Word Sense Disambiguation: The State of the Art

Nancy Ide*
Vassar College

Jean Véronis†
Université de Provence

Introduction

The automatic disambiguation of word senses has been an interest and concern since the earliest days of computer treatment of language in the 1950's. 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, man-machine communication, etc.; 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, 1949; Yngve, 1955; etc.).
- *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* a*s* 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 pre-defined 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 has long been recognized (see, for instance, Stone,

* Department of Computer Science, Vassar College, Poughkeepsie, New York 12604-0520, U.S.A. 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.

et al. 1966; Stone, 1969; Kelly and Stone, 1975; for a more recent discussion see Litowski, 1997) in order to include only those instances of a word in its proper sense.

- *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* 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 *et al.*, 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 *et al.,* 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 (for example, in French, changing *comte* to *comté*) (Yarowsky, 1994a and b), 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 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 1960's.

However, 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 WSD research and to consider the next steps that need to be taken in the field. To this end, this paper surveys the major, well-known approaches to WSD and considers the open problems and directions of future research.

Survey of WSD methods

In general terms, *word sense disambiguation* (WSD) 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 pre-defined 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;

etc.

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 recent 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), morpho-syntactic disambiguation accomplishes sense disambiguation, and therefore (especially since the development of reliable part-of-speech taggers), WSD work

has since 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.

1.1 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,* Weaver (1949) discusses the need for WSD in machine translation and outlines the basis of an approach to WSD which 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 ?"*

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 (1961), Koutsoudas and Korfhage (1956) on Russian, and Gougen-

heim and Michéa (1961), 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, focussed primarily on the translation of technical texts and in all cases dealing with texts from particular domains. Weaver's (1949) *Memorandum* discusses the role of the domain in sense disambiguation, making a point that was reiterated several decades later (Gale *et al.,* 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.*

Following directly from this observation, much effort in the early days of machine translation was devoted to the development of specialized dictionaries or "micro-glossaries" (Oswald, 1952, 1957; Oswald and Lawson, 1953; Oettinger, 1955; Dostert, 1955; Gould, 1957; Panov, 1960; etc.). Such micro-glossaries contain only the meaning of a given word relevant for texts in a particular domain of discourse; e.g., a micro-glossary 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 his *Memorandum* by noting the "tremendous amount of work [needed] in the logical structure of languages." 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, 1961—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.*

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 2000 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 *et al.,* 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).

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 large extent, forgotten until several decades later.

---

[1] For a detailed accounting of Masterman's methodology, see Wilks *et al.* (1996). Other researchers have discussed the use of thesauri for disambiguation in the context of early MT work, e.g. Gentilhomme and Tabory (1961).

1.2 AI-based methods

AI methods began to flourish in the early 1960's 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 which 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.

1.2.1 Symbolic methods. As mentioned above, semantic networks were developed in the late 1950's[2] and were immediately applied to the problem of representing word meanings. Masterman (1961), 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 work on semantic networks by Richens (1958), Quillian (1961, 1962a and b, 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* which contained information about words and their roles and relations to other words in individual sentences. For example, Hayes (1976, 1977a and b, 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

---

[2] Semantic networks derive from much earlier work on knowledge representation using graphs, such as Pierce's "existential graphs" (see Roberts,

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 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-d), 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 they feel is the human language understanding process: a coordination 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]

1973) and the graphs of the psychologist Selz (1913, 1922) which represent patterns of concepts and inheritance of properties.

[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.

Dahlgren's (1988) language understanding system includes a sense disambiguation component which 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, b; 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 important source of disambiguation information for nouns--another result which has been subsequently tested and noted.

1.2.2 Connectionist methods. Work in psycholinguistics in the 1960's and 70's 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 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 duration of events (second, min-

[4] Note, however, that, symbolic methods such as Quillian's implement propagation via mechanisms such as marker passing, whereas the neural network models which developed in the late 1970's and early 1980's 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.*

ute, 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.

1.3 Knowledge-based methods

The AI-based work of the 1970's and 80's 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 *et al.*, 1993). Work on WSD reached a turning point in the 1980's when large-scale lexical resources such as dictionaries, thesauri, and corpora became widely available. Efforts began to attempt 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.

1.3.1 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

[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,

during the 1980's involved attempts to automatically extract lexical and semantic knowledge bases from MRDs (Michiels *et al.*, 1980; Calzolari, 1984; Chodorow *et al.,* 1985; Markowitz *et al.,* 1986; Byrd *et al.,* 1987; Nakamura and Nagao, 1988; Klavans *et al.,* 1990; Wilks *et al.,* 1990; etc.). 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 and b), 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 the 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 which 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 definition: the presence or absence of a given word can radically alter the results. However, Lesk's method has served as the basis for most MRD-based disambiguation work that has followed.

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

---

1968). Urdang (1984) describes a similar enterprise during the same period at *Random House.*

[6] Lesk does not use this term.

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 *et al.,* 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 *et al.* (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). In many ways, the supple-

---

[7] Note that the assumptions underlying this method are very similar to Quillian's:

> *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[...]* (Quillian, 1968, 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.

mentary information in the *LDOCE,* and in particular the subject codes, are similar to those in a thesaurus, which, however, are 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.

1.3.2 Thesauri. Thesauri provide information about relationships among words, most notably synonymy. *Roget's International Thesaurus,* which was put into machine-tractable form in the 1950's[8] 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. Typically, each occurrence of the same word under different categories of the thesaurus represent 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 to discriminate the correct sense of verbs such as *inspire* (*to raise the spirits* vs. *to inhale, breathe in, sniff,* etc.), *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 ed.). A 100-word context of each word in the category is extracted from a corpus (the 1991 electronic text of *Grolier's Ency-*

---

[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 1950's; the Sedelow's (1969) work relied on a machine readable version of the 3rd Edition. *Roget's* is now widely available via anonymous ftp from various sites.

clopedia), 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 which is 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 3-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 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 to establish 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]

1.3.3 Computational lexicons. In the mid-1980's, several efforts began to construct 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 *et al.*, 1994; Macleod *et al.*, forthcoming), etc. 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) 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.*, 1996a and b).

---

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

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 hyponomy/hyperonymy, antonymy, meronymy, etc. 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 which 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 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 which 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

---

English (*LLOCE*) (Chen and Chang, in this issue).

[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 the most used lexical resource at present for disambiguation in English, and therefore determined that *WordNet* senses should form the basis for a common sense inventory (Kilgarriff, 1997).

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 which 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, b; 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, approached 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 ten *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 which capture regularities in sense creation, as for metonymy, meronymy, etc. As outlined in Buitelaar (1997), sense disambiguation in the generative context starts first with a semantic tagging which points to a complex knowledge representation reflecting all 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 *et al.*, 1995).

Viegas *et al.* (forthcoming) describe a similar approach to WSD undertaken in the context of their work on machine translation (see also Mahesh *et al.*, 1997a and b). They access a large syntactic and semantic lexicon which

provides detailed information about selectional restrictions, etc. 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.

