# Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art

Nancy Ide*
Vassar College

Jean Véronis†
Université de Provence

## 1. Introduction

The automatic disambiguation of word senses has been an interest and concern since the earliest days of computer treatment of language in the 1950s. Sense disambiguation is an "intermediate task" (Wilks and Stevenson 1996), which is not an end in itself, but rather is necessary at one level or another to accomplish most natural language processing tasks. It is obviously essential for language understanding applications, such as message understanding and man-machine communication; it is at least helpful, and in some instances required, for applications whose aim is not language understanding:

- *machine translation*: sense disambiguation is essential for the proper translation of words such as the French *grille*, which, depending on the context, can be translated as *railings, gate, bar, grid, scale, schedule*, etc. (see, for instance Weaver [1955], Yngve [1955]).
- *information retrieval and hypertext navigation*: when searching for specific keywords, it is desirable to eliminate occurrences in documents where the word or words are used in an inappropriate sense; for example, when searching for judicial references, it is desirable to eliminate documents containing the word *court* as associated with royalty, rather than with law (see, for instance, Salton [1968], Salton and McGill [1983], Krovetz and Croft [1992], Voorhees [1993], Schütze and Pedersen [1995]).
- *content and thematic analysis*: a common approach to content and thematic analysis is to analyze the distribution of predefined categories of words—i.e., words indicative of a given concept, idea, theme, etc.—across a text. The need for sense disambiguation in such analysis, in order to include only those instances of a word in its proper sense, has long been recognized (see, for instance, Stone et al. [1966], Stone [1969], Kelly and Stone [1975]; for a more recent discussion see Litowski [1997]).
- *grammatical analysis*: sense disambiguation is useful for part-of-speech tagging—for example, in the French sentence *L'étagère plie sous les livres* ('The shelf is bending under [the weight of] the books'), it is necessary to disambiguate the sense of *livres* (which can mean 'books' or 'pounds'

* Department of Computer Science, Vassar College, Poughkeepsie, New York 12604-0520. E-mail: ide@cs.vassar.edu
† Laboratoire Parole et Langage, ESA 6057 CNRS, Université de Provence, 29 Avenue Robert Schuman, 13621 Aix-en-Provence Cedex 1, France. E-mail: Jean.Veronis@lpl.univ-aix.fr

and is masculine in the former sense, feminine in the latter) to properly tag it as a masculine noun. Sense disambiguation is also necessary for certain syntactic analyses, such as prepositional phrase attachment (Jensen and Binot 1987; Whittemore, Ferrara, and Brunner 1990; Hindle and Rooth 1993), and, in general, restricts the space of competing parses (Alshawi and Carter 1994).

- *speech processing*: sense disambiguation is required for correct phonetization of words in speech synthesis, for example, the word *conjure* in *He conjured up an image* or in *I conjure you to help me* (Sproat, Hirschberg, and Yarowsky 1992; Yarowsky 1997), and also for word segmentation and homophone discrimination in speech recognition (Connine 1990; Seneff 1992).

- *text processing*: sense disambiguation is necessary for spelling correction (for example, to determine when diacritics should be inserted, such as in French, changing *comte* to *comté* [Yarowsky 1994a, 1994b]); for case changes (HE READ THE TIMES → *He read the Times*); for lexical access of Semitic languages (in which vowels are not written), etc.

The problem of word sense disambiguation (WSD) has been described as "AI-complete," that is, a problem which can be solved only by first resolving all the difficult problems in artificial intelligence (AI), such as the representation of common sense and encyclopedic knowledge. The inherent difficulty of sense disambiguation was a central point in Bar-Hillel's well-known treatise on machine translation (Bar-Hillel 1960), where he asserted that he saw no means by which the sense of the word *pen* in the sentence *The box is in the pen* could be determined automatically. Bar-Hillel's argument laid the groundwork for the ALPAC report (ALPAC 1966), which is generally regarded as the direct cause for the abandonment of most research on machine translation in the early 1960s.

At about the same time, considerable progress was being made in the area of knowledge representation, especially the emergence of semantic networks, which were immediately applied to sense disambiguation. Work on word sense disambiguation continued throughout the next two decades in the framework of AI-based natural language understanding research, as well as in the fields of content analysis, stylistic and literary analysis, and information retrieval. In the past ten years, attempts to automatically disambiguate word senses have multiplied, due, like much other similar activity in the field of computational linguistics, to the availability of large amounts of machine-readable text and the corresponding development of statistical methods to identify and apply information about regularities in this data. Now that other problems amenable to these methods, such as part-of-speech disambiguation and alignment of parallel translations, have been fairly thoroughly addressed, the problem of word sense disambiguation has taken center stage, and it is frequently cited as one of the most important problems in natural language processing research today.

Given the progress that has been recently made in WSD research and the rapid development of methods for solving the problem, it is appropriate at this time to stand back and assess the state of the field and to consider the next steps that need to be taken. To this end, this paper surveys the major, well-known approaches to word sense disambiguation and considers the open problems and directions of future research.

## 2. Survey of WSD Methods

In general terms, word sense disambiguation involves the association of a given word in a text or discourse with a definition or meaning (sense) which is distinguishable from other meanings potentially attributable to that word. The task therefore necessarily involves two steps: (1) the determination of all the different senses for every word relevant (at least) to the text or discourse under consideration; and (2) a means to assign each occurrence of a word to the appropriate sense.

Much recent work on WSD relies on predefined senses for step (1), including:

- a list of senses, such as those found in everyday dictionaries;
- a group of features, categories, or associated words (e.g., synonyms, as in a thesaurus);
- an entry in a transfer dictionary, which includes translations in another language;

The precise definition of a sense is, however, a matter of considerable debate within the community. The variety of approaches to defining senses has raised concern about the comparability of much WSD work, and given the difficulty of the problem of sense definition, no definitive solution is likely to be found soon (see Section 3.2). However, since the earliest days of WSD work, there has been general agreement that the problems of morpho-syntactic disambiguation and sense disambiguation can be disentangled (see, e.g., Kelly and Stone [1975]). That is, for homographs with different parts of speech (e.g., *play* as a verb and noun), morphosyntactic disambiguation accomplishes sense disambiguation, and therefore (especially since the development of reliable part-of-speech taggers), WSD work has focused largely on distinguishing senses among homographs belonging to the same syntactic category.

Step (2), the assignment of words to senses, is accomplished by reliance on two major sources of information:

- the *context* of the word to be disambiguated, in the broad sense: this includes information contained within the text or discourse in which the word appears, together with extra-linguistic information about the text, such as situation, etc.;
- *external knowledge sources*, including lexical, encyclopedic, etc. resources, as well as hand-devised knowledge sources, which provide data useful to associate words with senses.

All disambiguation work involves matching the context of the instance of the word to be disambiguated with either information from an external knowledge source (**knowledge-driven** WSD), or information about the contexts of previously disambiguated instances of the word derived from corpora (**data-driven** or **corpus-based** WSD). Any of a variety of **association methods** is used to determine the best match between the current context and one of these sources of information, in order to assign a sense to each word occurrence. The following sections survey the approaches applied to date.

### 2.1 Early WSD Work in MT

The first attempts at automated sense disambiguation were made in the context of machine translation (MT). In his famous memorandum (available mimeographed in

1949, but not printed until 1955) Weaver discusses the need for WSD in machine translation and outlines the basis of an approach to WSD that underlies all subsequent work on the topic:

> If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words.... But if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say $N$ words on either side, then if $N$ is large enough one can unambiguously decide the meaning of the central word.... The practical question is: "What minimum value of $N$ will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?" (1955, 20)

A well-known early experiment by Kaplan (1950) attempted to answer this question at least in part, by presenting ambiguous words in their original context and in a variant context providing one or two words on either side to seven translators. Kaplan observed that sense resolution given two words on either side of the word was not significantly better or worse than when given the entire sentence. The same phenomenon has been reported by several researchers since Kaplan's work appeared: e.g., Masterman (1962), Koutsoudas and Korfhage (1956) on Russian, and Gougenheim and Michéa (1961) and Choueka and Lusignan (1985) on French.

Reifler's (1955) "semantic coincidences" between a word and its context quickly became the determining factor in WSD. The complexity of the context, and in particular the role of syntactic relations, was also recognized; for example, Reifler (1955) says:

> Grammatical structure can also help disambiguate, as, for instance, the word *keep*, which can be disambiguated by determining whether its object is gerund (*He kept eating*), adjectival phrase (*He kept calm*), or noun phrase (*He kept a record*).

The goal of MT was initially modest, focused primarily on the translation of technical texts and in all cases dealing with texts from particular domains. Weaver discusses the role of the domain in sense disambiguation, making a point that was reiterated several decades later by Gale, Church, and Yarowsky (1992c):

> In mathematics, to take what is probably the easiest example, one can very nearly say that each word, within the general context of a mathematical article, has one and only one meaning. (1955, 20)

Following directly from this observation, much effort in the early days of machine translation was devoted to the development of specialized dictionaries or "microglossaries" (Oswald 1952, 1957; Oswald and Lawson 1953; Oettinger 1955; Dostert 1955; Gould 1957; Panov 1960). Such microglossaries contain only the meaning of a given word relevant for texts in a particular domain of discourse; e.g., a microglossary for the domain of mathematics would contain only the relevant definition of *triangle*, and not the definition of *triangle* as a musical instrument.

The need for knowledge representation for WSD was also acknowledged from the outset: Weaver concludes by noting the "tremendous amount of work [needed] in the logical structure of languages" (1995, 23). Several researchers attempted to devise

an "interlingua" based on logical and mathematical principles that would solve the disambiguation problem by mapping words in any language to a common semantic/conceptual representation. Among these efforts, those of Richens and Masterman eventually led to the notion of the "semantic network" (Richens [1958], Masterman [1962]; see Section 2.2.1); following on this, the first machine-implemented knowledge base was constructed from *Roget's Thesaurus* (Masterman 1957). Masterman applied this knowledge base to the problem of WSD: in an attempt to translate Virgil's *Georgics* by machine, she looked up, for each Latin word stem, the translation in a Latin-English dictionary and then looked up this word in the word-to-head index of *Roget's*. In this way, each Latin word stem was associated with a list of *Roget* head numbers associated with its English equivalents. The numbers for words appearing in the same sentence were then examined for overlaps. Finally, English words appearing under the multiply-occurring head categories were chosen for the translation.[1] Masterman's methodology is strikingly similar to that underlying much of the knowledge-based WSD accomplished recently (see Section 2.3).

It is interesting to note that Weaver's text also outlined the statistical approach to language analysis prevalent now, nearly fifty years later:

> This approach brings into the foreground an aspect of the matter that probably is absolutely basic—namely, the statistical character of the problem.... And it is one of the chief purposes of this memorandum to emphasize that statistical semantic studies should be undertaken, as a necessary primary step. (1955, 22)

Several authors followed this approach in the early days of machine translation (e.g., Richards 1953; Yngve 1955; Parker-Rhodes 1958). Estimations of the degree of polysemy in texts and dictionaries were made: Harper, working on Russian texts, determined the number of polysemous words in an article on physics to be approximately 30% (Harper 1957a) and 43% in another sample of scientific writing (Harper 1957b); he also found that Callaham's Russian-English dictionary provides, on average, 8.6 English equivalents for each Russian word, of which 5.6 are quasi-synonyms, thus yielding approximately three distinct English equivalents for each Russian word. Bel'skaja (1957) reports that in the first computerized Russian dictionary, 500 out of 2,000 words are polysemous. Pimsleur (1957) introduced the notion of levels of depth for a translation: level 1 uses the most frequent equivalent (e.g., German *schwer* = *heavy*), producing a text where 80% of the words are correctly translated; level 2 distinguishes additional meanings (e.g., *schwer* = *difficult*), producing a translation which is 90% correct; etc. Although the terminology is different, this is very similar to the notion of **baseline** tagging used in modern work (see, e.g., Gale, Church, and Yarowsky [1992b]).

A convincing implementation of many of these ideas was made several years later, paradoxically at the moment when MT began its decline. Madhu and Lytle (1965), working from the observation that domain constrains sense, calculated sense frequency for texts in different domains and applied a Bayesian formula to determine the probability of each sense in a given context—a technique similar to that applied in much later work and which yielded a similar 90% correct disambiguation result (see Section 2.4).

1 For a detailed accounting of Masterman's methodology, see Wilks, Slator, and Guthrie (1996). Other researchers have discussed the use of thesauri for disambiguation in the context of early MT work, e.g. Gentilhomme and Tabory (1960).

The striking fact about this early work on WSD is the degree to which the fundamental problems and approaches to the problem were foreseen and developed at that time. However, without large-scale resources, most of these ideas remained untested and to a large extent, forgotten, until several decades later.

## 2.2 AI-based Methods

AI methods began to flourish in the early 1960s and began to attack the problem of language understanding. As a result, WSD in AI work was typically accomplished in the context of larger systems intended for full language understanding. In the spirit of the times, such systems were almost always grounded in some theory of human language understanding that they attempted to model, and often involved the use of detailed knowledge about syntax and semantics to perform their task, which was exploited for WSD.

**2.2.1 Symbolic Methods.** As mentioned above, semantic networks were developed in the late 1950s and were immediately applied to the problem of representing word meanings.[2] Masterman (1962), working in the area of machine translation, used a semantic network to derive the representation of sentences in an interlingua comprised of fundamental language concepts; sense distinctions are implicitly made by choosing representations that reflect groups of closely related nodes in the network. She developed a set of 100 primitive concept types (THING, DO, etc.), in terms of which her group built a 15,000-entry concept dictionary, where concept types are organized in a lattice with inheritance of properties from superconcepts to subconcepts. Building on this and on work on semantic networks by Richens (1958), Quillian (1961, 1962a, 1962b, 1967, 1968, 1969) built a network that includes links among words (tokens) and concepts (types), in which links are labeled with various semantic relations or simply indicate associations between words. The network is created starting from dictionary definitions, but is enhanced by human knowledge that is hand-encoded. When two words are presented to the network, Quillian's program simulates the gradual activation of concept nodes along a path of links originating from each input word by means of **marker passing**; disambiguation is accomplished because only one concept node associated with a given input word is likely to be involved in the most direct path found between the two input words. Quillian's work informed later dictionary-based approaches to WSD (see Section 2.3.1).

Subsequent AI-based approaches exploited the use of **frames** containing information about words and their roles and relations to other words in individual sentences. For example, Hayes (1976, 1977a, 1977b, 1978) uses a combination of a semantic network and case frames. The network consists of nodes representing noun senses and links represented by verb senses; case frames impose IS-A and PART-OF relations on the network. As in Quillian's system, the network is traversed to find chains of connections between words. Hayes work shows that homonyms can be fairly accurately disambiguated using this approach, but it is less successful for other kinds of polysemy. Hirst (1987) also uses a network of frames and, again following Quillian, marker passing to find minimum-length paths of association between frames for senses of words in context in order to choose among them. He introduces "polaroid words," a mechanism which progressively eliminates inappropriate senses based on

2 Semantic networks derive from much earlier work on knowledge representation using graphs, such as Pierce's "existential graphs" (see Roberts [1973]) and the graphs of the psychologist Selz (1913, 1922) which represent patterns of concepts and inheritance of properties.

syntactic evidence provided by the parser, together with semantic relations found in the frame network. Eventually only one sense remains; however, Hirst reports that in cases where some word (including words other than the target) in the sentence is used metaphorically, metonymically, or in an unknown sense, the polaroids often end by eliminating all possible senses, and fail.

Wilks' preference semantics ([1968, 1969, 1973, 1975a, 1975b, 1975c, 1975d]; see the survey by Wilks and Fass [1990]), which uses Masterman's primitives, is essentially a case-based approach to natural language understanding and one of the first specifically designed to deal with the problem of sense disambiguation. Preference semantics specifies selectional restrictions for combinations of lexical items in a sentence that can be relaxed when a word with the preferred restrictions does not appear, thus enabling, especially, the handling of metaphor (as in *My car drinks gasoline*, where the restrictions on *drink* prefer an animate subject but allow an inanimate one). Boguraev (1979) shows that preference semantics is inadequate to deal with polysemous verbs and attempts to improve on Wilks' method by using a combination of evidence, including selectional restrictions, preferences, case frames, etc. He integrates semantic disambiguation with structural disambiguation to enable judgments about the semantic coherence of a given sense assignment. Like many other systems of the era, these systems are sentence-based and do not account for phenomena at other levels of discourse, such as topical and domain information. The result is that some kinds of disambiguation are difficult or impossible to accomplish.

