memory, as the memory trace justifying the generalization.

Denotational value. When a child learning a first language, or an older person learning a second language, first encounters utterances in that new language, nondenoting words are not marked in any uniform way. There is some evidence that various prosodic features are used in English and other languages to help the child, but in any case their denotational role must be learned.

Here a separate learning procedure is introduced to compute dynamically the denotational value of each word, with the limit being 1 for having a denotation and 0 for being nondenoting. Denoting words such as *nut* refer to elements of our categories; good examples of nondenoting words are definite and indefinite articles. Intuitively, only denoting words should acquire associations to elements of the environment, including possible actions, as represented internally. Several mean learning curves for denotational value are presented in Section 4.

Congruence. By using a concept of semantic congruence inspired by geometry, we simplify the grammars and at the same time permit direct comparisons across languages. The intuitive idea of such congruence is simple: two strings of a natural language are congruent when they have identical representations in the internal language.

Various strong and weak senses of congruence can be used to get varying degrees of closeness of meaning. The idea is not to be caught in the search for a single concept of synonymy, just as in modern geometry we are not caught in a single concept of congruence. In affine geometry, for example, there is a weaker sense of congruence than in Euclidean geometry, but it is also easy to get a commonsense notion of congruence that is stronger than the Euclidean one, namely congruence that requires sameness of orientation. Our use of congruence of meaning in this article is restricted. In particular, we only analyze congruence of nondenoting words.

Comprehension grammar. Most linguistic analysis is concerned with grammars detailed enough to produce natural utterances of the language being studied. A comprehension grammar, in contrast, as we characterize it here, can generate a superset of utterances. The rules are required only to lead to the correct semantic interpretation of an utterance of the language. Robots, like very young children, can easily have the capacity to understand language before they can produce it. Although it is difficult and subtle to collect accurate and anything-like-complete data on the comprehension grammar generated by very young children, the evidence is overwhelming that they comprehend much more than they can produce.

2.2 Background Assumptions

We state informally as background assumptions two essential aspects of any language learning device. First, there is the problem of how the internal representation of an utterance heard for the first time is generated by the learner. Second, at the other end of the comprehension process, so to speak, there is the problem of generating a semantic interpretation of a new utterance, but one that falls within the grammar and semantics already constructed by the learner.

In any complete theory, both of these processes require thorough formal analysis, but as will become clear, this analysis is not necessary for the framework of this article. We give only a schematic formulation here.

1. **Association by contiguity.** When a learner is presented a sentence that it cannot interpret, it associates the utterance to patterns in its contiguous environment whose internal representations may be, but are not necessarily, induced by the free or coerced actions of the learner.
2. **Comprehension-and-response axiom.** If a learner is presented a sentence, then using the associations and grammatical rules stored in long-term memory, the learner attempts to construct a semantic interpretation of the sentence and respond accordingly.

2.3 Internal Language

We use Lisp for the internal language of the study reported here, but in current ongoing work of machine learning of physics word problems we take the internal language to be a language for physical equations, close to what is ordinarily used in physics. In the present study the internal language is stored in memory prior to learning and does not undergo any change during learning.

The set of expressions of the internal language is specified by the grammar in Table 1 with lexical categories A_1, A_2, A_3, A_5 (= action), REL (= relation), PROP (= property), OBJ (= object property) and phrasal categories A (= action), S (= set of objects), O (= object), G (= region), and DIR (= direction). The lexicon of our internal language is given in Table 2. We refer to the elements of the lexical categories as **internal symbols**. The operations, such as fa_1 and fa_2 (read as **form action**), all have a straightforward procedural interpretation in a given robotic environment.

The English words used in Table 2 reflect an English lexicon, but the syntax of the internal language is Lisp, not English. Our categories closely match conventional linguistic categories: A corresponds to the category of a (imperative) sentence, A_1 to the category of transitive verbs, REL to the category of prepositions, PROP to the category of adjectives, OBJ to the category of common nouns, DIR to the category of adverbs, G to the category of prepositional phrases, O to the category of (determined) noun phrases, and S to the category of nominal groups. We chose, however, not to refer to these categories by their usual linguistic labels, since we think of them as semantic categories.

Table 1
Grammar of internal language.

I	A	$\rightarrow (fa_1\ A_1\ O)$
II	A	$\rightarrow (fa_2\ A_2\ G)$
III	A	$\rightarrow (fa_3\ A_3\ O\ G)$
IV	A	$\rightarrow (fa_5\ A_5\ DIR\ O)$
V	A	$\rightarrow (fa_5\ A_5\ O)$
VI	G	$\rightarrow (fr\ REL\ O)$
VII	DIR	$\rightarrow (fd\ REL)$
VIII	O	$\rightarrow (io\ S)$
IX	O	$\rightarrow (so\ S)$
X	S	$\rightarrow (fo\ PROP\ S)$
XI	S	$\rightarrow (fo\ OBJ\ *)$

Table 2
Lexicon of internal language.

Categories							Semantic Operations
OBJ	PROP	REL	A$_1$	A$_2$	A$_3$	A$_5$	
$screw	$large	$up	$get	$go	$put	$pick	*fa*$_1$ (form-action)
$nut	$medium	$on			$place		*fa*$_2$
$washer	$small	$into					*fa*$_3$
$hole	$square	$above					*fa*$_5$
$plate	$hexagonal	$to					*fr* (form-region)
$sleeve	$round	$behind					*fdir* (form-direction)
	$black						*io* (identify-object)
	$red						*so* (select-object)
	$gray						*fo* (form-object)

The grammar of the internal language would derive the following Lisp structure for the internal representation of the action corresponding to the English command *Get a screw*, where the asterisk $*$ refers to the set of objects present in a certain visual environment:

$$(fa_1\ \$get\ (so\ (fo\ \$screw\ *))). \tag{1}$$

Let $\gamma = (fo\ \$screw\ *)$. Then γ itself is the **minimal** Lisp expression in (1) containing only the internal symbol *$screw*, and $(so\ (fo\ \$screw\ *))$ is the **maximal** Lisp expression in (1) containing only the internal symbol *$screw*. We use this distinction later.

2.4 General Learning Axioms

We now turn to our learning axioms, which naturally fall into two groups: those for computations using working memory and those for changes in the state of long-term memory. We use a distinction about kinds of memory that is standard in psychological studies of human memory, but the details of our machine-learning process are not necessarily faithful to human learning of language, and we make no claim that they are. On the other hand, our basic processes of association, generalization, specification and rule-generation almost certainly have analogues in human learning, some better understood than others at the present time. In the general axioms formulated in this section we assume rather little about the specific language of the internal representation, although the examples that illustrate the axioms use the internal language described in the preceding section.