1.4 Corpus-based methods

1.4.1 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 & 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* which 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-60's interest in statistical treatment of language waned among linguists due to the trend in linguistics 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

---

[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]*

domains instead of general language. During the following ten to fifteen 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), the *Lancaster-Oslo-Bergen* (*LOB*) *Corpus* (Johansson, 1980), etc. 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 1980's. 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' results are encouraging (90% correct). Kelley and Stone (1975)'s work, which grew out of the Harvard "disambiguation project" for content analysis, is on a much larger scale; they extract KWIC concordances for 1800 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 1800 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', achieving 92% accuracy for gross homographic sense distinctions.

In the 1980's, 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 2000 concordance lines for five test words. Since then, s*upervised learning* from sense-tagged corpora has since been used by several researchers: Zernik (1990,

1991); Hearst (1991); Leacock *et al.* (1993); Gale *et al.* (1992d, 1993); Bruce and Wiebe (1994); Miller *et al.* (1994); Niwa and Nitta (1994); Lehman (1994); etc. 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.

1.4.2 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 or are being made: recently, 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 1000 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 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*) which includes a training phase during which each occurrence of a set of nouns[13] to be disambiguated is manually sense-tagged in several occurrences. 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 which avoids tagging each occurrence in the training corpus. Using letter fourgrams within a 1001 character window, his method (building on the vector-space model from information retrieval--see Salton *et al.*, 1975) first automatically clusters the words in the text, and 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. More seriously, it is not clear what the senses derived from the clusters correspond to (see

[13] This study involves only nouns.

for example Pereira *et al.*, 1993); and they are not in any case directly usable by other systems, since it is derived from the corpus itself.

Brown *et al.* (1991) and Gale *et al.* (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 *sentence* 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 *et al.* (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.

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.

1.4.3 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

[14] For example, Gale *et al.* (1993) remark that it is difficult to find any sense other than the financial sense for the word *bank* in the *Hansard Corpus*.

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]

S*moothing* 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 such 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 sub-parts 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 cited above (note 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 in to 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 *et al.* (1993) propose methods which 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 pre-tagged 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 *et al.,* 1993; Dagan *et al.,* 1994; Grishman and Sterling, 1993) exploit the same

---

[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).

idea of grouping observations for similar words, but without re-grouping 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 *et al.* (1993) give the following example: the pair (*chapter, describes*) does not appear in their corpus; however, *chapter* is similar to *book, introduction, section,* which are paired with *describes* in the corpus. On the other hand, the words similar to *book* are *books, documentation, manuals* (*op. cit.,* Fig. 1). Dagan *et al.*'s (1993) evaluation seems to show that similarity-based methods perform better than class-based methods. Karov and Edelman (in this issue) propose an extension to similarity-based methods by means of an iterative process at the learning stage, which gives 92% accurate results 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.

Open problems

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

1.5 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:

- The *bag of words* approach: here, 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 micro-context, topical context, and domain contributes to sense selection, but the relative role and importance of information from the different contexts and their inter-relations are not well understood. Very few

studies have used information of all three types, and the focus in much recent work is on micro-context alone. This is another area where systematic study is needed for WSD.

1.5.1 Micro-context. 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, 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 (in this issue) 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, in this issue), and to be in general less effective than methods which take other relations into consideration. However, as demonstrated in Yarowsky's (1992) work, the approach is cheaper than those which require 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 1949 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, dictionary-based WSD, as well as the more recent statistical, neural network, and symbolic machine learning, etc. 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 and b) examines different windows of micro-context, 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, *et al.* (in this issue) use a local window of ±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 is more workable in computational terms:

> *...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...

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 cooccurrences. 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 determined 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 (e.g., Hayes, 1977a and b; Wilks, 1973 and 1975b; Hirst, 1987) which relies on full

---

[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).

parsing, frames, semantic networks, the application of selectional preferences, etc. 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, etc.

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 and prepositional phrases and verb groups, and discards all other syntactic information. She examines items that are within plus/minus 3 phrase segments from the target and combines syntactic evidence with other kinds of evidence, such as capitalization. Yarowsky (1993) determined 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 (e.g., McRoy, 1992; Bruce and Wiebe, 1994; Leacock *et al.*, in this issue).

Evidence suggests that different kinds of disambiguation procedures are needed depending on syntactic category and the characteristics of the target word (Yarowsky, 1993, and Leacock, *et al.* in this issue)--an idea which is reminiscent of the word expert approach. However, to date there has been little systematic study 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.

1.5.2 Topical context. Topical context includes substantive words which co-occur with a given sense of a word, usually within a window of several sentences. Unlike micro-context, which has played a role in disambiguation work since the early 1950's, topical context has been less consistently used. Methods relying on topical context exploit redundancy in a text--that is, the repeated use of words which 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*, *ball*, etc. 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 *et al.* (1995) experiment with several statistical methods using a two-sentence window; Leacock *et al.* (1993, 1996) have similarly explored topical context for WSD. Gale *et al.* (1993),

looking at a ±50 word context, 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 micro-context 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 et al., 1992c). Leacock *et al.* (in this issue) 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 *et al.,* in this issue). 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 sub-topics. The most obvious way to divide a text is by sections (Brown and Yule, 1983), but this is only a gross division; sub-topics 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 sub-topics are beginning to emerge (see for example, Hearst, 1994; van der Eijk, 1994; Richmond *et al.*, 1997).

In this issue, Leacock *et al.* consider the role of micro-context vs. topical context, and attempt to assess the contribution of each. Their results indicate that for a statistical classifier, micro-context is superior to topical context as an indicator of sense. However, although a distinction is made between micro-context 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.

1.5.3 Domain. The use of domain for WSD is first evident in the micro-glossaries 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 the 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 et al'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 common, non-financial, sense. However, in the article THIRD WORLD, the word *intérêt* appears two times in each of these senses.

1.6 Sense division

1.6.1 The bank model. Because of their availability, most researchers in WSD work 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). 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 (e.g. Quine, 1960; Asprejan, 1974; Lyons, 1977; Weinrich, 1980; Cruse, 1986), but the lack of resolution over 2,000 years is striking.

1.6.2 Granularity. One of the foremost problems for WSD is to determine the appropriate degree of sense *granularity.* Several authors (e.g., Slator and Wilks, 1987) have remarked that the sense divisions one finds in dictionaries are

---

[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!

often too fine for the purposes of NLP work. Overly fine sense distinctions create practical difficulties for automated WSD: they introduce significant combinatorial effects (for example, Slator and Wilks (1987) note that the sentence phrase *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 (in this issue) propose an algorithm that combines senses in a dictionary (*LDOCE*) and link them to the categories of a thesaurus (*LLOCE*).

However, 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, e.g., 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, e.g., Cruse, 1986), and the points at which senses are combined or split can vary dramatically.

1.6.3 Senses or usages? The Aristotelian idea that words correspond to specific objects and concepts was displaced in the 20th century by the ideas of Saussure and others (Meillet, 1926; Hjemslev, 1953; Martinet, 1966; etc.). For Antoine Meillet :

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 :

Don't look for the meaning, but for the use.

Similar views are apparent in more recent theories of meaning, e.g., 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.

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 adopts 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 *et al.*, 1993). In this issue, Schütze continues in this vein and proposes a technique which avoids the problem of sense distinction altogether: he creates sense clusters from a corpus rather than rely on a pre-established sense list.

1.6.4 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 which 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 the play, 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.

