Chapter 3

MAKING SENSE ABOUT SENSE

Nancy Ide and Yorick Wilks
Vassar College; University of Sheffield

Abstract We suggest that the standard fine-grained division of senses and (larger) homographs by a lexicographer for use by a human reader may not be an appropriate goal for the computational WSD task. We argue that the level of sense-discrimination that NLP needs corresponds roughly to homographs, though we discuss psycholinguistic evidence that there are broad sense divisions with some etymological derivation (i.e. non-homographic) that are as distinct for humans as homographic ones and they may be part of the broad class of sense-divisions we seek to identify here. We link this discussion to the observation that major NLP tasks like MT and IR seem not to need independent WSD modules of the sort produced in the research field, even though they are undoubtedly doing WSD by other means. Our conclusion is that WSD should continue to focus on these broad discriminations, at which it can do very well, thereby possibly offering the close-to-100% success that most NLP seemingly requires, with the possible exception of very fine questions of target word choice in MT. This proposal can be seen as reorienting WSD to what it can actually perform at the standard success levels, but we argue that this, rather than some more idealized vision of sense inherited from lexicography, is what humans and machines can reliably discriminate.

1. INTRODUCTION

In Chapter 2, Kilgarriff identifies the source of the WSD “problem” as the attempt to assign one of several possible senses to a particular occurrence of a word in text—in particular, pre-defined sense lists provided in dictionaries and similar lexical resources. He goes on to suggest that the proper assignment of word senses requires a vast amount of lexical, syntactic, and pragmatic knowledge, together with generative procedures that can be exploited for every occurrence—a position reminiscent of the AI community's objections to statistical NLP two decades ago. At the same time, Kilgarriff gives a nod to "the important role" of pre-established lists of word senses for WSD, by which we assume he means that the identification of some limited number of broadly defined senses is useful in language processing applications. He seems to be suggesting, at least obliquely, that while lexicographers and linguists seek to represent word meaning in all its depth and complexity, NLP can provide some useful results by relying on far less. This is exactly right, but it begs the question of how much —or, more to the point, how little— information about word meaning is actually required to do something useful in NLP, given our current capabilities.

Interestingly, although this question should be pivotal for those engaged in the WSD activity, within the NLP community very little progress has been made toward answering it directly. Perhaps this results from aiming too high: for example, the organizers of Senseval-2 state that *"[Senseval's] underlying mission is to develop our understanding of the lexicon and language in general*" (Edmonds & Kilgarriff 2002:289). It is difficult to resist the temptation to answer the hard questions that have been debated by philosophers and linguists for millennia, rather than continue hard practical work within the considerable constraints on our current understanding of lexical semantics. But as Robert Amsler recently pointed out,

> I fear the state of our understanding of theoretical lexical semantics is about where astronomy was 2000 years ago. The theory or even the logical arguments as to what stars in the heavens (or the semantics of words) must be will be debated for years to come without affecting the work of those of us empirically measuring what is observable and predictable (*senseval-discuss*[1], August 27 2004).

Here we take a practical view of WSD, beginning with a reconsideration of the role of lexicographers in word-sense disambiguation as a computational task, as providers of both legacy material (dictionaries) and special test material for competitions like Senseval. We suggest that the standard fine-grained division of senses and (larger) homographs by a lexicographer for use by a human reader may not be an appropriate goal for the computational WSD task, and that the level of sense-discrimination that NLP needs corresponds roughly to homographs. We then consider psycholinguistic evidence that certain etymologically related (i.e., non-homographic) senses that are as distinct for humans as homographic ones which may be part of the broad class of sense-divisions required for NLP. We link this discussion to the observation that major NLP tasks like MT and IR seem not to need independent WSD modules of the sort produced in the research field, even though they are undoubtedly doing WSD by other means. We conclude by recommending that WSD focus on these broad discriminations, thereby reorienting WSD to what it can actually perform at the close-to-100% success rate that most NLP requires.

2. WSD AND THE LEXICOGRAPHERS

It is a truism of recent NLP that one should use machine learning techniques wherever appropriate, which in turn requires that training material be provided by the relevant experts, who will be translators in the case of machine translation (MT), and perhaps lexicographers in the case of WSD. This has been roughly the method pursued by the WSD Senseval competition, but there may be reasons for questioning it, by asking whether lexicographers are in fact the experts that NLP needs for WSD training and expert input.

Even raising this question can sound ungracious, in that there have been many fruitful intellectual and personal collaborations between NLPers and lexicographers, of which Church & Hanks (1990) is perhaps the best known. However, there is a serious point behind the question, and one motivated by the peculiar and indefinite nature of word-sense distinctions, right back to early observations that the sense distinctions you wish to make may depend on your purposes at any given moment (Wilks 1972).

That there is no absolutely right number of senses for a word is conceded by the fact that a publisher like Oxford University Press produces its major English dictionary in at least four sizes (Main, Shorter, Concise, Pocket) with a corresponding reduction in the number of senses for most words. But this is made more complex by the fact the senses in a shorter dictionary may not always be a subset of those in a longer one, but a different conceptualization of a word’s meanings. Hanks (1994) has noted that lexicographers can be distinguished as “lumpers” and “splitters”, where the latter prefer finer sense distinctions and the former prefer larger, more general, senses. And efforts to “map” senses between one dictionary and another, even if general senses are mapped to several finer-grained ones that they supposedly subsume, have shown that the correspondences are not always one-to-one (Ide & Véronis 1990).

However, whatever kind of lexicographer one is dealing with, one cannot be sure that their motivation and expertise is what is required for NLP, because their goal is and must be the explanation of meaning to one who does not know it, and it is not obvious that that is what NLP requires in the way of sense distinctions. This is not to question the line of research on the use of machine readable dictionaries in NLP that began at SDC with Olney et al. (1966) in the Sixties, and which blossomed with the availability of

LDOCE and other learner's dictionaries in the Eighties. It was always a research question whether MRDs would provide large-scale semantics effortlessly in the way optimists hoped. This possibility was questioned as early as (Ide & Véronis 1993) and perhaps it is now fairly clear that, although research with MRDs produced some useful artifacts, such as automatically generated hierarchies (Wilks et al. 1996), and indeed can be said to have started WSD as a subfield and task of NLP, their availability did not produce the revolution that had been hoped for.

None of the above is intended to express skepticism about the expert task of the lexicographer and his intuitions; the issue is whether the product of those intuitions —i.e. a classical dictionary— suits the needs of NLP in semantic analysis. That there has been dissention among lexicographers themselves over their output can be seen from Kilgarriff's published questionings, already touched on above, under titles like "*I don't believe in word senses*" (1997) as well as Hanks' reported musings that a dictionary could be published consisting entirely of examples of use. Any proponent of such a set of examples of use, as a proto-lexicon in itself, has to explain how it performs either the classic explanatory role of a dictionary for the layman, or the needs of the NLP researcher who is perfectly capable of finding his own corpora, which is all a set of usages could amount to. A set of usages may well guide a foreign learner, directly or by analogy, how to use a word, but they cannot show what it means, in the way that definitions do, whatever their other faults. As to the second need, there may well be additional constraints on a well-balanced corpus for experiments, but there is no reason to imagine the set of usages of a word provided by a lexicographer would constitute a balanced corpus, since such balance was not a consideration in the set's construction. Even if it were, there would be no particular reason to trust a lexicographer to balance a corpus for one, rather than a linguist or a computer algorithm.

