Bridging the gap between lexicon and corpus: convergence of formalisms

Adam Kilgarriff*
ITRI, University of Brighton
Adam.Kilgarriff@itri.bton.ac.uk

Abstract

I first consider the spectrum of lexical information from the semantic to the textual. A range of lexicons are classified according to where they sit on this scale. Lexicographic tools and WSD programs are included in the classification, and this is justified. There is currently a lacuna between the most text-oriented of the lexicographic approaches, and the most sophisticated of the data-driven ones. Lexical tuning requires that the lacuna be filled, so corpus data can flow into the lexicon. Following an analysis of similarities and differences between lexicographic tools and data-driven approaches, a strategy for bridging the gap is proposed.

1 A Spectrum of Lexicons

Lexicons look in two directions: towards the text, and towards semantics. When we look up a word or phrase encountered in a text to find the meaning, the textual orientation provides the input, the semantic one, the output. In a language generation system, the roles are reversed.

In principle lexicons could be very highly developed in relation to both the textual and the semantic orientation. In practice, for computational lexicons in particular, the emphasis tends to be on one or the other. (The MicroKosmos project is interesting in this regard. It has two teams, one working on the text-oriented lexicon, the other on the ontology (Viegas and Nirenburg, 1995).)

Thus lexicons can be placed on a spectrum according to where the emphasis lies: how 'surfacey' they are, as in Fig. 1. At the semantic end lie AI ontologies. Traditional native-speaker dictionaries such as the OED are somewhat closer to the text, and learner dictionaries, with their emphasis on grammatical and textual patterning, more so. (Since the advent of corpus lexicography in the early 1980s, the entire UK dictionary-publishing community has been steadily creeping in a text-ward direction, with the learners' dictionaries leading the way.) NLP lexicons which have been developed with parsing in mind, such as COMLEX or the ANLT lexicon, are further along a textual direction. Further still down this road are 'lexicons' of patterns as used in Information Extraction.

Then there is a gap; and then we move to data-driven artifacts like the decision lists Yarowsky uses for word sense disambiguation, Schütze's sense clusters, and Grefenstette's thesauri (Yarowsky, 1995; Schütze, 1998; Grefenstette, 1994). At the end point of the scale are sets of corpus citations for the word.

Underlying the analysis is the thesis that the lexical entry for a word is an abstraction from the occurrences of the word in the language. A higher level of abstraction takes us further away from the linguistic data, towards the semantic end of the spectrum.[1]

2 Correlated distinctions

There are several related but distinct dimensions along which lexicons can be analysed:

- **RHS/LHS**

  In a traditional dictionary, the textual orientation is the 'left hand side' of the dictionary entry, the semantic one, the 'right hand side'. Both can be complex. At a first glance, the LHS in a traditional dictionary is a simple headword, but on closer investigation, it becomes clear

[*] The work was undertaken under EPSRC grant GR/K/18931

[1] Investigators who view lexicons as primarily approximations to mental lexicons might find this perspective odd: here lurks an issue regarding the primacy of linguistic as against psychological data in the study of the lexicon. For discussion see (Kilgarriff, 1992, section 1.4.1).

