Natural Language Engineering 5 (3): 113–133. Printed in the United Kingdom

Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation

PHILIP RESNIK
Dept. of Linguistics and Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
resnik@umiacs.umd.edu

DAVID YAROWSKY
Dept. of Computer Science/CLSP
Johns Hopkins University
Baltimore, MD 21218
yarowsky@cs.jhu.edu

Abstract

Resnik and Yarowsky (1997) made a set of observations about the state of the art in automatic word sense disambiguation and, motivated by those observations, offered several specific proposals regarding improved evaluation criteria, common training and testing resources, and the definition of sense inventories. Subsequent discussion of those proposals resulted in SENSEVAL, the first evaluation exercise for word sense disambiguation (Kilgarriff and Palmer forthcoming). This article is a revised and extended version of our 1997 workshop paper, reviewing its observations and proposals and discussing them in light of the SENSEVAL exercise. It also includes a new in-depth empirical study of translingually-based sense inventories and distance measures, using statistics collected from native-speaker annotations of 222 polysemous contexts across 12 languages. These data show that monolingual sense distinctions at most levels of granularity can be effectively captured by translations into some set of second languages, especially as language family distance increases. In addition, the probability that a given sense pair will tend to lexicalize differently across languages is shown to correlate with semantic salience and sense granularity; sense hierarchies automatically generated from such distance matrices yield results remarkably similar to those created by professional monolingual lexicographers.

1 Introduction

Word sense disambiguation (WSD) is perhaps the great open problem at the lexical level of natural language processing. For English, at least, performance of state-of-the-art systems on other lexical tasks such as part-of-speech (POS) tagging and morphological analysis is respectable, if not perfect, and the dominant approaches (noisy channel models for tagging, two-level morphology) are by now well understood. These developments enable applications that can rely on accurate output from lexical analysis — for example, Davis (1996) improves performance in cross-language information retrieval using a constrained word-translation technique that relies on accurate part-of-speech analysis for query terms.

In contrast, although word sense ambiguity has been a central concern of natural language processing since the inception of the field (Weaver 1949), algorithms for word sense selection have not yet reached the level of a reliable enabling technology. Until fairly recently, evaluation of WSD algorithms on a small set of "interesting" cases was the norm, and few, if any, researchers had even attempted broad-coverage disambiguation. Prospects have changed, however, with the improved availability of common lexical resources (e.g. (Fellbaum 1998)), community-wide awareness of algorithms for exploiting large text corpora (Church and Mercer 1993), and the appearance of manually sense-tagged corpora (Landes *et al.* 1998; Ng and Lee 1996).

As a result of these developments, a SIGLEX Semantic Tagging Workshop was held in April, 1997, where we suggested a protocol for community-wide comparative analysis of word sense disambiguation techniques (Resnik and Yarowsky 1997). The proposal sparked a lively debate, and subsequent discussions led to the first evaluation exercise for word sense disambiguation, SENSEVAL (Kilgarriff and Palmer forthcoming), and a related evaluation (ROMANSEVAL) for Romance languages (Véronis 1998). This paper briefly reviews our observations and extends the presentation of our proposals, including additional discussion in light of the SENSEVAL exercise. We also include a new empirical study of our proposals regarding translingually motivated sense inventories and semantic distance measures.

2 Observations

Traditional evaluation for WSD is not standardized. In other natural language processing tasks such as POS tagging and parsing, evaluation has become fairly standardized, with most reported studies using common training and testing resources such as the Brown Corpus and Penn Treebank and fairly well accepted evaluation metrics. In contrast, apart from a few studies using common test suites (e.g. the 1993 Leacock et al. *line* data, shared by Lehman, 1994, Mooney, 1996 and others) there have traditionally been nearly as many WSD test suites as there are researchers in this field. As a consequence, it can be difficult to assess the state of the art.

The potential for WSD varies by task. As Stevenson and Wilks (1996) emphasize, WSD is not an end in itself, but rather an intermediate, enabling task. Among the most common language-related applications, speech recognition has seen little use for word senses, since equivalence classes of contexts (e.g. Bahl *et al.* 1983; Katz 1987) have a far better track record than smoothing based on classes of words (e.g. Brown *et al.* 1992). In information retrieval, even perfect word sense information may be of only limited utility (Krovetz and Croft 1992; Voorhees 1993), though NLP techniques do appear to show more promise in cross-language information retrieval than in monolingual retrieval (Doug Oard, personal communication). The potential for using word senses in high quality machine translation seems greater; for example, there is good reason to associate information about syntactic realizations of verb meanings with verb senses rather than verb tokens (Dorr and Jones 1996a, 1996b).

The field has narrowed down approaches, but only a little. In the area of POS tagging, the noisy channel model dominates (e.g. (Bahl and Mercer 1976; Jelinek 1985; Church 1988)), accompanied by transformational rule-based methods (Brill 1993) and grammatico-statistical hybrids (e.g. Tapanainen and Voutilainen 1994). There seems to be consensus on what makes POS tagging successful:

- The inventory of tags is small and fairly standard.
- Context outside the current sentence has little influence.
- The within-sentence dependencies are very local.
- Prior (decontextualized) probabilities dominate in many cases.
- The task can generally be accomplished successfully using only tag-level models without lexical sensitivities besides the priors.
- Standard annotated corpora of adequate size have long been available.

In contrast, approaches to WSD attempt to take advantage of many different sources of information (e.g. see McRoy 1992; Ng and Lee 1996; Bruce and Wiebe 1994; Wilks and Stevenson 1998); it seems possible to obtain benefit from sources ranging from local collocational clues (Yarowsky 1993) to membership in semantically or topically related word classes (Yarowsky 1992; Resnik 1993) to consistency of word usages within a discourse (Gale *et al.* 1992a); and disambiguation seems highly lexically sensitive, in effect requiring specialized disambiguators for each polysemous word. An up-to-date sampling of a wide range of methods can be found in the recent special issue of *Computational Linguistics* on WSD (Ide and Véronis 1998).

Adequately large sense-tagged data sets are difficult to obtain. Annotated data has facilitated recent advances in POS tagging, parsing, and other language processing subproblems. Unfortunately, of the few sense-annotated corpora currently available, virtually all are tagged collections of a single ambiguous word such as *line* or *tank*. The WordNet semantic concordance, SEMCOR (Miller *et al.* 1994), is an important and useful exception, providing the first large-scale, balanced data set for studying distributional properties of polysemy in English. However, its utility in supervised WSD is limited by its token-by-token sequential tagging methodology, yielding too few tagged instances of the large majority of polysemous words (typically fewer than 10 each). In addition, sequential tagging forces annotators to repeatedly refamiliarize themselves with the sense inventories of each word, slowing annotation speed and lowering intra- and inter-annotator agreement rates. The DSO corpus (Ng and Lee 1996), also having WordNet-based sense tags, is another potential resource, but it must be viewed with caution: measurements of agreement between DSO and SEMCOR are sufficiently low compared to SEMCOR inter-annotator agreement that, as Kilgarriff (1998, p. 583) comments, it is "impossible to regard [DSO] as a gold standard."