1.7 Evaluation

Among 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, vs. general texts where sense use may be more variable. It has

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

been noted that in a commonly-used corpus such as the *Wall Street Journal,* certain senses of typical test words such as *line*[21] are absent entirely. 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 which tend to appear more frequently in metaphoric, metonymic, etc. 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 assignment 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 et al. (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), *Text Retrieval Conferences* (e.g. Harman, 1993, 1995), etc. 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" which contributes to an overall task

---

[21] In particular, it has been pointed out that the common sense of *line* as in the sentence, *He gave me a line of bologna* is not present in the *WSJ* corpus.

such as information retrieval, machine translation, etc. 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 then being evaluated in isolation, results are evaluated in terms of their contribution to the overall performance of a system designed for a particular application (e.g., machine translation).

1.7.1 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 *et al.,* 1994; or *declarative evaluation*: Arnold *et al.,* 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 *et al.*, 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, treatment of metaphor, metonymy, etc. in evaluation, how to deal with words of differing degrees and types of polysemy, etc., 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 pre-defined 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 which 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 inter-language 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 which 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 orga-

nization. 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 vs. *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 (1995). 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.

1.7.2 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, speech recognition, etc. This approach (also called *adequacy evaluation*: Hirschman et Thompson, 1996; or *operational evaluation*: Arnold *et al.*, 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 pre-tagged corpus. Only the final result is taken into consideration, subjected to evaluation appropriate to the task at hand.

Methods for WSD have evolved largely independent 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 efforts 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 which combines search-by-word and search-by-sense.

It remains to be seen to what extent WSD can improve results in particular applications. 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.

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 the problem 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 a very few words, most often only nouns, and very frequently concern very 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 1990's have obviously improved on earlier results, but it appears that we may have reached near 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,[22] we have made an attempt to cover at least the major areas of work and sketch the broad lines of development in the field.

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 vari-

[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.

ous 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 tends to revisit many of the same issues over time as a result. 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 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 work on WSD.

References

Aarts, Jan (1990). "Corpus linguistics: An appraisal." In Hammesse, Jacqueline and Zampolli, Antonio (Eds.), *Computers in Literary and Linguistic Research,* Champion Slatkine, Paris-Genève, 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." *Proceedings of the 7th European Conference on Artificial Intelligence, ECAI'86,* July 1986, Brighton, United Kingdom, 222-235. Reprinted in Du Boulay, B., Hogg, D, Steels, L. (Eds.), *Advances in Artificial Intelligence II,* Elsevier, 403-416.

Adriaens, Geert (1989). "The parallel expert parser: a meaning-oriented, lexically guided, parallel-interactive model of natural language understanding." *International Workshop on Parsing Technologies,* Carnegie-Mellon University, 309-319.

Adriaens, Geert and Small, Steven L. (1988). "Word expert revisited in a cognitive science perspective." In Small, Steven; Cottrell, Garrison W.; and Tanenhaus, Michael K. (Eds.) (1988). *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence,* Morgan Kaufman, San Mateo, California, 13-43.

Ahlswede, Thomas E. (1992). "Issues in the Design of Test Data for Lexical Disambiguation by Humans and Machines." *Proceedings of the Fourth Midwest Artificial Intelligence and Cognitive Science Society Conference*, Starved Rock, Illinois, 112-116.

Ahlswede, Thomas E. (1993). "Sense Disambiguation Strategies for Humans and Machines." *Proceedings of the 9th Annual Conference on the New Oxford English Dictionary,* Oxford, England, September, 75-88.

Ahlswede, Thomas E. (1995). "Word Sense Disambiguation by Human Informants." *Proceedings of the Sixth Midwest Artificial Intelligence and Cognitive Society Conference,* Carbondale, Illinois, April 1995, 73-78.

Ahlswede, Thomas E. and Lorand, David (1993). "The Ambiguity Questionnaire: A Study of Lexical Disambiguation by Human Informants." *Proceedings of the Fifth Midwest Artificial Intelligence and Cognitive Society Conference,*

Chesterton, Indiana, 21-25.

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

Alshawi, Hiyan and Carter, David (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. Dissertation, University of Texas at Austin, Austin, Texas, 293pp.

Amsler, Robert A. and White, John S. (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, Texas.

Anderson, John Robert (1976). *Language, Memory, and Thought.* Lawrence Erlbaum and Associates, Hillsdale, New Jersey.

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

ARPA (1993). *Proceedings of the Fifth Message Understanding Conference*, Baltimore, Maryland, August 1993. 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." *Proceedings of the Third Annual Conference of the UW Center for the New OED,* Waterloo, Ontario, 17-36.

Atkins, Beryl T. S. and Levin, Beth (1988). "Admitting impediments." *Proceedings of the 4th Annual Conference of the UW Center for the New OED,* Oxford, United Kingdom.

Ayto, John R. (1983). "On specifying meaning." In Hartmann, R.R.K. (Ed.), *Lexicography: Principles and Practice,* Academic Press, London, 89-98.

Bahl, Lalit R. and Mercer, Robert L. (1976). "Part of speech assignment by a statistical decision algorithm." In *IEEE International Symposium on Information Theory,* Ronneby, 88-89.

Bar-Hillel, Yehoshua (1960). "Automatic Translation of Languages." In Alt, Franz; Booth, A. Donald and Meagher, R. E. (Eds), *Advances in Computers,* Academic Press, New York.

Barwise, Jon and Perry, John R. (1983). *Situations and attitudes*, MIT Press, Cambridge, Massachusetts.

Basili, Roberto; Della Rocca, Michelangelo; and Pazienza, Maria Tereza (1997). "Towards a bootstrapping framework for corpus semantic tagging." *ACL-SIGLEX Workshop* "*Tagging Text with Lexical Semantics: Why, What, and How?"* April 4-5, 1997, Washington, D.C., USA, 66-73.

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 Aitken, Adam J.; Bailey, Richard W., and Hamilton-Smith, Neil (Eds.) *The Computer and Literary Studies.* Edinburgh University Press, Edinburgh, United Kingdom, 103-112.

Bimbot, Frédéric; Chollet, Gérard; and Paoloni, A. (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*, 75-82.

Black, Ezra (1988). "An Experiment in Computational Discrimination of English Word 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.* Doctoral dissertation, Computer Laboratory, University of Cambridge, August 1979 [available as technical report 11].

Bookman, Lawrence A. (1987). "A Microfeature Based Scheme for Modelling Semantics." *Proceedings of the 10th International Joint Conference on Artificial Intelligence, IJCAI'87,* Milan, Italy, 611-614.

Braden-Harder, Lisa (1993). "Sense disambiguation using online dictionaries." In Jensen, Karen; Heidorn George E.; and Richardson, Stephen D.. (eds.). *Natural Language Processing: The PLNLP approach,* Kluwer Academic Publishers, Dordrecht, 247-261.

Briscoe, Edward J. (1991). "Lexical issues in natural language processing." In Klein, Ewan H. and Veltman, Frank (Eds.). *Natural Language and Speech.* [*Proceedings of the Symposium on Natural Language and Speech,* 26-27 November 1991, Brussels, Belgium.] Springer-Verlag, Berlin, 39-68.