These doubts about what lexicographers really have to offer NLP have been exacerbated by the realization that all successful WSD has operated at what, in LDOCE terms, we could call the homograph rather than the sense level. If we look at the results obtained by Yarowsky on small word sets (2000), probably the best known WSD results, they have all been at the ["crane"= bird or machine] level—a clear case of an LDOCE homograph. In some of the earliest reported large scale WSD (Cowie et al. 1992) it was clear that much better figures were obtained resolving to the LDOCE homograph, rather than the sense, level. Moreover, homograph distinctions do not require a lexicographer to locate them, since they are basically those that can be found easily in parallel texts in different languages, a point we shall return to below.

3. WSD AND SENSE INVENTORIES

With few exceptions, contemporary automatic WSD assigns sense labels[2] drawn from a pre-defined sense inventory to words in context. If lexicographers' output (i.e., dictionaries) is not a good source of sense inventories useful in NLP, where do we turn? For nearly a decade, the sense inventory used almost exclusively in WSD is the most recent version of WordNet (currently, version 2.0). In the late 1980's and early 90's, prior to the availability of WordNet, sense labels were often drawn from the few electronic dictionaries made available for computational linguistics research (LDOCE, Collins English, etc.). It is interesting to note that both during and before the hey-day of symbolic NLP in the 1970's and early 80's, word senses were more often represented by groups of features of varying kinds than by pre-defined inventories drawn from lexical resources; dictionaries and thesauri sometimes provided the starting point, but were frequently augmented by adding information from other sources, or by hand (for a fuller history, see Ide & Véronis 1998).

The problems for WSD arising from the use of the WordNet inventory are well-known. The most common complaints are that unlike traditional dictionaries, WordNet delineates different senses of a word on the basis of synset membership, and that the resulting distinctions are too fine-grained for WSD. At the same time, the community repeatedly acknowledged that for all its imperfections, WordNet has become a *de facto* standard because it is freely available for research. As a result, the European projects EuroWordNet (Vossen, 1998) and BalkaNet[3] created parallel wordnets for Western and Balkan languages, and several other wordnets are under development[4]. Whether or not calls for the development of better resources to support it are met, WordNet is likely to remain the benchmark sense inventory for WSD for the near future, at least. But the use of WordNet senses *per se* is not the root of the problem. Although it has been argued that using WordNet senses for WSD produces results worse than using senses from traditional dictionaries (Calzolari et al. 2002), the fact remains that pre-defined, enumerated sense lists from any source have proven to be problematic for WSD.

In recent Senseval exercises (see Chapter 4) and the discussions surrounding them, several fixes to what we can call, a bit unfairly, "the WordNet problem" have been proposed and in some cases implemented. The most often-cited obstacle to correct assignment of pre-defined senses concerns granularity: as early as 1993, Kilgarriff showed that human annotators cannot distinguish well between some of the finer-grained senses delineated in LDOCE (Kilgarriff 1993), and this fact has been re-established in numerous studies since then, at a the ceiling of ~80% inter-annotator agreement[5] (for English) reported in recent literature (see, e.g., Edmonds & Kilgarriff 2002). Senseval has addressed this problem by adopting a full or partial "coarse-grained" scoring scheme, where sub-senses are collapsed to their highest parent, and partial credit is given for identifying the parent of the correct sense. Collapsing finer-grained distinctions has been suggested repeatedly in the literature (e.g., Dolan 1994, Chen & Chang 1998, Palmer et al. submitted; see also Chapter 4) as a means to avoid the WordNet problem, but this again begs the question of the level at which to stop collapsing, which has so far not been thoroughly addressed by WSD researchers.

More extensive and radical proposals to improve WordNet have also been put forward, suggesting a major over-hauling of the lexicon by adding information such as selectional preferences and frequency information, as well as refining or improving the information it already contains by simplifying hierarchies, making senses mutually exclusive, deleting bad links and esoteric words, etc. (Hanks, 2003). While the suggested changes were not necessarily made with the aim of improving the resource for NLP specifically, they would certainly help. However, there seems to be little interest in (or perhaps, funding for) implementing changes to WordNet within the NLP community; despite its widespread use in NLP work, one sees very little in the literature about enhancing or extending WordNet to provide a better basis for automatically determining word senses.

There is of course a tradition that rejects the notion of a pre-defined inventory of senses altogether. One version, usually associated with Wierzbicka (1989) and, later, Pustejovsky (1995), is wholly linguistic; another approaches the problem of determining appropriate sense distinctions by using the kinds of information typically exploited in WSD (context, syntactic role, etc.) to identify groups of word occurrences that should, on these grounds, be regarded as representing a distinct sense (e.g., Schütze 1998; see also Chapter 6)[6]. This is a tradition that goes back to Karen Sparck Jones' thesis in the mid-Sixties (1986/1964). While at first glance this approach would seem to be an effort to adapt the answers to the questions rather than the other way around, at the very least it provides some insight into which sense distinctions we can reasonably make given the state of the art. Yet another approach uses cross-lingual correspondences to determine appropriate sense distinctions. Brown et al. (1990*)* and Dagan & Itai (1994) use translation equivalents as "sense tags" in parallel and comparable corpora rather than pre-defined senses. More recent work along this line extends to the claim that, for the purposes of NLP, the different senses of a word could be determined by considering only those distinctions that are lexicalized cross-linguistically (Dagan & Itai, 1994, Resnik & Yarowsky 1997a or b). Given that many ambiguities are preserved across languages, this approach demands examining translation equivalents in parallel texts from multiple languages, possibly languages spanning the various broad linguistic families to overcome arbitrary effects of joint inheritance. This idea was pursued in a series of studies (Ide 1998; Ide et al. 2001, 2002), where word occurrences in an English text were clustered based on their translation equivalents in parallel texts in seven languages from the Germanic, Romance, Slavic, and Finno-Ugric language families. The results showed that clusters produced automatically and based on translation equivalents agreed with clusters (i.e., groupings of occurrences deemed to be used in the same sense) produced by four human annotators at a level slightly below that of agreement among the annotators themselves (74% vs. 79%), but the clustering algorithm performed well enough to be considered a viable means to delineate senses. Other recent studies exploring this idea include Dyvik (1998, 2004), Resnik & Yarowsky (2000), Diab & Resnik (2002), Ng et al. (2003), and Tufis et al. (2004), with similar results.