A rather different approach to language understanding, which contains a substantial sense discrimination component, is the Word Expert Parser (Small 1980, 1983; Small and Reiger 1982; Adriaens 1986, 1987, 1989; Adriaens and Small 1988). The approach derives from the somewhat unconventional theory that human knowledge about language is organized primarily as knowledge about words rather than rules. Their system models what its authors feel is the human language understanding process: a co-ordination of information exchange among word experts about syntax and semantics as each determines its involvement in the environment under question. Each expert contains a **discrimination net** for all senses of the word, which is traversed on the basis of information supplied by the context and other word experts, ultimately arriving at a unique sense, which is then added to a semantic representation of the sentence. The well-known drawback of the system is that the word experts need to be extremely large and complex to accomplish the goal, which is admittedly greater than sense disambiguation.[3]

Dahlgren's (1988) language understanding system includes a sense disambiguation component that uses a variety of types of information: fixed phrases, syntactic information (primarily, selectional restrictions), and commonsense reasoning. The reasoning module, because it is computationally intensive, is invoked only in cases where the other two methods fail to yield a result. Although her original assumption was that much disambiguation could be accomplished based on paragraph topic, she found that half of the disambiguation was actually accomplished using fixed phrase and syntactic information, while the other half was accomplished using commonsense reasoning. Reasoning often involves traversing an ontology to find common ancestors for words in context; her work anticipates Resnik's (1993a, 1993b, 1995a) results by determining that ontological similarity, involving a common ancestor in the ontology, is a powerful disambiguator. She also notices that verb selectional restrictions are an

3 It is interesting to compare the word experts with the procedures of Kelly and Stone (1975), which similarly involve procedures for individual words, although their goal was only to disambiguate senses.

important source of disambiguation information for nouns—another result that has been subsequently tested and noted.

**2.2.2 Connectionist Methods.** Work in psycholinguistics in the 1960s and 1970s established that semantic priming—a process in which the introduction of a certain concept will influence and facilitate the processing of subsequently introduced concepts that are semantically related—plays a role in disambiguation by humans (see, e.g., Meyer and Schvaneveldt [1971]). This idea is realized in spreading activation models (see Collins and Loftus [1975]; Anderson [1976, 1983]), where concepts in a semantic network are activated upon use, and activation spreads to connected nodes. Activation is weakened as it spreads, but certain nodes may receive activation from several sources and be progressively reinforced. McClelland and Rumelhart (1981) added to the model by introducing the notion of inhibition among nodes, where the activation of a node might suppress, rather than activate, certain of its neighbors (see also Feldman and Ballard [1982]). Applied to lexical disambiguation, this approach assumes that activating a node corresponding to, say, the concept THROW will activate the "physical object" sense of *ball*, whose activation would in turn inhibit the activation of other senses of *ball*, such as "social event."

Quillian's semantic network, described above, is the earliest implementation of a spreading activation network used for word sense disambiguation. A similar model is implemented by Cottrell and Small (1983); see also Cottrell (1985). In both of these models, each node in the network represents a specific word or concept.[4] Waltz and Pollack (1985) and Bookman (1987) hand-encode sets of semantic "microfeatures," corresponding to fundamental semantic distinctions (animate/inanimate, edible/inedible, threatening/safe, etc.), characteristic durations of events (second, minute, hour, day, etc.), locations (city, country, continent, etc.), and other similar distinctions, in their networks. In Waltz and Pollack (1985), sets of microfeatures have to be manually primed by a user to activate a context for disambiguating a subsequent input word, but Bookman (1987) describes a dynamic process in which the microfeatures are automatically activated by the preceding text, thus acting as a short-term context memory. In addition to these local models (i.e., models in which one node corresponds to a single concept), distributed models have also been proposed (see, for example, Kawamoto [1988]). However, whereas local models can be constructed a priori, distributed models require a learning phase using disambiguated examples, which limits their practicality.

The difficulty of hand-crafting the knowledge sources required for AI-based systems restricted them to "toy" implementations handling only a tiny fraction of the language. Consequently, disambiguation procedures embedded in such systems are most usually tested on only a very small test set in a limited context (most often, a single sentence), making it impossible to determine their effectiveness on real texts. For less obvious reasons, many of the AI-based disambiguation results involve highly ambiguous words and fine sense distinctions (e.g., *ask, idea, hand, move, use, work,* etc.) and unlikely test sentences (*The astronomer married the star*), which make the results even less easy to evaluate in the light of the now-known difficulties of discriminating even gross sense distinctions.

4 Note, however, that, symbolic methods such as Quillian's implement propagation via mechanisms such as marker passing, whereas the neural network models developed in the late 1970s and early 1980s use numeric activation, inspired by the neural models of McCulloch and Pitts (1943) and Hebb's (1949) work on neurological development, which saw its first full development in Rosenblatt's (1958) "perceptrons."

### 2.3 Knowledge-based Methods

The AI-based work of the 1970s and 1980s was theoretically interesting but not at all practical for language understanding in any but extremely limited domains. A significant roadblock to generalizing WSD work was the difficulty and cost of hand-crafting the enormous amounts of knowledge required for WSD: the so-called "knowledge acquisition bottleneck" (Gale, Church, and Yarowsky 1993). Work on WSD reached a turning point in the 1980s when large-scale lexical resources, such as dictionaries, thesauri, and corpora, became widely available. Attempts were made to automatically extract knowledge from these sources (Sections 2.3.1 and 2.3.2) and, more recently, to construct large-scale knowledge bases by hand (Section 2.3.3). A corresponding shift away from methods based in linguistic theories and towards empirical methods also occurred at this time, as well as a decrease in emphasis on do-all systems in favor of "intermediate" tasks such as WSD.

**2.3.1 Machine-Readable Dictionaries.** Machine-readable dictionaries (MRDs) became a popular source of knowledge for language-processing tasks following Amsler's (1980) and Michiel's (1982) theses.[5] A primary area of activity during the 1980s involved attempts to automatically extract lexical and semantic knowledge bases from MRDs (Michiels, Mullenders, and Noël 1980; Calzolari 1984; Chodorow, Byrd, and Heidon 1985; Markowitz, Ahlswede, and Evens 1986; Byrd et al. 1987; Nakamura and Nagao 1988; Klavans, Chodorow, and Wacholder 1990; Wilks et al. 1990). This work contributed significantly to lexical semantic studies, but it appears that the initial goal—the automatic extraction of large knowledge bases—was not fully achieved: the only currently widely available large-scale lexical knowledge base (WordNet, see below) was created by hand. We have elsewhere demonstrated the difficulties of automatically extracting relations as simple as hyperonymy (Véronis and Ide 1991; Ide and Véronis 1993a, 1993b), in large part due to the inconsistencies in dictionaries themselves (well-known to lexicographers, cf. Atkins and Levin [1988], Kilgarriff [1994]) as well as the fact that dictionaries are created for human use, and not for machine exploitation.

Despite its shortcomings, the machine-readable dictionary provides a ready-made source of information about word senses and therefore rapidly became a staple of WSD research. The methods employed attempt to avoid the problems cited above by using the text of dictionary definitions directly, together with methods sufficiently robust to reduce or eliminate the effects of a given dictionary's inconsistencies. All of these methods (and many of those cited elsewhere in this paper) rely on the notion that the most plausible sense to assign to multiple co-occurring words is the one that maximizes the relatedness among the chosen senses.

Lesk (1986) created a knowledge base that associated with each sense in a dictionary a "signature"[6] composed of the list of words appearing in the definition of that sense. Disambiguation was accomplished by selecting the sense of the target word whose signature contained the greatest number of overlaps with the signatures of neighboring words in its context. The method achieved 50–70% correct disambiguation, using a relatively fine set of sense distinctions such as those found in a typical learner's dictionary. Lesk's method is very sensitive to the exact wording of each defi-

5 The first freely available machine-readable dictionaries were the *Merriam-Webster Seventh Collegiate Dictionary* and the *Merriam-Webster New Pocket Dictionary*, typed from printed versions under the direction of Olney and Ziff of the System Development Corporation in 1966–68 (Olney 1968). Urdang (1984) describes a similar enterprise during the same period at Random House.

6 Lesk does not use this term.

nition: the presence or absence of a given word can radically alter the results. However, Lesk's method has served as the basis for most subsequent MRD-based disambiguation work.

Wilks et al. (1990) attempted to improve the knowledge associated with each sense by calculating the frequency of co-occurrence for the words in definition texts, from which they derive several measures of the degree of relatedness among words. This metric is then used with the help of a vector method that relates each word and its context. In experiments on a single word (*bank*), the method achieved 45% accuracy on sense identification, and 90% accuracy on homograph identification. Lesk's method has been extended by creating a neural network from definition texts in the *Collins English Dictionary* (CED), in which each word is linked to its senses, which are themselves linked to the words in their definitions, which are in turn linked to their senses, etc. (Véronis and Ide 1990).[7] Experiments on 23 ambiguous words, each in six contexts (138 pairs of words), produced correct disambiguation, using the relatively fine sense distinctions in the CED, in 71.7% of the cases (three times better than chance: 23.6%) (Ide and Véronis 1990b); in later experiments, improving the parameters and only distinguishing homographs enabled a rate of 85% (vs. chance: 39%) (Véronis and Ide 1995). Applied to the task of mapping the senses of the CED and OALD for the same 23 words (59 senses in all), this method obtained a correct correspondence in 90% of the cases at the sense level, and 97% at the level of homographs (Ide and Véronis 1990a). Sutcliffe and Slater (1995) replicated this method on full text (samples from Orwell's *Animal Farm*) and found similar results (72% correct sense assignment, compared with a 33% chance baseline, and 40% using Lesk's method).

Several authors (for example, Krovetz and Croft [1989], Guthrie et al. [1991], Slator [1992], Cowie, Guthrie, and Guthrie [1992], Janssen [1992], Braden-Harder [1993], Liddy and Paik [1993]) have attempted to improve results by using supplementary fields of information in the electronic version of the *Longman Dictionary of Contemporary English* (LDOCE), in particular, the **box codes** and **subject codes** provided for each sense. Box codes include primitives such as ABSTRACT, ANIMATE, HUMAN, etc., and encode type restrictions on nouns and adjectives and on the arguments of verbs. Subject codes use another set of primitives to classify senses of words by subject (ECONOMICS, ENGINEERING, etc.). Guthrie et al. (1991) demonstrate a typical use of this information: in addition to using the Lesk-based method of counting overlaps between definitions and contexts, they impose a correspondence of subject codes in an iterative process. No quantitative evaluation of this method is available, but Cowie, Guthrie, and Guthrie (1992) improve the method using **simulated annealing** and report results of 47% for sense distinctions and 72% for homographs. The use of LDOCE box codes, however, is problematic: the codes are not systematic (see, for example, Fontenelle [1990]); in later work, Braden-Harder (1993) showed that simply matching box or subject codes is not sufficient for disambiguation. For example, in *I tipped the driver*, the codes for several senses of the words in the sentence satisfy the necessary constraints (e.g., *tip-money* + human object or *tip-tilt* + movable solid object).

7 Note that the assumptions underlying this method are very similar to Quillian's (1968):

> Thus one may think of a full concept analogically as consisting of all the information one would have if he looked up what will be called the "patriarch" word in a dictionary, then looked up every word in each of its definitions, then looked up every word found in each of these, and so on, continually branching outward... (p. 238).

However, Quillian's network also keeps track of semantic relationships among the words encountered along the path between two words, which are encoded in his semantic network; the neural network avoids the overhead of creating the semantic network but loses this relational information.

In many ways, the supplementary information in the LDOCE, and in particular the subject codes, is similar to that in a thesaurus, which, however, is more systematically structured.

Inconsistencies in dictionaries, noted earlier, are not the only and perhaps not the major source of their limitations for WSD. While dictionaries provide detailed information at the lexical level, they lack pragmatic information that enters into sense determination (see, e.g., Hobbs [1987]). For example, the link between *ash* and *tobacco*, *cigarette*, or *tray* in a network such as Quillian's is very indirect, whereas in the Brown corpus, the word *ash* co-occurs frequently with one of these words. It is therefore not surprising that corpora have become a primary source of information for WSD; this development is outlined below in Section 2.3.

**2.3.2 Thesauri.** Thesauri provide information about relationships among words, most notably synonymy. *Roget's International Thesaurus*, which was put into machine-tractable form in the 1950s and has been used in a variety of applications including machine translation (Masterman 1957), information retrieval (Sparck-Jones 1964, 1986), and content analysis (Sedelow and Sedelow [1969], see also Sedelow and Sedelow [1986, 1992]), also supplies an explicit concept hierarchy consisting of up to eight increasingly refined levels.[8] Typically, each occurrence of the same word under different categories of the thesaurus represents different senses of that word; i.e., the categories correspond roughly to word senses (Yarowsky 1992). A set of words in the same category are semantically related.

The earliest known use of *Roget's* for WSD is the work of Masterman (1957), described above in Section 2.1. Several years later, Patrick (1985) used *Roget's* to discriminate among verb senses, by examining semantic clusters formed by "e-chains" derived from the thesaurus (Bryan [1973, 1974]; see also Sedelow and Sedelow [1986]). He uses "word-strong neighborhoods," comprising word groups in low-level semicolon groups, which are the most closely related semantically in the thesaurus, and words connected to the group via chains. He is able, he claims, to discriminate the correct sense of verbs such as *inspire* (*to raise the spirits* vs. *to inhale, breathe in, sniff*, etc.), and *question* (*to doubt* vs. *to ask a question*) with high reliability. Bryan's earlier work had already demonstrated that homographs can be distinguished by applying a metric based on relationships defined by his chains (Bryan 1973, 1974). Similar work is described in Sedelow and Mooney (1988).

Yarowsky (1992) derives classes of words by starting with words in common categories in *Roget's* (4th edition). A 100-word context of each word in the category is extracted from a corpus (the 1991 electronic text of *Grolier's Encyclopedia*), and a mutual-information-like statistic is used to identify words most likely to co-occur with the category members. The resulting classes are used to disambiguate new occurrences of a polysemous word: the 100-word context of the polysemous occurrence is examined for words in various classes, and Bayes' Rule is applied to determine the class most likely to be that of the polysemous word. Since class is assumed by Yarowsky to represent a particular sense of a word, assignment to a class identifies the sense. He reports 92% accuracy on a mean three-way sense distinction. Yarowsky notes that his method is best for extracting topical information, which is in turn most successful for disambiguating nouns (see Section 3.1.2). He uses the broad category distinctions supplied by *Roget's*, although he points out that the lower-level information may provide

8 The work of Masterman (1957) and Sparck-Jones (1964) relied on a version of *Roget's* that was hand-punched onto cards in the 1950s; the Sedelows' (1969) work relied on a machine-readable version of the 3rd edition. *Roget's* is now widely available via anonymous ftp from various sites.

rich information for disambiguation. Patrick's much earlier study, on the other hand, exploits the lower levels of the concept hierarchy, in which words are more closely related semantically, as well as connections among words within the thesaurus itself; however, despite its promise this work has not been built upon since.

Like machine-readable dictionaries, a thesaurus is a resource created for humans and is therefore not a source of perfect information about word relations. It is widely recognized that the upper levels of its concept hierarchy are open to disagreement (although this is certainly true for any concept hierarchy), and that they are so broad as to be of little use in establishing meaningful semantic categories. Nonetheless, thesauri provide a rich network of word associations and a set of semantic categories potentially valuable for language-processing work; however, *Roget's* and other thesauri have not been used extensively for WSD.[9]

**2.3.3 Computational Lexicons.** In the mid-1980s, work began on the construction of large-scale knowledge bases by hand, for example, WordNet (Miller et al. 1990; Fellbaum forthcoming-a), CyC (Lenat and Guha 1990), ACQUILEX (Briscoe 1991), COMLEX (Grishman, Macleod, and Meyers 1994; Macleod, Grishman, and Myers, forthcoming). There exist two fundamental approaches to the construction of semantic lexicons: the **enumerative** approach, wherein senses are explicitly provided, and the **generative** approach, in which semantic information associated with given words is underspecified, and generation rules are used to derive precise sense information (Fellbaum, forthcoming-b).

*Enumerative Lexicons.* Among enumerative lexicons, WordNet (Miller et al. 1990; Fellbaum, forthcoming-a, forthcoming-b) is at present the best-known and the most utilized resource for word sense disambiguation in English. WordNet versions for several western and eastern European languages are currently under development (Vossen, forthcoming; Sutcliffe et al., *An Interactive Approach*, 1996, Sutcliffe et al., *IWNR*, 1996).

WordNet combines the features of many of the other resources commonly exploited in disambiguation work: it includes definitions for individual senses of words within it, as in a dictionary; it defines "synsets" of synonymous words representing a single lexical concept, and organizes them into a conceptual hierarchy,[10] like a thesaurus; and it includes other links among words according to several semantic relations, including hyponymy/hyperonymy, antonymy, and meronymy. As such, it currently provides the broadest set of lexical information in a single resource. Another, possibly more compelling, reason for WordNet's widespread use is that it is the first broad-coverage lexical resource that is freely and widely available; as a result, whatever its limitations, WordNet's sense divisions and lexical relations are likely to impact the field for several years to come.[11]

Some of the earliest attempts to exploit WordNet for sense disambiguation are in the field of information retrieval. Using the hyponomy links for nouns in WordNet, Voorhees (1993) defines a construct called a **hood** in order to represent sense categories, much as *Roget's* categories are used in the methods outlined above. A hood for a given word $w$ is defined as the largest connected subgraph that contains $w$. For each content

9 Other thesauri have been used for WSD, e.g., the German Hallig-Wartburg (see Schmidt [1988, 1991]) and the *Longman Lexicon of Contemporary English* (LLOCE) (Chen and Chang, this volume).

