Learning Words: Computers and Kids*

Peter M. Hastings† Steven L. Lytinen Robert K. Lindsay

Artificial Intelligence Laboratory
University of Michigan
Ann Arbor, MI 48109

Abstract

We present a computer-based model of acquisition of word meaning from context. The model uses semantic role assignments to search through a hierarchy of conceptual information for an appropriate meaning for an unknown word. The implementation of this approach has led to many surprising similarities with work in modelling human language acquisition. We describe the learning task and the model, then present an empirical test and discuss the relationships between this approach and the work in psycholinguistics.

Introduction

This paper describes a computational model of acquisition of lexical items from context. The learning task is defined as follows: given a set of natural language sentences in which a previously unknown lexical item appears, infer the syntactic class and the meaning (or meanings) of the word. We assume that the vast majority of other words appearing in the set of sentences are already known.

Our approach has been implemented as part of a unification-based natural language processing system called LINK [Lytinen, 1990]. LINK's grammar rules are quite similar in form to those used in PATR-II [Shieber, 1986]. We have incorporated semantic information into LINK's grammar, along the lines of HPSG [Pollard and Sag, 1987]. The integration of syntactic and semantic knowledge into the same grammar formalism is key to our system's ability to infer information about unknown words.

We are using LINK in two prototype applications involving relatively narrow domains (i.e. the necessary domain knowledge can be described fairly completely), but the textual input is entered by a large number of users and is therefore subject to wide variations in the terminology used. Our system is able to infer the meanings of many unknown words in these applications. The examples in this paper will be taken from one of these applications. The texts in this application describe sequences of actions to be performed on an assembly line.

In this paper, we will provide a sketch of our word-learning approach. In particular, we will focus on the acquisition of word meanings. The reader is referred to [Lytinen and Roberts, 1989] for a more detailed discussion of syntactic learning in LINK. We also present the results of an empirical test, in which our approach was used to learn the meanings of 22 undefined verbs from a corpus of 100 inputs from one of our application domains.

Our approach to the word-learning task was not developed with the modeling of human behavior in mind. The constraints of the learning task, however, guided the implementation to a state that closely resembles theoretical and empirical linguistic explanations of language acquisition in humans. We will discuss these relationships after the presentation of the empirical test.

The Learning Task

The LINK parser is often able to infer the syntactic category of an unknown word using grammatical constraints. Knowledge of the syntactic category allows LINK to make certain inferences about the semantic connections between the unknown word and other constituents of the sentence. This role-filling information is used in conjunction with a simple IS-A hierarchy in order to formulate hypotheses about the meaning of an undefined word. All of the semantic predicates defined in LINK's knowledge base are included in the hierarchy. Each concept definition includes a set of thematic roles or "slots" that can be (optionally or obligatorily) attached to the concept, as well as the type of concept which can fill each slot. The set of restrictions on fillers of slots for a concept must be at least as specific as the restrictions for its ancestors in the hierarchy (i.e. more general concepts). The ordering on generality of slot-filler constraints as well as other semantic information

*The research was funded in part by a grant from the Kellogg Foundation through the Presidential Initiatives Fund at the University of Michigan

†This author is also affiliated with EDS Center for Machine Intelligence, 2001 Commonwealth Boulevard, Suite 102, Ann Arbor, MI 48105

determines the structure of the semantic hierarchy.

Figure 1 presents a portion of the IS-A hierarchy for actions that is used in describing our assembly-line domain. Constraints on fillers of slots for actions are also represented in this figure. Slot-filling constraints on a concept are inherited from the concept's ancestors in the tree. For example, since GENERAL-FACTORY-ACTION requires an OBJECT that is a *FACTORY-OBJECT*, this restriction also implicitly holds for actions like *GET* and *INSPECT*. *RECORD-ACTION* is an example of a concept which makes a further restriction on a previously constrained slot. *RECORD*, the OBJECT of this action, must be a descendant of *FACTORY-OBJECT*.

LINK's concept hierarchy guides the process of learning word meanings. Initially, it is assumed that every concept in the hierarchy is a candidate hypothesis for the meaning of an unknown word. Example sentences can provide two types of restrictions on the set of candidate hypotheses. First, the unknown word may appear as the filler of a thematic role of another word, as in *Secure the flarge.* Because *flarge* is assigned as the direct object of *secure,* LINK's grammar suggests that it is the semantic OBJECT of *SECURE*. This condition places an upper bound on the generality of the word's meaning: *flarge* must be an AUTO-PART or one of its descendants in the hierarchy.