These "data-driven" approaches to determining word senses are philosophically in the good company of Halliday, Sinclair, Harris, and other major 20th century linguists, but on a practical level they seem unlikely to be used in NLP applications in the near future, if at all. The primary problem is that their implementation to produce a "full" sense inventory would require massive amounts of data, and even continuous re-computation as new data becomes available and languages evolve. Furthermore, it is not even clear that a usable, independent sense "list" could be produced by these means: for example, how would senses in such a list be labeled/distinguished so as to be meaningfully understood and used, without resorting to some sort of definition, such as one would find in a traditional dictionary? If cross-lingual distinctions are used as a basis, do we include any distinction that any language makes, or only the ones most or all languages make? For example, Romanian and Estonian have a special word for "*back of the head*", whereas in English the word *head* is generally used without further specification. In the phrase "*behind [one's] head*", *head* is translated as *kuklasse* (nominative: *kukal*) in Estonian and *ceafă* in Romanian, whereas in the phrase "*above [one's] head*", both Estonian and Romanian use a more general word for *head* (*pea* and *cap*, respectively) that corresponds to the English equivalent. Cross-linguistic data, then, suggests two "senses" to distinguish the concept of the back of the head from the head in general, but it is not clear whether the distinction should be made in sense labeling an English text, or if only the more general concept should be used even if the language being labeled makes the distinction (with language-specific refinements, as in EuroWordNet—see Peters et al. 1998).

Overall, then, no suitable sense inventory for general-purpose WSD has yet been identified or created. However, despite the questions noted above, the use of cross-lingual information to determine an inventory of sense distinctions useful for NLP seems to offer the best potential for developing a meaningful inventory for NLP applications. We return to this point later, in section 5.

4. NLP APPLICATIONS AND WSD

In his survey of WSD in NLP applications (Chapter 11), Resnik rightly points out that there is typically no explicit WSD phase in well-established applications such as monolingual information retrieval (IR) and machine translation (MT). MT remains the crucial and original NLP task, not just because of its age but because any NLP theory can almost certainly be expressed and tested in MT terms; moreover MT has undoubted and verifiable evaluation standards, in that it remains a task that can be evaluated outside any theory, simply because many people know what a translation is without any knowledge whatever of NLP or linguistics. That cannot be said of many classic NLP tasks, which require a great deal of skill and experience to evaluate, including WSD. Given that seniority of MT, we also know that tradition asserts firmly that WSD was one of the reasons early MT was not more successful, and this has been used as the justification for WSD since its inception: it would help MT. What we have to discuss and explain here is why the undoubted successes of WSD at the 95% level seem not to have so far materially assisted MT.

Martin Kay wrote somewhere long ago that, even if all the individual problems associated with MT were solved, including WSD and syntactic analysis, that fact alone might not raise the success level of MT substantially. The remark was in a paper that advocated human-aided MT, on the ground that pure MT seemed unlikely to succeed, a prediction that has turned out to be false. However, the remark about MT components now seems prescient. And again, it is worth asking why that is, if it is.

To answer it, we might look at the history of IR, a discipline of about the same maturity as MT. From its beginning, there have been those who argued that IR must need some WSD function to reduce the ambiguity of words in queries. One remembers here Bruce Croft’s dictum that, for any IR technique, there is some document collection for which it will improve retrieval. More seriously, Vossen (2001) and Stevenson & Clough (2004) have recently shown that WSD does seem to have a real role in cross-language IR. Nonetheless, the current prevailing view is that explicit WSD must be close to 100% accurate to improve monolingual IR (Krovetz & Croft 1992, Sanderson 1994), and therefore, for the long standard queries used in evaluations (as opposed to the short ambiguous queries sent to search engines), separate WSD modules seem to make little difference; it has even been argued that partially erroneous sense assignments from explicit WSD can degrade retrieval results (Voorhees 1999). This is certainly because the operation of an IR system, using as it normally does the overall context defined by the query, seems to perform WSD by indirect methods. So, the 100 terms in a classic (as in the U.S. TREC competition) query will effectively define a domain, and co-occurrence functions used in the retrieval ensure that associations of "inappropriate" senses of words in the query are eliminated in that process.

As for MT, it is a fact that most working MT systems, from SYSTRAN onwards to the present, do not have separate and identifiable WSD components, although they undoubtedly do a great deal of WSD somewhere. Does this suggest that some local functions are in fact doing WSD without being so named? Two different examples of systems suggest that this may be the case. Wilks & Stevenson (1998) have shown that, if the lexicon was arranged appropriately, a simple POS tagger could give 90% WSD. Appropriate lexical organization here meant the sort given in LDOCE where senses are grouped under main homographs and the homograph/sense clusters have all their members with a single part of speech. It is this last fact that allows a POS tagger to do so much sense discrimination at little or no computational cost: for instance if *bark* is tagged as a verb, then we know its sense is that of an animal (possibly human) vocalizing vociferously and need not concern ourselves at all with the ambiguity of that word as a noun.

This result is a serendipitous side effect of LDOCE's particular form of organization, but it does suggest something deeper about the extent to which sense distinction is not independent of part-of-speech distinction and how the latter can aid the discrimination of the former—i.e. without explicit WSD. Another example, quite different but pointing the same moral, is the generation component of the CANDIDE statistical MT system (Brown et al. 1990) where *prendre* has as its most frequent equivalent in bilingual texts the verb *take*. Yet, when translating "*prendre une decision*" CANDIDE is able to generate "*make a decision*" which is more common in US English, even though "*take a decision*" is not wrong. It does this because of the interaction of trigrams in the target language and bilingual associations. One could say that *prendre* is being disambiguated here but without its English alternatives ever being explicitly considered or compared by the system. The correct output is simply a by-product of the interaction of two very general statistical components.

In general, then, explicit WSD—as implemented in stand-alone systems such as those involved in the Senseval competitions—does not seem to play a role in the most prominent NLP applications.[7] We again have to ask ourselves, why not?

Before answering this question, it is useful to turn it around and ask, why is WSD generally treated as if it is an isolatable language processing step? The reasons would seem to be primarily historical. A "modular" view of language processing was firmly established in the mid-20th century by semioticians and structural linguists, who developed cognitive models that describe language understanding as an aggregative processing of various levels of information (syntax/semantics/pragmatics for the semioticians, morpho-phonological/syntactic/lexico-semantic for the structural linguists). This modular view was taken up by the earliest computational linguists, who treated the process of language understanding as a modular system of sub-systems that could be modeled computationally, and it has remained dominant (abetted by cognitive psychology and neuro-science) to this day. It is apparent in the design of "comprehensive" language processing systems, which invariably include multiple modules devoted to isolatable analytic steps, and it informed the "pipeline" approach to linguistic annotation introduced in the mid-90's (Ide & Véronis 1994) that has been implemented in major annotation systems[8] since then. In keeping with the modular approach, it is natural to treat disambiguation in the same way morpho-syntax and syntax were treated in the past: as a step in the language processing pipeline for which independent systems can be developed and tested, and which can then be integrated into more general language processing systems. As a result, for over 40 years considerable research activity has been devoted to the development and evaluation of stand-alone WSD systems, with techniques spanning the use of semantic and neural networks, hand-crafting of complex rules and semantic feature sets, exploitation of knowledge resources such as dictionaries, thesauri, and lexicons like WordNet, as well as the development of sophisticated statistical and machine learning techniques—despite the fact that these systems are rarely used as modules in language processing applications.