Notation. Concerning notation used in the axioms, we generally use Latin letters for sentences or their parts, whatever the natural language, and we use Greek letters to refer to internal representations of sentences or their parts. Turning now to specific notation, the letters $a, b, \ldots$ refer to words in a sentence, and the Greek letters $\alpha, \beta, \ldots$ refer to internal symbols. The symbol s refers to an entire sentence, and correspondingly σ to an entire internal representation. Grammatical forms—either sentential or term forms—are denoted by g or also $g(X)$ to show a category argument of a form; correspondingly the internal representations of a grammatical form are denoted by γ or $\gamma(X)$. We violate our Greek-Latin letter convention in the case of semantic categories or category variables X, X', Y, etc. We use the same category symbols in both grammatical forms and their internal representations.

To insure that the proper semantic meaning is carried from a natural language sentence to its internal representation, or vice versa, we index multiple occurrences of the same category in a given sentence and the corresponding occurrences in its internal representation. An example of this indexing is given later.

2.4.1 Axioms of Learning. The first set of axioms relates to computations using working memory.

Axiom 1.1 Probabilistic Association.
On any trial, let s be associated to σ, let a be in the set of words of s not associated to any internal symbol of σ, and let α be in the set of internal symbols not currently associated with any word of s. Then pairs (a, α) are sampled, possibly using the current denotational value, and associated, i.e. $a \sim \alpha$.

The probabilistic sampling in the case *Get the screw* could lead to the incorrect associations *get* $\sim$ *\$screw*, *the* $\sim$ *\$get* and no association for *screw*, for there are only two symbols to be associated to in the internal representation.

Axiom 1.2 Form Generalization.
If $g(g'_i) \sim \gamma(\gamma'_i)$, $g'_i \sim \gamma'_i$, and γ' is derivable from X, then $g(X_i) \sim \gamma(X_i)$, where i is the index of occurrence.

From the associations given after Axiom 1.1 we would derive the incorrect generalization:

$$\mathit{OBJ}\ A_1\ \mathit{screw} \sim (\mathit{fa}_1\ A_1\ (\mathit{io}\ (\mathit{fo}\ \mathit{OBJ}\ *))). \tag{2}$$

The correct one is:

$$A_1\ \mathit{the}\ \mathit{OBJ} \sim (\mathit{fa}_1\ (\mathit{io}\ (\mathit{fo}\ \mathit{OBJ}\ *))). \tag{3}$$

Axiom 1.3 Grammar-Rule Generation.
If $g \sim \gamma$ and γ is derivable from X, then $X \rightarrow g$.

Corresponding to Axiom 1.3, we now get the incorrect rule:

$$A \rightarrow \mathit{OBJ}\ A_1\ \mathit{screw}. \tag{4}$$

The correct one is:

$$A \rightarrow A_1\ \mathit{the}\ \mathit{OBJ}. \tag{5}$$

Axiom 1.4 Form Association.
If $g(g') \sim \gamma(\gamma')$ and g' and γ' have the corresponding indexed categories, then $g' \sim \gamma'$.

From (2), we get the incorrect form association:

$$\mathit{OBJ} \sim (\mathit{io}\ (\mathit{fo}\ \mathit{OBJ}\ *)). \tag{6}$$

The correct one—to be learned from more trials—is derived from (3):

$$\mathit{the\ OBJ} \sim (\mathit{io}\ (\mathit{fo}\ \mathit{OBJ}\ *)). \tag{7}$$

Axiom 1.5 Form Specification.
If $g(X_i) \sim \gamma(X_i)$, $g' \sim \gamma'$, and γ is derivable from X, then $g(g'_i) \sim \gamma(\gamma'_i)$.

As the inverse of Axiom 1.2 using the incorrect generalization given after Axiom 1.2, we use 1.5 to infer:

$$\mathit{Get\ the\ screw} \sim (\mathit{fa}_1\ \$\mathit{get}\ (\mathit{io}\ (\mathit{fo}\ \$\mathit{screw}\ *))).$$

Axiom 1.6 Content Deletion.
The content of working memory is deleted at the end of each trial.

All the axioms of the second set (outlined below) deal with changes in the state of long-term memory.

Axiom 2.1 Denotational Value Computation.
If at the end of trial n, a word a in the presented verbal stimulus is associated with some internal symbol α, then $d(a)$, the denotational value of a, increases and if a is not so associated, $d(a)$ decreases. Moreover, if a word a does not occur on a trial, then $d(a)$ stays the same unless the association of a to an internal symbol α is broken on the trial, in which case $d(a)$ decreases.

Because this axiom is conceptually less familiar, we give a more detailed example later.

Axiom 2.2 Form Factorization.
If $g \sim \gamma$ and g' is a substring of g that is already in long-term memory and g' and γ' are derivable from X, then g and γ are reduced to $g(X)$ and $\gamma(X)$. Also $g(X) \sim \gamma(X)$ is stored in long-term memory, as is the corresponding grammatical rule generated by Axiom 1.4.

We illustrate this axiom by a simple example, which seems complex because the premises take three lines, and we have two conclusions, an association, and the corresponding grammatical rule. Let:

$$\begin{array}{rrcl} g \sim \gamma: & A_1\ \mathit{the\ OBJ} & \sim & (\mathit{fa}_1\ A_1\ (\mathit{io}\ (\mathit{fo}\ \mathit{OBJ}\ *))) \\ g' \sim \gamma': & \mathit{the\ OBJ} & \sim & (\mathit{io}\ (\mathit{fo}\ \mathit{OBJ}\ *)) \\ X: & O & \rightarrow & \mathit{the\ OBJ} \\ \hline & A_1\ O & \sim & (\mathit{fa}_1\ A_1\ O) \\ & A & \rightarrow & A_1\ O \end{array}$$

Axiom 2.3 Form Filtering.
Associations and grammatical rules are removed from long-term memory at any time if they can be generated.

In the previous example, $g \sim \gamma$ can now be removed from long-term memory, and so can $A \rightarrow A_1$ *the OBJ* learned as an example of Axiom 1.3.