Brown, Gillian and Yule, George (1983). *Discourse analysis.* Cambridge Textbooks in Linguistics Series. Cambridge University Press, Cambridge, United Kingdom.

Brown, Peter F.; Della Pietra, Stephen; Della Pietra, Vincent J.; and Mercer Robert L. (1991). "Word sense disambiguation using statistical methods." *Proceedings of the 29th Annual Meeting of Association for Computational Linguistics,* Berkeley, California, 264-270.

Brown, Peter F.; Della Pietra, Vincent J.; deSouza, Peter V.; Lai, Jennifer C.; and Mercer Robert L. (1992). "Class-based *n*-gram models of natural language." *Computational Linguistics,* **18**(4), 467-479.

Bruce, Rebecca and Wiebe, Janyce (1994). "Word-sense Disambiguation Using Decomposable Models." *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics,* Las Cruces, New Mexico, 139-145.

Bryan, Robert M. (1973). "Abstract thesauri and graph theory applications to thesaurus research." In Sedelow, Sally Yeates (Ed.), *Automated Language Analysis, 1972-3.* University of Kansas Press, Lawrence, Kansas, 45-89.

Bryan, Robert M. (1974). "Modelling in thesaurus research." In Sedelow, Sally Yeates *et al.* (Ed.), *Automated Language Analysis, 1973-4.* University of Kansas Press, Lawrence, Kansas, 44-59.

Buitelaar, Paul (1997). "A lexicon for underspecified semantic tagging." *ACL-SIGLEX Workshop* "*Tagging Text with Lexical Semantics: Why, What, and How?"* April 4-5, 1997, Washington, D.C., 25-33.

Byrd, Roy J.; Calzolari, Nicoletta; Chodorov; Martin S.; Klavans, Judith L.; Neff, Mary S.; and Rizk, Omneya (1987). "Tools and methods for computational linguistics." *Computational Linguistics,* **13**(3/4), 219-240.

Calzolari, Nicoletta (1984). "Detecting patterns in a lexical data base." *Proceedings of the 10th International Conference on Computational Linguistics, COLING'84,* 2-6 July 1984, Stanford University, California, 170-173.

Chen, Jen Nan and Chang, Jason S. (1998). "Topical clustering of MRD senses based on information retrieval techniques." In this issue.

Chen, Stanley F. and Goodman, Joshua (1996). "An empirical

study of smoothing techniques for language modeling." *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics,* 24-27 June 1996, University of California, Santa Cruz, California, 310-318.

Chodorow, Martin S.; Byrd, Roy J.; and Heidorn, George E. (1985). "Extracting semantic hierarchies from a large online dictionary." *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics,* 8-12 July 1985, University of Chicago, Chicago, Illinois, 299-304.

Chomsky, Noam (1957). *Syntactic structures.* Mouton, The Hague.

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

Choueka, Yaacov, Goldberg, D. (1979). "Mechanical resolution of lexical ambiguity -- a combinatorial approach." *Proceedings of the International Conference on Literary and Linguistic Computing*, Israel, April 1979, Zvi Malachi (Ed.), The Katz Research Institute for Hebrew Literature, Tel-Aviv University, 149-165.

Choueka, Yaacov and Lusignan, Serge (1985). "Disambiguation by short contexts." *Computers and the Humanities,* 19, 147-158.

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

Collins, Allan M. and Loftus, Elisabeth F. (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 Altmann, Gerry T. (Ed.) *Cognitive models in speech processing.* The MIT Press. Cambridge, Massachussets, 540pp.

Cottrell, Garrison W., Small, Steven L. (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. Dissertation. Department of Computer Science, University of Rochester.

Cowie, Jim; Guthrie, Joe A.; and Guthrie, Louise (1992). "Lexical disambiguation using simulated annealing." *Proceedings of the 14th International Conference on Computational Linguistics, COLING'92,* 23-28 August, Nantes, France, vol. 1, 359-365.

Cruse, D. A. (1986). *Lexical Semantics.* Cambridge University Press. Cambridge, United Kingdom.

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

Dagan, Ido; Itai, Alon; and Schwall, Ulrike (1991). "Two languages are more informative than one." *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 18-21 June 1991, Berkeley, California, 130-137.

Dagan, Ido; Marcus, Shaul; and Markovitch, Shaul (1993). "Contextual word similarity and estimation from sparse data." *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics,* 22-26 June 1993, Columbus, Ohio.

Dagan, Ido; Peireira, Fernando and Lee, Lilian (1994). "Similarity-based estimation of word cooccurrence probabilities." *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics,* Las Cruces, New

Mexico, 272-278.

Dahlgren, Kathleen G. (1988). *Naive Semantics for Natural Language Understanding.* Kluwer Academic Publishers, Boston. 258pp.

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, 297pp.

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." *Proceedings of the 15th International Conference on Computational Linguistics, COLING'94,* 5-9 August 1994, Kyoto, Japan, 712-716.

Dostert, Leon E. (1955). "The Georgetown-I.B.M. experiment." In Locke, William N. and Booth, A. Donald (Eds.) (1955). *Machine translation of languages.* John Wiley & Sons, New York, 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, 441pp.

Engelson, Sean P. and Dagan, Ido (1996). "Minimizing manual annotation cost in supervised training from corpora." *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics,* 24-27 June 1996, University of California, Santa Cruz, California, 319-326.

Estoup, Jean-Baptiste (1907). *Gammes sténographiques.* Paris.

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

Fellbaum, Christiane, (Ed.) (forthcoming-a). *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, Massachusetts.

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

Fillmore, Charles J. and Atkins, Beryl T. S. (1991) Invited lecture presented at the *29th Annual Meeting of the Association for Computational Linguistics,* 18-21 June 1991, Berkeley, California.

Firth, J. R. (1957) "Modes of meaning." *Papers in Linguistics 1934-51.* Oxford University Press, Oxford, United Kingdom. 190-215.

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". *Proceedings of the 4th International Congress on Lexicography, EURALEX'90,* Benalmádena, Spain, 89-103.

Fries, Charles (1952). *The structure of English: An introduction to the construction of sentences.* Hartcourt & Brace, New York.

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

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

Gale, William A.; Church, Kenneth W.; and Yarowsky, David (1992b). "Estimating upper and lower bounds on the performance of word-sense disambiguation programns." *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics,* 28 June-2 July 1992, University of Delaware, Newark, Delaware, 249-256.

Gale, William A.; Church, Kenneth W.; and Yarowsky, David (1992c). "One sense per discourse." *Proceedings of the Speech and Natural Language Workshop,* San Francisco, Morgan Kaufmann, 233-37.

Gale, William A.; Church, Kenneth W.; and Yarowsky, David (1992d). "Work on statistical methods for word sense disambiguation." *Probabilistic Approaches to Natural Language: Papers from the 1992 AAAI Fall Symposium,* 23-25 October 1992, Cambridge, Massachusetts, 54-60.

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

Gentilhomme, Yves and Tabory, René (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 Michea, René (1961). "Sur la détermination du sens d'un mot au moyen du contexte." *La Traduction Automatique,* **2**(1), 16-17.

Gougenheim, Georges; Michea, René; Rivenc, Paul; and Sauvageot, Aurélien (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." *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI'77,* 172-178.