10 Note that the structure is not a perfect hierarchy since some of the synsets have more than one parent.

11 A recent workshop to set up common evaluations mechanisms for word sense disambiguation acknowledged the fact that, due to its availability, WordNet is, at present, the most used lexical resource for disambiguation in English, and therefore determined that WordNet senses should form the basis for a common sense inventory (Kilgarriff 1997).

word in a document collection, Voorhees computes the number of times each synset appears above that word in the WordNet noun hierarchy, which gives a measure of the expected activity (**global** counts); she then performs the same computation for words occurring in a particular document or query (**local** counts). The sense corresponding to the hood root for which the difference between the global and local counts is the greatest is chosen for that word. Her results, however, indicate that her technique is not a reliable method for distinguishing WordNet's fine-grained sense distinctions. In a similar study, Richardson and Smeaton (1994) create a knowledge base from WordNet's hierarchy and apply a semantic similarity function (developed by Resnik—see below) to accomplish disambiguation, also for the purposes of information retrieval. They provide no formal evaluation but indicate that their results are "promising."

Sussna (1993) computes a semantic distance metric for each of a set of input text terms (nouns) in order to disambiguate them. He assigns weights based on the relation type (synonymy, hyperonymy, etc.) to WordNet links, and defines a metric that takes account of the number of arcs of the same type leaving a node and the depth of a given edge in the overall "tree." This metric is applied to arcs in the shortest path between nodes (word senses) to compute semantic distance. The hypothesis is that for a given set of terms occurring near each other in a text, choosing the senses that minimize the distance among them selects the correct senses. Sussna's disambiguation results are demonstrated to be significantly better than chance. His work is particularly interesting because it is one of the few to date that utilizes not only WordNet's IS-A hierarchy, but other relational links as well.

Resnik (1995a) draws on his body of earlier work on WordNet, in which he explores a measure of semantic similarity for words in the WordNet hierarchy (Resnik 1993a, 1993b, 1995a). He computes the shared **information content** of words, which is a measure of the specificity of the concept that subsumes the words in the WordNet IS-A hierarchy—the more specific the concept that subsumes two or more words, the more semantically related they are assumed to be. Resnik contrasts his method of computing similarity to those which compute path length (e.g., Sussna 1993), arguing that the links in the WordNet taxonomy do not represent uniform distances (cf. Resnik 1995b). Resnik's method, applied using WordNet's fine-grained sense distinctions and measured against the performance of human judges, approaches human accuracy. Like the other studies cited here, his work considers only nouns.

WordNet is not a perfect resource for word sense disambiguation. The most frequently cited problem is the fine-grainedness of WordNet's sense distinctions, which are often well beyond what may be needed in many language-processing applications (see Section 3.2). Voorhees' (1993) hood construct is an attempt to access sense distinctions that are less fine-grained than WordNet's synsets, and less coarse-grained than the 10 WordNet noun hierarchies; Resnik's (1995a) method allows for detecting sense distinctions at any level of the WordNet hierarchy. However, it is not clear what the desired level of sense distinction should be for WSD (or if it is the same for all word categories, all applications, etc.), or if this level is even captured in WordNet's hierarchy. Discussion within the language-processing community is beginning to address these issues, including the most difficult one of defining what we mean by "sense" (see Section 3.2).

*Generative Lexicons.* Most WSD work to date has relied upon enumerative sense distinctions as found in dictionaries. However, there has been recent work on WSD which has exploited generative lexicons (Pustejovsky 1995), in which *related* senses (i.e., systematic polysemy, as opposed to homonymy) are not enumerated but rather are generated from rules that capture regularities in sense creation, as for metonymy, meronymy, etc.

As outlined in Buitelaar (1997), sense disambiguation in the generative context starts with a semantic tagging that points to a complex knowledge representation reflecting all of a word's systematically related senses, after which semantic processing may derive a discourse-dependent interpretation containing more precise sense information about the occurrence. Buitelaar (1997) describes the use of CORELEX for underspecified semantic tagging (see also Pustejovsky, Boguraev, and Johnston [1995]).

Viegas, Mahesh, and Nirenburg (forthcoming) describe a similar approach to WSD undertaken in the context of their work on machine translation (see also Mahesh et al. [1997] and Mahesh, Nirenburg, and Beale [1997]). They access a large syntactic and semantic lexicon that provides detailed information about constraints, such as selectional restrictions, for words in a sentence, and then search a richly connected ontology to determine which senses of the target word best satisfy these constraints. They report a success rate of 97%. Like CORELEX, both the lexicon and the ontology are manually constructed, and therefore still limited, although much larger than the resources used in earlier work. However, Buitelaar (1997) describes means to automatically generate CORELEX entries from corpora in order to create domain-specific semantic lexicons, thus demonstrating the potential to access larger-scale resources of this kind.

### 2.4 Corpus-based Methods

**2.4.1 Growth, Decline, and Re-emergence of Empirical Methods.** Since the end of the nineteenth century, the manual analysis of corpora has enabled the study of words and graphemes (Kaeding 1897–1898, Estoup 1902, Zipf 1935) and the extraction of lists of words and collocations for the study of language acquisition or language teaching (Thorndike 1921; Fries and Traver 1940; Thorndike and Lorge 1938, 1944; Gougenheim et al. 1956; etc.). Corpora have been used in linguistics since the first half of the twentieth century (e.g., Boas 1940; Fries 1952). Some of this work concerns word senses, and it is often strikingly modern: for example, Palmer (1933) studied collocations in English; Lorge (1949) computed sense frequency information for the 570 most common English words; Eaton (1940) compared the frequency of senses in four languages; and Thorndike (1948) and Zipf (1945) determined that there is a positive correlation between the frequency and the number of synonyms of a word, the latter of which is an indication of semantic richness (the more polysemous a word, the more synonyms it has).

A corpus provides a bank of **samples** that enable the development of numerical language models, and thus the use of corpora goes hand-in-hand with empirical methods. Although quantitative/statistical methods were embraced in early MT work, in the mid-1960s interest in statistical treatment of language waned among linguists due to the trend toward the discovery of formal linguistic rules sparked by the theories of Zellig Harris (1951) and bolstered most notably by the transformational theories of Noam Chomsky (1957).[12] Instead, attention turned toward full linguistic analysis and hence toward sentences rather than texts, and toward contrived examples and artificially limited domains instead of general language. During the following 10 to

12 Not all linguists completely abandoned the empirical approach at this time; consider, for instance, Pendergraft's (1967) comment:

> It would be difficult, indeed, in the face of today's activity, not to acknowledge the triumph of the theoretical approach, more precisely, of formal rules as the preferred successor of lexical and syntactic search algorithms in linguistic description. At the same time, common sense should remind us that hypothesis-making is not the whole of science, and that discipline will be needed if the victory is to contribute more than a haven from the rigors of experimentation (p. 313).

15 years, only a handful of linguists continued to work with corpora, most often for pedagogical or lexicographic ends (e.g., Quirk 1960; Michéa 1964). Despite this, several important corpora were developed during this period, including the Brown Corpus (Kucera and Francis 1967), the *Trésor de la Langue Française* (Imbs 1971), and the Lancaster-Oslo-Bergen (LOB) corpus (Johansson 1980). In the area of natural language processing, the ALPAC report (1966) recommended intensification of corpus-based research for the creation of broad-coverage grammars and lexicons, but because of the shift away from empiricism, little work was done in this area until the 1980s. Until then, the use of statistics for language analysis was almost the exclusive property of researchers in the fields of literary and humanities computing, information retrieval, and the social sciences. Within these fields, work on WSD continued, most notably in the Harvard "disambiguation project" for content analysis (Stone et al. 1966; Stone 1969), and also in the work of Iker (1974, 1975), Choueka and Dreizin (1976) and Choueka and Goldberg (1979).

In the context of the shift away from the use of corpora and empirical methods, the work of Weiss (1973) and Kelley and Stone (1975) on the automatic extraction of knowledge for word sense disambiguation seems especially innovative. Weiss (1973) demonstrated that disambiguation rules can be learned from a manually sense-tagged corpus. Despite the small size of his study (five words, a training set of 20 sentences for each word, and 30 test sentences for each word), Weiss's results are encouraging (90% correct). Kelley and Stone's (1975) work, which grew out of the Harvard "disambiguation project" for content analysis, is on a much larger scale; they extract KWIC concordances for 1,800 ambiguous words from a corpus of a half-million words. The concordances serve as a basis for the manual creation of disambiguation rules ("word tests") for each sense of the 1,800 words. The tests—also very sophisticated for the time—examine the target word context for clues on the basis of collocational information, syntactic relations with context words, and membership in common semantic categories. Their rules perform even better than Weiss's, achieving 92% accuracy for gross homographic sense distinctions.

In the 1980s, interest in corpus linguistics was revived (see, for example, Aarts [1990] and Leech [1991]). Advances in technology enabled the creation and storage of corpora larger than had been previously possible, enabling the development of new models most often utilizing statistical methods. These methods were rediscovered first in speech processing (e.g., Jelinek [1976]; see the overview by Church and Mercer [1993] and the collection of reprints by Waibel and Lee [1990]) and were immediately applied to written language analysis (e.g., in the work of Bahl and Mercer [1976], Debili [1977], etc.). For a discussion, see Ide and Walker (1992).

In the area of word sense disambiguation, Black (1988) developed a model based on decision trees using a corpus of 22 million tokens, after manually sense-tagging approximately 2,000 concordance lines for five test words. Since then, **supervised learning** from sense-tagged corpora has since been used by several researchers: Zernik (1990, 1991), Hearst (1991), Leacock, Towell, and Voorhees (1993), Gale, Church, and Yarowsky (1992d, 1993), Bruce and Wiebe (1994), Miller et al. (1994), Niwa and Nitta (1994), Lehman (1994), among others. However, despite the availability of increasingly large corpora, two major obstacles impede the acquisition of lexical knowledge from corpora: the difficulties of manually sense-tagging a training corpus, and data sparseness.

**2.4.2 Automatic Sense-Tagging.** Manual sense-tagging of a corpus is extremely costly, and, at present, very few sense-tagged corpora are available. Several efforts to create sense-tagged corpora have been or are being made: the Linguistic Data Consortium

distributes a corpus of approximately 200,000 sentences from the Brown Corpus and the *Wall Street Journal* in which all occurrences of 191 words are hand-tagged with their WordNet senses (see Ng and Lee [1996]). Also, the Cognitive Science Laboratory at Princeton has undertaken the hand-tagging of 1,000 words from the Brown Corpus with WordNet senses (Miller et al. 1993) (so far, 200,000 words are available via ftp), and hand-tagging of 25 verbs in a small segment of the *Wall Street Journal* (12,925 sentences), is also underway (Wiebe et al. 1997). However, these corpora are far smaller than those typically used with statistical methods.

Several efforts have been made to automatically sense-tag a training corpus via **bootstrapping** methods. Hearst (1991) proposed an algorithm (CatchWord) that includes a training phase during which each occurrence of a set of nouns to be disambiguated is manually sense-tagged in several occurrences.[13] Statistical information extracted from the context of these occurrences is then used to disambiguate other occurrences. If another occurrence can be disambiguated with certitude, the system automatically acquires additional statistical information from these newly disambiguated occurrences, thus improving its knowledge incrementally. Hearst indicates that an initial set of at least 10 occurrences is necessary for the procedure, and that 20 or 30 occurrences are necessary for high precision. This overall strategy is more or less that of most subsequent work on bootstrapping. Recently, a class-based bootstrapping method for semantic tagging in specific domains has been proposed (Basili et al. 1997).

Schütze (1992, 1993) proposes a method that avoids tagging each occurrence in the training corpus. Using letter fourgrams within a 1,001-character window, his method, building on the vector-space model from information retrieval (see Salton, Wong, and Yang [1975]), automatically clusters the words in the text (each target word is represented by a vector); a sense is then assigned manually to each cluster, rather than to each occurrence. Assigning a sense demands examining 10 to 20 members of each cluster, and each sense may be represented by several clusters. This method reduces the amount of manual intervention but still requires the examination of a hundred or so occurrences for each ambiguous word. A more serious issue for this method is that it is not clear what the senses derived from the clusters correspond to (see, for example Pereira, Tishby, and Lee [1993]); moreover, the senses are not directly usable by other systems, since they are derived from the corpus itself.

Brown et al. (1991) and Gale, Church, and Yarowsky, (1992a, 1993) propose the use of bilingual corpora to avoid hand-tagging of training data. Their premise is that different senses of a given word often translate differently in another language (for example, *pen* in English is *stylo* in French for its 'writing implement' sense, and *enclos* for its 'enclosure' sense). By using a parallel aligned corpus, the translation of each occurrence of a word such as *pen* can be used to automatically determine its sense. This method has some limitations, since many ambiguities are preserved in the target language (e.g., French *souris*—English *mouse*); furthermore, the few available large-scale parallel corpora are very specialized (for example, the Hansard corpus of Canadian Parliamentary Debates), which skews the sense representation.[14] Dagan, Itai, and Schwall (1991) and Dagan and Itai (1994) propose a similar method, but instead of a parallel corpus use two monolingual corpora and a bilingual dictionary. This solves, in part, the problems of availability and specificity of domain that plague the parallel corpus approach, since monolingual corpora, including corpora from diverse domains and genres, are much easier to obtain than parallel corpora.

13 This study involves only nouns.

14 For example, Gale, Church, and Yarowsky (1993) remark that it is difficult to find any sense other than the financial sense for the word *bank* in the Hansard corpus.

Other methods attempt to avoid entirely the need for a tagged corpus, such as many of those cited in the section below (e.g., Yarowsky [1992] who attacks both the tagging and data sparseness problems simultaneously). However, it is likely that, as noted for grammatical tagging (Mérialdo 1994), even a minimal phase of supervised learning improves radically on the results of unsupervised methods. Research into means to facilitate and optimize tagging is ongoing; for example, an optimization technique called **committee-based sample selection** has recently been proposed (Engelson and Dagan 1996), which, based on the observation that a substantial portion of manually tagged examples contribute little to performance, enables avoiding the tagging of examples that carry more or less the same information. Such methods are promising, although to our knowledge they have not been applied to the problem of lexical disambiguation.

**2.4.3 Overcoming Data Sparseness.** The problem of data sparseness, which is common for much corpus-based work, is especially severe for work in WSD. First, enormous amounts of text are required to ensure that all senses of a polysemous word are represented, given the vast disparity in frequency among senses. For example, in the Brown Corpus (one million words), the relatively common word *ash* occurs only eight times, and only once in its sense as *tree*. The sense *ashes = remains of cremated body*, although common enough to be included in learner's dictionaries such as the LDOCE and the OALD, does not appear, and it would be nearly impossible to find the dozen or so senses in many everyday dictionaries such as the CED. In addition, the many possible co-occurrences for a given polysemous word are unlikely to be found in even a very large corpus, or they occur too infrequently to be significant.[15]

**Smoothing** is used to get around the problem of infrequently occurring events, and in particular to ensure that non-observed events are not assumed to have a probability of zero. The best-known smoothing methods are that of Turing-Good (Good 1953), which hypothesizes a binomial distribution of events, and that of Jelinek and Mercer (1985), which combines estimated parameters on distinct subparts of the training corpus.[16] However, these methods do not enable distinguishing between events with the same frequency, such as the *ash-cigarette* and *ash-room* example given in footnote 15. Church and Gale (1991) have proposed a means to improve methods for the estimation of bigrams, which could be extended to co-occurrences: they take into account the frequency of the individual words that compose the bigram and make the hypothesis that each word appears independently of the others. However, this hypothesis contradicts hypotheses of disambiguation based on co-occurrence, which rightly assume that some associations are more probable than others.

**Class-based models** attempt to obtain the best estimates by combining observations of classes of words considered to belong to a common category. Brown et al. (1992), Pereira and Tishby (1992), and Pereira, Tishby, and Lee (1993) propose methods that derive classes from the distributional properties of the corpus itself, while other authors use external information sources to define classes: Resnik (1992) uses the taxonomy of WordNet; Yarowsky (1992) uses the categories of *Roget's Thesaurus*, Slator (1992) and Liddy and Paik (1993) use the subject codes in the LDOCE; Luk (1995) uses **conceptual sets** built from the LDOCE definitions. Class-based methods answer in part the problem of data sparseness and eliminate the need for pretagged

15 For example, in a window of five words to each side of the word *ash* in the Brown corpus, commonly associated words such as *fire, cigar, volcano*, etc., do not appear. The words *cigarette* and *tobacco* co-occur with *ash* only once, with the same frequency as words such as *room, bubble*, and *house*.

