The Ups and Downs of Lexical Acquisition

Peter M. Hastings
Artificial Intelligence Laboratory
1101 Beal Avenue
The University of Michigan
Ann Arbor, MI 48109
(313)763-9074
peter@umich.edu

Steven L. Lytinen
DePaul University
Dept. of Computer Science and Info. Systems
243 South Wabash Avenue
Chicago, IL 60604-2302
(312)362-6106
lytinen@cs.depaul.edu

Abstract

We have implemented an incremental lexical acquisition mechanism that learns the meanings of previously unknown words from the context in which they appear, as a part of the process of parsing and semantically interpreting sentences. Implementation of this algorithm brought to light a fundamental difference between learning verbs and learning nouns. Specifically, because verbs typically play the predicate role in English sentences, whereas nouns typically function as arguments, we found that different mechanisms were required to learn verbs and nouns. Because of this difference in usage, our learning algorithm formulates the most specific hypotheses possible, consistent with the data, for verb meanings, but the most general hypotheses possible for nouns. Subsequent examples may falsify a current hypothesis, causing verb meanings to be generalized and noun meanings to be made more specific. This paper describes the two approaches used to learn verbs and nouns in the system, and reports on the system's performance in substantial empirical testing.

Introduction

This paper describes the lexical acquisition system Camille (Contextual Acquisition Mechanism for Incremental Lexeme Learning (Hastings 1994)). Camille learns the lexical category and meaning of unknown words based on example sentences.

Acquisition systems are crucial to NLP systems that process real-world text. Because the complete range of the text cannot be specified, gaps in lexical knowledge are bound to occur. Such an occasion can either be disruptive for the NLP system, preventing it from processing the rest of the text, or the system can take advantage of the situation and learn something about the unknown word.

Camille is implemented as an extension of the LINK NLP system (Lytinen & Roberts 1989) which is a unification-based chart parser which integrates syntactic and semantic information. Unlike statistics-based acquisition mechanisms which require large corpora (Brent 1993; Church & Hanks 1990; Hindle 1990; Resnik 1992; Yarowsky 1992), Camille uses its domain knowledge when inferring the meaning of unknown words. The actual process of meaning inference, however, is not dependent on any particular domain hierarchy. It is a weak method that searches the hierarchy for an appropriate node for the meaning of a word.

By relying on this hierarchical knowledge structure, Camille not only gains representational and inferential power, but it also reveals an interesting fundamental principle of language. The search that Camille uses to identify the appropriate node in the semantic hierarchy for the meaning of an unknown word is data-driven; that is, the search is guided by the data provided by example sentences. Because different types of words tend to provide different data, we found that different search processes were required for different syntactic categories of words. In particular, because verbs typically fill the predicate role in English sentences, whereas nouns typically function as arguments, our learning algorithm formulates the most specific hypotheses possible, consistent with the data, for verb meanings, but the most general hypotheses possible for nouns. Subsequent examples may falsify a current hypothesis, causing the system to search up the hierarchy for verbs (i.e., generalize the hypothesis), but to search down the hierarchy for nouns (i.e., make the hypothesis more specific).

The next section describes the structure of Camille's semantic hierarchy, and the formal nature of the noun/verb dichotomy. The organization of the hierarchy and its constraints on noun learning is most apparent when Camille is faced with ambiguous nouns. The system's mechanism for inferring their meaning is described in the following section. The section after that describes the more difficult process of learning the meanings of verbs. After reviewing related work, the paper concludes with a discussion of Camille's limitations, other aspects of the system, and future work.

The Nature of the Knowledge

The knowledge representation for LINK consists of an inheritance hierarchy of domain-independent and domain-specific concepts. Figure 1 shows some of