The fact that different applications require different *degrees* of disambiguation is rarely considered in discussions of the application needs for WSD. In fact, IR and MT provide what may be the opposite ends of a continuum of WSD needs: IR typically demands "shallow" WSD, while MT may require more disambiguation precision to generate a translation that sounds more or less natural in the target language.[9] In fact, it appears that applications that need deeper linguistic analysis in general, may need finer-grained disambiguation. So, it follows that MT has exploited information gleaned from its more sophisticated linguistic processing to achieve more precise disambiguation, rather than turning to stand-alone WSD. IR, on the other hand, is virtually the only application that has seriously explored the use of stand-alone WSD, since the kind and level of disambiguation needed there is precisely what current WSD systems are good at.

The question is therefore not whether NLP applications such as IR and MT *need* WSD (they do), but rather, what degree of disambiguation they need and whether or not pre-defined sense inventories can provide it. We turn to this question in the next section.

5. WHAT LEVEL OF SENSE DISTINCTIONS DO WE NEED FOR NLP, IF ANY?

Dagan & Itai (1994) have long argued that sense distinctions roughly at the homograph level, where *crane* is a bird or a machine for lifting, are the ones actually used for most WSD and therefore those needed, by definition, for NLP. If we look a little more widely in the speculative literature on word sense, we see that the homograph-as-basic view has more support than at first appears: Wierzbicka (1989) is sometimes taken as having argued that there are no word senses, but only a basic sense for each word, a position held by Ruhl (1989) and, much earlier, by Antal (1965). However, Wierzbicka's position is more complex, in that she accepts homographs—what are often argued to be different words by linguists, and only masquerading, as it were, as the same word. One can see her in the tradition of those interested in the way a word extends its sense with time, while retaining a strong semantic link to its origin (which is precisely what homographic distinctions lack). In the AI/NLP world, this tradition has manifested itself as those who either want more compacted lexicons (e.g., Gazdar's DATR, Pustejovsky) or are interested in rules, or knowledge functions, by which sense lexicons extend (e.g., Givon, Wilks, Briscoe, Nirenburg —see Wilks & Catizone (2002) for a comparison of this class of systems). A similar approach is advocated by some linguists/lexicographers; for example, Nunberg (1979) argued that distinct senses should not be represented in the lexicon, but rather that pragmatic principles should be used to derive related senses from others.[10] This view is also evident in the psycho-linguistics literature: one theory of the mental lexicon holds that only a "core" meaning of a word is stored in memory, and polysemous extensions are computed on the fly from contextual features, using pragmatics and plausible reasoning (see e.g., Anderson & Ortony 1975, Caramazza & Grober, 1976).

The former group, as with Wierzbicka, tend to deny there is an extensive set of senses, even though there appears to be one in many dictionaries, while the latter group claim that some mechanism could recapitulate the apparent variety that time and usage have produced. These two variant positions may not be ultimately distinct, and can be parodied by the example "*She sat on her bicycle and rode away*" where, if a bicycle, has, say, 150 distinct parts one could perhaps argue that *bicycle* in that sentence is 150 ways ambiguous and needs resolving to *saddle* or *seat*. However, that position is obviously absurd; it would be far better to say that the word is simply vague, and that it is AI, knowledge bases and reasoning that should further resolve it, if that ever proved necessary, and not NLP or linguistics. To justify this, one could fall back on some form of Dagan's case: namely that every language will have a word for a bicycle and for each of its parts, but it is hard to imagine a language that would force the specification of a particular part in the example above —though, as we saw, in some specific and limited cases like the Romanian/Estonian *head*, such precision is forced.

In fact, homographs as strictly defined—i.e., etymologically unrelated words which through historical accident have the same "name", like the senses of *bank* and *calf* – are certainly not enough for WSD, since there are many instances where etymologically related senses are as distinct as homographs for most people. Take, for example, the word *paper:* in dictionaries that separate entries by homographs (most notoriously, the *Oxford English Dictionary*), the senses of *paper* that refer to sheets of material made from wood (as a "*sheet of paper*") and a newspaper (as a "*daily paper*") appear in the same entry and are therefore etymologically related. Other examples include words like *nail* (a finger nail vs. the metal object one drives with a hammer), *shower* (a rain shower vs. the stall in which one bathes), etc.[11] For such words, certain senses are as distinguishable as homographs, a fact that has been borne out in psycholinguistic experiments. For example, Klein & Murphy (2001) conducted experiments in which subjects were primed with a word in context in one sense and then presented with the same word in another context, reaction time for homographs was no less than reaction time for grossly polysemous words (e.g., *daily paper* vs. *wrapping paper*). This suggests that some senses of an ambiguous word, although not unrelated etymologically, are as distinct in the mind of the hearer as homographs, which in turn suggests that they may be just as relevant for NLP.

Some linguists (e.g., Lakoff 1987, Heine 1992, Malt et al. 1999) have proposed that polysemy develops via a chain of novel extensions to previously known senses, each building on its predecessors. This idea, and computational methods for it surveyed and discussed in Wilks & Catizone (2002), follows nicely on from proposals for the generative lexicon proposed by Pustejovsky (1995) and others, but adds the notion that at some point, senses diverge enough to deserve independent representation in the lexicon (either computational or mental). The problem, of course, is in identifying the point at which two senses become distinct enough to warrant separation for the purposes of NLP (or, for that matter, in dictionaries and the mental lexicon). Klein & Murphy (2002) extended their earlier study to involve more closely related senses, for instance senses for *paper* in WordNet such as sense 3 (newspaper as publication) vs. sense 7 (newspaper as a physical object) in order to address this question. Their results in this second slate of experiments lead them to several conclusions that have ramifications for automatic sense disambiguation. First, their results suggest that some of the different senses of polysemous[12] words are stored independently in memory, supporting the notion that some etymologically related words are as distinguishable as homographs. Second, they experimented with different categorical relations among senses similar to those outlined by Pustejovsky (1995), and determined that different senses of a polysemous word do not seem to correspond to a unified taxonomic, thematic, or *ad hoc* category, but rather that the types of relationships among senses are more or less random and unpredictable. This is bad news for proponents of the generative lexicon, because it means that rule sets for the online derivation of different senses of a given word cannot be determined in any systematic way. Furthermore, Klein and Murphy draw the conclusion that representation of a "core" sense (similar to a homograph) coupled with procedures to generate more refined meanings is inconsistent with their results; rather, they suggest that relational derivation of senses happens historically and/or during language acquisition, and once senses become sufficiently distinct, they are thereafter stored separately in the mental lexicon. This leads them to suggest a processing model for word meaning that they call "radical under-specification", in which a minimal, neutral placeholder is activated when a polysemous word is encountered (e.g., "*something called paper*" when "*The paper…*" is seen) and refined by later context.