Another potential source of sense-tagged data comes from parallel aligned bilingual corpora, where translation distinctions can provide a practical correlate to

Table 1. *Probability distributions assigned by four hypothetical systems*

Sense	System 1	2	3	4
(1) monetary (e.g. on a loan)	·47	·85	·28	1·00
(2) stake or share ⇐ *correct*	·42	·05	·24	·00
(3) benefit/advantage/sake	·06	·05	·24	·00
(4) intellectual curiosity	·05	·05	·24	·00

sense distinctions (e.g. French *droit* and *devoir* correspond to English *duty*/TAX versus *duty*/OBLIGATION). The availability and diversity of such corpora are increasing, offering the possibility of limitless "tagged" training data without manual annotation, and the World Wide Web represents another high-potential source of parallel text, with the added advantage that, unlike static corpora, text on the Web tracks the continuous evolution of languages and their lexicons (Resnik 1998; Resnik 1999). Given the data requirements for supervised learning algorithms and the current paucity of such data, we believe that unsupervised and minimally supervised methods offer the primary near-term hope for broad-coverage sense tagging.[1]

3 Proposals

3.1 A better evaluation criterion

Prior to SENSEVAL, the standard for evaluation of word sense disambiguation algorithms was the appealingly simple "exact match" criterion, or simple accuracy:

$$\%\ \text{correct} = 100 \times \frac{\#\ \text{exactly matched sense tags}}{\#\ \text{assigned sense tags}}$$

However, consider the context

(1) ... bought an **interest** in Lydak Corp. ...

and assume the existence of 4 hypothetical systems that assign the probability distributions in Table 1 to the 4 major senses of *interest*.

Each of the systems prefers the *incorrect* classification (sense 1) over the correct sense 2 (*a stake or share*). However, System 1 has been able to nearly rule out senses 3 and 4 and assigns reasonably high probability to the correct sense, but is given the same penalty as other systems that either have ruled out the correct sense (systems 2 and 4) or effectively claim ignorance (system 3).

If we intend to use the output of the sense tagger as input to another probabilistic

[1] In this context, we take "supervised learning" to mean algorithms requiring training on correctly sense-tagged text using a known inventory of senses, and "unsupervised" to refer to any method that does not require tagged training data; c.f. Schütze's (1998) use of the term "sense discrimination."

Table 2. *Illustration of cross-entropy calculation*

	System 1	2	3	4
$\Pr_{\mathcal{A}}(cs_i \mid w_i, \text{context}_i)$	·42	·05	·24	·00
$-log_2 \Pr_{\mathcal{A}}(cs_i \mid w_i, \text{context}_i)$	1·25	4·32	2·05	∞

system, such as a speech recognizer, topic classifier, or IR system, it is important that it yield probabilities that are as accurate and robust as possible. If the tagger is confident, it should assign high probability to its chosen classification. If it is less confident, but has effectively ruled out several options, the assigned probability distribution should reflect this, too.

Experience in the speech community suggests that *cross-entropy* (or its related measures, perplexity and Kullback-Leibler divergence) can measure how well a model assigns probabilities to its predictions. It is easily computed as

$$-\frac{1}{N}\sum_{i=1}^{N} \log_2 \Pr_{\mathcal{A}}(cs_i|w_i, \text{context}_i)$$

where N is the number of test instances and $\Pr_{\mathcal{A}}$ is the probability assigned by the algorithm $\mathcal{A}$ to the correct sense, cs_i of word w_i in context_i. Crucially, given the hypothetical case above, System 1 would get much of the credit for assigning high probability, even if not the highest, to the correct sense. Just as crucially, an algorithm would be penalized heavily for assigning very low probability to the correct sense,[2] as illustrated in Table 2. Optimal performance is achieved under this measure by systems that assign accurate probabilities, neither too conservative (System 3) nor too overconfident (Systems 2 and 4).

This evaluation measure need not *replace* exact match. However, a measure based on cross-entropy or perplexity would add a fairer test, especially for the common case where several fine-grained senses may be correct and it is nearly impossible to select exactly the sense chosen by the human annotator. A variant of the cross entropy measure without the log term ($\frac{1}{N}\sum_{i=1}^{N} \Pr_{\mathcal{A}}(cs_i|w_i, \text{context}_i)$) can be used to measure improvement in restricting and/or roughly ordering the possible classification set without excessive penalties for systems with poor or absent probability estimates. In the latter case, when the assigned tag is given probability 1 and all other senses probability 0, this variant is equivalent to exact match.[3]

[2] The extreme case of assigning 0 probability to the correct sense is given a penalty of ∞ by the cross-entropy measure.

[3] This variant of the cross entropy measure was suggested by Dan Melamed; expanded in (Melamed and Resnik submitted).

Table 3. *Example sense inventory and distance/cost matrix for* bank

				I.1a	I.1b	I.2	II.1	II.2	III
I	*Bank*	- REPOSITORY							
	I.1	Financial Bank							
		I.1a - the institution	I.1a	0	1	2	4	4	4
		I.1b - the building	I.1b	1	0	2	4	4	4
	I.2	General Supply/Reserve	I.2	2	2	0	4	4	4
II	*Bank*	- GEOGRAPHICAL	II.1	4	4	4	0	1	4
	II.1	Shoreline	II.2	4	4	4	1	0	4
	II.2	Ridge/Embankment	III	4	4	4	4	4	0
III	*Bank*	- ARRAY/GROUP/ROW							

3.2 Evaluation sensitive to semantic/communicative distance

Current WSD evaluation metrics also fail to take into account semantic/communicative distance between senses when assigning penalties for incorrect labels. This is most evident when word senses are nested or arranged hierarchically, as illustrated in Table 3, left. An erroneous classification between close siblings in the sense hierarchy should be given relatively little penalty, while misclassifications across homographs should receive a much greater penalty. A penalty matrix $distance(\text{subsense}_1, \text{subsense}_2)$ could capture taxonomic semantic distance, derived from a single semantic hierarchy such as WordNet, or be based on a weighted average of simple hierarchical distances from multiple sources such as sense/subsense hierarchies in several dictionaries. A very simple example of such a distance matrix for the *bank* sense hierarchy is given in Table 3, right.