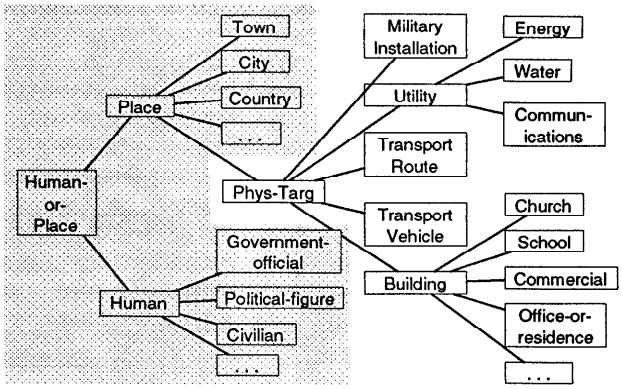

Figure 1: The pruned object tree

LINK's domain-specific object concepts from the Terrorism domain that served as the testing ground for ARPA's third and fourth Message Understanding Conferences (Sundheim 1992) (the shading will be explained later). The structure of the hierarchy forms an IS-A inheritance tree. Figure 2 shows some of the actions from the domain. Action concepts provide the relational structure that binds together the representation of the meaning of sentences. These concepts also constrain the types of arguments that can be attached as their slot-fillers (also included in fig 2).

The nodes in LINK's concept hierarchy serve as its basic units of meaning. Learning the meaning of an unknown word reduces to finding the appropriate node in the hierarchy — a graph search problem. To drive the search, the semantic constraints, which are normally used to limit attachment of slot-fillers to the Head verb, interact with the evidence provided by example sentences. But the interaction works in different ways for different classes of words. Nouns (as the Heads of noun phrases) normally serve as the slot-fillers of sentences and thus, as the items which are constrained. For example, in the sentence "Terrorists destroyed a flarge," the word "destroy" refers to the concept Destroy which has the constraint [Object = Phys-Targ]. When "flarge" is attached as the object of the verb, the constraint places an upper bound on its interpretation as shown in figure 1. The shaded-out nodes cannot be a valid interpretation of the meaning of "flarge".

For unknown verbs, however, the situation is quite different. Because they usually map to the actions in the domain, the verbs *apply* the constraints. Thus, the constraints place an upper bound on the interpretation of unknown verbs.[1] The shaded areas of figure 2 show the concepts that are ruled out for an example sentence like "Terrorists froobled the headquarters." It is important to note that this is not just an artifact of LINK's knowledge representation structure. It is due to a fundamental principle of language. Because actions serve as the relational elements of sentence structure, they are the only logical place for the constraints to reside.

Because of this dichotomy, Camille must have different strategies for learning verbs and learning nouns. They can be stated most succinctly as follows:

```
For nouns, choose the most general consistent
hypothesis.
For verbs, choose the most specific
hypothesis.
```

This difference is prescribed by the nature of the knowledge and it is consonant with psycholinguistic theories which maintain that humans treat verbs and nouns differently (Gentner 1978; Huttenlocher & Lui 1979; Graesser, Hopkinson, & Schmid 1987; Behrend 1990; Fernald & Morikawa 1993).

The implications of the noun-learning strategy are seen most clearly in the acquisition of ambiguous nouns as described in the next section. The following section describes the more difficult acquisition problem for verbs.

Learning Ambiguous Nouns

Word sense ambiguity has been a thorn in the side of NLP for a long time (Small & Cottrell 1988). The majority of the research on this issue has targetted methods for selecting the appropriate sense of an ambiguous word. For lexical acquisition, a different problem exists: how can a system recognize that a word has multiple senses and make a suitable definition?

If the system cannot learn ambiguous words, it will run into a parsing impasse. Consider two examples of the use of the word "lines" taken from the Terrorism corpus:

```
We have broken the defensive lines of the
enemy.
The Lempa River Hydroelectric Commission
reported that one of the country's main power
lines was out of service on 1 June because a
number of pylons were destroyed.
```

If the system does not know the word "lines" when it encounters the first sentence, it should infer a meaning like Military-Unit because within the domain, that is likely to be the target of Break. If the system cannot recognize ambiguity while processing the second sentence, it will either create an erroneous parse or fail altogether. Camille creates definitions for ambiguous

[1] Note that negative examples, for example, "You can't say 'Terrorists froobled the civilians"', would provide the opposing bound (upper for unknown verbs, lower for nouns). Then Mitchell's candidate-elimination approach (Mitchell 1977) to narrowing the hypothesis set might work. Unfortunately, negative examples are rare in human speech and non-existent in this and most other information extraction domains.