The second type of restriction that context may suggest is a filler for a thematic role of the unknown word, as in *Flarge the door.* In this case, LINK's unification grammar suggests that *DOOR* is the semantic OBJECT of *flarge.* Information about role-fillers of an unknown concept place a lower bound on the specificity of the concept: given that *DOOR* is the OBJECT, *flarge* can refer to concepts like *GENERAL-FACTORY-ACTION* and *ASSEMBLE*, but not to concepts like *FASTEN*, *REFILL*, or *TAPE-ACTION* (or any of its descendants) since a *DOOR* violates the restrictions that these concepts place on their OBJECTs.

Thus, two types of information are supplied by example sentences: information which provides a lower bound on the level in the hierarchy of the meaning of an unknown word, and information which provides an upper bound. This would suggest a least-commitment approach to learning, such as Mitchell's *candidate-elimination* algorithm [Mitchell, 1990]. Mitchell's algorithm used *version spaces* to represent the set of candidate hypotheses, and slowly narrowed the version space depending on the additional constraints provided by new examples. Unfortunately, in our word-learning task, often it is the case that particular kinds of words only appear in examples that provide one of the two types of restrictions. Nouns, which usually refer to things, almost always appear as role-fillers of actions or states; thus, examples only serve to limit the upper bound of the candidate hypotheses. Verbs, on the other hand, usually appear with role-fillers attached to them, and not as role-fillers themselves, since they refer to actions or states. Thus, examples only serve to place a lower bound on their candidate hypotheses. Thus, since examples only provide one of the two kinds of restrictions for many word classes, a least-commitment algorithm would not converge on a single hypothesis for the meaning of most unknown words.

Because of this, our algorithm is not a least-commitment algorithm. For nouns, we assume the most general candidate hypothesis is the correct one. Thus, the hypothesis for *Secure the flarge* is that *flarge* means *AUTO-PART*. In the case of verbs, the most specific candidate hypotheses are kept. From *flarge the nut,* then, *flarge* is assumed to mean *FASTEN* (since this concept requires a *NUT* as its OBJECT). A later example like *flarge the door* would eliminate the hypothesis *FASTEN*, since a *DOOR* cannot be its OBJECT. This would result in a generalization procedure which ascends the hierarchy and branches out until a concept (or concepts) whose constraints are satisfied by this set of slot-fillers (*DOOR* and *NUT*). The resulting set of hypotheses would then be *INSTALL*, *POSITION*, and *SECURE* since *DOOR* and *NUT* are both *AUTO-PART*s.

Limitations of This Technique

Several artifacts of the learning mechanism limit the sort of word definitions that can be learned. The first is the assumption that the representation of the ontology is complete, i.e. that every concept which is part of the domain is *a priori* represented by some node in the semantic hierarchy. This clearly limits the range of concepts which can be learned.

In addition, this technique relies solely on one type of information, the semantic constraints of role-fillers. While this information is sufficient to differentiate between many of the word meanings, large classes of words exist that require additional information to distinguish the members of the class from one another.

As mentioned above, the learning algorithm can not handle ambiguous words. In such cases, an apparent contradiction is found between competing hypotheses, and an over-general concept is then chosen. Some sort of mechanism is needed to determine whether a more general concept or a disjunctive mapping is justified in specific situations.

Finally, the learning algorithm (as we have described it so far) often does not converge on a single hypothesis for the meaning of a word, especially in the case of verbs. To see why this is true, consider again the example *Flarge the nut.* Intuitively it seems that the best hypothesis for the meaning of *flarge* is *FASTEN*, since only nuts can be fastened, and *FASTEN* is the only action in the hierarchy which can be done to only nuts. However, many other hypotheses cannot be eliminated as possibilities: *flarge* might mean *INSTALL*, since according to our hierarchy nuts can be installed, too. Given the hierarchy as it stands, no examples