Klein and Murphy's work, along with that of other psycholinguists, has ramifications for WSD. First of all, it suggests that there are some etymologically related senses that should be regarded as separate as homographs and could provide insight into which senses belong in this category. Unfortunately, the aim of Klein and Murphy's experiments is to provide evidence for separate representation of etymologically-related senses, rather than to identify which senses of a given word fall into this category and which do not. Therefore, their analysis provides no information concerning which senses might be regarded as the same and therefore collapsed into one, homograph-level sense for the purposes of WSD. This is also the case in other recent psycho-linguistic studies concerning word meaning (e.g., Rodd et al. 2002, 2004), which use pre-defined sense inventories as a point of departure without questioning the distinctness among multiple senses of the same homograph. Nonetheless, it is easy to imagine extending the methods and criteria used in psycho-linguistic studies of word meaning to determine the distinctness—in terms of the mental lexicon—of senses below the level of the homograph.

On the other hand, it is certainly possible that sufficiently separate senses can be identified using multi-lingual criteria—i.e., by identifying senses of the same homograph that have different translations in some significant number of other languages—as discussed in section 3. For example, the two senses of *paper* cited above are translated in French as *journal* and *papier,* respectively; similarly, the two etymologically-related senses of *nail* (fingernail and the metal object that one hammers) are translated as *ongle* and *clou*. At the same time, there is a danger in relying on cross-lingualism the basis of sense, since the same historical processes of sense "chaining" (Cruse 1986, Lakoff 1987) can occur in different languages. For example, the English *wing* and its equivalent *ala* in Italian have extended their original sense in the same way, from birds, to airplanes, to buildings and even to soccer positions. The Italian-English cross-corpus correlations of the two words would lead to the conclusion that both have a single sense, when in fact they have wide sense deviations approaching the homographic.

Another source of information concerning relevant sense distinctions is domain, as discussed in Chapter 10. If senses of a given word are distinguished by their use in particular domains, this could offer evidence that they are distinguishable at the homograph-like level. At the same time, senses that are *not* distinguished by domain—take, for example, the sense of *bank* as a financial institution vs. its sense as a building that houses a financial institution—might, for all practical purposes, be regarded as a single, homograph-level sense.

The psycholinguistic evidence also suggests that different kinds of evidence are needed to distinguish senses for different words. Experiments with a "multi-engine" WSD system (Stevenson & Wilks 1999) have already showed that the sense-discrimination of particular word-classes —usually part-of-speech classes like nouns or verbs— tended overwhelmingly to be carried out by a particular "engine" using a particular resource: for instance verbs and adjectives, but not nouns, were discriminated to a great degree by the selectional preferences loaded in from LDOCE, while the nouns tended to be discriminated by a combination of LDOCE definitions and thesaurus classes. None of this should be surprising, but it was confirmed strikingly by an overall machine learning algorithm which, in effect, decided for each word, which engine/resource best discriminated it. A further, less trivial, inference to be drawn from this result is that the different semantic resources used in WSD (thesauri, definitions, collocations etc.) are not, as some have suspected, merely different notations for the same semantic facts. Klein and Murphy's assertion that senses of a polysemous word are not unified by a common categorical relation suggests that these processing differences may extend to words of the same part-of-speech category as well, and even further, that the degree and nature of these relations depend not only on the word in question, but often varies for each pair of senses for that word. This notion could be taken to lead to a position similar to Kilgarriff's, that a vast array of knowledge about each word (similar to his "word sketches"—see Kilgarriff and Tugwell, 2001) is required for sense disambiguation; but at least for the purposes of NLP, another interpretation is possible.

If we accept that new senses of a given word develop historically through various relations, then we can also assume, based on the psycholinguistic evidence, that at some point a sense becomes distinct enough to be represented separately in the mental lexicon[13] and becomes as distinguishable from other senses of the same word as homographs are from one another. We would argue that these senses are discernable from context to the same degree as homographs, and therefore, WSD systems can achieve the same high degree of success in detecting them as for homographs. It is this level of sense distinction that Amsler referred to as "observable and predictable" in his comments to the Senseval discussion list, and, in our view, this is the only kind of sense distinction that stand-alone WSD should be concerned with. Senses that have not achieved this degree of distinction demand greater knowledge and resources to identify reliably, but in applications like MT that may need finer sense granularity, the results of deeper linguistic processing and knowledge is readily available to assist the disambiguation process.

To summarize, NLP applications, when they need WSD, seem to need homograph-level disambiguation, involving those senses that psycholinguists see as represented separately in the mental lexicon, are lexicalized cross-linguistically, or are domain-dependent. Finer-grained distinctions are rarely needed, and when they are, more robust and different kinds of processing are required. Lexicographers will necessarily continue to be concerned with the latter kind of sense distinction, as they must be; but for the purposes of NLP, work on the problem of WSD should focus on the broader distinctions that can be determined reliably from context.

6. WHAT NOW FOR WSD?

At present, WSD work is at a crossroads: systems have hit a reported ceiling of 70%+ accuracy (Edmonds & Kilgarriff 2002)[14], the source and kinds of sense inventories that should be used in WSD work is an issue of continued debate, and the usefulness of stand-alone WSD systems for current NLP applications is questionable.

The WSD community has grappled for years with the issue of sense distinctions because of its reliance on pre-defined sense inventories provided in mono-lingual dictionaries and similar reference materials. Such inventories are typically organized according to lexicographical principles, such as grouping senses on the basis of etymology and part of speech. Senses grouped according to these criteria are usually organized, either explicitly or implicitly, by frequency of use, and there is no other indication of the degree of distinguishability among them. Although WordNet is not the best example of a traditional dictionary, its organization is fairly typical; for example, if we stay with the *paper* example, WordNet gives us the following:

1. paper - (a material made of cellulose pulp derived mainly from wood or rags or certain grasses)
2. composition, paper, report, theme - (an essay (especially one written as an assignment); "he got an A on his composition")
3. newspaper, paper - (a daily or weekly publication on folded sheets; contains news and articles and advertisements; "he read his newspaper at breakfast")
4. paper - (a scholarly article describing the results of observations or stating hypotheses; "he has written many scientific papers")
5. paper - (medium for written communication; "the notion of an office running without paper is absurd")
6. newspaper, paper, newspaper publisher - (a business firm that publishes newspapers; "Murdoch owns many newspapers")
7. newspaper, paper - (a newspaper as a physical object; "when it began to rain he covered his head with a newspaper")

Clearly, sense 1 is far more distinguishable from sense 3 than sense 6 is, but in WSD experiments senses like these are usually considered to be distinct. A more intuitive list might collapse senses 1 and 5; 2 and 4; and 3, 6, and 7; yielding something like:

1. paper - material
2. paper - composition, article
3. paper - newspaper, publication, publisher

This is given as an example and not a scientifically determined set of senses, based in part on the fact that some other languages lexicalize these broad distinctions differently (e.g., in French, as *papier, article,* and *journal,* respectively). The WSD community has recently begun discussing “collapsing” senses that are more related (see Palmer et al. submitted, and also Chapter 4)—or at least, senses that WSD systems have difficulty distinguishing. This goes in the right direction, but it seems more appropriate to adopt a “top-down” rather than a “bottom-up” approach: that is, the starting point for WSD should be a bi-polar distinction, between homograph-level distinctions and “everything else”. The psycho-linguistic evidence supports this approach, by identifying senses that are, in psychological terms, represented separately in the mental lexicon; and it is in fact also indicated by the performance of current WSD systems, which show clearly superior results for disambiguating homographs—and, we would argue, would do so for all homograph-level distinctions if they were clearly identified.