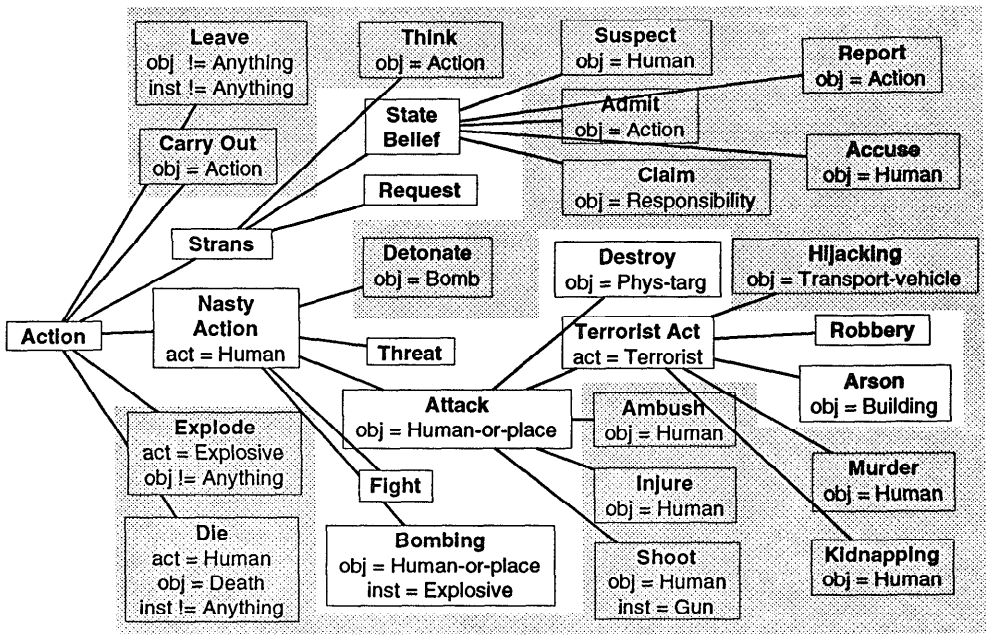

Figure 2: The pruned action tree

nouns through a simple extension of its noun-learning mechanism.

As previously stated, the constraints on actions provide an upper bound for the interpretation of unknown nouns. This provides the basis for a simple and elegant mechanism to acquire noun meanings. When an unknown noun is attached as the slot filler of a verb,[2] the unification procedure (because it returns the more specific concept) gives the representation of the meaning of that word the concept specified by the constraint. All Camille must do is to collect these induced definitions after the parse is complete.

When a word is ambiguous, the parser will try to unify incompatible concepts (Military-Unit and Electricity-Source in the example above). If the initial definition was inferred by Camille, however, it is marked as tentative. The unification procedure was extended to recognize such a situation and to infer a disjunctive definition for the word, for example (Military-Unit V Electricity-Source).

This mechanism was tested by removing the definitions of all 9 of the ambiguous nouns within the Terrorism domain: branch, charge, lines, others, plant, post, quarter, state, and system.[3] Although many of these words were not "targets" for the domain (i.e. they were not specified as interesting for the information extraction task), Camille, after processing 100 examples from the corpus which contained the words, created ambiguous definitions for five of the nine words: lines, others, post, state, and system.[4]

The scoring system used in the MUCs was adapted to facilitate evaluation of the empirical tests of Camille's lexical acquisition. The measures were defined as: Recall is the number of correct hypotheses[5] created by the system divided by the total number of undefined words. Precision is the number of correct concepts in the hypotheses divided by the total number of concepts generated. Accuracy is the number of correct hypotheses divided by the number of hypotheses generated.

The system hypothesized 5 out of 9 ambiguous definitions. Recall, counting the correct definitions, was 8 out of 18 possible definitions, or 44%. Precision and Accuracy were 8 out of 12, or 67%. As will be shown

[2]Camille's morphology component provides some indication of the lexical category of an unknown word. Consistent interpretations are entered into the parse. The application of syntactic constraints is usually sufficient to resolve the word's lexical category.

[3]Like the word "others", some additional words in the lexicon were vague. ((Lytinen 1988) also contains a discussion of dealing with vague versus ambiguous words.) "Others" was the only vague word tested because it occurred prominently in such examples as, "11 others were wounded."

[4]It also created single definitions for many other words that had been overlooked in the system development. For example the word "impunity" was inferred to be an Instrument-Object.

[5]In this paper, a hypothesis refers to a set of concepts that Camille generates as the tentative meaning of an unknown word.

in the next section, these scores are more descriptive for the larger verb-learning tests.