Axiom 2.4 Congruence Computation.
If w is a substring of g, w' is a substring of g' and they are such that

(a) $g \sim \gamma$ and $g' \sim \gamma$,

(b) g' differs from g only in the occurrence of w' in place of w,

(c) w and w' contain no words of high denotational value,

then $w' \approx w$ and the congruence is stored in long-term memory.

Using Axiom 2.4, reduction of the number of grammatical rules for a given natural language is further achieved by using congruence of meaning (Suppes 1973, 1991). Consider the following associations of grammatical forms:

$$\textit{die Schraube} \sim (\textit{io}\ (\textit{fo}\ \$\textit{screw}\ *)) \qquad (8)$$

$$\textit{der Schraube} \sim (\textit{io}\ (\textit{fo}\ \$\textit{screw}\ *)). \qquad (9)$$

Association (8) and (9) differ only with respect to the article. The article in (8) is in the nominative and accusative case, the article in (9) is in the genitive and dative case. What is important here is that there is no difference in the respective internal representations. We therefore call (8) **congruent** with (9) and collect the differing elements into a congruence class $[DA] = \{die, der\}$ where DA = definite article. This allows us to reduce the two grammatical forms (8) and (9) into one:

$$[\textit{DA}]\ \textit{Schraube} \sim (\textit{io}\ (\textit{fo}\ \$\textit{screw}\ *)). \qquad (10)$$

Notice that reduction by virtue of congruence is risky in the following way. We may lose information about the language to be learned. For instance, collapsing the gender distinction exhibited by the difference between (8) and (9) will make us incapable of distinguishing between the following sentences:

Steck die Schraube in das Loch (11)

Steck die Schraube in die Loch. (12)

Whereas (11) is grammatical, (12) is not. As long as our focus is on comprehension grammar, a command like (12) will probably not occur, but for purposes of production, congruence in its present form should not be used.

Axiom 2.5 Formation of Memory Trace.
The first time a form generalization, grammatical rule, or congruence is formed, the word associations on which the generalization, grammatical rule, or congruence is based are stored with it in long-term memory.

Using our original example after Axiom 1.3, the incorrect associations would be stored in long-term memory, but, with more learning, later deleted (2.6 (a)).

Axiom 2.6 Deletion of Associations.

(a) When a word in a sentence is given a new association, any prior association of that word is deleted from long-term memory.

(b) If $a \sim \alpha$ at the beginning of a trial, a appears in the utterance s given on that trial but α does not appear in the internal representation σ of s, then the association $a \sim \alpha$ is deleted from long-term memory.

(c) If no internal representation is generated from the occurrence of a sentence s, σ is then given as the correct internal representation, and if there are several words in s associated to an internal symbol α of σ such that the number of occurrences of these words is greater than the number of occurrences of α in σ, then these associations are deleted.

Axiom 2.7 Deletion of Form Association or Grammatical Rule.
If $a \sim \alpha$ is deleted, then any form generalization, grammatical rule, or congruence for which $a \sim \alpha$ is a memory trace is also deleted from long-term memory.

Of the thirteen axioms, only three need to be more specific to the study reported here. These three are Axiom 1.1 Probabilistic Association, Axiom 1.4 Form Association, and Axiom 2.1 Denotational Value Computations, which are given a more specific technical formulation in Section 2.5. Axiom 1.4 especially is given a much more detailed formulation.

2.5 Specialization of Certain Axioms and Initial Conditions

Axiom 1.1′ Probabilistic Association.
On any trial n, let s be associated to σ in accordance with Background Assumption 1, let A be the set of words of s not associated to any internal symbol of σ, let $d_n(a)$ be the current denotational value of each such a in A and let $\mathcal{A}$ be the set of internal symbols not currently associated with any word of s. Then

(i) an element α is uniformly sampled without replacement from $\mathcal{A}$,

(ii) at the same time, an element a is sampled without replacement from A with the sampling probability:

$$p(a) = \frac{d_n(a)}{\sum_A d_n(a)}.$$

(iii) The sampled pairs are associated, i.e. $a \sim \alpha$.

(iv) Sampling continues until either the set A or the set $\mathcal{A}$ is empty.

Due to the probabilistic nature of this procedure (Axiom 1.1) there are several possible outcomes. Consider, for example, *Get the screw*, which has the internal representation $(fa_1\ \$get\ (io\ (fo\ \$screw\ *)))$. The sampling process might generate any one of six different possible pairs, such as $get \sim \$screw$ and $screw \sim \$get$. Since there are three words occurring in the verbal command, there are in principle six ways to associate the three words of the command to the two symbols of the internal expression.

Axiom 1.4′ Form Association.
Let $g \sim \gamma$ at any step of an association computation on any trial.

(a) If X occurs in g and (*fo* X $*$) occurs in γ, then $X \sim$ (*fo* X $*$).

(b) If (i) wX is a substring of g with $g \sim \gamma$ such that $w = a$, which is a word with low denotational value, or if X is preceded by a variable, or is the first symbol of $g, w = \varepsilon$, the empty symbol, and (ii) $\gamma'(X)$ is the maximal Lisp form of γ containing the occurrence of X and no other occurrence of categories, then:

$$wX \sim \gamma'(X).$$

(c) If (i) $X_1 w_1 \cdots w_{m-1} X_m$ is a substring of g, where the $X_i, i = 1, \ldots, m$ are not necessarily distinct category names and w_i are substrings, possibly empty, or words that have no association to internal symbols on the given trial, and (ii) $\gamma'(X_{\pi(1)}, \ldots, X_{\pi(m)})$ is the minimal Lisp form of γ containing $X_{\pi(1)}, \ldots, X_{\pi(m)}$, then:

$$X_1 w_1 \cdots w_{m-1} X_m \sim \tau(X_{\pi(1)}, \ldots, X_{\pi(m)}),$$

where π is a permutation of the numbers $1, \ldots, m$.

To show how Axiom 1.4′ works, assume we have arrived at the following association of grammatical forms:

$$A_1 \textit{ the PROP OBJ} \sim (\textit{fa}_1\ A_1\ (\textit{io}\ (\textit{fo PROP}\ (\textit{fo OBJ}\ *)))) \tag{13}$$

which could be obtained as a generalization, for instance, from the command *Get the red screw* with the words correctly associated.