In fact, there are good reasons to suggest that WSD should focus on a top-down approach to sense *distinction* rather than sense *determination*. Klein and Murphy's notion of "radical under-specification" implies such a model for human processing, by stipulating that disambiguation starts with only the most general of concepts when an ambiguous word is encountered, and proceeds by refining meaning as additional context is provided. For example, when "*the paper…*" is seen or heard, we can imagine that if the remainder of the sentence is "…was picked up at the corner newsstand", the reader will make the homograph-level distinction and determine that here, *paper* refers to a newspaper. More importantly, only the homograph-level distinction needs to be made: no choice between the "newspaper-as-physical-object" and "information source" senses of *paper* (senses 3 and 7 in the WordNet list above) is necessary—that is, there is no need to choose one of these senses and explicitly eliminate the other. Even if the discourse emphasizes one of the two possibilities, both are likely to exist in the mind as a single encompassing concept that has not (yet) been torn apart. We can hypothesize, then, that sense "disambiguation" is really a process of step-wise sense refinement that progressively distinguishes "sub-senses" as needed for understanding. We argue that there is rarely a need to make distinctions below the homograph-like level for understanding, human or automated; and in the unusual circumstance where it becomes necessary to explicitly throw one of the sub-senses away, we can expect there to be contextual clues that will enable both humans and machines to do so.

Based on all of this, our recommendations for WSD work in the near future are, first, to focus attention on identifying the homograph-level sense distinctions that are necessary for NLP. The obvious sources of this information are cross-lingual and psycho-linguistic evidence, together with domain information. Cross-lingual evidence provides inventory-free distinctions based solely on translation equivalents, but will demand further work to acquire sufficient parallel data in order to overcome problems such as parallel sense chaining (as mentioned in the previous section) and mono-lingual synonymy. It will also require determining the number and types (in terms of representatives of different language families, etc.) of languages needed to ensure that all relevant distinctions are captured. At the same time, some threshold must be determined so that fine distinctions made by one or only a few languages, and/or which are highly culture dependent (e.g., different ways to greet a person depending on one's relation to that person, or the time of day), are not included for the general WSD task (although they certainly need to be retained for the purposes of MT).

To gather psycho-linguistic evidence, further experimentation will be required, since research in this area has been focused on developing psychological models of language processing and has not directly addressed the problem of identifying those senses that are distinct enough to warrant, in psychological terms, a separate representation in the mental lexicon. Also, psycho-linguistic experiments currently rely on pre-defined sense inventories from traditional dictionaries, thereby providing sense distinctions *a priori* rather than seeking to determine which distinctions are sufficiently independent. Collaboration between the WSD and psycho-linguistic communities could enable experimentation with "inventory-free" distinctions, and provide valuable results for WSD as well as theories of the mental lexicon.

Another source of information about relevant sense distinctions is corpus evidence—that is, senses of a given word that "fall out" due to differences in their patterns of usage. For example, the Corpus Pattern Analysis (CPA) project (Hanks & Pustejovsky, to appear) is currently compiling a lexicon of verb patterns based on randomly chosen samples from the British National Corpus, consisting of syntactic frame information coupled with semantic types and roles. While the information in such a lexicon may provide a compendium of sense-distinguishing characteristics that are detectable using relatively sophisticated NLP techniques and accompanying resources, it does not follow that the granularity of the identified senses is what is needed for most NLP applications. It will, however, certainly be informative to compare sense distinctions identified by projects such as CPA with those identified on the basis of domain, cross-lingual evidence, and psycho-linguistic experiments.

Our second recommendation is to shift the focus of work on WSD to enhancing stand-alone systems in order to achieve near-100% accuracy for homograph-level distinctions. As we have argued above, disambiguation at the homograph-level is sufficient for IR, MT, and other NLP applications, and robust WSD systems that deliver accurate results at this level are potentially more useful for NLP applications than existing systems have so far proved to be. For example, Sanderson (1994) argued against the use of existing WSD systems for IR based on his observation that inaccurate WSD can negatively impact results. Likely, other NLP applications such as MT could profit from accurate WSD at this level as well.

As a final note, we point out that while concern with sense distinctions at levels finer than the homograph may not be appropriate at this point for WSD research aimed at contributing to NLP applications, it is still a matter of interest for lexicographers and certainly valuable to "develop our understanding of the lexicon and language in general". It may also be relevant for MT systems that seek to generate natural-sounding prose —for example, several alternative translations for *recur* exist in French (*se reproduire, revenir, se retrouver, réapparaître, se représenter*); to generate a natural-sounding translation, additional knowledge and/or reasoning may be applied to determine the nature of the verb's agent (*l'événement se reproduit, l'idée se retrouve, la maladie réapparaît, le problème se represent*)—see, for example, Edmonds & Hirst (2002), who have explored means to choose among near-synonyms in order to produce natural-sounding prose. This type of lexical refinement, however, is primarily the work of lexicography, AI, and knowledge engineering, and should be left to specialized modules outside the scope of mainstream WSD.

7. CONCLUSION

Our conclusions could seem both pessimistic and optimistic for WSD. They are optimistic in that, if something on the order of homograph distinctions are the level of WSD we need for NLP, then we have pretty good techniques for achieving that; and the data may be relatively easily obtained from multilingual corpora, and that we do not really need the expertise of lexicographers to help us in that task. They may also be considered pessimistic, in that it may be that many NLP systems do not require a separate WSD module at the level of granularity attempted by current systems, and that therefore much of the WSD work of the last decade has been wasted in presenting it as a separate task —however useful it has been as a hothouse of techniques. Given that evaluating WSD, as a free-standing, independent task has been so expensive and time-consuming, this discovery may be a relief all round. But this does not mean that work on stand-alone WSD is finished, by any means. There still remains the considerable task of identifying the homograph-level distinctions that are useful for NLP, since they are not explicitly identified as such in any existing resource. The WSD community therefore has work to do, and should now turn itself to the task.

NOTES

[1] http://listserv.hum.gu.se/mailman/listinfo/senseval-discuss

[2] We include here not only sense labels derived from sense inventories such as WordNet, traditional dictionaries, and thesauri, but also "concept labels" such as EuroWordNet's inter-lingual index (ILI), "semantic annotations" as used in, say, Information Extraction systems, as well as codings used in interlingual MT systems.