The importance of the ambiguity mechanism to the noun/verb dichotomy is that it highlights the difference between the conservative and liberal approaches to meaning inference. The conservative approach selects the concept specified by the verb's constraint because it is consistent with the data. The liberal approach searches under that concept for a more specific node (perhaps one which is not already the label of some other word).[6]

As described in the next section, when learning verb meanings, Camille must take a liberal approach, favoring the most specific hypotheses, in order to get usable, falsifiable hypotheses. For learning ambiguous nouns, Camille must use the the conservative approach. If the system used the liberal approach and later encountered a conflicting use of the noun, Camille would not know if it had found an ambiguous word, or if it had made a wrong initial guess about the referent of the word. This produces the two-part strategy described above.

Learning Verbs

As previously mentioned, verbs tend to play the role of the predicate in language. Thus, they serve to organize the overall semantic structure of a sentence, with arguments such as the subject and direct object attaching to them in various "slots." This makes verbs both more important and more difficult to learn, since a sentence with an unknown verb is missing its head concept.

As with nouns, Camille learns verb meanings by searching through the concept hierarchy for an appropriate concept. Because the knowledge representation imposes a lower bound on the interpretation of unknown verbs, the system must either settle for an overly general hypothesis (for example: Action but not Hijacking or Kidnapping) or inductively set its own upper bound. In order to increase the usability and the falsifiability of its hypotheses, Camille takes the latter approach.

To learn nouns, the system merely applied the constraints from the actions to the unknown slot fillers. Because verbs refer to the actions, however, the system cannot know which constraints apply. It must therefore infer the meaning of an unknown verb by comparing the slot fillers that are attached to it with the constraints of the various action concepts. Camille does this incrementally, adjusting the definition as each slot filler is attached, and as each example of the word's use is processed.

As when it learns nouns, the system initially places a default definition into the parse structure for an unknown verb and gives it the default meaning Action. As each slot filler is attached, Camille checks which descendants of the current meaning hypothesis have constraints that are compatible with the slot filler. For example, with the sentence, "Terrorist froobled the headquarters", "headquarters" is initially attached as the Object of "froobled". All of the non-shaded nodes in figure 2 have constraints which are consistent with this Object. Because Camille wants to induce an upper bound on this hypothesis set, it eliminates from consideration all but the most specific members of this set. That is, if any node in the set is the parent of another node in the set, the parent is eliminated. To make the set even more specific, the distance in the hierarchy between the slot-filler concept and the constraint concept is computed for each concept, and only the closest matches are kept in the hypothesis set. For example, Arson's Object constraint is Building which is the parent of Headquarters and therefore has a distance of one. Human-or-Place, the Object constraint for Attack has a distance of four from Headquarters, so Attack is removed from consideration. This process is repeated as each slot filler is attached for this and future sentences. After each sentence is processed, Camille stores new or modified word definitions in the lexicon.

By trimming down the hypothesis set as described, Camille would infer the single concept Arson as the meaning of "frooble". Note that other concepts (Attack and Bombing, for example) are consistent with the evidence, but these concepts would not be as easily disconfirmed. For example if the system encountered the sentence, "Terrorists froobled the pedestrians", the Arson hypothesis would be disconfirmed but not the others. This is a key to Camille's success in learning word meanings. By choosing the most specific concepts, Camille makes the most falsifiable hypotheses. Thus further examples will be more likely to conflict with an initial hypothesis, invoking the generalization procedure. This procedure searches the hierarchy starting at the current hypothesis until a concept is found which has constraints that do not conflict with all of the slot fillers that have been encountered. If another example of the the unknown word does not conflict with the initial hypothesis, the falsifiability of that hypothesis increases the likelihood that it was correct.

To empirically test Camille's verb-learning mechanism, 50 sentences were randomly selected from the corpus. The definitions of the 17 verbs from those sentences were removed from the lexicon. The average length of the sentences was 24 words, and the average number of repetitions of each unknown word was 2.7. After processing the sentences, Camille had produced 15 hypotheses of which 7 were correct (i.e. the hypothesis set included a correct concept). The average number of concepts per hypothesis was 2.5. This resulted in scores of 41% Recall, 19% Precision, and 47% Accuracy.[7] For comparison, the average of six runs in

[6] A psycholinguistic theory, Mutual Exclusivity (Markman 1991), suggests that children use a similar approach to "fill gaps" in their lexical knowledge and thereby reduce the computational complexity of their early lexical acquisition.