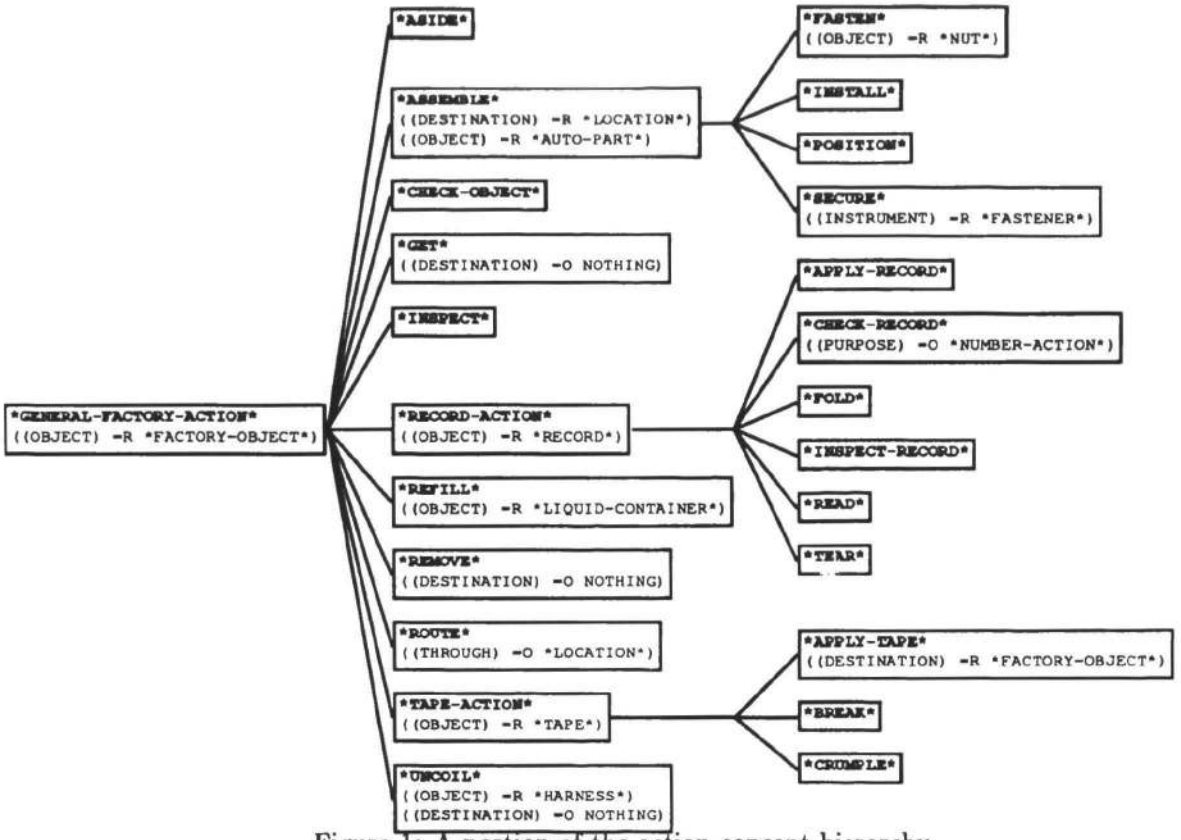

Figure 1: A portion of the action concept hierarchy

can be given which will eliminate all other candidate hypotheses (assuming *flarge* really does mean *FASTEN*), since nothing which meets the restrictions on the slots of *FASTEN* will violate any of the restrictions on the slots of these other candidates.

To remedy this problem, our algorithm ranks the list of candidate hypotheses according to how *tightly* each candidate's constraints on slots match with the actual slot fillers found in the examples. For the example *flarge the nut,* *FASTEN* is the highest-ranked candidate hypothesis for the meaning of *flarge,* since its restriction on the OBJECT slot exactly matches the OBJECT of *flarge* in the example sentence.

In addition, the final set of meaning hypotheses for an unknown word is checked at the end of the parse to see if all of the required slots for each hypothesis is present in the parse. Thus for an example like *flarge the door handle*, where *INSTALL*, *POSITION*, and *SECURE* would be hypotheses based on the filler of the OBJECT slot, *SECURE* would be eliminated from consideration because its required instrument slot is missing in the sentence.

The Empirical Test

To evaluate the effectiveness of our approach, an empirical test was conducted. In the test, a set of 100 example inputs from one of our application corpora was chosen at random. The test corpus consisted of short descriptions of sequences of actions to be performed on an assembly line. This corpus was chosen for the test because we had already developed extensive sets of grammar rules and lexical entries for it.

We were particularly interested in evaluating our algorithm's performance on learning verbs, since they presented the largest challenge. For the test, we removed the definitions of all of the verbs that appeared in the 100 examples from LINK's lexicon. There were 22 verbs in this set of examples. We then ran the system on the 100 examples and inspected the definitions of the verbs to see whether the system had inferred their meanings correctly. Table 1 presents 2 typical verbs from the set of examples as well as the sentences in which they appeared.