[3] http://www.ceid.upatras.gr/Balkanet/publications.htm

[4] See http://www.globalwordnet.org/gwa/wordnet_table.htm

5. A problem we do not address but which must occur to many readers is that, in the case of WSD in particular, claimed and tested success rates in the 90%+ range are strikingly higher than the inter-annotator agreement level of 80%+, and to some this is a paradox. The answer may simply be that the better machine learning systems in fact simulate the better, more sensitive, discriminators and that the low agreement figure reflects the relative difficulty of the task, rather than some inherent level of vagueness in the material. We all know some people are better lexicographers than others, and this is not a "democratic" task like speaking a language. No other explanation seems to fit the experimental data.
6. The applicability of this approach is not limited to WSD: Hanks (2000) outlines a method by which lexicographers can determine sense distinctions for inclusion in traditional dictionaries by iteratively clustering concordance lines judged to represent the use of a given word in the same sense.
7. See Chapter 11 for a comprehensive review of the role of WSD in IR and MT.
8. For example, MULTEXT (Ide & Véronis, 1994), LT XML (McKelvie et al. 1998), GATE (Cunningham 2002), and ATLAS (Bird et al. 2000).
9. In fact, it is almost certainly the case that the degree of disambiguation required for MT depends on the word in question (more ambiguous words, especially those often used metaphorically such as *hard* and *run*, may demand more analysis to disambiguate) as well as the target language and its similarity to the source language, both etymologically and structurally.
10. This approach is in contrast to that of other lexicologists such as Zgusta (1971), who argue for representing each distinguishable sense.
11. A list of 175 polysemous words of this type and their most common different senses is given in (Durkin & Manning 1989).
12. Klein and Murphy's conception of polysemy is defined primarily through examples, and does not seem to rely on a pre-defined sense inventory (although in their 2001 article they mention the use of the *OED* for determining homographs).
13. Note that there is no psycholinguistic evidence that the links among derived senses are themselves stored.
14. Of course, statistics such as these depend on the assumption that the criteria used—in this case, identification of WordNet sense distinctions—are good ones.

REFERENCES

Anderson, Richard C. & Andrew Ortony. 1975. ``On Putting Apples into Bottles—A Problem of Polysemy". Cognitive Psychology, 7, 167–180.

Antal, László. 1965. Content, Meaning and Understanding. The Hague: Mouton.

Bird, Steven, David Day, John Garofolo, John Henderson, Christophe Laprun & Mark Liberman. 2000. ``ATLAS: A Flexible and Extensible Architecture for Linguistic Annotation". Proceedings of the Second International Language Resources and Evaluation Conference (LREC), May, 2000, Athens, Greece, 1699-1706. Paris: European Language Resources Association.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J.Della Pietra, Frederick Jelinek, John Lafferty, Robert Mercer & Paul Roosin. 1990. ``A Statistical Approach to Machine Translation". Computational Linguistics, 16:2, 79-85.

Calzolari, Nicoletta, Claudia Soria, Francesca Bertagna, & Francesco Barsotti. 2002. ``Evaluating lexical resources using Senseval". In Philip Edmonds & Adam Kilgariff (eds.), Special Issue on Evaluating Word Sense Disambiguation Systems, Natural Language Engineering, Cambridge: Cambridge Univ. Press, 8:4, 375-390.

Caramazza, Alfonso & Ellen Grober. 1976. ``Polysemy and the Structure of the Subjective Lexicon". Georgetown University Round Table on Languages and Linguistics. Semantics: Theory and application, ed. by C. Rameh. Washington, DC: Georgetown Univ. Press, 181–206.

Chen, Jen Nan & Jason S. Chang. 1998. ``Topical clustering of MRD senses based on information retrieval techniques". In Nancy Ide & Jean Véronis (eds.), Special Issue on Word Sense Disambiguation, Computational Linguistics, 24:1, 61-95.

Church, Kenneth W. & Patrick Hanks. 1990. ``Word Association Norms, Mutual Information, and Lexicography". Computational Linguistics, 16 :1, 22-29.

Cowie, James, Joe Guthrie & Louise Guthrie. 1992. ``Lexical disambiguation using simulated annealing. Proceedings of the 14th International Conference on Computational Linguistics (COLING92), Nantes, France, 359-365.

Cruse, David. 1986. Lexical semantics. Cambridge: Cambridge University Press.

Cunningham, H.amish. 2002. ``GATE, A General Architecture for Text Engineering". Computers and the Humanities, 36:2, 223-254.

Dagan, Ido & Alon Itai. 1994. ``Word Sense Disambiguation Using a Second Language Monolingual Corpus". Computational Linguistics, 20:4, 563-596.

Diab, Mona & Philip Resnik. 2002. ``An Unsupervised Method for Word Sense Tagging using Parallel Corpora". Proceedings of the 40th Meeting of the Association for Computational Linguistics (ACL-02), Philadelphia.

Dolan, William. 1994. ``Word Sense Ambiguation: Clustering Related Senses". Proceedings□of COLING'94, 712-716.

Durkin, Kevin & Jocelyn Manning. 1989. ``Polysemy and the Subjective Lexicon: Semantic Relatedness and the Salience of Intraword Senses". Journal of Psycholinguistic Research, 18,577–612.

Dyvik, Helge. 1998. ``Translations as Semantic Mirrors". Proceedings of the Workshop on Multilinguality in the Lexicon II, The 13th biennial European Conference on Artificial Intelligence (ECAI 98), Brighton, UK, 24-44.

Dyvik, Helge. 2004. ``Translations as semantic mirrors: From parallel corpus to Wordnet". Language and Computers, 1, 311-326.

Edmonds, Philip & Graeme Hirst. 2002. ``Near-synonymy and lexical choice. Computational Linguistics, 28:2), 105-144.

Edmonds, Philip & Adam Kilgarriff. 2002. ``Introduction to the special issue on evaluating word sense disambiguation systems." Journal of Natural Language Engineering, 8(4), 279-291.

Hanks, Patrick. 1994. personal communication.

Hanks, Patrick. 2000. ``Do Word Meanings Exist?". Computers and the Humanities, 34:2, 205-215.

Hanks, Patrick. 2003. ``WordNet: What is to be Done?". Panel presentation at Prague Workshop □on Lexico-Semantic Classification and Tagging Linguistic and Knowledge-Based Foundations, Existing Schemes and Taxonomies, and Possible Applications, December 8-9, Prague. Available at http://ckl.mff.cuni.cz/events/lexsem/hanks-panel.pdf.

Hanks, Patrick & James Pustejovsky. To appear. ``A Pattern Dictionary for Natural Language Processing". Revue Française de Langue Appliquée.

Heine, Bernd. 1992. ``Grammaticalization Chains". Studies in Language, 16, 335–368.

Ide, Nancy. 1998. ``Cross-lingual Sense Determination: Can it work?". Computers and the Humanities, 34:1-2, 223-34.

Ide, Nancy, Tomaz Erjavec & Dan Tufiş. 2001. ``Automatic Sense Tagging Using Parallel Corpora". Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, pages 212-219.

Ide, Nancy, Tomaz Erjavec & Dan Tufiş. 2002. ``Sense Discrimination with Parallel Corpora". Proceedings of the SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, 56-60.