Penalties could also be based on general pairwise *functional communicative distance*: errors between subtle sense differences would receive little penalty while gross errors likely to result in misunderstanding would receive a large penalty. Such distances could be based on psycholinguistic data or models, such as experimentally derived estimates of similarity or confusability (e.g. (Miller and Charles 1991; Resnik forthcoming)). They could be based on a given task; for example, in speech synthesis penalizing only sense distinction errors corresponding to pronunciation distinctions (e.g. bass-/bæs/ vs. bass-/beɪs/). For machine translation, only sense differences lexicalized differently in the target language would be penalized, with the penalty proportional to communicative distance. Distances based on the weighted percentage of all languages that lexicalize two subsenses differently are proposed in detail in Section 3.5. In general such a distance matrix could support arbitrary communicative cost/penalty functions, dynamically changable according to task.

There are several ways in which such a (hierarchical) distance penalty weighting could be utilized along with a cross-entropy measure. The simplest is to minimize mean distance/cost between assigned sense (as_i) and correct sense (cs_i) over all N examples as an independent figure of merit:

$$\frac{1}{N}\sum_{i=1}^{N} \text{distance}(cs_i, as_i)$$

However, one could also use a metric such as the following, which measures efficacy of probability assignment in a manner that penalizes probabilities assigned to incorrect senses weighted by the communicative distance/cost between that incorrect sense and the correct one:

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{S_i}\mathrm{distance}(cs_i, s_j) \times \Pr_{\mathcal{A}}(s_j|w_i, \mathrm{context}_i)$$

where for any test example i, we consider all S_i senses (s_j) of word w_i, weighting the probability mass assigned by the classifier $\mathcal{A}$ to incorrect senses ($\Pr_{\mathcal{A}}(s_j|w_i, \mathrm{context}_i)$) by the communicative distance or cost of that misclassification.[4]

Melamed and Resnik (submitted) proposed a variation of these ideas for the SENSEVAL exercise, which used the HECTOR dictionary (Atkins 1993), organized in a fashion similar to Table 3, as its sense inventory. One innovation of that proposal was a scheme for the distribution of probability across levels of the sense hierarchy, accommodating selection of higher-level nodes (e.g. homograph-level distinctions) rather than bottom-level senses by human annotators and by disambiguation systems. Another innovation was an extension to handle test instances for which which multiple "correct" sense tags are identified, interpreting such multiple taggings disjunctively.

In SENSEVAL, scoring was done a number of different ways, varying the assumed level of granularity (bottom-level versus higher-level senses), the assumption of unique tags underlying the probabilistic scoring proposal, and the treatment of multiple correct tags. The Melamed-Resnik scoring was adopted as one of the set, but other scores were computed according to different assumptions — for example, interpreting multiple correct tags conjunctively rather than disjunctively, thus penalizing systems whenever they failed to include *all* the human-assigned sense tags for a test instance. Computing the "score of reference" for systems in SENSEVAL, instances assigned multiple tags by the human annotators were excluded from the test set, reducing it by about 15%. In retrospect, there appears to have been some confusion as to whether multiple human-assigned sense tags were intended to have been interpreted conjunctively or disjunctively; presumably this will be resolved by clearer specifications in future evaluations.

In practice, it appears that most of the SENSEVAL systems provided categorical responses (whether a single tag or multiple tags) rather than a probability distribution, and SENSEVAL scoring more closely resembles the traditional exact match criterion than it does some variant of cross-entropy.

3.3 A framework for common evaluation and test set generation

Supervised and unsupervised sense disambiguation methods have different needs regarding system development and evaluation. Although unsupervised methods may

[4] Although this function enumerates over all S_i senses of w_i, because $\mathrm{distance}(cs_i, cs_i) = 0$ this function only penalizes probability mass assigned to incorrect senses for the given example. Note that in the special case of sense tagging without probability estimates (all are either 0 or 1), this formula is equivalent to the previous one (simple mean distance or cost minimization).

1. Collect a very large (e.g., N = 1 billion words), diverse unannotated corpus.
2. Select a sense inventory (e.g. WordNet, LDOCE) with respect to which algorithms will be evaluated (see Section 3.4).
3. Pick a subset of $R < N$ (e.g., 100 million) words of unannotated text, and release it to the community as a training set.
4. Pick a smaller subset of $S < R < N$ (e.g., 10 million) words of text as the source of the test set. Generate the test set as follows:
   (a) Select a set of M (e.g., 100) ambiguous words that will be used as the basis for the evaluation, *without* revealing what those words will be.
   (b) For each of the M words, annotate all available instances of that word in the test corpus. Make sure each annotator tags all instances of a *single* word, e.g. using a concordance tool, as opposed to going through the corpus sequentially.
   (c) For each of the M words, compute evaluation statistics using individual annotators against other annotators.
   (d) For each of the M words, go through the cases where annotators disagreed and make a consensus choice, by vote if necessary.
5. Instruct participants in the evaluation to "freeze" their code; that is, from this point onwards, no changes may be made.
6. Have each participating algorithm do WSD on the full S-word test corpus.
7. Evaluate the performance of each algorithm considering *only* instances of the M words that were annotated as the basis for the evaluation. Compare exact match, cross-entropy, and inter-judge reliability measures (e.g. Cohen's κ) using inter-annotator results as an upper bound.
8. Release this year's S-word test corpus as a *development* corpus for those algorithms that require supervised training, so they can participate from now on, being evaluated in the future via cross-validation.
9. For next year's evaluation, go back to Step 3..

Fig. 1. Protocol for common evaluation and test set generation

be evaluated (with some limitations) by a sequentially tagged corpus such as SEMCOR (with a large number of polysemous words represented but with few examples of each), supervised methods require much larger data sets to provide adequate training and testing material. The protocol in Figure 1 satisfies the needs of both supervised and unsupervised tagging research; it served with some modification as the basis for the SENSEVAL exercise.

There are a number of advantages to this paradigm, in comparison with simply trying to annotate large corpora with word sense information.

First, it combines an emphasis on broad coverage with the advantages of evaluating on a limited set of words, as is done traditionally in the WSD literature. Step 4.a can involve any desired criteria (frequency, level of ambiguity, part of speech, etc.) to narrow down to a set of candidate words, and then employ random selection among those candidates. At the same time, it avoids a common criticism of studies based on evaluation using small sets of words, namely that there is not enough attention given to scalability. In this evaluation paradigm, *all* algorithms must be able to sense tag *all* words in the corpus meeting specified criteria, because there is no way to know in advance which words will be used to compute the figure(s) of merit.