16 See the survey of methods in Chen and Goodman (1996).

data. However, there is some information loss with these methods because the hypothesis that all words in the same class behave in a similar fashion is too strong. For example, *residue* is a hypernym of *ash* in *WordNet*; its hyponyms form the class {*ash, cotton(seed) cake, dottle*}. Obviously the members of this set of words behave very differently in context: *volcano* is strongly related to *ash*, but has little or no relation to the other words in the set.

**Similarity-based methods** Dagan, Marcus, and Markovitch 1993, Dagan, Pereira, and Lee 1994, and Grishman and Sterling 1993 exploit the same idea of grouping observations for similar words, but without regrouping them into fixed classes. Each word has a potentially different set of similar words. Like many class-based methods, such as Brown et al. (1992), similarity-based methods exploit a similarity metric between patterns of co-occurrence. Dagan, Marcus, and Markovitch (1993) give the following example: the pair (*chapter, describes*) does not appear in their corpus; however, *chapter* is similar to *book, introduction,* and *section,* which are paired with *describes* in the corpus. On the other hand, the words similar to *book* are *books, documentation,* and *manuals* (see their Figure 1). Dagan, Marcus, and Markovitch's (1993) evaluation seems to show that similarity-based methods perform better than class-based methods. Karov and Edelman (this volume) propose an extension to similarity-based methods by means of an iterative process at the learning stage, which gives results that are 92% accurate on four test words—approximately the same as the best results cited in the literature to date. These results are particularly impressive given that the training corpus contains only a handful of examples for each word, rather than the hundreds of examples required by most methods.

## 3. Open Problems

We have already noted various problems faced in current WSD research related to specific methodologies. Here, we discuss issues and problems that all approaches to WSD must face and suggest some directions for further work.

### 3.1 The Role of Context

Context is the only means to identify the meaning of a polysemous word. Therefore, all work on sense disambiguation relies on the context of the target word to provide information to be used for its disambiguation. For data-driven methods, context also provides the prior knowledge with which current context is compared to achieve disambiguation.

Broadly speaking, context is used in two ways:

- *Bag-of-words approach*: context is considered as words in some window surrounding the target word, regarded as a group without consideration for their relationships to the target in terms of distance, grammatical relations, etc.
- *Relational information*: context is considered in terms of some relation to the target, including distance from the target, syntactic relations, selectional preferences, orthographic properties, phrasal collocation, semantic categories, etc.

Information from microcontext, topical context, and domain contributes to sense selection, but the relative roles and importance of information from the different contexts, and their interrelations, are not well understood. Very few studies have used

information of all three types, and the focus in much recent work is on microcontext alone. This is another area where systematic study is needed for WSD.

**3.1.1 Microcontext.** Most disambiguation work uses the local context of a word occurrence as a primary information source for WSD. Local or "micro" context is generally considered to be some small window of words surrounding a word occurrence in a text or discourse, from a few words of context to the entire sentence in which the target word appears.

Context is very often regarded as all words or characters falling within some window of the target, with no regard for distance, syntactic structure, or other relations. Early corpus-based work, such as that of Weiss (1973) used this approach; spreading activation and dictionary-based approaches also do not usually differentiate context input on any basis other than occurrence in a window. Schütze's vector space method (this volume) is a recent example of an approach that ignores adjacency information. Overall, the bag-of-words approach has been shown to work better for nouns than for verbs (cf. Schütze, this volume), and to be in general less effective than methods that take other relations into consideration. However, as demonstrated in Yarowsky's (1992) work, the approach is cheaper than those requiring more complex processing and can achieve sufficient disambiguation for some applications. We examine below some of the other parameters.

*Distance.* It is obvious from the quotation in Section 2.1 from Weaver's memorandum that the notion of examining a context of a few words around the target to disambiguate has been fundamental to WSD work since its beginnings: it has been the basis of WSD work in MT, content analysis, AI-based disambiguation, and dictionary-based WSD, as well as the more recent statistical, neural network, and symbolic machine learning, approaches. However, following Kaplan's early experiments (Kaplan 1950), there have been few systematic attempts to answer Weaver's question concerning the optimal value of $N$. A notable exception is the study of Choueka and Lusignan (1985), who verified Kaplan's finding that 2-contexts are highly reliable for disambiguation, and even 1-contexts are reliable in 8 out of 10 cases. However, despite these findings, the value of $N$ has continued to vary over the course of WSD work more or less arbitrarily.

Yarowsky (1993, 1994a, 1994b) examines different windows of microcontext, including 1-contexts, $k$-contexts, and words pairs at offsets $-1$ and $-2$, $-1$ and $+1$, and $+1$ and $+2$, and sorts them using a log-likelihood ratio to find the most reliable evidence for disambiguation. Yarowsky makes the observation that the optimal value of $k$ varies with the kind of ambiguity: he suggests that local ambiguities need only a window of $k = 3$ or 4, while semantic or topic-based ambiguities require a larger window of 20–50 words (see Section 3.1.2). No single best measure is reported, suggesting that for different ambiguous words, different distance relations are more efficient. Furthermore, because Yarowsky also uses other information (such as part of speech), it is difficult to isolate the impact of window-size alone. Leacock, Chodorow, and Miller (this volume) use a local window of $\pm 3$ open-class words, arguing that this number showed best performance in previous tests.

*Collocation.* The term "collocation" has been used variously in WSD work. The term was popularized by J. R. Firth in his 1951 paper "Modes of meaning": "One of the meanings of *ass* is its habitual collocation with an immediately preceding *you silly*...." He emphasizes that collocation is not simple co-occurrence but is "habitual"

or "usual."[17] Halliday's (1961) definition of collocation as "the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur at $n$ removes (a distance of $n$ lexical items) from an item $x$, the items $a$, $b$, $c$..." is more workable in computational terms.

Based on this definition, a **significant collocation** can be defined as a syntagmatic association among lexical items, where the probability of item $x$ co-occurring with items $a$, $b$, $c$ ... is greater than chance (Berry-Rogghe 1973). It is in this sense that most WSD work uses the term. There is some psychological evidence that collocations are treated differently from other co-occurrences. For example, Kintsch and Mross (1985) show that priming words that enter frequent collocations with test words (i.e., *iron-steel*, which they call **associative context**) activate these test words in lexical decision tasks. Conversely, priming words that are in the **thematic context** (i.e., relations determined by the situation, scenario, or script such as *plane-gate*) do not facilitate the subjects' lexical decisions (see also Fischler [1977], Seidenberg et al. [1982], De Groot [1983], Lupker [1984]).

Yarowsky (1993) explicitly addresses the use of collocations in WSD work, but admittedly adapts the definition to his purpose as "the co-occurrence of two words in some defined relation." As noted above, he examines a variety of distance relations, but also considers adjacency by part of speech (e.g., first noun to the left). He determines that in cases of binary ambiguity, there exists one sense per collocation, that is, in a given collocation, a word is used with only one sense with 90–99% probability.

*Syntactic Relations.* Earl (1973) used syntax exclusively for disambiguation in machine translation. In most WSD work to date, syntactic information is used in conjunction with other information. The use of selectional restrictions weighs heavily in AI-based work that relies on full parsing, frames, semantic networks, the application of selectional preferences, etc. (Hayes 1977a, 1997b; Wilks 1973 and 1975b; Hirst 1987). In other work, syntax is combined with frequent collocation information: Kelley and Stone (1975), Dahlgren (1988), and Atkins (1987) combine collocation information with rules for determining, for example, the presence or absence of determiners, pronouns, noun complements, as well as prepositions, subject-verb and verb-object relations.

More recently, researchers have avoided complex processing by using shallow or partial parsing. In her disambiguation work on nouns, Hearst (1991) segments text into noun phrases, prepositional phrases, and verb groups, and discards all other syntactic information. She examines items that are within $\pm 3$ phrase segments from the target and combines syntactic evidence with other kinds of evidence, such as capitalization. Yarowsky (1993) determines various behaviors based on syntactic category; for example, that verbs derive more disambiguating information from their objects than from their subjects, adjectives derive almost all disambiguating information from the nouns they modify, and nouns are best disambiguated by directly adjacent adjectives or nouns. In recent work, syntactic information most often is simply part of speech, used invariably in conjunction with other kinds of information (McRoy 1992; Bruce and Wiebe 1994; Leacock, Chodorow, and Miller, this volume).

Evidence suggests that different kinds of disambiguation procedures are needed depending on the syntactic category and other characteristics of the target word (Yarowsky 1993; Leacock, Chodorow, and Miller, this volume)—an idea reminiscent of the word expert approach. However, to date there has been little systematic study

17 Later, several attempts were made to define the term more precisely in the framework of modern linguistic theory. See, for example, Haas (1966), Halliday (1961, 1966), Lyons (1966), McIntosh (1966), Sinclair (1966), van Buren (1967).

of the contribution of different information types for different types of target words. It is likely that this is a next necessary step in WSD work.

**3.1.2 Topical Context.** Topical context includes substantive words that co-occur with a given sense of a word, usually within a window of several sentences. Unlike microcontext, which has played a role in disambiguation work since the early 1950s, topical context has been less consistently used. Methods relying on topical context exploit redundancy in a text—that is, the repeated use of words that are semantically related throughout a text on a given topic. Thus, *base* is ambiguous, but its appearance in a document containing words such as *pitcher*, and *ball* is likely to isolate a given sense for that word (as well as the others, which are also ambiguous). Work involving topical context typically uses the bag-of-words approach, in which words in the context are regarded as an unordered set.

The use of topical context has been discussed in the field of information retrieval for several years (Anthony 1954; Salton 1968). Recent WSD work has exploited topical context: Yarowsky (1992) uses a 100-word window, both to derive classes of related words and as context surrounding the polysemous target, in his experiments using *Roget's Thesaurus* (see Section 2.3.2). Voorhees, Leacock, and Towell (1995) experiment with several statistical methods using a two-sentence window; Leacock, Towell, and Voorhees (1993, 1996) have similarly explored topical context for WSD. Gale, Church, and Yarowsky (1993), looking at a context of ±50 words, indicate that while words closest to the target contribute most to disambiguation, they improved their results from 86% to 90% by expanding context from ±6 (a typical span when only microcontext is considered) to ±50 words around the target. In a related study, they make a claim that for a given discourse, ambiguous words are used in a single sense with high probability ("one sense per discourse") (Gale, Church, and Yarowsky 1992c). Leacock, Chodorow, and Miller (this volume) challenge this claim in their work combining topical and local context, which shows that both topical and local context are required to achieve consistent results across polysemous words in a text (see also Towell and Voorhees, this volume). Yarowsky's (1993) study indicates that while information within a large window can be used to disambiguate nouns, for verbs and adjectives the size of the usable window drops off dramatically with distance from the target word. This supports the claim that both local and topical context are required for disambiguation, and points to the increasingly accepted notion that different disambiguation methods are appropriate for different kinds of words.

Methods utilizing topical context can be ameliorated by dividing the text under analysis into subtopics. The most obvious way to divide a text is by sections (Brown and Yule 1983), but this is only a gross division; subtopics evolve inside sections, often in unified groups of several paragraphs. Automatic segmentation of texts into such units would obviously be helpful for WSD methods that use topical context. It has been noted that the repetition of words within successive segments or sentences is a strong indicator of the structure of discourse (Skorochod'ko 1972; Morris 1988; Morris and Hirst 1991); methods exploiting this observation to segment a text into subtopics are beginning to emerge (see, for example, Hearst [1994], van der Eijk [1994], Richmond, Smith, and Amitay [1997]).

In this volume, Leacock, Chodorow, and Miller consider the role of microcontext vs. topical context and attempt to assess the contribution of each. Their results indicate that for a statistical classifier, microcontext is superior to topical context as an indicator of sense. However, although a distinction is made between microcontext and topical context in current WSD work, it is not clear that this distinction is meaningful. It may be more useful to regard the two as lying along a continuum, and to consider the role

and importance of contextual information as a function of distance from the target.

**3.1.3 Domain.** The use of domain for WSD is first evident in the microglossaries developed in early MT work (see Section 2.1). The notion of disambiguating senses based on domain is implicit in various AI-based approaches, such as Schank's script approach to natural language processing (Schank and Abelson 1977), which matched words to senses based on the context or "script" activated by the general topic of the discourse. This approach, which activates *only* the sense of a word relevant to the current discourse domain, demonstrates its limitations of this approach when used in isolation; in the famous example *The lawyer stopped at the bar for a drink*, the incorrect sense of *bar* will be assumed if one relies only on the information in a script concerned with law.[18]

Gale, Church, and Yarowsky's (1992c) claim for one sense per discourse is disputable. Dahlgren (1988) observes that domain does not eliminate ambiguity for some words: she remarks that the noun *hand* has 16 senses (or so) and retains 10 of them in almost any text. The influence of domain likely depends on factors such as the type of text (how technical the text is, etc.), the relation among the senses of the target word (strongly or weakly polarized, common vs. specialized usage, etc.). For example, in the French *Encyclopaedia Universalis*, the word *intérêt* ("interest") appears 62 times in the article on INTEREST—FINANCE, in all cases in its financial sense; the word appears 139 times in the article INTEREST—PHILOSOPHY AND HUMANITIES in its common, nonfinancial, sense. However, in the article THIRD WORLD, the word *intérêt* appears two times in each of these senses.

### 3.2 Sense Division

**3.2.1 The Bank Model.** Most researchers in WSD are currently relying on the sense distinctions provided by established lexical resources, such as machine-readable dictionaries or WordNet (which uses the OALD's senses), because they are widely available. The dominant model in these studies is the "bank" model, which attempts to extend the clear delineation between *bank-money* and *bank-riverside* to all sense distinctions. However, it is clear that this convenient delineation is by no means applicable to all or even most other words. Although there is some psychological validity to the notion of senses (Simpson and Burgess 1988; Jorgensen 1990), lexicographers themselves are well aware of the lack of agreement on senses and sense divisions (see, for example, Malakhovski [1987], Robins [1987], Ayto [1983], Stock [1983]). The problem of sense division has been an object of discussion since antiquity: Aristotle[19] devoted a section of his *Topics* to this subject in 350 B.C. Since then, philosophers and linguists have continued to discuss the topic at length (see Quine [1960], Asprejan [1974], Lyons [1977], Weinrich [1980], Cruse [1986]), but the lack of resolution over 2,000 years is striking.

**3.2.2 Granularity.** One of the foremost problems for WSD is to determine the appropriate degree of sense granularity. Several authors (for example, Slator and Wilks [1987]) have remarked that the sense divisions one finds in dictionaries are often too fine for the purposes of NLP work. Overly fine sense distinctions create practical difficul-

18 An interesting development based on Schank's approach is described in Granger (1977), where he utilizes information in scripts and conceptual dependency representations of sentences to determine the meaning of entirely unknown words encountered in text. The approach, which examines domain and contextual evidence to determine meaning, is similar to that employed in much AI-based work on disambiguation.

19 One of the reviewers for this special issue remarked humorously that if Aristotle had had a PC, he would have probably worked on word sense disambiguation!

ties for automated WSD: they introduce significant combinatorial effects (for example, Slator and Wilks [1987] note that the sentence *There is a huge envelope of air around the surface of the earth* has 284,592 different potential combined sense assignments using the moderately-sized LDOCE); they require making sense choices that are extremely difficult, even for expert lexicographers; and they increase the amount of data required for supervised methods to unrealistic proportions. In addition, the sense distinctions made in many dictionaries are sometimes beyond those which human readers themselves are capable of making. In a well-known study, Kilgarriff (1992, 1993) shows that it is impossible for human readers to assign many words to a unique sense in LDOCE (see, however, the discussion in Wilks [forthcoming]). Recognizing this, Dolan (1994) proposes a method for "ambiguating" dictionary senses by combining them to create grosser sense distinctions. Others have used the grosser sense divisions of thesauri such as *Roget's*; however, it is often difficult to assign a unique sense, or even find an appropriate one among the options (see, for example, Yarowsky [1992]). Chen and Chang (this volume) propose an algorithm that combines senses in a dictionary (LDOCE) and links them to the categories of a thesaurus (LLOCE).

Combining dictionary senses does not solve the problem. First of all, the degree of granularity required is task dependent. Only homograph distinction is necessary for tasks such as speech synthesis or restoration of accents in text, while tasks such as machine translation require fine sense distinctions—in some cases finer than what monolingual dictionaries provide (see, for example, ten Hacken [1990]). For example, the English word *river* is translated as *fleuve* in French when the river flows into the ocean, and otherwise as *rivière*. There is not, however, a strict correspondence between a given task and the degree of granularity required. For example, as noted earlier, the word *mouse*, although it has two distinct senses (animal, device), translates into French in both cases to *souris*. On the other hand, for information retrieval the distinction between these two senses of *mouse* is important, whereas it is difficult to imagine a reason to distinguish *river* (sense *fleuve*) - *river* (sense *rivière*). Second, and more generally, it is unclear when senses should be combined or split. Even lexicographers do not agree: Fillmore and Atkins (1991) identify three senses of the word *risk* but find that most dictionaries fail to list at least one of them. In many cases, meaning is best considered as a continuum along which shades of meaning fall (see, for example, Cruse [1986]), and the points at which senses are combined or split can vary dramatically.