Ide, Nancy & Jean Véronis. 1990. ``Mapping Dictionaries: A Spreading Activation Approach". Proceedings of the 6th Annual Conference of the Centre for the New Oxford English Dictionary, Waterloo, Ontario, 52-64.

Ide, Nancy & Jean Véronis. 1993. ``Extracting Knowledge Bases from Machine-Readable Dictionaries : Have We Wasted Our Time?". Proceedings of the First International Conference on Building and Sharing of Very Large-Scale Knowledge Bases (KB&KS'93), Tokyo, Japan, 257-266.

Ide, Nancy & Jean Véronis. 1994. ``MULTEXT: Multilingual Text Tools and Corpora. Proceedings of the 15th International Conference on Computational Linguistics, COLING'94, Kyoto, Japan, 588-92.

Ide, Nancy & Jean Véronis. 1998. ``Word Sense Disambiguation: The State of the Art". Computational Linguistics, 24:1, 1-40.

Kilgarriff, Adam. 1993. ``Dictionary word sense distinctions: An enquiry into their nature". Computers and the Humanities, 26:356-387

Kilgarriff, Adam. 1997. ``I Don't Believe in Word Senses". Computers and the Humanities. 31(2. `` 91-113.

Kilgarriff, Adam & David Tugwell. 2001. ``WASP-Bench: an MT Lexicographers' Workstation Supporting State-of-the-art Lexical Disambiguation". Proceedings of MT Summit VII, Santiago de Compostela, 187-190.

Klein, Devorah & Gregory Murphy. 2001. ``The Representation of Polysemous Words". Journal of Memory and Language 45, 259–82.

Klein, Devorah & Gregory Murphy. 2002. ``Paper Has Been My Ruin: Conceptual Relations of Polysemous Words:. Journal of Memory and Language 47, 548-70.

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

Lakoff, George. 1987. Women, Fire, and Dangerous Things. Chicago: University of Chicago Press.

Palmer, Martha, Hoa Dang & Christiane Fellbaum, ``Making Fine-Grained and Coarse-Grained Sense Distinctions, Both Manually and Automatically". Journal of Natural Language Engineering, submitted.

Malt, Barbara C., Steven A. Sloman, Silvia Gennari, Meiyi Shi & Yuan Wang. 1999. Knowing vs. Naming: Similarity and the Linguistic Categorization of Artifacts. Journal of Memory and Language, 40, 230–262.

McKelvie, David, Chris Brew & Henry Thompson. 1998. ``Using SGML as a Basis for Data-Intensive Natural Language Processing". Computers and the Humanities, 31:5, 367-388.

Ng, Hwee Tou, Bin Wang & Yee Seng Chan. 2003. ``Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study". Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, pp 455-462.

Nunberg, Geoffrey. 1979. ``The Non-Uniqueness Of Semantic Solutions: Polysemy. Linguistics and Philosophy, 3, 143–184.

Olney, John, Carter Revard & Panl Ziff. 1966. ``Some Monsters in Noah's Ark". Research Memorandum, Systems Development Corp., Santa Monica. CA.

Peters, Wim, Piek Vossen, Pedro Diez-Orzas & Geert Adrians. 1998. ``Cross-Linguistic Alignment Of Wordnets With An Inter-Lingual Index". Computers and the Humanities, 32:2-3, 221-51.

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

Resnik, Philip & David Yarowsky. 1997a. ``Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation". Natural Language Engineering, 5:2, 113-133.

Resnik, Philip & David Yarowsky. 1997b. ``A Perspective On Word Sense Disambiguation Methods and Their Evaluation". proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? April 4-5, 1997, Washington, D.C., USA, 79-86.

Resnik, Philip & David Yarowsky. 2000. ``Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation". Natural Language Engineering, 5:2, 113-133.

Rodd, Jennifer, M. Gareth Gaskell & William Marslen-Wilson. 2002. ``Making Sense Of Semantic Ambiguity: Semantic Competition in Lexical Access". Journal of Memory and Language, 46, 245-266.

Rodd, Jennifer, M. Gareth Gaskell & William Marslen-Wilson. 2004. ``Modelling the effects of semantic ambiguity in word recognition". Cognitive Science, 28, 89-104.

Ruhl, Charles. 1989. On Monosemy: A Study in Linguistic Semantics. Albany: State Univ. of New York Press.

Sanderson, Mark. 1994. ``Word Sense Disambiguation and Information Retrieval". Proceedings of the 17th ACM SIGIR Conference, 142-151.

Schutze. Hinrich. 1998. Automatic word sense discrimination. Computational Linguistics, 24:1, 97-124.

Sparck Jones, Karen. 1986/1964. Synonymy and Semantic Classification. Edinburgh: Edinburgh University Press.

Stevenson, Mark & Paul Clough. 2004. ``EuroWordNet as a Resource for Cross-Language Information Retrieval". Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-04), May 2004.

Stevenson, Mark & Yorick Wilks. 1999. ``Combining Weak Knowledge Sources for Sense Disambiguation". Proceedings of the International Joint Conference for Artificial Intelligence (IJCAI-99), Stockholm, 884-889.

Tufiş, Dan, Radu Ion & Nancy Ide. 2004. ``Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering, and Aligned WordNets". Proceedings of the 20th International Conference on Computational Linguistics,(COLING2004), Geneva, 1312-1318.

Voorhees, Ellen. 1999. ``Natural Language Processing and Information Retrieval". Maria Teresa Pazienza, ed., Information Extraction: Towards Scalable, Adaptable Systems, Germany: Springer, 32-48.

Vossen Piek. 2001. ``Extending, Trimming and Fusing WordNet for Technical Documents". Proceedings of the Workshop on WordNet and Other Lexical Resources Applications, Extensions and Customizations, North American Association for Computational Linguistics (NAACL-2001), June 2001, Pittsburgh, USA.

Vossen, Piek, ed. 1998. EuroWordNet: A Multilingual Database With Lexical Semantic Networks. Amsterdam: Kluwer.

Wierzbicka, Anna. 1989. ``Semantic Primitives and Lexical Universals". Quaderni di Semantica, X:1, 103-121.

Wilks, Yorick. 1972. Grammar, Meaning and the Machine Analysis of Language. London and Boston: Routledge.

Wilks, Yorick & Roberta Catizone. 2002. ``What is Lexical Tuning?". Journal of Semantics, 156-169.

Wilks, Yorick, & Mark Stevenson. 1998. ``The Grammar of Sense: Using Part-of-Speech Tags as a First Step in Semantic Disambiguation". Journal of Natural Language Engineering 4:2, 74-87.

Wilks, Yorick, Brian Slator & Louise Guthrie. 1996. Electric Words: Dictionaries, Computers and Meanings. Cambridge, MA: MIT Press.

Yarowsky, David. 2000. ``Hierarchical Decision Lists for Word Sense Disambiguation". Computers and the Humanities, 34:2, 179-186.

Zgusta, Ladislav. 1971. Manual of Lexicography. The Hague: Mouton.