From Axiom 1.4′(a), we may infer:

$$\textit{OBJ} \sim (\textit{fo OBJ}\ *). \tag{14}$$

From Axiom 1.4′(b), we infer:

$$\textit{PROP OBJ} \sim (\textit{fo PROP}\ (\textit{fo OBJ}\ *)). \tag{15}$$

From Axiom 1.4′(c), we infer:

$$\textit{the PROP OBJ} \sim (\textit{io}\ (\textit{fo PROP}\ (\textit{fo OBJ}\ *))). \tag{16}$$

Using Grammar-Rule Generation (Axiom 1.3), and the grammar of the internal language (Table 1), we infer from (14) and Rule XI of Table 1:

$$S \rightarrow \textit{OBJ}. \tag{17}$$

From (15), Rule X of Table 1 and Form Generalization (Axiom 1.2):

$$\textit{PROP S} \sim (\textit{fo PROP S}), \tag{18}$$

and finally from Grammar-Rule Generation (Axiom 1.3):

$$S \rightarrow \textit{PROP S} \tag{19}$$

as a rule of English grammar. We also derive from (16), (15) and the internal grammar using Axiom 1.2:

$$\textit{the } S \sim (\textit{io } S) \tag{20}$$

and then again by Grammar-Rule Generation:

$$O \rightarrow \textit{the } S \tag{21}$$

as a rule for our English grammar.

Before the introduction of the axioms in Section 2.4, we promised to give an example of indexing of categories to preserve meaning. Such indexing can be avoided for the restricted corpora here, but is needed for more general purposes. Here is an example from our corpus showing how it works. Consider the sentence:

Put the nut on the screw.

The correct grammatical form and associated internal representation would, with indexing, look like this:

$$A_3 \textit{ the OBJ1 REL the OBJ2} \sim (\textit{fa}_3\ A_3\ (\textit{io}\ (\textit{fo OBJ1} *))(\textit{fr REL}\ (\textit{io}\ (\textit{fo OBJ2} *))))$$

where postscript numerals are used for indexing *OBJ*.

Axiom 2.1′ Denotational Value Computations.
If at the end of trial n a word a in the presented verbal stimulus is associated with some internal symbol α of the internal representation σ of s, then:

$$d_{n+1}(a) = (1-\theta)d_n(a) + \theta,$$

and if a is not associated with some denoting internal symbol α of the internal representation:

$$d_{n+1}(a) = (1-\theta)d_n(a).$$

Moreover, if a word a does not occur on trial n, then:

$$d_{n+1}(a) = d_n(a),$$

unless the association of a to an internal symbol α is broken on trial n, in which case:

$$d_{n+1}(a) = (1-\theta)d_n(a).$$

To show how the computation of denotational value (Axiom 2.1) works, let us consider further the associations given are *get* $\sim$ *\$screw*, *the* $\sim$ *\$get*. Let us further assume that at the end of this trial:

$$\begin{aligned} d(\textit{get}) &= 0.900 \\ d(\textit{screw}) &= 0.950 \\ d(\textit{the}) &= 0.700. \end{aligned}$$

On the next trial the verbal command is:

Get the nut.

As a result, we end this trial with:

$$get \sim \$get, \; nut \sim \$nut$$

and with the association of *the* deleted (Axiom 2.6 (a)). Using $\theta = 0.03$, as we usually do, we now have:

$$d(get) = 0.903$$
$$d(the) = 0.679.$$

After, let us say, three more occurrences of *the* without any association being formed, the denotational value would be further reduced to 0.620. If the command *Get the sleeve* is given and *sleeve* has not previously occurred and *get* ~ *\$get*, then we may see how *sleeve* has a higher probability of being associated to *\$sleeve* than *the*. For under the hypotheses given, the sampling probabilities for *the* and *sleeve* would be:

$$p(the) = \frac{d(the)}{d(the) + d(sleeve)} = \frac{0.620}{0.620 + 1} = 0.383$$

and:

$$p(sleeve) = \frac{d(the)}{d(the) + d(sleeve)} = \frac{1}{1.620} = 0.617.$$

The dynamical computation of denotational value continues after initial learning even when no mistakes are being made. As a consequence high-frequency words such as *a* and *the* in English and *ba* in Chinese have their denotational values approach zero rather quickly, as can be seen from the learning curves in Figures 1 and 2 in Section 4. (From a formal point, it is useful to define a word as **nondenoting** if its asymptotic denotational value is zero, or, more realistically, below a certain threshold.)

This particular linear learning model with two parameters, $d_1(a)$, and θ, could easily be replaced by more elaborate alternatives.

Initial conditions. At the beginning of trial 1, the association relation $\sim$, the congruence relation $\approx$ and the set of grammatical rules is empty. Moreover, the initial denotational value $d_1(a)$ is the same for all words a.

3. The Corpora

To test our system, we applied it to corpora of ten different languages. These languages are: English, Dutch, German, French, Spanish, Catalan, Russian, Chinese, Korean, and Japanese. The size of our corpora varied from 400 to 440 sentences. The corpora in the ten languages cover an almost identical set of internal structures. They could not be made completely identical for the following reason: an internal-language word that was translated by one word in one language, say $\mathcal{L}$, might require two or more words (depending on context) in another language $\mathcal{L}'$. As a consequence, $\mathcal{L}'$ might not be learnable from the same corpus of 400 sentences that suffice for language $\mathcal{L}$. To arrive at a complete learning of all the words, we therefore either added sentences as, e.g. in Spanish, or removed sentences, as in Japanese.

The most important variation requirement on the various corpora was that two words of a given language that were intended to correspond to two internal symbols must not always co-occur if the correct meaning were to be learned. For example, if *nimm* and *Schraube* only occurred in the single command *Nimm die Schraube!* there would be no assurance that the intended associations *Schraube* ~ *\$screw* and *nimm* ~ *\$get* would ever be learned, no matter how many learning trials there were.

We also want to note that in the Japanese case we deleted all the sentences translating *$above* as an internal symbol (ten sentences), because in Japanese *above* and *on* are expressed by the same word *ue.*

Despite careful instruction of our translators, we are not sure whether the translations sound natural in all cases and would in fact be used in a robotic working environment. To check this, however, would go far beyond what can be done by standard translation methods and requires field studies of language use in a working environment. In cases of a lexical gap, we used very simple devices. For example, French has no single idiomatic word for the nonspatial meaning of the English adjective *medium,* so we used the circumlocutionary phrase *de taille moyenne.* In some cases we avoided technical language where a single word consists of two morphemes and expresses two internal denoting symbols. As might be expected, this occurs frequently in German. For example, *Rundschraube* expresses the idea of *round screw* that is the property *round* and the object *screw.*

In the case of Catalan, we set in advance the denotational value for a few words as different from the initial value of 1. We did so because there were many nondenoting words, in our sense, and these nondenoting words uniformly co-occurred with certain action verbs, so that within our limited corpus the intuitively correct denotational value could be learned only with a probability less than one. We therefore set the initial denotational value for the words *d* and *de* at 0.05.