Grishman, Ralph and Sterling, John (1993). "Smoothing of automatically generated selectional constraints." *Human Language Technology.* Morgan Kaufmann, 254-259.

Grishman, Ralph; MacLeod, Catherine; and Meyers, Adam (1994). "COMLEX syntax: Building a computational lexicon." *Proceedings of the 15th International Conference on Computational Linguistics, COLING'94,* 5-9 August 1994, Kyoto, Japan, 268-272.

Guthrie, Joe A.; Guthrie, Louise; Wilks, Yorick; and Aidinejad, Homa (1991). "Subject-dependent co-occurrence and word sense disambiguation." *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics,* 18-21 June 1991, Berkeley, California, 146-152.

Haas, W. (1966) "Linguistic relevance." In Bazell, C.E. *et al.* (Eds.), *In Memory of J.R. Firth,* Longman, London, 116-48.

Halliday, M.A.K. (1961) "Categories of the theory of grammar." *Word,* **17, *241-92.***

Halliday, M.A.K. (1966) "Lexis as a linguistic level." In Bazell, C.E. *et al.* (Eds.), *In Memory of J.R. Firth,* Longman, London, 148-63.

Harman, Donna (Ed.) (1993). *National Institute of Standards and Technology Special Publication No. 500-207 on the First Text Retrieval Conference (TREC-1)*, Washington, DC, 1993. National Institute of Standards and Technology, U.S. Department of Commerce, U.S. Government Printing Office.

Harman, Donna (Ed.) (1995). Special Issue on The Second Text Retrieval Conference (TREC-2). *Information Processing and Management,* **31**(3).

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, xv-384 pp.

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. *Proceedings of the 5th International. Joint Conference on Artificial Intelligence,* Cambridge, Massachusetts, 99-107.

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." *Proceedings of the 7th Annual Conf. of the University of Waterloo Centre for the New OED and Text Research,* Oxford, United Kingdom, 1-19.

Hearst, Marti. A. (1994). "Multiparagraph segmentation of expository text." *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics,* Las Cruces, New Mexico, 9-16.

Hebb, Donald O. (1949). *The organisation of behavior: A neuropsychological approach.* John Wiley & Sons, New York.

Hindle, Donald, and Rooth, Mats (1993). "Structural Ambiguity and lexical relations." *Computational Linguistics,* **19**(1), 103-120.

Hirschman, Lynette and Thomson, Henry S. (1996). "Overview of evaluation in speech and natural language processing." In Cole, Ronald A. (Ed.), *Survey of the State of the Art in Human Language Technology.* Section 13.1. URL: http://www.cse.ogi.edu/CSLU/ HLTsurvey/

Hirst, Grame (1987). *Semantic interpretation and the resolution of ambiguity.* Studies in Natural Language Processing. Cambridge University Press, Cambridge, United Kingdom, 263pp.

Hjemslev, Louis (1953). *Prolegomena to a theory of language.* Translated from Danish. Indiana University, Bloomington, Indiana.

Hobbs, Jerry R. (1987). "World knowledge and word meaning." *Proceedings of the Third Workshop on Theoretical Issues in Natural Language Processing, TINLAP-3.* Las Cruces, New Mexico, 20-25.

Ide, Nancy and Véronis, Jean (1990a). "Very large neural networks for word sense disambiguation." *Proceedings of the 9th European Conference on Artificial Intelligence, ECAI'90,* Stockholm, 366-368.

Ide, Nancy and Véronis, Jean (1990b). "Mapping dictionaries: A spreading activation approach," *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary,* Waterloo, 52-64.

Ide, Nancy and Véronis, Jean (1993a). "Refining taxonomies extracted from machine-readable dictionaries." In Hockey, Susan and Ide, Nancy (Eds.) *Research in Humanities Computing II,* Oxford University Press, 145-59.

Ide, Nancy and Véronis, Jean (1993b). "Knowledge extraction from machine-readable dictionaries: an evaluation." *Third International EAMT Workshop "Machine Translation and the Lexicon",* Heidelberg (Germany), April 1993. Published in Steffens, Petra (Ed.) (1995), *Machine Translation*

and the Lexicon, Lecture Notes in Artificial Intelligence 898, Springer-Verlag, Berlin, 19-34.

Ide, Nancy and Walker, Donald (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, 313-19.

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 Leitner, Gerhard (Ed.). *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 Mercer, Robert L. (1985). "Probability distribution estimation from sparse data." *IBM Technical Disclosure Bulletin,* 28, 2591-2594.

Jensen, Karen and Binot, Jean-Louis (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, 18pp, November 1950. [Published as: Kaplan, Abraham (1955). "An experimental study of ambiguity and context." *Mechanical Translation,* **2**(2), 39-46.]

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

Karov, Yael and Edelman, Shimon (1998). "Similarity-based word sense disambiguation". In this issue.

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

Kilgarriff, Adam (1992). *Polysemy.* Ph. D. Thesis. University of Sussex, United Kingdom.

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)." *Proceedings of the 6th International Congress on Lexicography, EURALEX'94,* Amsterdam, Holland, 101-106.

Kilgarriff, Adam (1997). "Evaluation of word sense disambiguation programs." *SALT Club Workshop "Evaluation in Speech and Language Technology",* Sheffield University, Sheffield, United Kingdom, 17-18 June 1997.

Kilgarriff, Adam (forthcoming). "I don't believe in word senses." To appear in *Computers and the Humanities.*

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

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

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

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

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

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

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

Leacock, Claudia; Towell, Geoffrey; and Voorhees, Ellen M. (1996) "Towards building contextual representations of word senses using statistical models." In Boguraev, Branimir and Pustejovsky, James (Eds.), *Corpus Processing for Lexical Acquisition,* MIT Press, Cambridge, Massachusetts, 97-113.

Leacock, Claudia; Miller, George A.; and Chodorow, Martin (1998). "Using corpus statistics and WordNet relations for sense identification". In this issue.

Lehman, Jill Fain (1994). "Toward the essential nature of statistical knowledge in sense resolution." *Proceedings of the 12th International Conference on Artificial Intelligence, AAAI'94,* 31 July - 4 August 1994, Seattle, Washington, 734-741.

Leech, Geoffrey (1991). "The state of the art in corpus linguistics." In Aijmer, K., Altenberg, B. (Eds.), *English Corpus Linguistics.* Longman, London, 8-29.

Lenat, Douglas B. and Guha, Ramanathan V., (1990). *Building large knowledge-based systems.* Addison-Wesley, Reading, Massachusetts.

Lesk, Michael (1986). "Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone." *Proceedings of the 1986 SIGDOC Conference,* Toronto, Canada, June 1986, 24-26.

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

Litowski, Kenneth C. (1997). "Desiderata for tagging with WordNet sysnsets or MCAA categories." *ACL-SIGLEX Workshop "Tagging Text with Lexical Semantics: Why, What, and How?"* April 4-5, 1997, Washington, D.C., USA, 12-17.

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." *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics,* Cambridge Massachusetts, 181-188.

Lupker, Stephen J. (1984). "Semantic priming without association: A second look." *Journal of Verbal Learning and Ver-*

bal Behavior, **23**(6), 709-733.