Second, the process avoids some problems that arise in using exhaustively annotated corpora for evaluation. By focusing on a relatively small set of polysemous words, much larger data sets for each word can be produced. This focus will also allow more attention to be paid to selecting and vetting comprehensive and robust sense inventories, including detailed specifications and definitions for each. Furthermore, by having annotators focus on one word at at time using concordance software, the initial level of consistency is likely to be far higher than that obtained by a process in which one jumps from word to word to word by going sequentially through a text, repeatedly refamiliarizing oneself with different sense inventories at each word. Finally, by computing inter-annotator statistics blindly and *then* allowing annotators to confer on disagreements, a cleaner test set can be obtained without sacrificing trustworthy upper bounds on performance.

Third, the experience of the Penn Treebank and other annotation efforts has demonstrated that it is difficult to select and freeze a comprehensive tag set for the entire vocabulary in advance. Studying and writing detailed sense tagging guidelines for each word is comparable to the effort required to create a new dictionary. By focusing on only 100 or so polysemous words per evaluation, the annotating organization can afford to do a multi-pass study of and detailed tagging guidelines for the sense inventory present in the data for each target word. This would be prohibitively expensive to do for a full vocabulary. Also, by utilizing different sets of words in each evaluation, such factors as the level of detail and the sources of the sense inventories may change without worrying about maintaining consistency with previous data.

Fourth, both unsupervised and supervised WSD algorithms are better accommodated in terms of the data available. Unsupervised algorithms can be given very large quantities of training data: since they require no annotation the value of R can be quite large. And although supervised algorithms are typically plagued by sparse data, this approach will yield much larger training and testing sets per word.

The SENSEVAL exercise adopted some though not all aspects of this protocol. The diverse, balanced corpus (Step 1) was a 17M word pilot for the British National Corpus, which has since reached a size of over 100M words. The selected sense inventory (Step 2) was the HECTOR database, constructed by selecting a sample of words and sense-tagging all instances of them in the corpus — thus, as suggested above, the sense-tagging process provided feedback for refinement of the sense inventory itself.

Because the evaluation exercise included both supervised and unsupervised systems, the initial distribution of training materials included *tagged* rather than untagged data (contrary to Step 3) for a set of 29 target words; otherwise, however, the creation of the test set proceeded largely as specified in Step 4, for a set of 34 target words and a test set of 8448 instances, and systems were frozen in advance of the release of test data, as specified in Step 5.

The greatest departure of the SENSEVAL exercise from the protocol described above was the requirement that systems perform WSD on all words in a test corpus (Steps 6 and 7), with their developers remaining ignorant of which words were to be used for scoring. Instead, participating systems were grouped into categories, making it possible to do within-group comparisons of systems that disambiguated only the words in the test set, and a separate within-group comparison for those that disambiguated all content words appearing in the test collection. Comparison across groups indicates, not surprisingly, that the highest performance was obtained by systems using supervised training to learn classifiers specifically tuned to the words in the test set.

3.4 A multilingual sense inventory for evaluation

One of the most vexed issues in applied lexical semantics is how to define word senses. Although we certainly do not propose a definitive answer to that question, we suggest here a general purpose criterion that can be applied to existing sources of word senses in a way that, we suggest, makes sense both for target applications and for evaluation, and is compatible with the major sources of available training and test data.

The essence of the proposal is to restrict a word sense inventory to distinctions that are typically *lexicalized cross-linguistically.* This cuts a middle ground between restricting oneself to homographs within a single language, which tends toward a very coarse-grained distinction, and an attempt to express all the fine-grained distinctions made in a language, as found in monolingual dictionaries. In practice the idea would be to define a set of target languages (and associated bilingual dictionaries), and then to require that any sense distinction be realized lexically in a minimum subset of those languages. This would eliminate many distinctions that are arguably better treated as regular polysemy. For example, *table* can be used to refer to both a physical object and a group of people:

(1) a. The waiter put the food on the table.
 b. Then he told another table their food was almost ready.
 c. He finally brought appetizers to the table an hour later.

In German, for example, the two meanings can actually be lexicalized differently (*Tisch* vs. *Tischrunde*). However, as such sense distinctions are typically conflated into a single word in most languages, and because even German can use *Tisch* in both cases, one could plausibly argue for a common sense inventory for evaluation that conflates these meanings.

A useful reference source for both training and evaluation would be a table linking sense numbers in established lexical resources (such as WordNet or LDOCE) with these crosslinguistic translation distinctions, such as Table 4. A comparable mapping could readily be extracted semi-automatically from bilingual dictionaries or EuroWordNet (Bloksma *et al.* 1996). We note that the table follows many lexical resources, such as the original WordNet, in being organized at the top level according to parts of speech. This seems to us a sensible approach to take for sense inventories, since POS tagging accomplishes much of the work of semantic disambiguation, at least at the level of homographs (Stevenson and Wilks 1996).

Although cross-linguistic divergence is a significant problem, and 1-1 translation maps do not exist for all sense-language pairs, this table suggests how *multiple* parallel bilingual corpora can be used to yield sets of training data covering different subsets of the English sense inventory, that in aggregate may yield tagged data for all given sense distinctions when any one language alone may not be adequate.

For example, a German-English parallel corpus could yield tagged data for senses

Table 4. *Mapping between cross-linguistic sense labels and established lexicons*

Target Word	WordNet Sense #	English description	Spanish	French	German	Italian	Japanese
interest (noun)	1	monetary (e.g. on loan)	interés, rédito	intérêt	Zinsen	interesse	rishi, risoku
	2	stake/share	interés, participación	intérêt participation	Anteil	interesse	riken
	3,4	intellectual curiosity	interés,	intérêt	Interesse	interesse	kanshin, kyōmi
	5	benefit, advantage	provecho, interés, beneficio	intérêt	Interesse	interesse	rieki
drug (noun)	1a	medicine	medicamento, droga	medicament	Medikament, Arzheimittel	medicina	kusuri
	1b	narcotic	narcótica droga	drogue	Drogue, Rauschgift	droga	mayaku
bank (noun)	1	shoreline	ribera, orilla	banc, rive	Ufer	sponda,riva	kishi
	2	embankment	loma, cuesta	talus, terasse	Erdwall	muccio	teibō
	3	financial inst.	banco	banque	Bank	banca	ginkō
	4	supply/reserve	banco	banque	Bank	banca	ginkō
	5	bank building	banco	banque	Bank	banca	ginkō
	6	array/row	hilera, batería	rang, batterie	Reihe	batteria	retsu
fire (t. verb)	1	dismiss from job	despedir, echar	renvoyer	feuern	licenziare	kubi ni shimasu
	2	arouse, provoke	excitar, enardecer	enflammer, animer	beflügeln entzünden	accendere infiammare	kōfun saseru
	4	discharge weapn	disparar	lâcher	abfeuern	sparare	happō s.
	5	bake pottery	cocer	cuire	brennen	cuocere	yaku

1 and 2 for *interest*, and the presence of certain Spanish words (provecho, beneficio) aligned with *interest* in a Spanish-English corpus will tag some instances of sense 5, with a Japanese-English aligned corpus potentially providing data for the remaining sense distinctions. In some cases it will not be possible to find any language (with adequate on-line parallel corpora) that lexicalizes some subtle English sense distinction differently, but this may be evidence that the distinction is regular or subtle enough to be excluded or handled by other means. Section 3.5 provides empirically-observed examples of such cases.