To obtain a successful learning performance on a corpus, it is often useful to make some modifications in the corpus. In the final tests of a theory such changes are, of course, undesirable. The changes we did make are described below.

The introduction of *pick up* created many problems because of the special nature of this verb and relation in English. In many languages the notion of picking something up is expressed by a single action verb with no additional preposition required. This created an important problem: individual words in the corpus of a given natural language sometimes denoted more than one internal symbol. Our solution (admittedly artificial but the only really artificial solution we had to adopt), was to split such words into two parts:

French: *ramasse* into *ra* and *masse*

Spanish: *recoge* into *rec* and *oge*

Catalan: *recull* into *re* and *cull*

Russian: *podnimi* into *pod* and *nimi*

Korean: *cip&ela* into *cip* and *&ela*

Japanese: *toriagero* into *tori* and *agero*

This absence of isomorphism of the semantic categories of words across languages is not surprising, as has been observed in Bowerman (1996) and Choi and Bowerman (1991). What is identified as the same action and different relations in one language may be identified as different actions and the same relation in another language.

4. Empirical Results

In summarizing our results in this section, we first present some learning results, followed by the congruence classes of nondenoting words in our sense. We then use these abstract classes to simplify and facilitate the summary table of semantically based grammatical rules of comprehension generated from the ten languages.

Table 3
Comprehension lexica for ten comparable corpora.

	Words	Nondenoting Words	Symbols with More Than 1 Association
English	28	2	0
Dutch	38	3	2
German	49	10	9
French	42	7	5
Spanish	40	7	6
Catalan	42	8	8
Russian	75	0	16
Chinese	33	8	0
Korean	31	4	1
Japanese	31	4	1

Lexicon. In Table 3 we specify for each natural language the number of words learned. We count as different words different inflected forms of what, according to dictionary usage, is the same word. As might be expected, Russian has the largest number of internal symbols with more than one association, just because of its rich inflectional patterns.

Learning. A natural question is whether or not the order of presentation of the sentences in the corpora was fixed in advance to facilitate learning. The answer is that the order was not fixed. For each language the sentences presented were chosen randomly without replacement from the corpus of approximately 400 sentences.

Compared to most neural-net rates of learning, the learning was rapid. In each of the ten languages, one cycle through the entire corpus was sufficient to produce a comprehension grammar that was intuitively correct. In contrast, and somewhat paradoxically, even for a corpus of only 400 sentences, the standard mean learning curves, theoretically based on random sampling of sentences are computationally not feasible, as we proved in Suppes, Liang, and Böttner (1992). In this same earlier paper, based on some very large runs—in fact, mean learning curves computed from up to 10,000 sample paths—we conjectured that the mean learning rate for the kind of corpora we have studied is polynomically bound. The learning of the ten languages studied in this paper quite clearly supports the conjecture for our theoretical framework of learning.

In Figures 1, 2, and 3 we show mean learning curves for the denotational value of words for English, Chinese and German. The averaging in this case is over the total number of denoting or nondenoting words in a given corpus. The number of nondenoting words in the three languages is 2, 8 and 10, respectively, as also shown in Table 3. As would be expected, for the three languages the rate of learning the denotational value of nondenoting words is inversely proportional to their number, an argument from pure frequency of occurrence in the corpora. This is not the whole story, as can be seen by comparing the Chinese and German mean curves, even though the number of nondenoting words is very close.

Congruence classes. In general in order to compare the rules being used by languages, we consolidated across languages as much as possible. The most important extension of congruence classes across languages was to introduce the empty word ϵ, so that when no nondenoting word appeared in a particular place, a given language could

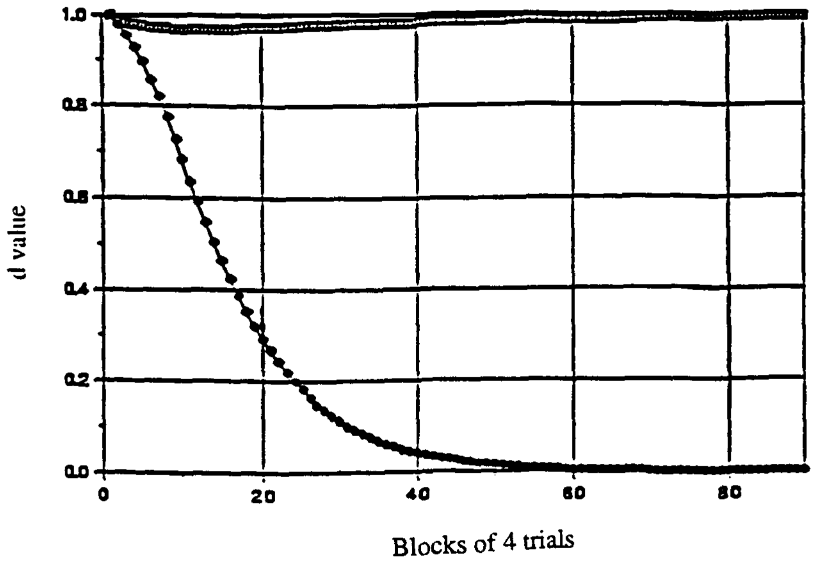

Figure 1
Mean denotational learning curves for English. The upper curve is for denoting words, with an asymptote of 1, and the lower curve is for nondenoting words with an asymptote of 0.