[7] Camille was also tested in another domain which con-

which meaning assignments were generated randomly from a weighted distribution produced scores of 22% Recall, 10% Precision, and 23% Accuracy.

Related Work

Other systems have concentrated on the acquisition of specific kinds of words. Granger noted the importance and difficulty of acquiring verbs in his description of Foul-Up (Granger 1977) which used heuristic methods to learn verbs based on the prepositions in a sentence. Zernik's Rina (Zernik 1987) concentrated on learning verb-particle combinations using interactive training and extensive domain knowledge. Unfortunately, neither was evaluated on real-world data. The extent of special-purpose knowledge that these systems required would have made that extremely difficult to do.

Salveter, Selfridge, and Siskind have developed cognitive models which perform lexical acquisition (Salveter 1979; Selfridge 1986; Siskind 1990). These systems are interesting from the psychological point of view, but they each focus on such a limited acquisition task as to render them inapplicable to real-world processing.

On the other hand, Cardie's and Riloff's systems (Cardie 1993; Riloff 1993) were specifically oriented toward the processing of real-world texts. Cardie's case-based system, MayTag, did not infer meanings for verbs though. Riloff's AutoSlog learned what amounted to pattern-based production rules. One rule, for example, matched on some subject noun phrase followed by the passive tense of "kidnap" and then assigned the subject to the victim slot of a database form which described the text. These rules could be viewed as definitions for the words. But the system knew so little about the words that it required separate rules for active and gerund uses of the same word. It also required a separate set of rules for related words like "abduct". AutoSlog created a large set of rules which required filtering by a human user. Both AutoSlog and MayTag were batch systems which performed one-shot learning.

Although the scores reported above for Camille's performance are significantly lower than the hit rates reported by Cardie's system, which was also set within an information extraction task, Cardie's scores were combined scores of all different lexical categories, and, as mentioned previously, MayTag made no concept hypotheses for verbs.

Camille's approach to lexical acquisition is incremental so its processing and storage requirements are minimized. The system learns automatically from example sentences so it does not require guidance from a human trainer. Camille doesn't need additional knowledge sources. It uses only the knowledge that is present for standard parsing.

Limitations and Future Work

An obvious limitation of the system as it is described here is that it assumed that every aspect of meaning about the domain was *a priori* represented in the concept hierarchy. This conflicts with our intuitions that lexical and concept learning interact, at least to some extent. Another aspect of Camille's implementation partially addresses this limitation, allowing the addition of object nodes. Because Camille has no other window on the world than its linguistic input, however, learning action concepts is a much more difficult problem and will be left to future research.

The basic Camille approach does have some weaknesses. The production of large sets of concepts in hypotheses was not completely mitigated by the elimination of less-specific concepts. Many sets of concepts remain that are indistinguishable based only on the use of slot fillers. The full implementation of Camille also includes a mechanism which uses scripts (Schank & Abelson 1977; Cullingford 1977) to further refine hypotheses.

The learning procedure is sensitive to noisy input. Because it uses an inductive procedure, Camille assumes that if one of its hypotheses conflicts with subsequent evidence, then the original guess was incorrect and the hypothesis should be altered. Noise can be produced by a number of sources, most commonly incomplete parses and ungrammatical input. The domains on which Camille has been tested contain mostly grammatical text. The Terrorism corpus was so complex, however, that it caused great difficulty for the parser, and incorrect or incomplete parses were common. (Camille always produces definitions for unknown verbs that it encounters. The fact that it created no definitions for 2 of the 17 in the test set signifies that no parses or parse fragments containing these words were passed to Camille.) Noisy input can cause Camille to infer that a word takes a larger range of slot-fillers. As a result, the system will make an overly general hypothesis for a word's meaning. One approach to handling noise is suggested by the Camille's mechanism which handles ambiguous words. The implementation of this addition is left to future research.

Because Camille was implemented with the goal of using only the knowledge that LINK requires for parsing, it is unable to make certain inferences about word meaning. The representation for action concepts describes only their names, their IS-A relationships to each other, and their constraints on slot fillers. Although the script mechanism allows Camille to make inferences based on sequences of actions, the system has no knowledge of the results of actions, their causes, or what goals they might achieve. The addition of such

tained much simpler sentences (average length: 4.3 words). Scores in this domain were considerably higher: Recall 71%, Precision 22%, and Accuracy 76%. As discussed below, the complexity of the test sentences in the Terrorism domain considerably decreased Camille's ability to learn because it received noisy data.

knowledge would enhance Camille's learning abilities, but it would also impose an additional resource requirement.