Lyons, John (1966) "Firth's theory of meaning." In Bazell, C.E. et al. (Eds.). *In Memory of J.R. Firth,* Longman, London, 288-302.

Lyons, John (1977). *Semantics*. Cambridge University Press, Cambridge, England.

Macleod, Catherine, Grishman, Ralph, Meyers, Adam (forthcoming). "A large syntactic dictionary for natural language processing." To appear in *Computers and the Humanities.*

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

Mahesh, Kavi; Nirenburg, Sergei; Beale, Stephen; Viegas, Evelyne; Raskin, Victor; and Onyshkevych, Boyan (1997a). "Word Sense Disambiguation: Why statistics when we have these numbers?" *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation,* 23-25 July 1997, Santa Fe, New Mexico, 151-159.

Mahesh, Kavi; Nirenburg, Sergei; and Beale, Stephen (1997b). "If you have it, flaunt it: Using full ontological knowledge for word sense disambiguation." *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation,* 23-25 July 1997, Santa Fe, New Mexico, 1-9.

Malakhovski, L. V. (1987). "Homonyms in English dictionaries." In Burchfield, R. W. (Ed.), *Studies in Lexicography.* Oxford University Press, Oxford, United Kingdom, 36-51.

Markowitz, Judith; Ahlswede, Thomas; and Evens, Martha (1986). "Semantically significant patterns in dictionary definitions." *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics,* New York, 112-119.

Martinet, André (1950). "Review of Nida, E., 'Morphology : The descriptive analysis of words'", *Word,* **6**(1), 84-87.

Masterman, Margaret (1957). "The thesaurus in syntax and semantics." *Mechanical Translation,* 4, 1-2.

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

McClelland, James L. and Rumelhart, David E. (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 Pitts, Walter (1943). "A logical calculus of the ideas imminent in nervous activity." *Bulletin of Mathematical Biophysics,* 5, 115-133.

McIntosh, A. (1966). "Patterns and ranges." *Papers in General, Descriptive, and Applied Linguistics.* Longman, London, 183-99.

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.* Vol. 1. Champion, Paris, 351pp. (2nd édition).

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

Meyer, David E. and Schvaneveldt, Roger W. (1971). "Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations." *Journal of Experimental Psychology,* **90**(2), 227-34.

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, Archibal; Mullenders, Jacques; and Noël, Jacques

(1980). "Exploiting a large database by Longman." *Proceedings of the 8th International Conference on Computational Linguistics, COLING'80,* Tokyo, Japan, 374-382.

Michiels, Archibal (1982). *Exploiting a large dictionary data base.* Doctoral dissertation, Université de Liège, Liège, Belgique.

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

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

Miller, George A.; Chodorow, Martin; Landes, Shari; Leacock, Claudia; and Thomas, Robert G. (1994). "Using a semantic concordance for sense identification." *ARPA Workshop on Human Language Technology*, Plainsboro, New Jersey, March 1994, 240-243.

Miller, George A.; Leacock, Claudia ; Tengi, Randee ; and Bunker, Ross (1993). "A semantic concordance." *Proceedings of the 3rd DARPA Worshop on Human Language Technology,* Plainsboro, New Jersey, March 1993, 303-308.

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 Hirst, Graeme (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 Nagao, Makoto (1988). "Extraction of semantic information from an ordinary English dictionary and its evaluation." *Proceedings of the 12th International Conference on Computational Linguistics, COLING'88,* 22-27 August 1988, Budapest, Hungary, 459-464.

Ng, Hwee Tou and Lee, Hian Beng (1996). "Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach." *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics,* 24-27 June 1996, University of California, Santa Cruz, California, 40-47.

Niwa, Yoshiki and Nitta, Yoshihiko (1994). "Coocurrence vectors from corpora vs distance vectors from dictionaries." *Proceedings of the 15th International Conference on Computational Linguistics, COLING'94,* 5-9 August 1994, Kyoto, Japan, 304-309.

Oettinger, Anthony G. (1955). "The design of an automatic Russian-English technical dictionary." In Locke, William N. and Booth, A. Donald (Eds.) (1955). *Machine translation of languages.* John Wiley & Sons, New York, 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, California, October 1968.

Oswald, Victor A. Jr. (1952). "Microsemantics." Communication presented at the first *M.I.T. conference on Mechanical Translation,* 17-20 June 1952. Mimeographed, 10 pp. [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 Dostert, Leon E. (Ed.), *Research in Machine Translation,* Georgetown University Press, Washington, D.C., 63-69.

Oswald, Victor A. Jr. and Lawson, Richard H. (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, Kansas.

Pendergraft, Eugene (1967) “Translating languages”. In Borko, Harold (Ed.), *Automated Language Processing.* John Wiley and Sons, New York.

Pereira, Fernando and Tishby, Naftali (1992). “Distributional similarity, phase transitions and hierarchical clustering.” *Working notes of the AAAI Symposium on Probabilistic Approaches to Natural Language,* October 1992, Cambridge, Massachusetts, 108-112.

Pereira, Fernando; Tishby, Naftali; and Lee, Lilian (1993). “Distributional clustering of English.” *Proceeedings of the 31st Annual Meeting of the Association for Computational Linguistics,* 22-26 June 1993, Ohio State University, Colombus, Ohio, 183-190.

Pimsleur, P. (1957). “Semantic frequency counts.” *Mechanical Translation,* **4**(1-2), 11-13.

Pustejovsky, James (1995). *The Generative Lexicon.* The MIT Press, Cambridge, Massachusetts.

Pustejovsky, James; Boguraev, Bran; and Johnston, Michael (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.” Communication presented at the colloquium *Semantic problems in natural language.* September 1961. King’s College, Cambridge University, Cambridge, United Kingdom.

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, M.I.T., August 1962.

Quillian, M. Ross (1967). “Word concepts: A theory and simulation of some basic semantic capabilities.” *Behavioral Science,* 12, 410-30.

Quillian, M. Ross (1968). “Semantic memory.” In Minsky, M. (Ed.), *Semantic Information Processing,* MIT Press, 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,* The MIT Press, Cambridge, Massachusetts.

Quirk, Randolph (1960). “Towards a description of English usage.” *Transactions of the Philological Society,* 40-61.

Rayner, Keith and Morris, R. K. (1991). “Comprehension processes in reading ambiguous sentences: reflections from eye movements.” In Simpson, G. (Ed.). *Understanding Word and Sentence.* North-Holland, Amsterdam, 175-198.

Reifler, Erwin (1955). The mechanical determination of meaning. In Locke, William N. and Booth, A. Donald (Eds.), *Machine translation of languages.* John Wiley & Sons, New York, 136-164.

Resnik, Philip (1992). “WordNet and distributional analysis: a class-based approach to statistical discovery.” *AAAI Workshop on Statistically-based Natural Language Processing Techniques,* San Jose, California, 48-56.

Resnik, Philip (1993a). *Selection and information: A class-based approach to lexical relationships.* Ph.D. dissertation, University of Pennsylvania. Also University of Pennsylvania Technical Report 93-42.

Resnik, Philip (1993b). "Semantic classes and syntactic ambiguity." *ARPA Workshop on Human Language Technology,* 278-83.