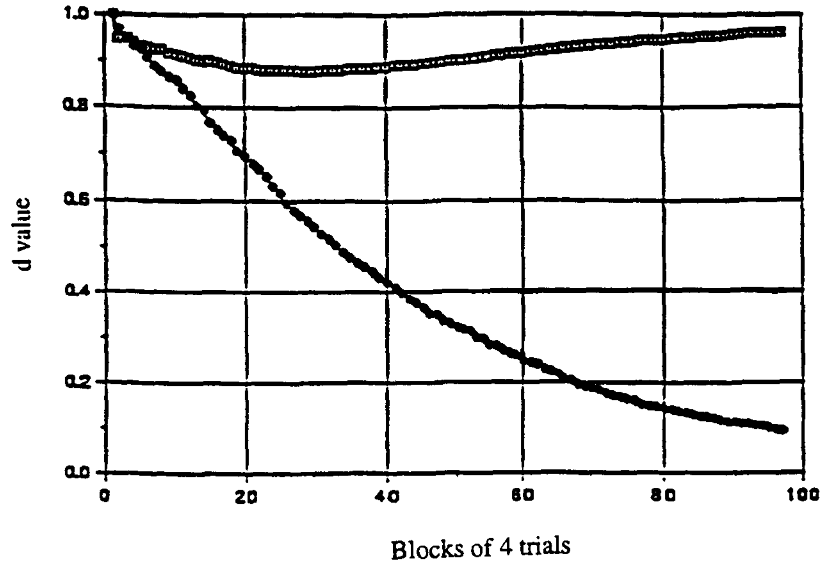

Figure 2
Mean denotational learning curves for Chinese.

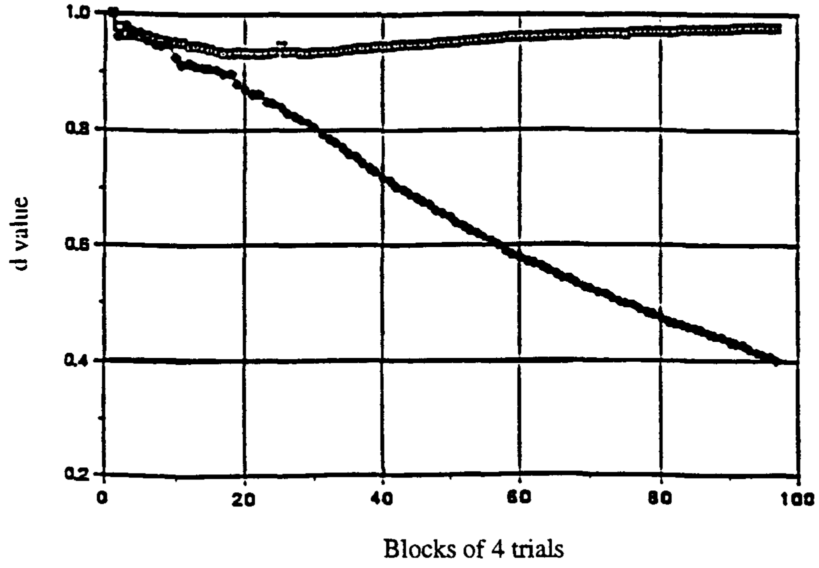

Figure 3
Mean denotational learning curves for German.

still be included in the group of languages using that rule. This has as a consequence that the grammatical rules were differentiated only by the order of occurrence of the semantic categories in the rule and not by the appearance of nondenoting words. In other words, two rules that have exactly the same semantic categories appearing in exactly the same order, independent of the appearance of nondenoting words, are treated as the same rule. (See Table 5). It is our feeling that this kind of congruence reduction is desirable in order to get a real semantic comparison of the comprehension grammars generated for different languages. The occurrence of ϵ in Table 4 indicates that the language used the same grammatical rule as another language, but with no nondenoting word occurring.

A consequence of what has just been said is that a particular grammatical rule with congruence notation may permit an instantiation in a given language of a grammatical form not instantiated in the corpus itself. For the purposes of comprehension as opposed to production, this does not lead to difficulties.

There are a number of descriptive remarks to be made about Table 4. Congruence class γ_1 is made up of various forms of definite articles, including in the case of Russian, and Japanese, the null word ϵ, because these two languages do not standardly use a definite article. The same remark applies to class γ_2 for indefinite articles. In introducing classes γ_1 and γ_2 across languages we are admittedly introducing a coarse-grained but useful clustering of nondenoting words that function rather similarly in different languages. More refinements of congruence of meaning and use could lead to the division of γ_1, and also γ_2, into several smaller classes.

The class γ_3 is specific to French because several words are required to express the idea of *medium* as in *medium screw*, namely the phrase *de taille moyenne*. We recognize the resulting analysis is a distortion of the natural semantics of the French. The distortion arises from the English lexical bias, but not grammatical bias, built into our fixed set of denoting words in the internal language.

Table 4
Congruence Classes. The languages are, from left to right, English, Dutch, German, French, Spanish, Catalan, Russian, Chinese, Korean, and Japanese.

	E	*D*	*G*	*F*	*S*	*C*	*R*	*Ch*	*K*	*J*
γ_1	*the*	*het* *de*	*die* *das* *der* *dem* *den*	*le* *l* *la*	*el* *l* *la*	*el* *l* *la*	ϵ	*nage* *zhege*	*ku* ϵ	ϵ ϵ
γ_2	*a*	*een*	*ein* *eine* *einem* *einen* *einer*	*un* *une*	*un* *una*	ϵ	ϵ	*yige*	*han* ϵ	–
γ_3	–	–	–	*de taille* ϵ	ϵ	ϵ	–	–	–	–
γ_4	ϵ	ϵ	ϵ	ϵ	–	–	ϵ	ϵ	ϵ	*no* ϵ
γ_5	–	–	–	–	–	–	–	ϵ	ϵ	*no tokoro* *no*
γ_6	ϵ	ϵ	ϵ	*de* ϵ	*de* ϵ	*d* *de* ϵ	ϵ	ϵ	–	–
γ_7	–	–	–	–	–	–	–	–	–	*o*
γ_8	–	–	–	–	–	–	–	*ba*	ϵ	ϵ
γ_9	–	–	–	–	–	–	–	–	*ul* *lul*	*o*
γ_{10}	–	–	–	–	–	–	–	–	ϵ	*ni*
γ_{11}	–	–	–	–	–	–	–	*zai* ϵ	–	–
γ_{12}	–	–	–	–	–	–	–	*chao* ϵ	ϵ	ϵ
γ_{13}	–	–	–	–	–	–	–	*na4* *na4li* ϵ	ϵ	*ni*
γ_{14}	ϵ	ϵ	ϵ	ϵ	*junto* ϵ	*a l* ϵ	ϵ	–	–	–

Given that Chinese is not an inflected language, as German is, the number of nondenoting words in Table 4 is large. We cannot discuss in detail all of these particles. The particle *ba* occurs before noun phrases describing direct objects of actions. The particle *zai* occurs before noun phrases describing objects of relations. We note that the grammar for Chinese given in Table 5 has the following two rules:

$$A \rightarrow [\gamma_8]\ O\ A_3\ [\gamma_{11}]\ G$$

and:

$$G \rightarrow REL\ [\gamma_6]\ O$$

But to put *zai* before a relation word is not correct Chinese. This superset causes no problem for comprehension, and the correct instances are when ϵ is used rather than *zai*. Remarks very similar to those for *zai* apply to *chao*. The difference in their use reflects a semantic distinction we do not need for comprehension but would need for production. The particle *zai* is used mainly in connection with relations of position such as *on* and *under*. In contrast *chao* is used in connection with the direction of motion.