Table 4 is not intended for direct use in machine translation. Also note that when two word senses are in a cell they are not necessarily synonyms. In some cases they realize differences in meaning or contextual usage that are salient to the target language. However, at the level of sense distinction given in the table, they correspond to the same word senses in English and the presence of either in an aligned bilingual corpus will indicate the same English word sense.

Monolingual sense tagging of another language such as Spanish would yield a similar map, such as distinguishing the senses of the Spanish word *dedo*, which can mean *finger* or *toe*. Either English or German could be used to distinguish these senses, but not Italian or French, which share the same sense ambiguity.

It would also be helpful for Table 4 to include alignments between multiple monolingual sense representations, such as COBUILD sense numbers, LDOCE tags, or WordNet synsets, to support the sharing and leveraging of results between multiple systems. This highlights an existing problem, of course: different sense inventories lead to different algorithmic biases. For example, WordNet as a sense inventory would tend to bias an evaluation in favor of algorithms that take advantage of taxonomic structure; LDOCE might bias in favor of algorithms that can take advantage of topical/subject codes, and so forth. Unfortunately we have no solution to propose for the problem of which representation (if any) should be the ultimate standard; we anticipate that discussion of the use of HECTOR in SENSEVAL will shed some light on this issue.[5]

3.5 A translingual empirical study of sense inventories and measures

This section presents an empirical investigation of the proposals outlined in Sections 3.2 and 3.4. Specifically, it will define a translingually motivated distance measure for word senses, and show how this can be used to generate an empirically motivated sense inventory and cost matrix. This measure will also be used to evaluate the HECTOR sense inventory used in the SENSEVAL framework. To this end, 21 native speakers of 12 diverse languages annotated 222 words in context, randomly selected from the SENSEVAL round-2 training set.[6] Each of the examples had an associated HECTOR sense tag, but these were hidden from the annotators. For 180 of the sentences, annotators were asked to select a single preferred translation of the English SENSEVAL word in context in their native language, and give the uninflected root form of that word.[7] An example of this tagging environment is given in Table 5, with the annotator's response in the boxes on the left.

Table 5. *Example of the free annotation task for a Japanese annotator*

Japanese Translation	Word In Context	HECTOR Sense (hidden)
`bando`	West Country folk jazz **band** Red Jasper will be	I.1
`haba`	cope with quite a narrow **band** of frequencies .	II.2.1
`suji`	of obsidian , except for a **band** of turquoise around	II.2.3
`ichidan`	fiend who with his rag-tag **band** of followers, obtains	I.2
`ittai`	under-populated. In a wide **band** of west Africa ,	II.1.3
`ichidan`	are preparing to repel a **band** of gypsies who have	I.2

[5] The SIGLEX'99 workshop on "Standardizing Lexical Resources" (University of Maryland, June 1999) focuses on standardization of lexical resources and performance-preserving mappings between existing resources.

[6] Basque, Japanese, Korean, Chinese, Turkish, Hungarian, Romanian, Greek, Hindi, Arabic, Spanish, and Swedish native speakers, all at a high level of English proficiency.

[7] Nancy Ide proposed a similar cross-lingual annotation and clustering effort using native speakers in her Herstmonceux SENSEVAL presentation (Ide forthcoming).

Table 6. *Example of the pairwise annotation task for a Turkish annotator*

Turkish Translation	Word In Context	HECTOR Sense (hidden)
`topluluk`	marvelous jazz and blues **band** .	I.1
`bant`	hand he bent the flexible **band** around the bird's leg	II.2.1
`bant`	hand he bent the flexible **band** around the bird's leg	II.2.1
`serit`	of obsidian , except for a **band** of turquoise around	II.2.3

The remaining 52 examples consisted of *paired* SENSEVAL sentences exhibiting two different senses of a single word. The granularity of the sense difference varied from different top-level (homograph) sense numbers to different subsenses of the same major sense, as illustrated in Table 6. Additional details are given in Section 3.5.2.

Annotators were asked specifically on the pairwise test to identify if there was any word pair in their language that distinguished the two meanings, i.e. a translation for word 1 that could not be used for word 2, and a translation for word 2 that could not be used for word 1. Thus these pairwise annotations attempted to elicit directly whether a lexical distinction existed in the tagger's native language sufficient to separate the two meanings (and hence would be usable as cross-lingual sense labels for this particular sense distinction).

3.5.1 A cross-linguistic measure of sense difference

One measure of the significance of a particular sense difference $sense_i/sense_j$ in a given inventory is the probability that these two senses will be lexicalized differently in some language L, or more formally, $P_L(\text{different-lexicalization}|sense_i, sense_j)$.

One can estimate this probability directly from the pairwise data shown in Table 6 by presenting several $sense_i/sense_j$ pairs and measuring the percentage that are lexicalized differently in language L. Although this directly addresses the question, data collection costs limit pairwise enumeration to a relatively small subset of the possible sense pairs. Thus this measure is primarily useful for computing aggregate values such as the average probability for a given granularity of sense ambiguity.

The second method of estimating $P_L(\text{different-lexicalization}|sense_i, sense_j)$ is based on the other part of the data set, where annotators simply gave their preferred translation for a randomly ordered set of examples covering several instances of all the target HECTOR word senses. The probability of different lexicalization can be averaged over all possible pairings of $sense_i/sense_j$ examples, as follows:

$$P_L(\text{different-lexicalization}|sense_i, sense_j) = \frac{1}{|sense_i||sense_j|} \sum_{\substack{x \in \{sense_i \ examples\},\\ y \in \{sense_j \ examples\}}} translation[x, L] == translation[y, L].$$

Essentially this computes the likelihood of an arbitrary pairing of examples of $sense_i$ and $sense_j$ in the data being labelled with the same translation in language L. It is a weaker estimate of the probability that language L lexicalizes the distinction between $sense_i$ and $sense_j$ in that annotators were not told to use distinguishing words if they exist, nor would this be possible as they were not considering specific pairings. They may have choosen to use the same word for two subtly different meanings even though another word pair may exist that can capture the meaning difference. Nevertheless, this measure does capture the tendency for the *preferred* word choice in language L to lexicalize a given English/HECTOR sense distinction. This measure can also be computed over all pairs $sense_i/sense_j$, not merely the selected subset given in the experiment. For these two reasons the measure is of practical merit.