Resnik, Philip (1995a). "Disambiguating Noun Groupings with Respect to WordNet Senses." *Proceedings of the Third Workshop on Very Large Corpora,* Cambridge, Massachusetts, 54-68.

Resnik, Philip (1995b). "Using information content to evaluate semantic similarity in a taxonomy.", *Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI'95.* Montréal, 448-53.

Resnik, Philip and Yarowsky, David (1997a). "Evaluating automatic semantic taggers." *ACL-SIGLEX Workshop* "*Tagging Text with Lexical Semantics: Why, What, and How?"* April 4-5, 1997, Washington, D.C., 91.

Resnik, Philip and Yarowsky, David (1997b). "A perspective on word sense disambiguation methods and their evaluation." *ACL-SIGLEX Workshop* "*Tagging Text with Lexical Semantics: Why, What, and How?"* April 4-5, 1997, Washington, D.C., 79-86.

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

Richardson, Ray and Smeaton, Alan F. (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), 144-47.

Richmond, Korin; Smith, Andrew; and Amitay, Einat (1997). "Detecting Subject Boundaries Within Text: A Language Independent Statistical Approach." *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2),* Brown University, Providence, Rhode Island, August 1-2, 47-54.

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

Robins , R. H. (1987). "Polysemy and the lexicographer." In Burchfield, R. W. (Ed.), *Studies in Lexicography.* Oxford University Press, Oxford, United Kingdom, 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 McGill, M. (1983). *Introduction to Modern Information Retrieval.* McGraw-Hill, New York.

Salton, Gerard; Wong, A.; and Yang, C. S. (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 Croft, W. Bruce and van Rijsbergen, C. J. (Eds.), *Proceedings of the 17th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval,* Las Vegas, 161-175.

Schank, Roger C. and Abelson, Robert P. (1977). *Scripts, Plans, Goals and Understanding.* Lawrence Erlbaum, Hillsdale, New Jersey.

Schmidt, Klaus M. (1988). "Der Beitrag der begriffsorientierten Lexicographie zue systematischen Erfassung von Sprachwandel und das Begriffwörterbuch zur wdh. Epik." In Bachofer, W. (ed.). *Mittelhochdeutches Wörterbuch in der Diskussion.* Tubingen; Max Niemeyer, 25-49.

Schmidt, Klauss M. (1991). "Ein databanksystem für das Begriffwörterbuch Mittelhochdeutcher Epik und Fortschritte bie der automatischen Disambiguierung." In Gärtner, K., Sappler, P., Trauth, M. (Eds.). *Maschinelle Verarbeitung altdeutscher Text IV.* Tubingen; Max Niemeyer, 192-204.

Schütze, Hinrich (1992). "Dimensions of meaning." *Proceedings of Supercomputing'92.* IEEE Computer Society Press, Los Alamitos, California. 787-796.

Schütze, Hinrich (1993). "Word space." In Hanson, Stephen J.; Cowan, Jack D.; and Giles, C. Lee (Eds.) *Advances in Neural Information Processing Systems 5,* Morgan Kauffman, San Mateo, California, 5, 895-902.

Schütze, Hinrich (1998). "Automatic word sense discrimination". In this issue.

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

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

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

Sedelow, Sally Yeates and Sedelow, Walter. A. Jr. (1986). "Thesaural knowledge representation." *Proceedings of the University of Waterloo Conference on Lexicology.* Waterloo, Ontario, 29-43.

Sedelow, Sally Yeates and Sedelow, Walter. A. Jr. (1992) "Recent Model-Based and Model-Related Studies of a Large-Scale Lexical Resource (Roget's Thesaurus)." *Proceedings of the 14th International Conference on Computational Linguistics, COLING'92,* 23-28 August, Nantes, France, 1223-1227.

Seidenberg, Mark S.; Tanenhaus, Michael K.; Leiman, James M.; and Bienkowski, Marie A. (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 Burgess, Curt (1989). "Implications of lexical ambiguity resolution for word recognition and comprehension." In Small, Steven; Cottrell, Garrison W.; and Tanenhaus, Michael K. (Eds.) (1988). *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence,* Morgan Kaufman, San Mateo, California, 271-288.

Sinclair, John (1966). "Beginning the study of lexis." In Bazell, C.E. et al. (Eds.). *In Memory of J.R. Firth.* Longman, London, 410-31.

Sinclair, John (Ed.) (1987). *Looking up: An account of the COBUILD project in lexical computing.* Collins, London, 182pp.

Skorochod'ko, E.F. (1972). "Adaptative methods of automatic abstracting and indexing." In Freiman, C.V. (Ed.), *Information Processing 71: Proceedings of the IFIP Congress 71,* North Holland Publishing Company, 1179-1182.

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

Slator, Brian M. and Wilks, Yorick A. (1987). "Towards semantic structures forom dictionary entries." *Proceedings of the 2nd Annual Rocky Mountain Conference on Artificial*

Intelligence. 17-19 June 1987, Boulder, Colorado, 85-96.

Small, Steven L. (1980). *Word expert parsing: a theory of distributed word-based natural language understanding.* Doctoral dissertation, Department of Computer Science, University of Maryland, September 1980. [available as technical report 954].

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

Small, Steven L.; Cottrell, Garrison, Tanenhaus, Michael K. (1988) (Eds.). *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence,* Morgan Kaufman, San Mateo, California, 518pp.

Small, Steven L. and Rieger Charles (1982), "Parsing and comprehencing with word experts (a theory and its realization)." In Lenhert Wendy and Ringle Martin (Eds.), *Strategies for Natural Language Processing,* Lawrence Erlbaum and Associates, Hillsdale, New Jersey, 89-147.

Sparck Jones, Karen (1964). *Synonymy and semantic classification.* Ph. D. thesis, University of Cambridge, Cambridge, England.

Sparck Jones, Karen (1986). *Synonymy and semantic classification.* Edinburgh, Edinburgh University Press, England.

Sparck Jones, Karen (à paraître). "What is the role of NLP in Text Retrieval?" In Strzalkowski, Tomek (Ed.) *Natural Language Information Retrieval,* Text, Speech and Language Technology Series. Kluwer Academic Publishers, Dordrecht.

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

Stock, Penelope F. (1983). "Polysemy." *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 Gerbner, George; Holsti, Ole, Krippendorf, Klaus; Paisley, William J.; and Stone, Philip J. (Eds.), *The Analysis of Communication Content,* John Wiley and Sons, New York, 199-221.

Stone, Philip J.; Dunphy, Dexter C.; Smith, Marshall S.; and Ogilvie, Daniel M. (Eds.) (1966). *The General Inquirer: A Computer Approach to Content Analysis,* The MIT Press, Cambridge, Massachusetts, 651pp.

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

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

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

Sutcliffe, Richard F. E.; O Sullivan, D.; Polikarpov, A. A.; Kuzmin, L. A.; McElligott, A.; and Véronis, J. (1996b). "IWNR - Extending A Public Multilingual Taxonomy to Russian." *Proceedings of the Workshop "Multilinguality in the Lexicon," AISB Second Tutorial and Workshop Series,* 31 March - 2 April 1996, University of Sussex, Brighton,

United Kingdom, 14-25.