**3.2.3 Senses or usages?** The Aristotelian idea that words correspond to specific objects and concepts was displaced in the twentieth century by the ideas of Saussure and others (Meillet [1926], Hjemslev [1953], Martinet [1960], etc.). For Antoine Meillet, for example, the sense of a word is defined only by the average of its linguistic uses. Wittgenstein takes a similar position in his *Philosophische Utersuchungen*[20] in asserting that there are no senses, but only usages:

> "For a large class of cases—though not for all—in which we employ the word 'meaning' it can be defined thus: the meaning of a word is its use in the language" (1953, Sect. 43).

Similar views are apparent in more recent theories of meaning, for example, Bloomfield (1933) and Harris (1954), for whom meaning is a function of distribution; and in Barwise and Perry's (1953) situation semantics, where the sense or senses of a word are seen as an abstraction of the role that it plays systematically in the discourse.

20 Note that Wittgenstein had first defended the Aristotelian view in his *Tractatus*.

The COBUILD project (Sinclair 1987) adopts this view of meaning by attempting to anchor dictionary senses in current usage by creating sense divisions on the basis of **clusters** of citations in a corpus. Atkins (1987) and Kilgarriff (forthcoming) also implicitly adopt the view of Harris (1954), according to which each sense distinction is reflected in a distinct context. A similar view underlies the class-based methods cited in Section 2.4.3 (Brown et al. 1992; Pereira and Tishby 1992; Pereira, Tishby, and Lee 1993). In this volume, Schütze continues in this vein and proposes a technique that avoids the problem of sense distinction altogether: he creates sense clusters from a corpus rather than relying on a pre-established sense list.

**3.2.4 Enumeration or generation?** The development of generative lexicons (Pustejovsky 1995) provides a view of word senses that is very different from that of almost all WSD work to date. The enumerative approach assumes an a priori, established set of senses that exist independent of context—fundamentally the Aristotelian view. The generative approach develops a discourse-dependent representation of sense, assuming only underspecified sense assignments until context is taken into account, and bears closer relation to distributional and situational views of meaning.

Considering the difficulties of determining an adequate and appropriate set of senses for WSD, it is surprising that little attention has been paid to the potential of the generative view in WSD research. As larger and more complete generative lexicons become available, there is merit to exploring this approach to sense assignment.

### 3.3 Evaluation

Given the variety in the studies cited throughout the previous survey, it is obvious that it is very difficult to compare one set of results, and consequently one method, with another. The lack of comparability results from substantial differences in test conditions from study to study. For instance, different types of texts are involved, including both highly technical or domain-specific texts where sense use is limited and general texts where sense use may be more variable. It has been noted that in a commonly used corpus such as the *Wall Street Journal*, certain senses of typical test words such as *line* are absent entirely.[21] When different corpora containing different sense inventories and very different levels of frequency for a given word and/or sense are used, it becomes futile to attempt to compare results.

Test words themselves differ from study to study, including not only words whose assignment to clearly distinguishable senses varies considerably or which exhibit very different degrees of ambiguity (e.g., *bank* vs. *line*), but also words across different parts of speech and words that tend to appear more frequently in metaphoric, metonymic, and other nonliteral usages (e.g., *bank* vs. *head*). More seriously, the criteria for evaluating the correctness of sense assignment vary. Different studies employ different degrees of sense granularity (see Section 3.2 above), ranging from identification of homographs to fine sense distinctions. In addition, the means by which correct sense assignment is finally judged are typically unclear. Human judges must ultimately decide, but the lack of agreement among human judges is well documented: Amsler and White (1979) indicate that while there is reasonable consistency in sense assignment for a given expert on successive sense assignments (84%), agreement is significantly lower among experts. Ahlswede (1995) reports between 63.3% and 90.2% agreement among judges on his *Ambiguity Questionnaire*; when faced with on-line sense assign-

21 For example, the common sense of *line* as in the sentence, *He gave me a line of bologna*, is not present in the *Wall Street Journal* corpus.

ment in a large corpus, agreement among judges is far less, and in some cases worse than chance (see also Ahlswede [1992, 1993], Ahlswede and Lorand [1993]). Jorgensen (1990) found the level of agreement in her experiment using data from the Brown Corpus to be about 68%.

The difficulty of comparing results in WSD research has recently become a concern within the community, and efforts are underway to develop strategies for evaluation of WSD. Gale, Church, and Yarowsky (1992b) attempt to establish lower and upper bounds for evaluating the performance of WSD systems; their proposal for overcoming the problem of agreement among human judges in order to establish an upper bound provides a starting point, but it has not been widely discussed or implemented. A recent discussion at a workshop sponsored by the ACL Special Interest Group on the Lexicon (SIGLEX) on "Evaluating Automatic Semantic Taggers" (Resnik and Yarowsky [1997a]; see also Resnik and Yarowsky [1997b], Kilgarriff [1997]) has sparked the formation of an evaluation effort for WSD (SENSEVAL), in the spirit of previous evaluation efforts such as the ARPA-sponsored Message Understanding Conferences (e.g., ARPA [1993]), and Text Retrieval Conferences (e.g. Harman [1993, 1995]). SENSEVAL will see its first results at a subsequent SIGLEX workshop to be held at Herstmonceux Castle, England in September, 1998.

As noted above, WSD is not an end in itself but rather an "intermediate task" that contributes to an overall task such as information retrieval or machine translation. This opens the possibility of two types of evaluation for WSD work (using terminology borrowed from biology): *in vitro* evaluation, where WSD systems are tested independent of a given application, using specially constructed benchmarks; and evaluation *in vivo*, where, rather than being evaluated in isolation, results are evaluated in terms of their contribution to the overall performance of a system designed for a particular application, such as machine translation.

**3.3.1 Evaluation In Vitro.** In vitro evaluation, despite its artificiality, enables close examination of the problems plaguing a given task. In its most basic form, this type of evaluation (also called variously *performance evaluation*: Hirschman and Thompson [1996]; *assessment*: Bimbot, Chollet, and Paoloni [1994]; or *declarative evaluation*: Arnold, Sadler, and Humphreys [1993]) involves comparison of the output of a system for a given input, using measures such as **precision** and **recall**. SENSEVAL currently envisages this type of evaluation for WSD results. Alternatively, in vitro evaluation can focus on study of the behavior and performance of systems on a series of test suites representing the range of linguistic problems likely to arise in attempting WSD (*diagnostic evaluation*: Hirschman and Thompson [1996]; or *typological evaluation*: Arnold, Sadler, and Humphreys 1993). Considerably deeper understanding of the factors involved in the disambiguation task is required before appropriate test suites for typological evaluation of WSD results can be devised. Basic questions such as the role of part of speech in WSD, the treatment of metaphor, metonymy, and the like in evaluation, and how to deal with words of differing degrees and types of polysemy, must first be resolved. SENSEVAL will likely take us a step closer to this understanding; at the least, it will force consideration of what can be meaningfully regarded as an isolatable sense distinction and provide some measure of the distance between the performance of current systems and a predefined standard.

The in vitro evaluation envisaged for SENSEVAL demands the creation of a manually sense-tagged reference corpus containing an agreed-upon set of sense distinctions. The difficulties of attaining sense agreement, even among experts, have already been outlined. Resnik and Yarowsky (1997b) have proposed that for WSD evaluation,

it may be practical to retain only those sense distinctions that are lexicalized cross-linguistically. This proposal has the merit of being immediately usable, but in view of the types of problems cited in the previous section, systematic study of interlanguage relations will be required to determine its viability and generality. At present, the apparent best source of sense distinctions is assumed to be on-line resources such as LDOCE or WordNet, although the problems of utilizing such resources are well known, and their use does not address issues of more complex semantic tagging that goes beyond the typical distinctions made in dictionaries and thesauri.

Resnik and Yarowsky (1997b) also point out that a binary evaluation (correct/incorrect) for WSD is not sufficient, and propose that errors be penalized according to a distance matrix among senses based on a hierarchical organization. For example, failure to identify homographs of *bank* (which would appear higher in the hierarchy) would be penalized more severely than failure to distinguish *bank* as an institution from *bank* as a building (which would appear lower in the hierarchy). However, despite the obvious appeal of this approach, it runs up against the same problem of the lack of an established, agreed-upon hierarchy of senses. Aware of this problem, Resnik and Yarowsky suggest creating the sense distance matrix based on results in experimental psychology such as Miller and Charles (1991) or Resnik (1995b). Even ignoring the cost of creating such a matrix, the psycholinguistic literature has made clear that these results are highly influenced by experimental conditions and the task imposed on the subjects (see, for example, Tabossi [1989, 1991], Rayner and Morris [1991]); in addition, it is not clear that psycholinguistic data can be of help in WSD aimed toward practical use in NLP systems.

In general, WSD evaluation confronts difficulties of criteria that are similar to, but orders of magnitude greater than, those facing other tasks such as part-of-speech tagging, due to the elusive nature of semantic distinctions. It may be that at best we can hope to find practical solutions that will serve particular needs; this is considered more fully in the next section.

**3.3.2 Evaluation In Vivo.** Another approach to evaluation is to consider results insofar as they contribute to the overall performance in a particular application, such as machine translation, information retrieval, or speech recognition. This approach (also called *adequacy evaluation*: Hirschman and Thompson [1996]; or *operational evaluation*: Arnold, Sadler, and Humphreys [1993]), although it does not assure the general applicability of a method nor contribute to a detailed understanding of problems, does not demand agreement on sense distinctions or the establishment of a pretagged corpus. Only the final result is taken into consideration, subjected to evaluation appropriate to the task at hand.

Methods for WSD have evolved largely independently of particular applications, especially in the recent past. It is interesting to note that few, if any, systems for machine translation have incorporated recent methods developed for WSD, despite the importance of WSD for MT noted by Weaver almost 50 years ago. The most obvious effort to incorporate WSD methods into larger applications is in the field of information retrieval, and the results are ambiguous: Krovetz and Croft (1992) report only a slight improvement in retrieval using WSD methods; Voorhees (1993) and Sanderson (1994) indicate that retrieval degrades if disambiguation is not sufficiently precise. Sparck-Jones (forthcoming) questions the utility of any NLP technique for document retrieval. On the other hand, Schütze and Pedersen (1995) show a marked improvement in retrieval (14.4%) using a method that combines search-by-word and search-by-sense.

It remains to be seen to what extent WSD can improve results in particular ap-

plications. However, if meaning is largely a function of use, it may be that the only relevant evaluation of WSD results is achievable in the context of specific tasks.

## 4. Summary and Conclusion

Work on automatic WSD has a history as long as automated language processing generally. Looking back, it is striking to note that most of the problems and the basic approaches to solving them were recognized at the outset. Since so much of the early work on WSD is reported in relatively obscure books and articles across several fields and disciplines, it is not surprising that recent authors are often unaware of it. What is surprising is that in the broad sense, relatively little progress seems to have been made in nearly 50 years. Even though much recent work cites results at the 90% level or better, these studies typically involve very few words, most often only nouns, and frequently concern only broad sense distinctions.

In a sense, WSD work has come full circle, returning most recently to empirical methods and corpus-based analyses that characterize some of the earliest attempts to solve the problem. With sufficiently greater resources and enhanced statistical methods at their disposal, researchers in the 1990s have obviously improved on earlier results, but it appears that we may nearly have reached the limit of what can be achieved in the current framework. For this reason, it is especially timely to assess the state of WSD and consider, in the context of its entire history, the next directions of research. This paper is an attempt to provide that context, at least in part, by bringing WSD into the perspective of the past 50 years of work on the topic. While we are aware that much more could be added to what is presented here, we have made an attempt to cover at least the major areas of work and sketch the broad lines of development in the field.[22]

Of course, WSD is problematic in part because of the inherent difficulty of determining or even defining word sense, and this is not an issue that is likely to be solved in the near future. Nonetheless, it seems clear that current WSD research could benefit from a more comprehensive consideration of theories of meaning and work in the area of lexical semantics. One of the obvious stumbling blocks in much recent WSD work is the rather narrow view of sense that comes hand-in-hand with the attempt to use sense distinctions in everyday dictionaries, which cannot, and are not intended to, represent meaning in context. A different sort of view, one more consistent with current linguistic theory, is required; here, we see the recent work using generative lexicons as providing at least a point of departure.

Another goal of this paper is to provide a starting point for the growing number of researchers working in various areas of computational linguistics who want to learn about WSD. There is renewed interest in WSD as it contributes to various applications, such as machine translation and document retrieval. WSD as "intermediate task," while interesting in its own right, is difficult and perhaps ultimately impossible to assess in the abstract; incorporation of WSD methods into larger applications will therefore hopefully inform and enhance future work.

Finally, if a lesson is to be learned from a review of the history of WSD, it is that research can be very myopic and, as a result, tends to revisit many of the same issues over time. This is especially true when work on a problem has been cross-disciplinary. There is some movement toward more merging of research from various areas, at

22 There are several important topics we have not been able to treat except in a cursory way, including lexical semantic theory, work in psycholinguistics, and statistical methods and results from literary and linguistic analysis.

least as far as language processing is concerned, spurred by the practical problems of information access that we are facing as a result of rapid technological development. Hopefully, this will contribute to further progress on WSD.

**References**


Aarts, Jan. 1990. Corpus linguistics: An appraisal. In Jacqueline Hammesse and Antonio Zampolli, editors, *Computers in Literary and Linguistic Research*. Champion Slatkine, Paris-Geneve, pages 13–28.

Adriaens, Geert. 1986. Word expert parsing: A natural language analysis program revised and applied to Dutch. *Leuvensche Bijdragen*, 75(1):73–154.

Adriaens, Geert. 1987. WEP (word expert parsing) revised and applied to Dutch. In *Proceedings of the 7th European Conference on Artificial Intelligence, ECAI'86*, pages 222–235, Brighton, United Kingdom, July. Reprinted in B. Du Boulay, D. Hogg, L. Steels, editors, *Advances in Artificial Intelligence II*, pages 403–416, Elsevier.

Adriaens, Geert. 1989. The parallel expert parser: A meaning-oriented, lexically guided, parallel-interactive model of natural language understanding. In *Proceedings of the International Workshop on Parsing Technologies*, pages 309–319, Carnegie-Mellon University.

Adriaens, Geert and Steven L. Small. 1988. Word expert revisited in a cognitive science perspective. In Steven Small, Garrison W. Cottrell, and Michael K. Tanenhaus, editors, *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence*. Morgan Kaufman, San Mateo, CA, pages 13–43.

Ahlswede, Thomas E. 1992. Issues in the Design of Test Data for Lexical Disambiguation by Humans and Machines. In *Proceedings of the Fourth Midwest Artificial Intelligence and Cognitive Science Society Conference*, pages 112–116, Starved Rock, IL.

Ahlswede, Thomas E. 1993. Sense Disambiguation Strategies for Humans and Machines. In *Proceedings of the 9th Annual Conference on the New Oxford English Dictionary*, pages 75–88, Oxford, UK, September.

Ahlswede, Thomas E. 1995. Word Sense Disambiguation by Human Informants. In *Proceedings of the Sixth Midwest Artificial Intelligence and Cognitive Society Conference*, pages 73–78, Carbondale, IL, April.

Ahlswede, Thomas E. and David Lorand. 1993. The Ambiguity Questionnaire: A Study of Lexical Disambiguation by Human Informants. In *Proceedings of the Fifth Midwest Artificial Intelligence and Cognitive Society Conference*, pages 21–25, Chesterton, IN.

ALPAC. 1966. *Language and Machine: Computers in Translation and Linguistics*. National Research Council Automatic Language Processing Advisory Committee, Washington, DC.

Alshawi, Hiyan and David Carter. 1994. Training and scaling preference functions for disambiguation. *Computational Linguistics*, 20(4):635–648.

Amsler, Robert A. 1980. *The Structure of the Merriam-Webster Pocket Dictionary*. Ph. D. thesis, University of Texas at Austin, Austin, TX.

Amsler, Robert A. and John S. White. 1979. Development of a computational methodology for deriving natural language semantic structures via analysis of machine-readable dictionaries. Final report on NSF project MCS77-01315. University of Texas at Austin, Austin, TX.

Anderson, John Robert. 1976. *Language, Memory, and Thought*. Lawrence Erlbaum and Associates, Hillsdale, NJ.

Anderson, John Robert. 1983. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22(3):261–95.

Anthony, Edward. 1954. An exploratory inquiry into lexical clusters. *American Speech*, 29(3):175–180.