Table 1: Typical verbs and example sentences

secure	*Secure rr/dr hndl w/2 nuts* (secure right-rear door handle ...)
	Secure hrns to rsb w/2 int clips (Secure harness to right-side bolster with 2 int-clips)
get	*Get inspection record*
	At bench, get manifest
	Get lock cylinder kit
	Get driver

A representative set of verbs along with their inferred meanings is presented in table 2. In table 3, the verbs are grouped according the quality of the result achieved. The 17 verbs in group 1 (77% of the total set) were learned to the maximum extent possible given the amount of knowledge that exists in the system. For 7 of these words, the correct meaning was the top-ranked hypothesis. For the others, the correct meaning is included among a set of hypotheses that are indistinguishable using only the role-filler constraints. For example, the actions *INSTALL* and *POSITION* are both defined as requiring an *AUTO-PART* for an object. Without additional information, there is no way to distinguish between these concepts. Thus, both of the concepts remain as hypotheses for the meanings of *install* and *position* at the end of the test run. Verbs of this type are counted as having been successfully learned in our test results.

Group 2 contains verbs that were ambiguous, i.e. that referred to two or more nodes in the semantic hierarchy. As stated above, the algorithm currently has no way of successfully handling such words.

The verbs in group 3 were the victims of shortcomings in the implementation. *Allow* always occurs with a sentential object, e.g. *Allow to load paper to printers.* This causes difficulty for the learning algorithm since it can only handle one word at a time (notice that *load* doesn't show up in the results). The word *preload* was only found in one sentence in this test set, so the hypothesis was overly specific.

The results of this test suggest that a large portion of the meaning of unknown words can be inferred automatically using only very basic conceptual information about the domain.

Table 2: Sample results of test run

Verb	Ordered meaning hypotheses
check	*ASIDE* *CHECK-OBJECT* *GET* *INSPECT* *LOAD* *LUBRICATE* *OPEN* *PLACE* *REMOVE* *REPAIR* *RESTOCK* *ROUTE* *TOSS*
crumple	*BREAK* *CRUMPLE*
fasten	*FASTEN*
install	*INSTALL* *POSITION*
preload	*SECURE*
reach	*REACH*
uncoil	*UNCOIL*

Related Computer Models

Similar efforts at using machine learning techniques in lexical acquisition were reported in [Zernik, 1987]. Zernik described his approach as using a version space technique to learn phrasal lexicon rules. However, Zernik's system receives feedback from a teacher in the form of user-supplied "contexts" that explain what the input means. It is not clear if Zernik's approach can be adapted to a situation in which feedback is not available.

Table 3: Grouping of verbs in test results

Group 1	aside, break, crumple, fasten, fold, get, install, position, reach, remove, return, route, secure, step, toss, uncoil, walk
Group 2	apply, check, place
Group 3	allow, preload

Selfridge's CHILD program [1986] used contextual information to provide constraints on definitions of undefined words in much the same way as our system does for nouns. However, CHILD learned from only one example, and could not further refine meanings based on subsequent examples.

Jacobs and Zernik [1988] describe the RINA system, in which a task very similar to our word-learning task is performed. RINA examines large corpora, extracting many examples of a given unknown word. Although they do not describe their algorithm in detail, it appears from examples discussed in the paper that word meaning acquisition in RINA is driven more heavily by discourse context than in LINK.

Relationships to Developmental Psycholinguistics

Although this model was developed solely to allow efficient use of the limited information available to the natural language processing system, some of the challenges we faced in the development of the system bear a striking resemblance to issues brought up in the psycholinguistic literature. This suggests that these challenges are not unique to computational models but are inherent difficulties in language learning in general. Some of these issues are discussed below.

The No-Negative-Evidence Problem

When children learn language, they must induce the structure of the language and the meanings of the words relying almost entirely on examples of utterances which are *within* the language. They don't have the benefit of negative evidence to help them in their learning task. This lack of discriminating information makes the learning process computationally very complex, yet children do learn language. The Subset Principle was described in [Berwick, 1985] as one way that children could reduce the complexity of the learning task. This principle suggests that children have a hierarchical mental representation of languages ordered on the specificity of the grammars. When learning syntax, children first hypothesize the most specific grammar that accounts for the input in order to avoid overgeneralization.