3.5.2 Sensitivity of cross-lingual lexicalization differences to sense granularity

Before considering specific polysemous words, we examine the general effect of sense granularity on the tendency of word senses to be lexicalized differently across languages.

The paired sense data can be classified as one of four levels of similarity: the Roman-numeraled homograph level (band-I (group) vs. band-II (ring)), the major sense level (band-I.1 (music group) vs band-I.2 (other group)), the subsense level (which we arbitrarily use to refer to the distance between a general sense number such as I.1 and its specialization (I.1.2)), and finally the subsubsense level (such as between I.1.1 and I.1.2).[8] Computing average $P_L(\text{different-lexicalization}|sense_i, sense_j)$ broken down by granularity yields the following table:

Table 7. *Sense lexicalization probabilities based on the pairwise sense annotations*

	Average $P_L(\text{different-lexicalization}\|sense_i, sense_j)$		
Level of Granularity	All Languages	Indo-European Languages	Non-IndoEuoropean Languages
Homograph Level	.95	.94	.96
Major sense level	.78	.64	.85
Subsense level	.72	.59	.82
Subsubsense level	.52	.39	.62
Avg. of all levels	.74	.64	.81

Note that for homographs, 95% of all observed pairings were given different translations.[9] In contrast, at the finer subsubsense level only 52% of the given

[8] The full sense inventories and examples utilized for these data are available at `http://www.cs.jhu.edu/~yarowsky/nle/inventories.html`.

[9] Indeed the sole case where a homograph-level distinction was translated the same in a non-Indo-European language was for a single annotator in Japanese, who gave the musical band and bird-leg-band senses of band the same translation, *bando*, which is actually a case of polysemy inherited from English through two independent borrowings.

pairs were translated differently. This suggests that homograph-level distinctions are broadly salient and tend to be treated consistently as separate words across languages, while subsubsense distinctions appear to be less salient in that separate lexicalizations for these similar concepts have not evolved in the majority of the studied languages.

There also appear to be interesting differences in granularity effects between Indo-European and non-Indo-European languages. Both tend to strongly lexicalize homograph-level distinctions at nearly equal probability, but for the finer sense distinctions, many of the Indo-European languages tend to exhibit parallel ambiguities to English and differently lexicalize the more subtle meaning distinctions at a lower probability than more distantly related languages. This suggests the important practical implication that if parallel bilingual corpora are to be used for assigning monolingual sense tags, languages more distantly related to English will tend to be more effective at differently labelling the finer sense ambiguities.

3.5.3 Correlation between pair-based annotation and free annotation

As we have observed, pair-based annotation produces a more direct measure of the ability of languages to differently lexicalize specific sense distinctions, while free annotation of unpaired examples achieves broader coverage at the risk of giving the same translation for a pair of examples where a pair of adequately distinguishing words may well exist in the target language. However, for all but the finest subsubsense level, these two different measuring strategies tend to yield results that are closely correlated. Table 8 is the analog to Table 7 above, but based on free translation rather than pairwise annotation. The correlation coefficient between the all-languages columns in the two tables exceeds $r = .99$. This suggests that the free annotations of translations on average tend to capture the same general distinguishing capacity for word senses as an explicit pairwise analysis of specific sense differences. This indicates that at least at the coarser levels of sense granularity, the statistics utilized in this approach may be collected adequately from bilingual corpora produced by human translation.

Table 8. *Sense lexicalization probabilities based on the free translation annotations*

	Average $P_L(\text{different-lexicalization}\mid sense_i, sense_j)$		
Level of Granularity	All Languages	Indo-European Languages	Non-IndoEuoropean Languages
Homograph Level	.95	.94	.96
Major sense level	.74	.69	.80
Subsense level	.68	.58	.78
Subsubsense level	.44	.38	.50
Avg. of all levels	.70	.65	.76

3.5.4 Correlation between language distance and tendency to lexicalize differently

Table 9 lists the mean probability that a given language differently lexicalizes an English sense distinction in the HECTOR inventory, averaged over the 4 different levels of sense granularity. There appears to be a strong association between language distance from English and this mean probability value, further refining the differences in distinguishing strength observed between Indo-European (IE) and non-Indo-European (NI) languages.

Table 9. *Mean probablility that a language L will differently lexicalize an English sense ambiguity, correlated with language family distance*

Language	Avg. P_L	# Taggers	Language	Avg. P_L	# Taggers
NI - Basque	0.885	1	IE - Romanian	0.667	3
NI - Japanese	0.856	4	IE - Greek	0.635	2
NI - Korean	0.846	1	IE - Hindi	0.558	2
NI - Chinese	0.808	3	NI - Arabic	0.538	1
NI - Turkish	0.692	1	IE - Spanish	0.500	1
NI - Hungarian	0.692	1	IE - Swedish	0.461	1

One implication of these results for machine translation is that for relatively similar languages, such as Spanish-English, the importance of word sense disambiguation is apparently lower, given that approximately 50% of the sense distinctions noted by lexicographers need not be resolved due to parallel polysemy in the target language, while for more distant languages from English such as Japanese, 86% of the monolingual sense distinctions also corresponded to translation distinctions and hence need resolving for MT. Nevertheless, both values are arguably high enough to warrant some form of word sense disambiguation for lexical choice in MT systems.

3.5.5 A cross-lingually motivated definition for cost matrices

Section 3.2 discusses the advantages of evaluating sense taggers via a matrix of semantic distance and/or the communicative cost of confusing two senses. A very natural measure of this semantic distance is the mean probability that the two senses will be lexicalized differently in a second language, which we have already argued is an indication of the salience of a sense distinction and clearly correlates directly with error rate in lexical choice.[10]

We can define a single-language-specific cost function as:

$$Cost(sense_i, sense_j, L) = P_L(\text{different-lexicalization}|sense_i, sense_j)$$

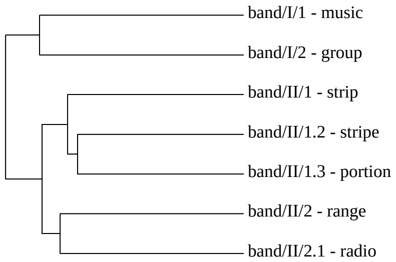
or a multi-lingual cost function:

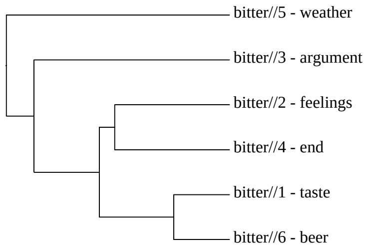
$$Cost(sense_i, sense_j) = \frac{1}{|Languages|} \sum_{L \in Languages} P_L(\text{diff-lexicalization}|sense_i, sense_j)$$

[10] Senses lexicalized differently in a target language tend to yield translation errors when confused.