Arnold, Doug, Louisa Sadler, and R. Lee Humphreys. 1993. Evaluation: An assessment. *Machine Translation*, 8(1-2):1–24. Special issue on evaluation of MT systems.

ARPA. 1993. *Proceedings of the Fifth Message Understanding Conference*, Baltimore, MD, August. Morgan Kaufmann.

Asprejan, Jurij D. 1974. Regular polysemy. *Linguistics*, 142:5–32.

Atkins, Beryl T. S. 1987. Semantic ID tags: Corpus evidence for dictionary senses. In *Proceedings of the Third Annual Conference of the UW Center for the New OED*, pages 17–36, Waterloo, Canada.

Atkins, Beryl T. S. and Beth Levin. 1988. Admitting impediments. In *Proceedings of the 4th Annual Conference of the UW Center for the New OED*, Oxford, UK.

Ayto, John R. 1983. On specifying meaning. In R. R. K. Hartmann, editor, *Lexicography:*

*Principles and Practice*. Academic Press, London, pages 89–98.
Bahl, Lalit R. and Robert L. Mercer. 1976. Part of speech assignment by a statistical decision algorithm. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 88–89, Ronneby.
Bar-Hillel, Yehoshua. 1960. Automatic Translation of Languages. In Franz Alt, A. Donald Booth, and R. E. Meagher, editors, *Advances in Computers*. Academic Press, New York.
Barwise, Jon and John R. Perry. 1983. *Situations and Attitudes*. MIT Press, Cambridge, MA.
Basili, Roberto, Michelangelo Della Rocca, and Maria Tereza Pazienza. 1997. Towards a bootstrapping framework for corpus semantic tagging. In *Proceedings of the ACL-SIGLEX Workshop "Tagging Text with Lexical Semantics: Why, What, and How?"* pages 66–73, Washington, DC, April.
Bel'skaja, Izabella K. 1957. Machine translation of languages. *Research*, 10(10).
Berry-Rogghe, Godelieve. 1973. The computation of collocations and their relevance to lexical studies. In Adam J. Aitken, Richard W. Bailey, and Neil Hamilton-Smith, editors, *The Computer and Literary Studies*. Edinburgh University Press, Edinburgh, UK, pages 103–112.
Bimbot, Frédéric, Gérard Chollet, and A. Paoloni. 1994. Assessment methodology for speaker identification and verification systems: An overview of SAM-A Esprit project 6819–Task 2500. In *Proceedings of the ESCA Workshop on Automatic Speaker Recognition Identification and Verification*, pages 75–82.
Black, Ezra. 1988. An experiment in computational discrimination of englishword senses. *IBM Journal of Research and Development*, 32(2):185–194.
Bloomfield, Leonard. 1933. *Language*. Holt, New York.
Boas, Franz. 1940. *Race, Language and Culture*. Macmillan, New York.
Boguraev, Branimir. 1979. *Automatic Resolution of Linguistic Ambiguities*. Ph.D. thesis, Computer Laboratory, University of Cambridge, August. (Available as Technical Report 11.)
Bookman, Lawrence A. 1987. A microfeature based scheme for modelling semantics. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence, IJCAI'87*, pages 611–614, Milan, Italy.
Braden-Harder, Lisa. 1993. Sense disambiguation using on-line dictionaries. In Karen Jensen, George E. Heidorn, and Stephen D. Richardson, editors, *Natural Language Processing: The PLNLP Approach*. Kluwer Academic Publishers, Dordrecht, pages 247–261.
Briscoe, Edward J. 1991. Lexical issues in natural language processing. In Ewan H. Klein and Frank Veltman, editors, *Natural Language and Speech. Proceedings of the Symposium on Natural Language and Speech*, pages 39–68, Springer-Verlag, Berlin.
Brown, Gillian and George Yule, 1983. *Discourse Analysis*. Cambridge Textbooks in Linguistics Series. Cambridge University Press, Cambridge, UK.
Brown, Peter F., Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting*, pages 264–270, Berkeley, CA. Association for Computational Linguistics.
Brown, Peter F., Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based *n*-gram models of natural language. *Computational Linguistics*, 18(4):467–479.
Bruce, Rebecca and Janyce Wiebe. 1994. Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting*, pages 139–145, Las Cruces, NM. Association for Computational Linguistics.
Bryan, Robert M. 1973. Abstract thesauri and graph theory applications to thesaurus research. In Sally Yeates Sedelow, editor, *Automated Language Analysis, 1972–3*. University of Kansas Press, Lawrence, KS, pages 45–89.
Bryan, Robert M. 1974. Modelling in thesaurus research. In Sally Yeates Sedelow et al., editor, *Automated Language Analysis, 1973–4*. University of Kansas Press, Lawrence, KS, pages 44–59.
Buitelaar, Paul. 1997. A lexicon for underspecified semantic tagging. In *Proceedings of the ACL-SIGLEX Workshop "Tagging Text with Lexical Semantics: Why, What, and How?"*, pages 25–33, Washington, DC, April.
Byrd, Roy J., Nicoletta Calzolari, Martin S. Chodorov, Judith L. Klavans, Mary S. Neff, and Omneya Rizk. 1987. Tools and methods for computational linguistics. *Computational Linguistics*, 13(3/4):219–240.
Calzolari, Nicoletta. 1984. Detecting patterns in a lexical data base. In *Proceedings of the 10th International Conference on Computational Linguistics, COLING'84*, pages 170–173, Stanford University, CA, July.

Chen, Stanley F. and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting*, pages 310–318, University of California, Santa Cruz, CA, June.

Chodorow, Martin S., Roy J. Byrd, and George E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting*, pages 299–304, University of Chicago, Chicago, IL. Association for Computational Linguistics.

Chomsky, Noam. 1957. *Syntactic Structures*. Mouton, The Hague.

Choueka, Yaacov, and F. Dreizin. 1976. Mechanical resolution of lexical ambiguity in a coherent text. In *Proceedings of the International Conference on Computational Linguistics, COLING'76*.

Choueka, Yaacov and D. Goldberg. 1979. Mechanical resolution of lexical ambiguity—A combinatorial approach. In Zvi Malachi, editor, *Proceedings of the International Conference on Literary and Linguistic Computing*, pages 149–165, The Katz Research Institute for Hebrew Literature, Tel-Aviv University, Israel, April.

Choueka, Yaacov and Serge Lusignan. 1985. Disambiguation by short contexts. *Computers and the Humanities*, 19:147–158.

Church, Kenneth W. and William A. Gale. 1991. A comparison of enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer, Speech and Language*, 5:19–54.

Church, Kenneth W. and Robert L. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

Collins, Allan M. and Elisabeth F. Loftus. 1975. A spreading activation theory of semantic processing. *Psychological Review*, 82(6):407–428.

Connine, Cynthia. 1990. Effects of sentence context and lexical knowledge in speech processing. In Gerry T. Altmann, editor, *Cognitive models in speech processing*. MIT Press, Cambridge, MA.

Cottrell, Garrison W. and Steven L. Small. 1983. A connectionist scheme for modelling word sense disambiguation. *Cognition and Brain Theory*, 6:89–120.

Cottrell, Garrisson W. 1985. *A Connectionist Approach to Word-Sense Disambiguation*. Ph.D. thesis. Department of Computer Science, University of Rochester.

Cowie, Jim, Joe A. Guthrie, and Louise Guthrie. 1992. Lexical disambiguation using simulated annealing. In *Proceedings of the 14th International Conference on Computational Linguistics, COLING'92*, volume 1, pages 359–365, Nantes, France, August.

Cruse, D. A. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge, UK.

Dagan, Ido and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

Dagan, Ido, Alon Itai, and Ulrike Schwall. 1991. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting*, pages 130–137, Berkeley, CA. Association for Computational Linguistics.

Dagan, Ido, Shaul Marcus, and Shaul Markovitch. 1993. Contextual word similarity and estimation from sparse data. In *Proceedings of the 31st Annual Meeting*, Columbus, OH, June. Association for Computational Linguistics.

Dagan, Ido, Fernando Peireira, and Lilian Lee. 1994. Similarity-based estimation of word cooccurrence probabilities. In *Proceedings of the 32nd Annual Meeting*, pages 272–278, Las Cruces, NM. Association for Computational Linguistics.

Dahlgren, Kathleen G. 1988. *Naive Semantics for Natural Language Understanding*. Kluwer Academic Publishers, Boston.

Debili, Fathi. 1977. *Traitements syntaxiques utilisant des matrices de précédence fréquentielles construites automatiquement par apprentissage*. Thèse de Docteur-Ingénieur, Université de Paris VII, U.E.R. de Physique.

De Groot, Annette M. B. 1983. The range of automatic spreading activation in word priming. *Journal of Verbal learning and Verbal Behavior*, 22(4):417–436.

Dolan, William B. 1994. Word sense ambiguation: Clustering related senses. In *Proceedings of the 15th International Conference on Computational Linguistics, COLING'94*, pages 712–716, Kyoto, Japan, August.

Dostert, Leon E. 1955. The Georgetown-I.B.M. experiment. In William N. Locke and A. Donald Booth, editors, *Machine Translation of Languages*. John Wiley & Sons, New York, pages 124–135.

Earl, Lois L. 1973. Use of word government in resolving syntactic and semantic ambiguities. *Information Storage and Retrieval*, 9:639–664.

Eaton, Helen S. 1940. *Semantic Frequency List for English, French, German and Spanish.* Chicago University Press, Chicago.

Engelson, Sean P. and Ido Dagan. 1996. Minimizing manual annotation cost in supervised training from corpora. In *Proceedings of the 34th Annual Meeting*, pages 319–326, University of California, Santa Cruz, CA. Association for Computational Linguistics.

Estoup, Jean-Baptiste. 1902. *Gammes sténographiques.* Paris.

Feldman, Jerome A. and Dana H. Ballard. 1982. Connectionist models and their properties. *Cognitive Science*, 6(3):205–254.

Fellbaum, Christiane, editor. Forthcoming-a. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.

Fellbaum, Christiane. Forthcoming-b. The organization of verbs and verb concepts in a semantic net. In Patrick Saint-Dizier, editor, *Predicative Forms in Natural Language and Lexical Knowledge Bases.* Text, Speech and Language Technology Series. Kluwer Academic Publishers, Dordrecht.

Fillmore, Charles J. and Beryl T. S. Atkins. 1991. Invited lecture. Presented at the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, June.

Firth, J. R. 1951. Modes of meaning. *Papers in Linguistics 1934–51*, pages 190–215, Oxford University Press, Oxford, UK.

Fischler, Ira. 1977. Semantic facilitation without association in a lexical decision task. *Memory and Cognition*, 5(3):335–339.

Fontenelle, Thierry. 1990. Automatic extraction of lexical-semantic relations from dictionary definitions. In *Proceedings of the 4th International Congress on Lexicography, EURALEX'90*, pages 89–103, Benalmádena, Spain.

Fries, Charles. 1952. *The Structure of English: An Introduction to the Construction of Sentences.* Harcourt & Brace, New York.

Fries, Charles and Aileen Traver. 1940. *English Word Lists: A Study of their Adaptability and Instruction.* American Council of Education, Washington, DC.

Gale, William A., Kenneth W. Church, and David Yarowsky. 1992a. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 101–112.

Gale, William A., Kenneth W. Church, and David Yarowsky. 1992b. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th Annual Meeting*, pages 249–256, University of Delaware, Newark, DE, July. Association for Computational Linguistics.

Gale, William A., Kenneth W. Church, and David Yarowsky. 1992c. One sense per discourse. In *Proceedings of the Speech and Natural Language Workshop*, pages 233–237, San Francisco, Morgan Kaufmann.

Gale, William A., Kenneth W. Church, and David Yarowsky. 1992d. Work on statistical methods for word sense disambiguation. In *Probabilistic Approaches to Natural Language: Papers from the 1992 AAAI Fall Symposium*, pages 54–60, Cambridge, MA, October.

Gale, William A., Kenneth W. Church, and David Yarowsky. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439.

Gentilhomme, Yves and René Tabory. 1960. Le problème des vraies polysémies et la méthode du paramètre conceptuel. *La Traduction Automatique*, 1(1):9–14.

Good, Irwin J. 1953. The population frequencies of species and the distribution of population parameters. *Biometrika*, 40(3/4):237–264.

Gougenheim, Georges and René Michéa. 1961. Sur la détermination du sens d'un mot au moyen du contexte. *La Traduction Automatique*, 2(1):16–17.

Gougenheim, Georges, René Michéa, Paul Rivenc, and Aurélien Sauvageot. 1956. *L'élaboration du français élémentaire.* Didier, Paris.

Gould, R. 1957. Multiple correspondence. *Mechanical Translation*, 4(1/2):14–27.

Granger, Richard. 1977. FOUL-UP: A program that figures out meanings of words from context. In *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI'77*, pages 172–178.

Grishman, Ralph, Catherine MacLeod, and Adam Meyers. 1994. COMLEX syntax: Building a computational lexicon. In *Proceedings of the 15th International Conference on Computational Linguistics, COLING'94*, pages 268–272, Kyoto, Japan, August.

Grishman, Ralph and John Sterling. 1993. Smoothing of automatically generated selectional constraints. In *Human Language Technology.* Morgan Kaufmann, pages 254–259.

Guthrie, Joe A., Louise Guthrie, Yorick Wilks, and Homa Aidinejad. 1991. Subject-dependent co-occurrence and word sense disambiguation. In *Proceedings of the 29th Annual Meeting*, pages 146–152, Berkeley, CA, June. Association for

Computational Linguistics.
Haas, W. 1966. Linguistic relevance. In C. E. Bazell et al., editors, *In Memory of J. R. Firth*. Longman, London, pages 116–148.
Halliday, M. A. K. 1961. Categories of the theory of grammar. *Word*, 17:241–292.
Halliday, M. A. K. 1966. Lexis as a linguistic level. In C. E. Bazell et al., editors, *In Memory of J. R. Firth*. Longman, London, pages 148–163.
Harman, Donna, editor. 1993. National Institute of Standards and Technology Special Publication No. 500-207 on the First Text Retrieval Conference (TREC-1), Washington, DC. National Institute of Standards and Technology, U.S. Department of Commerce, U.S. Government Printing Office.
Harman, Donna, editor. 1995. *Information Processing and Management*, 31(3). Special Issue on The Second Text Retrieval Conference (TREC-2).
Harper, Kenneth E. 1957a. Semantic ambiguity. *Mechanical Translation*, 4(3):68–69.
Harper, Kenneth E. 1957b. Contextual analysis. *Mechanical Translation*, 4(3):70–75.
Harris, Zellig S. 1951. *Methods in Structural Linguistics*. The University of Chicago Press, Chicago.
Harris, Zellig S. 1954. Distributional structure. *Word*, 10:146–162.
Hayes, Philip J. 1976. A process to implement some word-sense disambiguation. Working paper 23. Institut pour les Etudes Sémantiques et Cognitives, Université de Genève.
Hayes, Philip J. 1977a. On semantic nets, frames and associations. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, pages 99–107, Cambridge, MA.
Hayes, Philip J. 1977b. *Some Association-based Techniques for Lexical Disambiguation by Machine*. Doctoral dissertation, Département de Mathématiques, Ecole Polytechnique Fédérale de Lausanne.
Hayes, Philip J. 1978. Mapping input into schemas. Technical Report 29, Department of Computer Science, University of Rochester.
Hearst, Marti A. 1991. Noun homograph disambiguation using local context in large corpora. In *Proceedings of the Seventh Annual Conference of the Centre for the New OED and Text Research: Using Corpora*, pages 1–22, Oxford, UK.
Hearst, Marti. A. 1994. Multiparagraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting*, pages 9–16, Las Cruces, NM. Association for Computational Linguistics.
Hebb, Donald O. 1949. *The Organisation of Behavior: A Neuropsychological Approach*. John Wiley & Sons, New York.
Hindle, Donald and Mats Rooth. 1993. Structural Ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.
Hirschman, Lynette and Henry S. Thomson. 1996. Overview of evaluation in speech and natural language processing. In Ronald A. Cole, editor, *Survey of the State of the Art in Human Language Technology*. Section 13.1. URL: http://www.cse.ogi.edu/CSLU/HLTsurvey/
Hirst, Graeme. 1987. *Semantic Interpretation and the Resolution of Ambiguity*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, UK.
Hjemslev, Louis. 1953. *Prolegomena to a Theory of Language*. Translated from Danish. Indiana University, Bloomington, IN.
Hobbs, Jerry R. 1987. World knowledge and word meaning. In *Proceedings of the Third Workshop on Theoretical Issues in Natural Language Processing, TINLAP-3*, pages 20–25, Las Cruces, NM.
Ide, Nancy and Jean Véronis. 1990a. Very large neural networks for word sense disambiguation. In *Proceedings of the 9th European Conference on Artificial Intelligence, ECAI'90*, pages 366–368, Stockholm.
Ide, Nancy and Jean Véronis. 1990b. Mapping dictionaries: A spreading activation approach. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary*, pages 52–64, Waterloo, Canada.
Ide, Nancy and Jean Véronis. 1993a. Refining taxonomies extracted from machine-readable dictionaries. In Susan Hockey and Nancy Ide, editor, *Research in Humanities Computing II*. Oxford University Press, pages 145–159.
Ide, Nancy and Jean Véronis. 1993b. Knowledge extraction from machine-readable dictionaries: An evaluation. Presented at the Third International EAMT Workshop "Machine Translation and the Lexicon," Heidelberg, Germany, April. In *Machine Translation and the Lexicon*. See Steffens 1995.
Ide, Nancy and Donald Walker. 1992. Common methodologies in humanities computing and computational linguistics. *Computers and the Humanities*, 26(5/6):327–331.
Iker, H. P. 1974. SELECT: A computer program to identify associationally rich words for content analysis. I. Statistical

results. *Computers and the Humanities*, 8:313–319.