We are faced with a similar problem in our model of meaning acquisition. The lack of negative evidence about word meanings as well as the nature of the role-filler constraints provides a lower bound on the set of hypotheses, but no upper bound. Thus we are forced to choose the most specific hypotheses to be able to learn from a training set consisting of only positive examples.

Bowerman [1983] presents a model of how children deal with the no-negative-evidence problem in learning verb meanings. She describes a method in which children could use syntactic information to, in effect, subcategorize verbs according to aspects of their meanings (e.g. causation). Bowerman suggests that additional discriminatory information such as this can be used as pseudo-negative evidence in that children can make predictions about word usage from syntactic clues. The violation of their assumptions provides the negative evidence that makes the learning process less computationally overwhelming.

In our model, we try to find the most specific, falsifiable hypotheses. If a later example has a slot-filler that violates our original hypothesis, we choose one that can accommodate both the old and the new slot-fillers.

Syntactic and Semantic Bootstrapping

Gleitman [1990] detailed a mechanism called "syntactic bootstrapping" that children might use to guide their search for meanings of verbs through the space of possible meanings that could be inferred from the immediate context. She suggested that children as young as 17 months have strong capabilities for recognizing syntactic distinctions and using them to constrain the meanings of verbs they are learning. For example, children who had no prior knowledge of the word *flex* were shown two videos, one of Big Bird and the Cookie Monster crossing and uncrossing their own arms, and another with one of them crossing the arms of the other. When the sentences *Big Bird is flexing with the Cookie Monster* and *Big Bird is flexing Cookie Monster* were broadcast through a speaker, the children showed a definite preference for the "syntactically congruent screen", i.e. the video that was showing the action that was being described, even though they had no semantic knowledge of the meaning of *flex*. Gleitman argued that without such a constraining mechanism, the task of word learning would be computationally infeasible. But while her approach relies solely on the syntactic structure of the sentence to yield semantic clues, our approach combines use of syntactic and semantic information (but no external context) to generate hypotheses.

In Shatz' [1987] description of a similar bootstrapping mechanism, she gives an example of a 4-year-old who said "I pricked my finger" after she had stuck herself with a needle, and then asked, "What does prick mean?" This suggests that children learning language can use their limited knowledge of the context in which a word is used to develop a partial hypothesis for the meaning of that word, just as our system incrementally refines inferred meanings over multiple examples.

Later Language Acquisition

Although our approach presents many similar properties to some aspects of children's language acquisition, it cannot be seriously considered a model of the learning process of children because of the assumption that the system's domain knowledge is complete at the time of word meaning acquisition. In this sense, the model is more similar to human language acquisition that is done later in life. Two examples of this are Genie and second language learners.

As Curtiss explains [Curtiss, 1982], Genie, during her developmental years, was deprived of all of the linguistic input that children usually receive. She was also partially deprived of information about the world. She still had information about her own surroundings, however, and presumably the maturation of her cognitive abilities gave her a much more developed (though still quite limited) conceptual representation for the world than, say, a 2-year-old would have. But Genie didn't know the words that went with the concepts she knew. Because of this, Genie's task of learning language is very similar to the one that our model is faced with. Unfortunately, Genie's linguistic deprivation during her "sensitive years" appears to have rendered her syntactic ability permanently limited. Although Genie has done quite well in acquiring the meanings of words, there are still noticeable deficits. In the face of this computer model and the work on syntactic bootstrapping, it is easy to see why she would have difficulties in learning. A large part of the information that constrains the word-learning process is unavailable to her.

The learning of a second language is another case where a fully developed conceptual representation exists when word learning takes place. Unfortunately, the second language acquisition literature tends to concentrate on teaching methods and problems, and not on psychological or linguistic theories of the processes involved. One example of the former that leans toward the latter is Cornell's description [1985] of the difficulties of teaching second-language learners the meanings of phrasal verbs (verb-particle pairs). He cites many reasons for these difficulties, among them the subtle differences between meanings for these verbs, and varying syntactic constraints. Unfortunately, our model doesn't contain the answers to these problems either, since we're still trying to learn the gross differences in meanings of words in our limited domain. Cornell does give us motivation, however, stating, "Presumably what is needed is a computer intelligent enough to scan a corpus and recognize phrasal groupings and assign meanings to them."

As mentioned above, our model of language learning was not developed for the purpose of simulating lan-

guage acquisition in humans. If the similarities found between our model and the psycholinguistic models are more than coincidence, however, then our model will provide a valuable testbed for the computational evaluation of language theories.