Rules. An overview of the resulting grammars for the ten languages is given in Table 5. In the first column we list the production rules in the order of internal-language production rules (cf. Table 1). In the next columns we list the languages in the following order: English, Dutch, German, French, Spanish, Catalan, Russian, Chinese, Korean, Japanese. To each internal production rule corresponds in general a set of language specific rules. In the case of Rules VII and XI the set is a unit set for the obvious reason that no variation can arise in our limited internal language. Clearly in a more general setting, in the case of Rule VII, for example, relations could have modifying properties.

The most important contrast between the Indo-European and Asian languages is that in the Indo-European languages the imperative verb expressing action is usually at the beginning of an utterance, but in the Asian languages it is usually in final position. See, for example, the two rules derived from Rule II. The first is for the seven Indo-European languages and the second for the three Asian languages. Similar remarks hold for the three rules derived from Rule III.

This well-known contrast between the two groups of languages leads to a more systematic question about the rules given in Table 5. Does the set of rules corresponding to each rule of the internal language exhaust the possible permutations of the order of the semantic categories? Surprisingly the answer is affirmative except for the set generated by Rule III, which has only three members rather than six.

However, reflecting only on the three languages controlled as native speakers by the authors of this article, we can give simple examples within the vocabulary and conceptual framework of our various corpora exhibiting two of the missing permutations:

	A_3	G			O	
E:	*Put*	*near*	*the*	*washer*	*a*	*screw.*
G:	*Leg*	*neben*	*die*	*Scheibe*	*eine*	*Schraube.*

	G			A_3	O	
E:	*Near*	*the*	*washer*	*put*	*a*	*screw.*
G:	*Neben*	*die*	*Scheibe*	*leg*	*eine*	*Schraube.*
Ch:	*Zai*	*nage*	*dianquan*	*fujin fang*	*yige*	*luosiding*

Table 5
Comprehension grammars for ten comparable corpora.

	Production Rules	*E*	*D*	*G*	*F*	*S*	*C*	*R*	*Ch*	*K*	*J*
I	$A \rightarrow A_1 + O$	+	+	+	+	+	+	+			
	$A \rightarrow [\gamma_8] + O + [\gamma_9] + A_1$								+	+	+
II	$A \rightarrow A_2 + [\gamma_{14}] + G$	+	+	+	+	+	+	+			
	$A \rightarrow [\gamma_{12}] + G + [\gamma_{13}] + A_2$								+	+	+
III	$A \rightarrow A_3 + O + G$	+	+	+	+	+	+	+			
	$A \rightarrow [\gamma_8] + O + A_3 + [\gamma_{11}] + G$								+		
	$A \rightarrow O + [\gamma_9] + G + [\gamma_{10}] + A_3$									+	+
IV	$A \rightarrow A_5 + DIR + O$	+									
	$A \rightarrow A_5 + O + DIR$		+								
	$A \rightarrow DIR + O + A_5$			+							
	$A \rightarrow [\gamma_8] + O + [\gamma_9] + A_5 + DIR$								+	+	
	$A \rightarrow DIR + A_5 + O$				+	+	+	+			
	$A \rightarrow O + [\gamma_7] + DIR + A_5$										+
V	$A \rightarrow A_5 + O$	+	+	+	+			+	+		
	$A \rightarrow O + [\gamma_7] + A_5$										+
VI	$G \rightarrow REL + [\gamma_6] + O$	+	+	+	+	+	+	+	+		
	$G \rightarrow O + [\gamma_5] + REL$								+	+	+
VII	$DIR \rightarrow REL$	+	+	+	+	+	+	+	+	+	+
VIII	$O \rightarrow [\gamma_2] + S$	+	+	+	+	+	+	+	+	+	+
	$O \rightarrow [\gamma_7] + S$										+
IX	$O \rightarrow [\gamma_1] + S$	+	+	+	+	+	+	+	+	+	+
	$O \rightarrow [\gamma_7] + S$										+
X	$S \rightarrow PROP + [\gamma_4] + S$	+	+	+	+			+	+	+	+
	$S \rightarrow S + [\gamma_3] + PROP$				+	+	+				
XI	$S \rightarrow OBJ$	+	+	+	+	+	+	+	+	+	+

For the third missing permutation, we obtained from our Korean informant the following example:

	G			O		A3
K:	*ku*	*nasapati*	*yephey*	(han)	*nasa-lul*	*nohala.*
	the	washer	near	(one)	screw	put

The close correspondence between Tables 1 and 5 is, as a matter of principle, misleading. Although the grammatical rules of Table 5 are derived via Axiom 1.3 directly from the primitive grammatical rules of the internal language, as given in Table 1, this need not be the case. The larger corpus that is a superset of the one studied here for English, Chinese, and German has examples requiring derived internal grammatical rules in applying Axiom 1.3. A German example of this phenomenon, taken from Suppes, Böttner, and Liang (1995), is:

$$A \rightarrow A_4 \ [einen] \ S \ D \ [der] \ PROP \ ist.$$

An instance of this rule is:

Heb eine Mutter hoch, die nicht groß ist.

5. Related Work and Unsolved Problems

By far the most extensive research on language learning has been on children's learning of their first language. Rather extensive theoretical treatments are to be found in Wexler and Culicover (1980) and in Pinker (1984, 1989). Some of this work is related to ours, but not in a central way in terms of the details of theoretical ideas. For example, Wexler and Culicover assume the learner already has a deep context-free grammar of the language in question, and learning is centered on the learning of transformations. In contrast, we begin with no grammar of the language to be learned. We make no detailed claims of the relevance of our work to children's learning, but connections undoubtedly exist at a certain level.