Conclusion

The task of lexical acquisition for Camille reduces to searching for an appropriate node in the domain representation. This abstraction of the task reveals an important distinction between learning nouns and learning verbs. The constraints on actions provide a natural upper bound on the interpretation of unknown object labels. For action labels, no such upper bound exists. Thus, in order for Camille to make useful inferences about verb meanings, it must inductively limit its search space. Camille does this by choosing the most readily falsifiable hypotheses. This gives Camille the best chance for correcting its mistakes. Thus the system uses a two-part strategy to quickly converge on an appropriate hypothesis for many unknown words.

References

Behrend, D. 1990. The development of verb concepts: Children's use of verbs to label familiar and novel events. *Child Development* 61:681–696.

Brent, M. 1993. Surface cues and robust inference as a basis for the early acquisition of subcategorization frames. *Lingua.* in press.

Cardie, C. 1993. A case-based approach to knowledge acquisition for domain-specific sentence analysis. In *Proceedings of the* 11^{th} *National Conference on Artificial Intelligence*, 798–803.

Church, K., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16.

Cullingford, R. 1977. *Organizing World Knowledge for Story Understanding by Computer.* Ph.D. Dissertation, Yale University, New Haven, CT.

Fernald, A., and Morikawa, H. 1993. Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child Development* 64:637–656.

Gentner, D. 1978. On relational meaning: The acquisition of verb meaning. *Child Development* 49:988–998.

Graesser, A.; Hopkinson, P.; and Schmid, C. 1987. Differences in interconcept organization between nouns and verbs. *Journal of Memory and Language* 26:242–253.

Granger, R. 1977. Foul-up: A program that figures out meanings of words from context. In *Proceedings of Fifth International Joint Conference on Artificial Intelligence.*

Hastings, P. 1994. *Automatic Acquisition of Word Meaning from Context.* Ph.D. Dissertation, University of Michigan, Ann Arbor, MI.

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

Huttenlocher, J., and Lui, F. 1979. The semantic organization of some simple nouns and verbs. *Journal of verbal learning and verbal behavior* 18:141–162.

Lytinen, S., and Roberts, S. 1989. Unifying linguistic knowledge. AI Laboratory, Univ of Michigan, Ann Arbor, MI 48109.

Lytinen, S. 1988. Are vague words ambiguous? In Small, S., and Cottrell, G., eds., *Lexical Ambiguity Resolution.* San Mateo, CA: Morgan Kaufmann Publishers. 109–128.

Markman, E. 1991. The whole object, taxonomic, and mutual exclusivity assumptions as initial constraints on word meanings. In Byrnes, J. P., and Gelman, S. A., eds., *Perspectives on language and thought: Interrelations in development.* Cambridge: Cambridge University Press.

Mitchell, T. 1977. Version spaces: A candidate elimination approach to rule learning. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, 305–309.

Resnik, P. 1992. A class-based approach to lexical discovery. In *Proceedings of the* 30^{th} *Annual Meeting of the Association for Computational Linguistics*, 327–329.

Riloff, E. 1993. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the* 11^{th} *National Conference on Artificial Intelligence*, 811–816.

Salveter, S. 1979. Inferring conceptual graphs. *Cognitive Science* 3:141–166.

Schank, R., and Abelson, R. 1977. *Scripts, plans, goals, and understanding.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Siskind, J. 1990. Acquiring core meanings of words. In *Proceedings of the* 28^{th} *Annual Meeting of the Association for Computational Linguistics*, 143–156.

Small, S., and Cottrell, G., eds. 1988. *Lexical Ambiguity Resolution.* San Mateo, CA: Morgan Kaufmann Publishers.

Sundheim, B. 1992. Overview of the fourth message understanding evaluation and conference. In *Proceedings of the Fourth Message Understanding Conference.* San Mateo, CA: Morgan Kaufmann Publishers.

Yarowsky, D. 1992. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings, COLING-92.*

Zernik, U. 1987. Strategies in language acquisitions: Learning phrases from examples in context. Technical Report UCLA-AI-87-1, UCLA.