Table 10. *Translingually generated distance matrices for* **band** *and* **bitter**

	I/1	I/2	II/1	II/1.2	II/1.3	II/2	II/2.1
band/I/1 (music)	0	0.857	0.885	0.943	0.979	0.962	0.943
band/I/2 (group)	0.857	0	0.995	0.969	1.000	0.865	0.961
band/II/1 (strip)	0.885	0.995	0	0.740	0.729	0.847	0.844
band/II/1.2 (stripe)	0.943	0.969	0.740	0	0.698	0.833	0.750
band/II/1.3 (portion)	0.979	1.000	0.729	0.698	0	0.778	0.729
band/II/2 (range)	0.962	0.865	0.847	0.833	0.778	0	0.771
band/II/2.1 (radio)	0.943	0.961	0.844	0.750	0.729	0.771	0





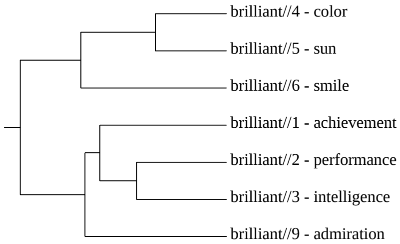
	1	2	3	4	5	6
bitter//1 (taste)	0	0.576	0.875	0.549	0.896	0.250
bitter//2 (feelings)	0.576	0	0.788	0.514	0.882	0.583
bitter//3 (argument)	0.875	0.787	0	0.725	0.879	0.875
bitter//4 (end)	0.549	0.514	0.725	0	0.875	0.583
bitter//5 (weather)	0.896	0.882	0.879	0.875	0	0.896
bitter//6 (beer)	0.250	0.583	0.875	0.583	0.896	0

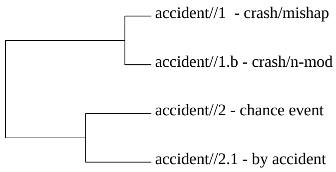
This estimates the pairwise cost of confusing HECTOR $sense_i$ and $sense_j$ based on the tendency of the language to use different words for the two meanings. If the meaning distinction has a high probability of being lexicalized in many languages, then this provides some evidence that the distinction is important. If few or no human languages lexicalize this meaning distinction, this may be considered evidence that the distinction is less salient or has lower cost of ambiguity.

Tables 10 and 11 show distance matrices computed using the multi-lingual cost function above, based on the free translation rather than pairwise annotation method of computing $P_L(\text{different-lexicalization}|sense_i, sense_j)$. Finer sense distinctions clearly have lower pairwise costs than coarser distinctions under this measure in the given examples. To help visualize these matrices better, we have applied a hierarchical agglomerative clustering procedure using maximal linkage (Duda and Hart 1973), yielding automatically derived sense trees that optimize between-cluster distance. These trees (also shown in Tables 10 and 11) are based exclusively on the free-tagging of the preferred translations of randomly ordered examples in context, and the HECTOR sense numbers were not utilized in any way in the clustering procedure. Yet these induced trees precisely mirror the sense hierarchy given by the HECTOR lexicographers, at not only the homograph level but down to the subsubsense level as well.

Table 11. *Translingually generated distance matrices for* **brilliant** *and* **accident**

	1	2	3	4	5	6	9
brilliant//1 (achievement)	0	0.537	0.570	0.781	0.756	0.850	0.637
brilliant//2 (performance)	0.537	0	0.405	0.862	0.866	0.929	0.625
brilliant//3 (intelligence)	0.570	0.405	0	0.844	0.856	0.900	0.613
brilliant//4 (color)	0.781	0.863	0.844	0	0.320	0.656	0.766
brilliant//5 (sun)	0.756	0.866	0.856	0.320	0	0.630	0.750
brilliant//6 (smile)	0.850	0.929	0.900	0.656	0.630	0	0.792
brilliant//9 (admiration)	0.637	0.625	0.613	0.766	0.750	0.792	0





		1	1.b	2	2.1
accident//1	(crash/mishap)	0	0.18	0.97	0.98
accident//1.b	(crash/n-mod)	0.18	0	0.97	0.96
accident//2	(chance event)	0.97	0.97	0	0.45
accident//2.1	(by accident)	0.98	0.96	0.45	0

Interestingly, many of the HECTOR sense inventories are quite flat (such as for the adjectives *bitter* and *brilliant*), exhibiting only a single non-hierarchical list of major numbered senses. However, the sense trees derived using the translingual cost matrix show a quite natural hierarchical clustering of these meanings, such as recognizing that *bitter//1 (taste)* and *bitter//6 (beer)* are quite similar (only 25% probability of being lexicalized differently across languages). Also, note that the "radiant" senses of brilliant (4=color,5=sun,6=smile) are clustered together while the achievement/accomplishment/intelligence senses (1,2,3,9) are also clustered together in a natural hierarchy. This suggests that hierarchical clustering based on the probability of differential lexicalization across languages may have additional merit in superimposing an empirically motivated sense hierarchy on flat sense inventories.

4 Conclusions

The most important of our observations about the state of the art in word sense disambiguation is that it is still a difficult, open, and interesting problem, on which the field has not typically reached consensus. We have made several suggestions that we believe will help assess progress and advance the state of the art. In summary:

- We proposed that the accepted standard for WSD evaluation include a cross-entropy like measure that tests the accuracy of the probabilities assigned to sense tags and offers a mechanism for assigning partial credit.

- We suggested a paradigm for common evaluation that combines the benefits of traditional "interesting word" evaluations with an emphasis on broad coverage and scalability.
- We outlined a criterion that should help in determining a suitable sense inventory to use for comparison of algorithms, compatible with both hierarchical sense partitions and multilingually motivated sense distinctions.

These proposals have in large part been put into practice by the first SENSEVAL exercise, yielding an impressive array of new comparative data on the performance of sense disambiguation systems, insights into the nature of the problem, and fresh debates over the process of evaluation.

We also presented a substantial exploration of the relationship between monolingual sense inventories and translation distinctions across languages. Specifically, we measured the probability of English monolingual sense distinctions in the HECTOR database being lexicalized differently across 12 widely diverse languages, studied at several levels of sense granularity. This measure has been shown to correlate with monolingual sense distance, and thus may be effective as the basis of a semantic distance or cost matrix for sense disambiguation evaluation. New sense hierarchies automatically generated from these matrices using hierarchical agglomerative clustering also yield results remarkably similar to those created by the HECTOR monolingual lexicographers. These parallel structures suggest that the lexicographer's intuitions regarding sense distance and clustering closely resemble empirically measured distances in cross-lingual data, providing further evidence for the plausibility of these monolingual sense hierarchies.