Iker, H. P. 1975. SELECT: A computer program to identify associationally rich words for content analysis. II. Substantive results. *Computers and the Humanities*, 9:3–12.

Imbs, Paul. 1971. *Trésor de la Langue Française. Dictionnaire de la langue du XIXè et du XXè siècles (1989–1960)*. Editions du Centre National de la Recherche Scientifique, Paris.

Janssen, Sylvia. 1992. Tracing cohesive relations in corpora samples using dictionary data. In Gerhard Leitner, editor, *New Directions in English Language Corpora*, Mouton de Gruyter, Berlin.

Jelinek, Frederick. 1976. Continuous speech recognition by statistical methods. *IEEE*, 64(4):532–556.

Jelinek, Frederick and Robert L. Mercer. 1985. Probability distribution estimation from sparse data. *IBM Technical Disclosure Bulletin*, 28:2591–2594.

Jensen, Karen and Jean-Louis Binot. 1987. Disambiguating prepositional phrase attachments by using on-line dictionary definitions. *Computational Linguistics*, 13(3/4):251–260.

Johansson, Stig. 1980. The LOB corpus of British English texts: Presentation and comments. *ALLC Journal*, 1(1):25–36.

Jorgensen, Julia. 1990. The psychological reality of word senses. *Journal of Psycholinguistic Research*, 19:167–190.

Kaeding, F. W. 1897–1898. *Häufigkeitswörterbuch der deutschen Sprache. Festgestellt durch Arbeitsausschüss der deutschen Stenographie-System*. Selbstverlag, Steglitz bei Berlin.

Kaplan, Abraham. 1950. An experimental study of ambiguity and context. Mimeographed. (Published 1955 in *Mechanical Translation*, 2(2):39-46.)

Kawamoto, Alan H. 1988. Distributed representations of ambiguous words and their resolution in a connectionist network. In Steven Small, Garrison W. Cottrell, and Michael K. Tanenhaus, editors, *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence*. Morgan Kaufman, San Mateo, CA, pages 195–228.

Kelly, Edward F. and Philip J. Stone. 1975. *Computer Recognition of English Word Senses*. North-Holland, Amsterdam.

Kilgarriff, Adam. 1992. *Polysemy*. Ph. D. thesis. University of Sussex, UK.

Kilgarriff, Adam. 1993. Dictionary word sense distinctions: An enquiry into their nature. *Computers and the Humanities*, 26:365–387.

Kilgarriff, Adam. 1994. The myth of completeness and some problems with consistency (the role of frequency in deciding what goes in the dictionary). In *Proceedings of the 6th International Congress on Lexicography, EURALEX'94*, pages 101–106, Amsterdam, Holland.

Kilgarriff, Adam. 1997. Evaluation of word sense disambiguation programs. *SALT Club Workshop "Evaluation in Speech and Language Technology,"* Sheffield University, Sheffield, UK, June.

Kilgarriff, Adam. Forthcoming. I don't believe in word senses. *Computers and the Humanities*.

Kintsch, Walter and Ernest F. Mross. 1985. Context effects in word identification. *Journal of Memory and Language*, 24(3):336–349.

Klavans, Judith, Martin Chodorow, and Nina Wacholder. 1990. From dictionary to knowledge base via taxonomy. In *Proceedings of the 6th Conference of the UW Centre for the New OED*, pages 110–132, Waterloo, Canada.

Koutsoudas, Andreas K. and R. Korfhage. 1956. M.T. and the problem of multiple meaning. *Mechanical Translation*, 2(2):46–51.

Krovetz, Robert and William Bruce Croft. 1989. Word sense disambiguation using machine-readable dictionaries. In *Proceedings of the 12th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, SIGIR'89*, pages 127–136, Cambridge, MA.

Krovetz, Robert and William Bruce Croft. 1992. Lexical Ambiguity and Information Retrieval. *ACM Transactions on Information Systems*, 10(2):115–141.

Kucera, Henri and Winthrop N. Francis. 1967. *Computational Analysis of Present-Day American English*, Brown University Press, Providence.

Leacock, Claudia, Geoffrey Towell, and Ellen Voorhees. 1993. Corpus-based statistical sense resolution. In *Proceedings of the ARPA Human Language Technology Workshop*, San Francisco, Morgan Kaufman.

Leacock, Claudia, Geoffrey Towell, and Ellen M. Voorhees. 1996. Towards building contextual representations of word senses using statistical models. In Branimir Boguraev and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*. MIT Press, Cambridge, MA, pages 97–113.

Leech, Geoffrey. 1991. The state of the art in

corpus linguistics. In K. Aijmer and B. Altenberg, editors, *English Corpus Linguistics*. Longman, London, pages 8–29.

Lehman, Jill Fain. 1994. Toward the essential nature of statistical knowledge in sense resolution. In *Proceedings of the 12th International Conference on Artificial Intelligence, AAAI'94*, pages 734–741, Seattle, Washington, July/August.

Lenat, Douglas B. and Ramanathan V. Guha. 1990. *Building Large Knowledge-based Systems*. Addison-Wesley, Reading, MA.

Lesk, Michael. 1986. Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC Conference*, pages 24–26, Toronto, Canada, June.

Liddy, Elisabeth D. and Woojin Paik. 1993. Statistically-guided word sense disambiguation. In *Proceedings of the AAAI Fall Symposium Series*, pages 98–107.

Litowski, Kenneth C. 1997. Desiderata for tagging with WordNet sysnsets or MCAA categories. In *Proceedings of the ACL-SIGLEX Workshop "Tagging Text with Lexical Semantics: Why, What, and How?"* pages 12–17, Washington, DC, April.

Lorge, Irving. 1949. *Semantic Content of the 570 Commonest English Words*. Columbia University Press, New York.

Luk, Alpha K. 1995. Statistical sense disambiguation with relatively small corpora using dictionary definitions. In *Proceedings of the 33rd Annual Meeting*, pages 181–188. Cambridge, MA. Association for Computational Linguistics.

Lupker, Stephen J. 1984. Semantic priming without association: A second look. *Journal of Verbal Learning and Verbal Behavior*, 23(6):709–733.

Lyons, John. 1966. Firth's theory of meaning. In C. E. Bazell et al., editors, *In Memory of J. R. Firth*. Longman, London, pages 288–302.

Lyons, John. 1977. *Semantics*. Cambridge University Press, Cambridge, UK.

Macleod, Catherine, Ralph Grishman, and Adam Meyers. Forthcoming. A large syntactic dictionary for natural language processing. *Computers and the Humanities*.

Madhu, Swaminathan and Dean W. Lytle. 1965. A figure of merit technique for the resolution of non-grammatical ambiguity. *Mechanical translation*, 8(2):9–13.

Mahesh, Kavi, Sergei Nirenburg, and Stephen Beale. 1997. If you have it, flaunt it: Using full ontological knowledge for word sense disambiguation. In *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 1–9, Sante Fe, NM, July.

Mahesh, Kavi, Sergei Nirenburg, Stephen Beale, Evelyne Viegas, Victor Raskin, and Boyan Onyshkevych. 1997. Word Sense Disambiguation: Why statistics when we have these numbers? In *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 151–159, Santa Fe, NM, July.

Malakhovski, L. V. 1987. Homonyms in English dictionaries. In R. W. Burchfield, editor, *Studies in Lexicography*. Oxford University Press, Oxford, UK, pages 36–51.

Markowitz, Judith, Thomas Ahlswede, and Martha Evens. 1986. Semantically significant patterns in dictionary definitions. In *Proceedings of the 24th Annual Meeting*, pages 112–119, Association for Computational Linguistics.

Martinet, André. 1960. Eléments de linguistique générale. Armand Colin, Paris.

Masterman, Margaret. 1957. The thesaurus in syntax and semantics. *Mechanical Translation*, 4:1–2.

Masterman, Margaret. 1962. Semantic message detection for machine translation, using an interlingua. In *1961 International Conference on Machine Translation of Languages and Applied Language Analysis*. Her Majesty's Stationery Office, London, pages 437–475.

McClelland, James L. and David E. Rumelhart. 1981. An interactive activation of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, 88:375–407.

McCulloch, Warren S. and Walter Pitts. 1943. A logical calculus of the ideas imminent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

McIntosh, A. 1966. Patterns and ranges. In *Papers in General, Descriptive, and Applied Linguistics*. Longman, London, pages 183–199.

McRoy, Susan W. 1992. Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, 18(1):1–30.

Meillet, Antoine. 1926. *Linguistique historique et linguistique générale*. Volume 1. Second edition. Champion, Paris.

Mérialdo, Bernard. 1994. Tagging text with a probabilistic model. *Computational Linguistics*, 20(2):155–172.

Meyer, David E. and Roger W.

Schvaneveldt. 1971. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2):227–234.

Michéa, René. 1964. Les vocabulaires fondamentaux. *Recherche et techniques nouvelles au service de l'enseignement des langues vivantes*, Université de Strasbourg, Strasbourg, 21–36.

Michiels, Archibald, Jacques Mullenders, and Jacques Noël. 1980. Exploiting a large database by Longman. In *Proceedings of the 8th International Conference on Computational Linguistics, COLING'80*, pages 374–382, Tokyo, Japan.

Michiels, Archibald. 1982. *Exploiting a Large Dictionary Data Base*. Ph.D. thesis, Université de Liège, Liège, Belgium.

Miller, George A., Richard T. Beckwith, Christiane D. Fellbaum, Derek Gross, and Katherine J. Miller. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Miller, George A., Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 240–243, Plainsboro, NJ.

Miller, George A., Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pages 303–308, Plainsboro, NJ, March.

Miller, George A. and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Morris, Jane. 1988. Lexical cohesion, the thesaurus, and the structure of text. Technical Report CSRI 219, Computer Systems Research Institute, University of Toronto, Toronto, Canada.

Morris, Jane and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Nakamura, Jun-Ichi and Makoto Nagao. 1988. Extraction of semantic information from an ordinary English dictionary and its evaluation. In *Proceedings of the 12th International Conference on Computational Linguistics, COLING'88*, pages 459–464, Budapest, Hungary, August.

Ng, Hwee Tou and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach. In *Proceedings of the 34th Annual Meeting*, University of California, Santa Cruz, CA, June. Association for Computational Linguistics.

Niwa, Yoshiki and Yoshihiko Nitta. 1994. Coocurrence vectors from corpora vs distance vectors from dictionaries. In *Proceedings of the 15th International Conference on Computational Linguistics, COLING'94*, pages 304–309, Kyoto, Japan, August.

Oettinger, Anthony G. 1955. The design of an automatic Russian-English technical dictionary. In William N. Locke and A. Donald Booth, editors, *Machine Translation of Languages*. John Wiley & Sons, New York, pages 47–65.

Olney, John C. 1968. To all interested in the Merriam-Webster transcripts and data derived from them. Technical Report L-13579, System Development Corporation, Santa Monica, CA, October.

Oswald, Victor A. Jr. 1952. Microsemantics. Presented at the first M.I.T. conference on Mechanical Translation, 17–20 June 1952. Mimeographed. (Available on microfilm at M.I.T., Papers on Mechanical Translation, roll 799.)

Oswald, Victor A. Jr. 1957. The rationale of the idioglossary technique. In Leon E. Dostert, editor, *Research in Machine Translation*. Georgetown University Press, Washington, DC, pages 63–69.

Oswald, Victor A. Jr. and Richard H. Lawson. 1953. An idioglossary for mechanical translation. *Modern Language Forum*, 38(3/4):1–11.

Palmer, H. 1933. *Second Interim Report on English Collocations*. Institute for Research in English Teaching, Tokyo.

Panov, D. 1960. La traduction mécanique et l'humanité. *Impact*, 10(1):17–25.

Parker-Rhodes, Arthur F. 1958. The use of statistics in language research. *Mechanical Translation*, 5(2):67–73.

Patrick, Archibald B. 1985. *An Exploration of Abstract Thesaurus Instantiation*. M. Sc. thesis, University of Kansas, Lawrence, KS.

Pendergraft, Eugene. 1967. Translating languages. In Harold Borko, editor, *Automated Language Processing*. John Wiley & Sons, New York.

Pereira, Fernando and Naftali Tishby. 1992. Distributional similarity, phase transitions and hierarchical clustering. *Working Notes of the AAAI Symposium on Probabilistic Approaches to Natural Language*, pages 108–112, Cambridge, MA, October.

Pereira, Fernando, Naftali Tishby, and Lilian Lee. 1993. Distributional clustering of

English. In *Proceeedings of the 31st Annual Meeting*, pages 183–190, Ohio State University, Colombus, OH, June. Association for Computational Linguistics.

Pimsleur, P. 1957. Semantic frequency counts. *Mechanical Translation*, 4(1–2):11–13.

Pustejovsky, James. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

Pustejovsky, James, Branimir Boguraev, and Michael Johnston. 1995. A core lexical engine: The contextual determination of word sense. Technical Report, Department of Computer Science, Brandeis University.

Quillian, M. Ross. 1961. A design for an understanding machine. Presented at the Semantic Problems in Natural Language colloquium, King's College, Cambridge University, Cambridge, UK, September.

Quillian, M. Ross. 1962a. A revised design for an understanding machine. *Mechanical Translation*, 7(1):17–29.

Quillian, M. Ross. 1962b. A semantic coding technique for mechanical English paraphrasing. Internal memorandum of the Mechanical Translation Group, Research Laboratory of Electronics, MIT, August.

Quillian, M. Ross. 1967. Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, 12:410–430.

Quillian, M. Ross. 1968. Semantic memory. In M. Minsky, editor, *Semantic Information Processing*. MIT Press, Cambridge, MA, pages 227–270.

Quillian, M. Ross. 1969. The teachable language comprehender: A simulation program and theory of language. *Communications of the ACM*, 12(8):459–476.

Quine, Willard V. 1960. *Word and Object*. MIT Press, Cambridge, MA.

Quirk, Randolph. 1960. Towards a description of English usage. *Transactions of the Philological Society*, pages 40–61.

Rayner, Keith and R. K. Morris. 1991. Comprehension processes in reading ambiguous sentences: Reflections from eye movements. In G. Simpson, editor, *Understanding Word and Sentence*. North-Holland, Amsterdam. pages 175–198.

Reifler, Erwin. 1955. The mechanical determination of meaning. In William N. Locke and A. Donald Booth, editors, *Machine Translation of Languages*. John Wiley & Sons, New York, pages 136–164.

Resnik, Philip. 1992. WordNet and distributional analysis: A class-based approach to statistical discovery. In *Proceedings of the AAAI Workshop on Statistically-based Natural Language Processing Techniques*, pages 48–56. San Jose, CA.

Resnik, Philip. 1993a. Selection and Information: A Class-based Approach to Lexical Relationships. Ph. D. thesis, University of Pennsylvania. Also University of Pennsylvania Technical Report 93-42.

Resnik, Philip. 1993b. Semantic classes and syntactic ambiguity. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 278–283.

Resnik, Philip. 1995a. Disambiguating noun groupings with respect to WordNet senses. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 54–68, Cambridge, MA.

Resnik, Philip. 1995b. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI'95*, pages 448–453, Montreal, Canada.

Resnik, Philip and David Yarowsky. 1997a. Evaluating automatic semantic taggers. In *Proceedings of the ACL-SIGLEX Workshop "Tagging Text with Lexical Semantics: Why, What, and How?"*, page 91, Washington, DC, April.

Resnik, Philip and David Yarowsky. 1997b. A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of the ACL-SIGLEX Workshop "Tagging Text with Lexical Semantics: Why, What, and How?"*, pages 79–86, Washington, DC, April.

Richards, I. A. 1953. Towards a theory of translation. In *Studies in Chinese Thought*, University of Chicago Press, Chicago.

Richardson, Ray and Alan F. Smeaton. 1994. Automatic word sense disambiguation in a KBIR application. Working paper CA-0595, School of Computer Applications, Dublin City University, Dublin, Ireland.