Future Work

There are many ways in which our algorithm can be extended. First, the algorithm as it currently stands uses only information about semantic dependencies that the parser is able to identify between words in example sentences. It should be able take advantage of other information available from the examples, such as the syntactic constructions used with an unknown word, additional semantic contextual information, and so on. The use of such additional information would enhance the similarity between this approach and syntactic bootstrapping.

Second, the assumption that a word must map directly to a unique concept in the hierarchy is not a realistic one. Many words are ambiguous, and thus refer to two or more nodes in the hierarchy. Even an unambiguous word's meaning may not correspond exactly to an already existing node in the hierarchy. In fact the mutual exclusivity (contrast) assumption, described in [Markman, in press, Clark, 1989], suggests that children learning word meanings are biased against two words having the same meaning. Our system should be able to use a similar bias by entertaining disjunctive hypotheses for word meanings, and should also be able to consider "splitting" a node in the hierarchy (similar to the approach in [Winston, 1975]), so that a word can refer to a new subconcept. In addition, a mechanism could be added to the system to check for words that refer to particular concepts. If a concept already has a referent word, it can be skipped when looking for a meaning for an unknown word.

Finally, we will continue to examine the related issues found in the psycholinguistic literature and explore methods of incorporating these theoretical and experimental results into our computational model. Hopefully these relationships will allow us to make our model more efficient and more relevant to human learning.

References

[Berwick, 1985] R. Berwick. *The Acquisition of Syntactic Knowledge.* MIT Press, 1985.

[Bowerman, 1983] M. Bowerman. How do children avoid constructing an overly general grammar in the absence of feedback about what is not a sentence? In *Proceedings of Research on Childrens Language Development*, volume 22, 1983.

[Clark, 1989] E. Clark. On the logic of constrast. *Journal of Child Language*, 15:317–335, 1989.

[Cornell, 1985] A. Cornell. Realistic goals in teaching and learning phrasal verbs. *International review of Applied Linguistics in Language Teaching*, 23(4):269–280, 1985.

[Curtiss, 1982] S. Curtiss. Developmental dissociations of language and cognition. In L. K. Obler & L. Menn, editor, *Exceptional language and linguistics*, pages 285–312. Academic Press, New York, 1982.

[Gleitman, 1990] L. Gleitman. The structural sources of verb meaning. *Language Acquisition*, I(1):3–55, 1990.

[Jacobs and Zernik, 1988] P. Jacobs and U. Zernik. Acquiring lexical knowledge from text: A case study. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, pages 739–744, 1988.

[Lytinen and Roberts, 1989] Steven L. Lytinen and Susan N. Roberts. Unifying linguistic knowledge. AI Laboratory, Univ of Michigan, Ann Arbor, MI 48109, 1989.

[Lytinen, 1990] Steven L. Lytinen. Robust processing of terse text. In *Proceedings of the 1990 AAAI Symposium on Intelligent Text-based Systems*, pages 10–14, Stanford, CA, 1990.

[Markman, in press] E. Markman. The whole object, taxonomic, and mutual exclusivity assumptions as initial constraints on word meanings. In J. P. Byrnes & S. A. Gelman, editor, *Perspectives on language and cognition: Interrelations in development.* Cambridge University Press, New York, in press.

[Mitchell, 1990] T. Mitchell. Version spaces: A candidate elimination approach to rule learning. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pages 305–309, 1990.

[Pollard and Sag, 1987] C. Pollard and I. Sag. *An Information-based Syntax and Semantics.* CLSI, Palo Alto, CA, 1987.

[Selfridge, 1986] Mallory Selfridge. A computer model of child language learning. *Artificial Intelligence*, 29:171–216, 1986.

[Shatz, 1987] M. Shatz. Bootstrapping operations in child language. *Children's Language*, 6:1–22, 1987.

[Shieber, 1986] S. Shieber. *An Introduction To Unification-Based Approaches To Grammar.* CSLI, Stanford, CA, 1986.

[Winston, 1975] Patrick Henry Winston. Learning structural descriptions from examples. In Patrick Henry Winston, editor, *The Psychology of Computer Vision.* McGraw-Hill Book Company, New York, NY, 1975.

[Zernik, 1987] Uri Zernik. How do machine language paradigms fare in language acquisition. In *Proceedings of the Fourth International Workshop on Machine Learning*, Los Altos, CA, 1987. Morgan Kaufmann.