Tabossi, Patricia (1989). "What's in a context?" In Gorfein, D. (Ed.). *Resolving Semantic Ambiguity,* Springer-Verlag, New York, 25-39.

Tabossi, Patricia (1991). "Understanding words in context." In Simpson, G. (Ed.). *Understanding Word and Sentence.* North-Holland, Amsterdam, 1-22.

ten Hacken, Pius (1990). "Reading distinction in machine translation." *Proceedings of the 12th International Conference on Computational Linguistics, COLING'90,* 20-25 August 1990, Helsinki, Finland, vol. 2, 162-166.

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 Lorge, Irving (1938). *Semantic counts of English Words,* Columbia University Press, New York.

Thorndike, Edward L. and Lorge, Irving (1944). *The Teacher's Word Book of 30,000 Word.,* Columbia University Press, New York.

Towell, Geoffrey and Voorhees, Ellen (1998). "Disambiguating highly ambiguous words." In this issue.

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),* August 1994, Kyoto, Japan, 143-159.

Véronis, Jean and Ide, Nancy (1990). "Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries." *13th International Conference on Computational Linguistics, COLING'90,* Helsinki, Finland, vol. 2, 389-394.

Véronis, Jean and Ide, Nancy (1991). "An assessment of information automatically extracted from machine readable dictionaries." *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics,* Berlin, Germany, 227-232.

Véronis, Jean and Ide, Nancy (1995). "Large Neural Networks for the Resolution of Lexical Ambiguity." In Saint-Dizier, Patrick; Viegas, Evelyne (Eds.) *Computational Lexical Semantics.* Natural Language Processing Series, Cambridge University Press, Cambridge, United Kingdom, 251-269.

Viegas, Evelyne; Mahesh, Kavi; and Nirenburg, Sergei (forthcoming). "Semantics in action." In Saint-Dizier, Patrick (Ed.). *Predicative Forms in Natural Language and Lexical Knowledge Bases.* Text, Speech and Language Technology Series. Kluwer Academic Publishers, Dordrecht.

Voorhes, Ellen M. (1993). "Using WordNet to disambiguate word senses for text retrieval." *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 27 June-1 July 1993, Pittsburgh, Pennsylvania, 171-180.

Voorhees, Ellen M., Claudia Leacock, and Geoffrey Towell (1995). "Learning context to disambiguate word senses." In Thomas Petsche; Stephen José Hanson; and Jude Shavlik, eds., *Computational Learning Theory and Natural Learning Systems.* MIT Press, Cambridge, Massachusetts.

Vossen, Piek (forthcoming). "Introduction to EuroWordNet." To appear in a Special Issue of *Computers and the Humanities* on *EuroWordNet.*

Waibel, Alex and Lee, Kai-Fu (Eds.) (1990). *Readings in Speech Recognition.* Morgan Kaufmann, San Mateo, California.

Waltz, David L. and Pollack, Jordan B. (1985). "Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation." *Cognitive Science,* 9, 51-74.

Weinreich, Uriel (1980). *On semantics.* University of Pennsylvania Press, 128pp.

Weiss, S. (1973). Learning to disambiguate. *Information Storage and Retrieval,* 9, 33-41.

Weaver, Warren (1949). *Translation.* Mimeographed, 12 pp., July 15, 1949. Reprinted in Locke, William N. and Booth, A. Donald (1955) (Eds.), *Machine translation of languages.* John Wiley & Sons, New York, 15-23.

Weibe, Janyce; Maples, Julie; Duan, Lee; and Bruce, Rebecca (1997). "Experience in WordNet sense tagging in the Wall Street Journal." *ACL-SIGLEX Workshop* "*Tagging Text with Lexical Semantics: Why, What, and How?"* April 4-5, 1997, Washington, D.C., 8-11.

Weiss, Stephen (1973). "Learning to disambiguate." *Information Storage and Retrieval,* 9,

Whittemore, Greg; Ferrara, Kathleen; and Brunner, Hans (1990). "Empirical studies of predictive powers of simple attachment schemes for post-modifier prepositional phrases." *Proceedings of the 28th Annual Meeting of Association for Computational Linguistics,* 6-9 June 1990, Pittsburgh, Pennsylvania, 23-30.

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 Schank, Roger and Colby, Kenneth (Eds.). *Computer Models of Thought and Language,* San Francisco: W H Freeman, 114-151.

Wilks, Yorick A. (1975a). "Primitives and words." *Proceedings of the Interdisciplinary Workshop on Theoretical Issues in Natural Language Processing,* Cambridge, MA, June 1975, 42-45.

Wilks, Yorick A. (1975b). "Preference semantics." In Keenan, E. L. III (Ed.), *Formal Semantics of Natural Language,* Cambridge University Press, 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." To appear in *Computers and the Humanities.*

Wilks, Yorick A.; Fass, Dan (1990). *Preference semantics: A family history.* Report MCCS-90-194, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico.

Wilks, Yorick A.; Fass, Dan; Guo, Cheng-Ming; MacDonald, James E.; Plate, Tony; and Slator, Brian A. (1990). "Providing Machine Tractable Dictionary Tools." In Pustejovsky, James (Ed.), *Semantics and the Lexicon.* MIT Press, Cambridge, Massachusetts.

Wilks, Yorick A.; Slator, Brian A.; and Guthrie, Louise M. (1996). *Electric words: Dictionaries, computers, and meanings.* A Bradford Book. The MIT Press, Cambridge, Massachusetts, 289pp.

Wilks, Yorick and Stevenson, Mark (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, United Kingdom.

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,* 23-28 August,

Nantes, France, 454-460.

Yarowsky, David (1993). "One sense per collocation." *Proceeding of ARPA Human Language Technology Workshop,* Princeton, New Jersey, 266-271.

Yarowsky, David (1994a). "Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French." *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics,* Las Cruces, New Mexico, 88-95.

Yarowsky, D. (1994b). "A comparison of corpus-based techniques for restoring accents in Spanish and French text." *Proceedings of the 2nd Annual Workshop on Very Large Text Corpora.* Las Cruces, 19-32.

Yarowsky, David (1995). "Unsupervised word sense disambiguation rivaling supervised methods." *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics,* 26-30 June 1995, Cambridge, Massachusetts, 189-196.

Yarowsky, David (1997). "Homograph disambiguation in text-to-speech synthesis." In van Santen, Jan T. H.; Sproat, Richard; Olive, Joseph P.; and Hirschberg, Julia. *Progress in Speech Synthesis.* Springer-Verlag, New York, 157-172.

Yngve, Victor H. (1955). "Syntax and the problem of multiple meaning." In Locke, William N. and Booth, A. Donald (Eds.), *Machine translation of languages.* John Wiley & Sons, New York, 208-226.

Zernik, Uri (1990). "Tagging word senses in a corpus: the needle in the haystack revisited." In Jacobs, P. (Ed.), *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." *Proceedings of Intelligent Text and Image Handling, RIAO'91,* Barcelona, Spain, 567-585.

Zipf, George K. (1935). *The Psycho-biology of language: An Introduction to Dynamic Biology,* MIT Press, Cambridge, Massachusetts.

Zipf, George K. (1945). "The meaning-frequency relationship of words." *Journal of General Psychology,* 33, 251-266.