Richens, Richard H. 1958. Interlingual machine translation. *Computer Journal*, 1(3):144–147.

Richmond, Korin, Andrew Smith, and Einat Amitay. 1997. Detecting Subject Boundaries Within Text: A Language Independent Statistical Approach. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, EMNLP-2*, pages 47–54, Brown University, Providence, RI, August.

Roberts, D. D. 1973. *The Existential Graphs of Charles S. Pierce*, Mouton, The Hague.

Robins, R. H. 1987. Polysemy and the

lexicographer. In R. W. Burchfield, editor, *Studies in Lexicography*. Oxford University Press, Oxford, UK, pages 52–75.

Rosenblatt, Frank. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408.

Salton, Gerard. 1968. *Automatic Information Organization and Retrieval*. McGraw-Hill, New York.

Salton, Gerard and M. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Salton, Gerard, A. Wong, and C. S. Yang. 1975. A vector space for information retrieval. *Communications of the ACM*, 18(11):613–620.

Sanderson, Mark. 1994. Word sense disambiguation and information retrieval. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 161–175, Las Vegas.

Schank, Roger C. and Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum, Hillsdale, NJ.

Schmidt, Klaus M. 1988. Der Beitrag der begriffsorientierten Lexicographie zue systematischen Erfassung von Sprachwandel und das Begriffwörterbuch zur wdh. Epik. In W. Bachofer, editor, *Mittelhochdeutches Wörterbuch in der Diskussion*. Max Niemeyer, Tubingen, pages 25–49.

Schmidt, Klaus M. 1991. Ein databanksystem für das Begriffwörterbuch Mittelhochdeutcher Epik und Fortschritte bie der automatischen Disambiguierung. In K. Gärtner, P. Sappler, and M. Trauth, editors, *Maschinelle Verarbeitung altdeutscher Text IV*. Max Niemeyer, Tubingen, pages 192–204.

Schütze, Hinrich. 1992. Dimensions of meaning. In *Proceedings of Supercomputing'92*, pages 787–796. Los Alamitos, CA.

Schütze, Hinrich. 1993. Word space. In Stephen J. Hanson, Jack D. Cowan, and C. Lee Giles, editors, *Advances in Neural Information Processing Systems 5*. Morgan Kauffman, San Mateo, CA, pages 895–902.

Schütze, Hinrich and Jan Pedersen. 1995. Information retrieval based on word senses. In *Proceedings of SDAIR'95*, Las Vegas, Nevada, April.

Sedelow, Sally Yeates and Donna Weir Mooney. 1988. Knowledge retrieval from domain-transcendent expert systems: II. Research results. In *Proceedings of the American Society for Information Science (ASIS) Annual Meeting*, pages 209–212, Knowledge Industry Publications, White Plains, New York.

Sedelow, Sally Yeates and Walter. A. Sedelow Jr. 1969. Categories and procedures for content analysis in the humanities. In George Gerbner, Ole Holsti, Klaus Krippendorf, William J. Paisley, and Philip J. Stone, editors, *The Analysis of Communication Content*. John Wiley & Sons, New York, pages 487–499.

Sedelow, Sally Yeates and Walter. A. Sedelow Jr. 1986. Thesaural knowledge representation. In *Proceedings of the University of Waterloo Conference on Lexicology*, pages 29–43, Waterloo, Canada.

Sedelow, Sally Yeates and Walter. A. Sedelow Jr. 1992. Recent model-based and model-related studies of a large-scale lexical resource (Roget's Thesaurus. In *Proceedings of the 14th International Conference on Computational Linguistics, COLING'92*, pages 1223–1227, Nantes, France, August.

Seidenberg, Mark S., Michael K. Tanenhaus, James M. Leiman, and Marie A. Bienkowski. 1982. Automatic access of the meaning of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology*, 14(4):489–537.

Selz, O. 1913. *Über die Gesetze des Georneten Denkberlaufs*, Spemman, Stuttgart.

Selz, O. 1922. *Zue Psychologie des produktive Denkens un des Irrtums*, Friedrich Cohen, Bonn.

Seneff, Stephanie. 1992. TINA, A natural language system for spoken language applications. *Computational Linguistics*, 18(1):61–86.

Simpson, Greg B. and Curt Burgess. 1989. Implications of lexical ambiguity resolution for word recognition and comprehension. In Steven Small, Garrison W. Cottrell, and Michael K. Tanenhaus, editors, *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence*. Morgan Kaufman, San Mateo, CA, pages 271–288.

Sinclair, John. 1966. Beginning the study of lexis. In C. E. Bazell et al., editors, *In Memory of J. R. Firth*. Longman, London, pages 410–431.

Sinclair, John, editor. 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing*. Collins, London.

Skorochod'ko, E. F. 1972. Adaptative methods of automatic abstracting and

indexing. In C. V. Freiman, editor, *Information Processing 71: Proceedings of the IFIP Congress 71*, pages 1179–1182, North Holland Publishing Company.

Slator, Brian M. 1992. Sense and preference. *Computer and Mathematics with Applications*, 23(6/9):391–402.

Slator, Brian M. and Yorick A. Wilks. 1987. Towards semantic structures from dictionary entries. In *Proceedings of the 2nd Annual Rocky Mountain Conference on Artificial Intelligence*, pages 85–96, Boulder, CO.

Small, Steven L. 1980. *Word Expert Parsing: A Theory of Distributed Word-based Natural Language Understanding*. Ph.D. thesis, Department of Computer Science, University of Maryland, September. Available as Technical Report 954.

Small, Steven L. 1983. Parsing as cooperative distributed inference. In Margaret King, editor, *Parsing Natural Language*. Academic Press, London.

Small, Steven L., Garrison W. Cottrell, and Michael K. Tanenhaus, editors. 1988. *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence*. Morgan Kaufman, San Mateo, CA.

Small, Steven L. and Charles Rieger. 1982. Parsing and comprehending with word experts (a theory and its realization). In Wendy Lenhert and Martin Ringle, editors, *Strategies for Natural Language Processing*. Lawrence Erlbaum and Associates, Hillsdale, NJ, pages 89–147.

Sparck-Jones, Karen. 1964. *Synonymy and Semantic Classification*. Ph.D. thesis, University of Cambridge, Cambridge, UK.

Sparck-Jones, Karen. 1986. *Synonymy and Semantic Classification*. Edinburgh, Edinburgh University Press, UK.

Sparck-Jones, Karen. Forthcoming. What is the role of NLP in Text Retrieval? In Tomek Strzalkowski, editor, *Natural Language Information Retrieval*. Text, Speech and Language Technology Series. Kluwer Academic Publishers, Dordrecht.

Sproat, Richard, Julia Hirschberg, and David Yarowsky. 1992. A corpus-based synthesizer. In *Proceedings of the International Conference on Spoken Language Processing*, Banff, Alberta, Canada, October.

Steffens, Petra, editor. 1995. *Machine Translation and the Lexicon*. Lecture Notes in Artificial Intelligence 898. Springer-Verlag, Berlin.

Stock, Penelope F. 1983. Polysemy. In *Proceedings of the Exeter Lexicography Conference*, 131–140.

Stone, Philip J. 1969. Improved quality of content analysis categories: Computerized-disambiguation rules for high-frequency English words. In George Gerbner, Ole Holsti, Klaus Krippendorf, William J. Paisley, and Philip J. Stone, editors, *The Analysis of Communication Content*. John Wiley & Sons, New York, pages 199–221.

Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie, editors. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.

Sussna, Michael. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Base Management, CIKM'93*, pages 67–74, Arlington, VA.

Sutcliffe, Richard F. E., A. McElligott, D. O'Sullivan, A. A. Polikarpov, L. A. Kuzmin, G. O'Neill, and J. Véronis. 1996. An Interactive Approach to the Creation of a Multilingual Concept Ontology for Language Engineering. In *Proceedings of the Workshop "Multilinguality in the Software Industry," European Conference on Artificial Intelligence, ECAI'96*, Budapest University of Economics, Budapest, Hungary, August.

Sutcliffe, Richard F. E., D. O'Sullivan, A. A. Polikarpov, L. A. Kuzmin, A. McElligott, and J. Véronis. 1996. IWNR—Extending a public multilingual taxonomy to Russian. In *Proceedings of the Workshop "Multilinguality in the Lexicon," AISB Second Tutorial and Workshop Series*, pages 14–25, University of Sussex, Brighton, UK, March/April.

Sutcliffe, Richard F. E. and Bronwyn E. A. Slater. 1995. Disambiguation by association as a practical method: Experiments and findings. *Journal of Quantitative Linguistics*, 2(1):43–52.

Tabossi, Patricia. 1989. What's in a context? In D. Gorfein, editor, *Resolving Semantic Ambiguity*. Springer-Verlag, New York, pages 25–39.

Tabossi, Patricia. 1991. Understanding words in context. In G. Simpson, editor, *Understanding Word and Sentence*. North-Holland, Amsterdam, pages 1–22.

ten Hacken, Pius. 1990. Reading distinction in machine translation. In *Proceedings of the 12th International Conference on Computational Linguistics, COLING'90*, volume 2, pages 162–166, Helsinki, Finland, August.

Thorndike, Edward L. 1921. *A Teacher's Word*

*Book.* Columbia Teachers College, New York.
Thorndike, Edward L. 1948. On the frequency of semantic changes in modern English. *Journal of General Psychology,* 66:319–327.
Thorndike, Edward L. and Irving Lorge. 1938. *Semantic counts of English Words,* Columbia University Press, New York.
Thorndike, Edward L. and Irving Lorge. 1944. *The Teacher's Word Book of 30,000 Words.* Columbia University Press, New York.
Urdang, Laurence. 1984. A lexicographer's adventures in computing. *Datamation,* 30(3):185–194.
van Buren, P. 1967. Preliminary aspects of mechanisation in lexis. *CahLex,* 11:89–112; 12:71–84.
van der Eijk, Pim. 1994. Comparative discourse analysis of parallel texts. *Second Annual Workshop on Very Large Corpora (WVLC2),* pages 143–159, Kyoto, Japan, August.
Véronis, Jean and Nancy Ide. 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th International Conference on Computational Linguistics, COLING'90,* volume 2, pages 389–394, Helsinki, Finland.
Véronis, Jean and Nancy Ide. 1991. An assessment of information automatically extracted from machine readable dictionaries. In *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics,* pages 227–232, Berlin, Germany.
Véronis, Jean and Nancy Ide. 1995. Large neural networks for the resolution of lexical ambiguity. In Patrick Saint-Dizier and Evelyne Viegas, editors, *Computational Lexical Semantics.* Natural Language Processing Series. Cambridge University Press, Cambridge, United Kingdom, pages 251–269.
Viegas, Evelyne, Kavi Mahesh, and Sergei Nirenburg. Forthcoming. Semantics in action. In Patrick Saint-Dizier, editor, *Predicative Forms in Natural Language and Lexical Knowledge Bases.* Text, Speech and Language Technology Series. Kluwer Academic Publishers, Dordrecht.
Voorhees, Ellen M. 1993. Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pages 171–180, Pittsburgh, PA, June/July.
Voorhees, Ellen M., Claudia Leacock, and Geoffrey Towell. 1995. Learning context to disambiguate word senses. In Thomas Petsche, Stephen José Hanson, and Jude Shavlik, editors, *Computational Learning Theory and Natural Learning Systems.* MIT Press, Cambridge, MA.
Vossen, Piek. Forthcoming. Introduction to EuroWordNet. To appear in a Special Issue of *Computers and the Humanities on EuroWordNet.*
Waibel, Alex and Kai-Fu Lee, editors. 1990. *Readings in Speech Recognition.* Morgan Kaufmann, San Mateo, CA.
Waltz, David L. and Jordan B. Pollack. 1985. Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science,* 9:51–74.
Weaver, Warren. 1955. *Translation.* In William N. Locke and A. Donald Booth, editors, *Machine Translation of Languages.* John Wiley & Sons, New York, pages 15–23. (Reprint of mimeographed version, 1949.)
Weibe, Janyce, Julie Maples, Lee Duan, and Rebecca Bruce. 1997. Experience in WordNet sense tagging in the Wall Street Journal. In *Proceedings of the ACL-SIGLEX Workshop "Tagging Text with Lexical Semantics: Why, What, and How?",* pages 8–11, Washington, DC, April.
Weinreich, Uriel. 1980. *On Semantics.* University of Pennsylvania Press.
Weiss, S. 1973. Learning to disambiguate. *Information Storage and Retrieval,* 9:33–41.
Whittemore, Greg, Kathleen Ferrara, and Hans Brunner. 1990. Empirical studies of predictive powers of simple attachment schemes for post-modifier prepositional phrases. In *Proceedings of the 28th Annual Meeting of Association for Computational Linguistics,* pages 23–30, Pittsburgh, PA, June.
Wilks, Yorick A. 1968. On-line semantic analysis of English texts. *Mechanical Translation,* 11(3–4):59–72.
Wilks, Yorick A. 1969. Getting meaning into the machine. *New Society,* 361:315–317.
Wilks, Yorick A. 1973. An artificial intelligence approach to machine translation. In Roger Schank and Kenneth Colby, editors, *Computer Models of Thought and Language.* W. H. Freeman, San Francisco, pages 114–151.
Wilks, Yorick A. 1975a. Primitives and words. In *Proceedings of the Interdisciplinary Workshop on Theoretical Issues in Natural Language Processing,* pages 42–45, Cambridge, MA, June.
Wilks, Yorick A. 1975b. Preference semantics. In E. L. Keenan III, editor, *Formal Semantics of Natural Language.*

Cambridge University Press, pages 329–348.
Wilks, Yorick A. 1975c. An intelligent analyzer and understander of English. *Communications of the ACM*, 18(5):264–274.
Wilks, Yorick A. 1975d. A preferential, pattern-seeking semantics for natural language inference. *Artificial Intelligence*, 6:53–74.
Wilks, Yorick A. Forthcoming. Senses and texts. *Computers and the Humanities*.
Wilks, Yorick A. and Dan Fass. 1990. Preference semantics: A family history. Report MCCS-90-194, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.
Wilks, Yorick A., Dan Fass, Cheng-Ming Guo, James E. MacDonald, Tony Plate, and Brian A. Slator. 1990. Providing machine tractable dictionary tools. In James Pustejovsky, editor, *Semantics and the Lexicon*. MIT Press, Cambridge, MA.
Wilks, Yorick A., Brian A. Slator, and Louise M. Guthrie. 1996. *Electric Words: Dictionaries, Computers, and Meanings*. A Bradford Book. MIT Press, Cambridge, MA.
Wilks, Yorick and Mark Stevenson. 1996. The grammar of sense: Is word sense tagging much more than part-of-speech tagging? Technical Report CS-96-05, University of Sheffield, Sheffield, UK.
Wittgenstein, Ludwig. 1953. Philosophical Investigations. Translated by G. E. M. Anscombe. Basil Blackwell, Oxford.
Yarowsky, David. 1992. Word sense disambiguation using statistical models of Roget's categories trained on large corpora. *Proceedings of the 14th International Conference on Computational Linguistics, COLING'92*, pages 454–460, Nantes, France, August.
Yarowsky, David. 1993. One sense per collocation. In *Proceedings of ARPA Human Language Technology Workshop*, pages 266–271, Princeton, NJ.
Yarowsky, David. 1994a. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In *Proceedings of the 32nd Annual Meeting*, pages 88–95, Las Cruces, NM. Association for Computational Linguistics.
Yarowsky, David. 1994b. A comparison of corpus-based techniques for restoring accents in Spanish and French text. In *Proceedings of the 2nd Annual Workshop on Very Large Text Corpora*, pages 19–32, Las Cruces, NM.
Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting*, pages 189–196, Cambridge, MA, June. Association for Computational Linguistics.
Yarowsky, David. 1997. Homograph disambiguation in text-to-speech synthesis. In Jan T. H. van Santen, Richard Sproat, Joseph P. Olive, and Julia Hirschberg, editors, *Progress in Speech Synthesis*. Springer-Verlag, New York, pages 157–172.
Yngve, Victor H. 1955. Syntax and the problem of multiple meaning. In William N. Locke and A. Donald Booth, editors, *Machine Translation of Languages*. John Wiley & Sons, New York, pages 208–226.
Zernik, Uri. 1990. Tagging word senses in a corpus: The needle in the haystack revisited. In P. Jacobs, editor, *Text-based Intelligent Systems: Current Research in Text Analysis, Information Extraction and Retrieval*. GE Research and Development Center, Schenectady, New York.
Zernik, Uri. 1991. Train1 vs. Train2: Tagging word senses in a corpus. In *Proceedings of Intelligent Text and Image Handling, RIAO'91*, pages 567–585, Barcelona, Spain.
Zipf, George K. 1935. *The Psycho-biology of Language: An Introduction to Dynamic Biology*. MIT Press, Cambridge, MA.
Zipf, George K. 1945. The meaning-frequency relationship of words. *Journal of General Psychology*, 33:251–266.