References

S. Atkins. (1993). Tools for computer-aided lexicography: the Hector project. In *Papers in Computational Lexicography: COMPLEX '93*, Budapest.

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

L. Bahl, F. Jelinek, and R. Mercer. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190.

L. Bloksma, P. Díez-Orzas and P. Vossen. (1996). User Requirements and Functional Specification of the EuroWordNet Project. `http://www.let.uva.nl/~ewn`

E. Brill. (1993). *A Corpus-Based Approach to Language Learning*. Ph.D. thesis, Computer and Information Science, University of Pennsylvania.

P. Brown, V. Della Pietra, P. deSouza, J. Lai, and R. Mercer. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–480.

R. Bruce and J. Wiebe. (1994). Word-sense disambiguation using decomposable models. In *Proceedings of ACL '96*, Las Cruces, NM., pp. 139-146.

K. Church. (1988). A stochastic parts program and noun phrase parser for unrestricted texts. In *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas, pp. 136–143.

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

M. Davis. (1996). New Experiments In Cross-Language Text Retrieval at NMSU's Computing Research Lab, In E. M. Voorhees and D. K. Harman (eds.), The Fifth

Text REtrieval Conference (TREC-5), NIST Special Publication 500-238, Department of Commerce, National Institute of Standards and Technology, pp. 447–454. http://trec.nist.gov/pubs/trec5/papers/nmsu.davis.paper.ps

B. Dorr and D. Jones. (1996a). Acquisition of semantic lexicons: Using word sense disambiguation to improve precision. In *Proceedings of the SIGLEX Workshop on Breadth and Depth of Semantic Lexicons*, Santa Cruz, CA.

B. Dorr and D. Jones. (1996b). Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues. In *Proceedings of the International Conference on Computational Linguistics*, Copenhagen, Denmark.

R. Duda and P. Hart. (1973). *Pattern Classification and Scene Analysis*. Wiley: New York.

C. Fellbaum, ed. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

C. Fellbaum, J. Grabowski, and S. Landes. (1998). Performance and Confidence in a Semantic Annotation Task. In Fellbaum, ed. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

W. Francis and H. Kučera. (1982). *Frequency Analysis of English Usage*. Houghton Mifflin.

W. Gale, K. Church, and D. Yarowsky. (1992a). One sense per discourse. *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pp. 233-237.

W. Gale, K. Church, and D. Yarowsky. (1992b). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439.

N. Ide. (forthcoming). Cross-lingual sense determination: Can it work? *Computers and the Humanities*, 33(4-5).

N. Ide. (forthcoming). Parallel Translations as Sense Discriminators. To appear in *Proceedings of SIGLEX'99*.

N. Ide and J. Véronis. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24:1–40.

F. Jelinek. (1985). Markov source modeling of text generation. In J. Skwirzinski, editor, *Impact of Processing Techniques on Communication*. Dordrecht.

S. Katz. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(3):400–401.

A. Kilgarriff. (1998). SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs. In Proc. LREC, Granada, May 1998, pp. 581–588.

A. Kilgarriff and M. Palmer, Eds. (forthcoming). Special double issue on SENSEVAL, *Computers and the Humanities*, 33:4-5.

R. Krovetz and W. B. Croft. (1992). Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141.

S. Landes, C. Leacock, and R. Tengi. (1998). Building Semantic Concordances. In C. Fellbaum, ed. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

C. Leacock, G. Towell and E. Voorhees. (1993). Corpus-based statistical sense resolution. In *Proceedings, ARPA Human Language Technology Workshop*, pp. 260–265.

J. Lehman. (1994). Toward the essential nature of statistical knowledge in sense resolution. In *Proceedings of the 12th National Conference on Artificial Intelligence*, pp. 734–471.

R. Mandala, T. Takenobu, and T. Hozumi. (1998). The use of WordNet in information retrieval. In S. Harabagiu (ed.), Usage of WordNet in Natural Language Processing Systems, COLING/ACL-98 Workshop, Montreal, August 16, 1998, pp. 31–37.

D. Melamed and P. Resnik. (submitted). Evaluation of sense disambiguation given hierarchical tag sets. Extends http://www.itri.brighton.ac.uk/events/senseval/mr.asc.

R. Mooney. (1996). Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, pp. 82-91.

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

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

G. Miller, M. Chodorow, S. Landes, C. Leacock, and R. Thomas. (1994). Using a semantic concordance for sense identification. In *Proceedings of the ARPA Human Language Technology Workshop*, San Francisco. Morgan Kaufmann, pp. 240–243.

H. Ng and H. Lee. (1996). Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Society for Computational Linguistics*, Santa Cruz, CA, pp. 40–47.

P. Resnik. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships.* Ph.D. thesis, University of Pennsylvania. `ftp://ftp.cis.upenn.edu/pub/ircs/tr/93-42.ps.Z`

P. Resnik. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95).* `cmp-lg/9511007`.

P. Resnik. (1998). Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text, In *Proceedings of AMTA-98*, October, pp. 72-82. `http://umiacs.umd.edu/~resnik/pubs/amta98.ps.gz`

P. Resnik. (1999). Mining the Web for Bilingual Text. To appear in *Proceedings of ACL'99*, College Park, MD, June.

P. Resnik. (forthcoming) Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. To appear in *Journal of Artificial Intelligence Research.*

P. Resnik and D. Yarowsky. (1997). A perspective on word sense disambiguation methods and their evaluation, in Marc Light (ed.) *ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, April, Washington, D.C., pp. 79-86.

R. Richardson, A. Smeaton, and J. Murphy. (1994). Using WordNet as a knowledge base for measuring semantic similarity between words. Working Paper CA-1294, Dublin City University, School of Computer Applications, Dublin, Ireland. `ftp://ftp.compapp.dcu.ie/pub/w-papers/1994/CA1294.ps.Z`

H. Schütze. (1998) Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124.

M. Stevenson and Y. Wilks. (1996). The grammar of sense: Is word-sense tagging much more than part-of-speech tagging? `cmp-lg/9607028`.

P. Tapanainen and A. Voutilainen. (1994). Tagging accurately – don't guess if you know. In *Proceedings of ANLP '94.*

E. Voorhees. (1993). Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval.* Pittsburgh, pp. 171–180.

W. Weaver. (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, pp. 15–23.

Y. Wilks and M. Stevenson. (1998). Word sense disambiguation using optimised combinations of knowledge sources. In *Proceedings of COLING/ACL-98*, Montreal, pp. 1398–1092.

D. Yarowsky. (1992). Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING-92*, pp. 454–460, Nantes.

D. Yarowsky. (1993). One sense per collocation. In *Proceedings of the ARPA Human Language Technology Workshop*, Morgan Kaufmann, pp. 266-271.

D. Yarowsky. (1995) Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA, pp. 189-196.