In the past decade or so there has been a relatively small number of articles or books on machine learning of natural language. Langley and Carbonell (1987) provide an excellent overview of research up to the date of their publication. Compared to other research in this area, ours is more semantically than syntactically driven. This semantic commitment is also evident in the recent work of Feldman et al. (1990, 1994) and Siskind (1990, 1992, 1994), which is also the work closest to our own. Feldman et al. (1990) describe in direct and simple terms their original idea. First, the learning system is presented pairs of pictures and true natural language statements about the pictures. Second, the system is to learn the language well enough to determine whether or not a new sentence is true of the accompanying picture. Feldman et al.'s (to appear) approach to language learning separates the learning of the grammar from the learning of the lexical concepts. The grammar is learned by use of Bayesian inference over a set of possible grammars and model merging. Siskind's original work (1992), his dissertation, was in the context of naive physics, but focused also on the algorithms children may use in learning language. This work is continued in Siskind (1994), but with any assumption of prior language knowledge eliminated. The concentration is on lexical acquisition via possible internal representations of meaning. Although Siskind (1994) concentrates entirely on lexical meaning, his seven-step procedure, which constitutes a learning algorithm, bears resemblance at the top level but not in detail to our procedure. Siskind (1991) has a concept that is certainly different but similar in certain respects to our concept of denotational value equal to 0 in the limit for nondenoting words. His ideas are, however, not probabilistic, and he does not present any learning curves. He does offer in his concept of temperature (1994) a treatment of homonymy, which we do not.

In spite of our considering ten languages, the present test of our theory must be regarded as very preliminary in character. The theory needs to be extended to solve a variety of pressing unsolved problems. We restrict ourselves to four problems, but it is a mark of the still-primitive character of theoretical developments in this area, ours and others, that any informed reader can quickly double or triple this list. In our view, large scale experimentation is premature for the kind of theory we are developing until more conceptual problems are solved:

- **Learning of Anaphora.** Even very restricted and sometimes rather artificial uses of natural language are usually saturated with uses of anaphora. Physics word problems are a good example.

- **Learning temporal features.** The ordinary use of language marks temporal sequences of events in a rich variety of ways. Much systematic discourse in science and technology requires continual time and tense distinctions that must be learned.
- **Learning multiple meanings.** Troublesome examples already exist in robotic use, e.g., *screw* as both a noun and a verb in English. Much more exotic are the many meanings of *washer* in English: ring of metal (our robotic case), a machine for washing, a raccoon, and so forth.
- **Learning concepts as well as words.** It can well be claimed that concepts should be learned, not just words that stand for them. At present our theory offers nothing in this respect, although we have begun some systematic work involving multivariate network models (Suppes and Liang, to appear).

Serious applications of our theory must await serious progress on these and other problems.

Acknowledgments

We would like to thank our informants for providing us with linguistic data: Paul Meyer (Dutch), Michael Donio (French), Lydia Sanchez (Spanish), David Casacuberta (Spanish and Catalan), Irene Hayrapetian (Russian), Atsushi Sekino (Japanese), Hye-won Choi, and Yookyung Kim (Korean). Detailed criticisms by several anonymous referees have led us to reorganize the presentation and clarify the exposition.

References

Bowerman, Melissa. 1996. The origin of children's spatial semantic categories: Cognitive vs. linguistic determinants. In *Rethinking Linguistic Relativity*, edited by John J. Gumperz and Stephen C. Levinson, editors, pages 145–176. Cambridge University Press.

Choi, Soonja, and Melissa Bowerman. 1991. Learning to express motion events in English and Korean: The influence of language-specific lexicalization patterns. *Cognition* 41: 83–121.

Feldman, Jerome A., George Lakoff, Andreas Stolcke, and Susan Hollbach Weber. 1990. Miniature Language Acquisition: A touchstone for Cognitive Science. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, pages 686–693. MIT, Cambridge, MA.

Feldman, Jerome A., George Lakoff, David Bailey, Srini Narayanan, Terry Regier, and Andreas Stolcke. To appear. L_0—The First Four Years. *AI Review.*

Kittredge, Richard and John Lehrberger. 1982. *Sublanguage. Studies of Language in Restricted Semantic Domains.* De Gruyter.

Langley, Pat and Jaime G. Carbonell. 1987. Language Acquisition and Machine Learning. In *Mechanisms of Language Acquisition*, edited by Brian MacWhinney, editor, pages 115–155. Erlbaum.

Pinker, Steven. 1984. *Language Learnability and Language Development.* Harvard University Press.

Pinker, Steven. 1989. *Learnability and Cognition.* MIT Press.

Siskind, Jeffrey Mark. 1990. Acquiring core meanings of words, represented as Kackendoff-style conceptual structures, from correlated streams of linguistic and non-linguistic input. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 143–156, Pittsburgh, PA.

Siskind, Jeffrey Mark. 1991. Dispelling myths about Language Bootstrapping. *AAAI Spring Symposium Workshop on Machine Learning of Natural Language and Ontology.*

Siskind, Jeffrey Mark. 1992. *Naive Physics, Event Perception, Lexical Semantics, and Language Acquisition.* Ph.D. dissertation, M.I.T.

Siskind, Jeffrey Mark. 1994. Lexical Acquisition in the Presence of Noise and Homonymy. In *Proceedings of the Twelfth*

National Conference on Artificial Intelligence, AAAI-94, Vol. I, pages 760–766.

Suppes, Patrick. 1973. Congruence of meaning. In *Proceedings and Addresses of the American Philosophical Association* 46: 21–38.

Suppes, Patrick. 1991. *Language for Humans and Robots.* Blackwell.

Suppes, Patrick, Michael Böttner, and Lin Liang. 1995. Comprehension grammars generated from machine learning of natural languages. *Machine Learning* 19(2): 133–152. (Published in preliminary form in *Proceedings of the Eighth Amsterdam Colloquium, December 17–20, 1991*, Paul Dekker and Martin Stokhof, editors, pages 93–112. Institute for Logic, Language and Computation, University of Amsterdam, 1992.)

Suppes, Patrick and Lin Liang. To appear. Concept Learning Rates and Transfer Performance of Several Multivariate Neural Network Models. In *Progress in Mathematical Psychology*, Cornelia Dowling, Fred Roberts, and Peter Theuns, editors.

Suppes, Patrick, Lin Liang, and Michael Böttner. 1992. Complexity issues in robotic machine learning of natural language. In *Modeling Complexity Phenomena*, Lui Lam and Vladimir Naroditsky, editors, pages 102–127. Springer.

Wexler, Kenneth and Peter W. Culicover. 1980. *Formal Principles of Language Acquisition.* MIT Press.