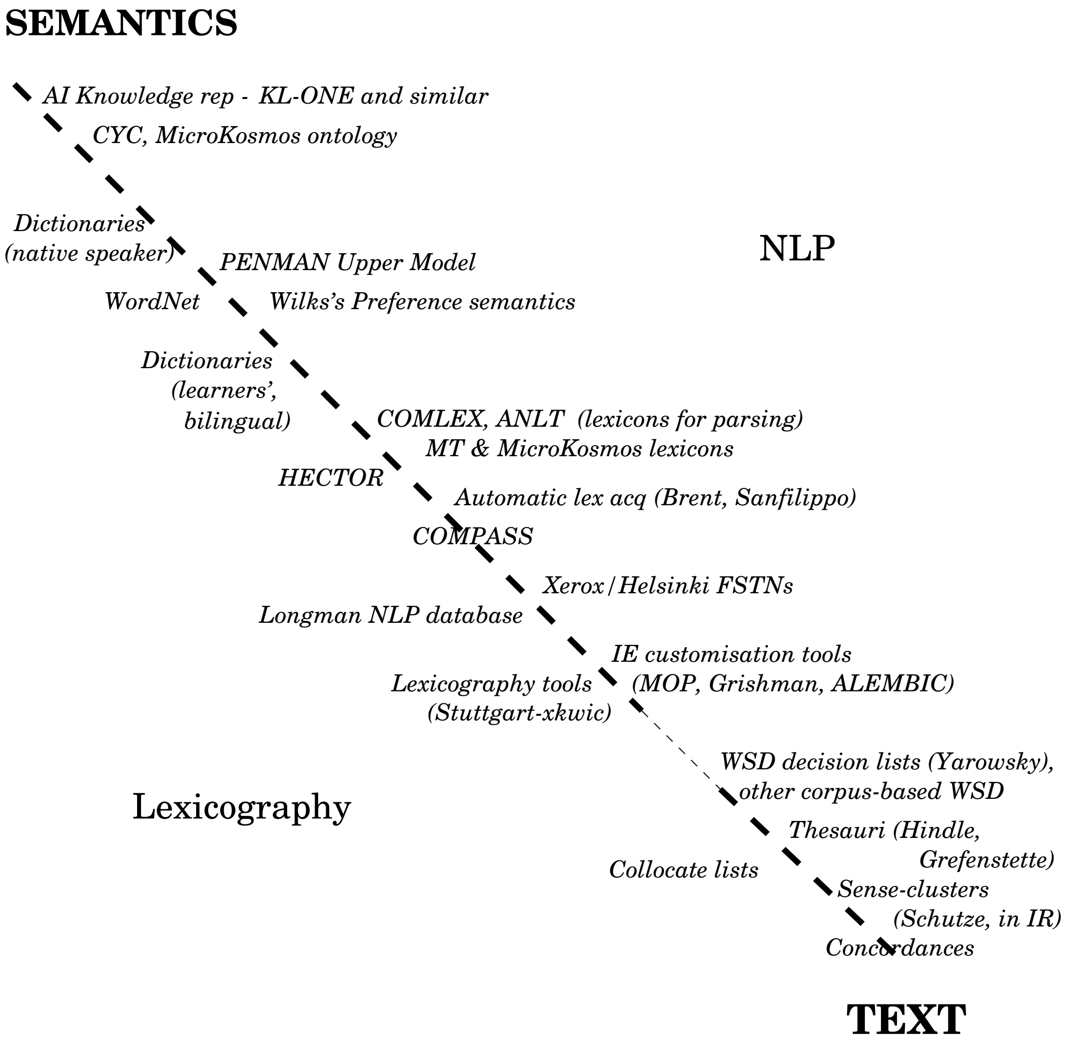

Figure 1: Lexical resources on the text/semantics spectrum

that there is more to it. Fixed and semi-fixed phrases, collocates, domain labels like *medical* or *Mil. Hist.*, all serve to help the user find the appropriate match between text and dictionary. In NLP systems this goes further. Data in the lexicon for purposes of word sense disambiguation, for example, is LHS data, there to obtain a correct match so that the appropriate RHS is picked up. In a lexicon such as COMLEX, the subcategorisation patterns simultaneously allow matching with the text and provide the syntactic-semantic key whcih permits further analysis of the text. For NLP lexicons the LHS/RHS distinction often fails to apply.

- **automatic/manual**

  There is more potential for automation of lexicons towards the 'text' end of the scale. However the two considerations do not always coincide: methods such as Brent (1993)'s for automatic acquisition of stativity and subcategorisation information for verbs aim to classify according to a pre-defined set of possibilities which are more remote from the text, and higher in their level of abstraction, than the kinds of information about grammatical patterning, selection restrictions and common collocates in the (manually produced) Longman database of common verbs, or DELIS lexical entries (Heid and Krüger, 1996).

- **empiricism/rationalism**

  Empiricists will find themselves more at ease at the text end of the scale, where more of the warts of the language are preserved. Rationalists will be more at ease at the semantics end, where the impurities need not be heeded and general theories can capture the common behaviour. Empiricists will tend to find the semantics end too unconnected to any data to ground it, whereas rationalists will find the frequency facts that dominate the text end arbitrary, intractable and irritating.

- **what commonly happens/what can happen**

  The previous point is related to a dilemma that a working lexicographer encounters every working day: at what level of generality should a word's behaviour be described? Consider the verb *glue.* Anything can be glued to anything, so at one level there should be no limitation on the object (beyond a possible constraint to 'physical', but even that is difficult since we may speak of, eg, gluing theories together). But on another level, we know (and a large corpus such as the BNC confirms) that most gluing is of components[2] and this is salient information about the behaviour of the verb which is likely to be of use to a human dictionary user or an NLP system (where, e.g., it is trying to determine a PP attachment).

  Dictionaries generally have a range of techniques for indicating everything on the spectrum from "only appears with this object", via "usually occurs with this or some other item similar in meaning" to "can occur with anything but commonly occurs with items in this category". The repertoire of devices at the dictionary editor's disposal, for use in definitions or examples, are: brackets, slashes, ellipsis-dots and *etc.*. [3]

  Note the interactions of frequency and semantic similarity in the problem. I illustrate with a thought experiment: first, the objects of verbal *glue* are measured according to their semantic similarity to the most 'central' object, say *part.* Then we find the frequencies of all words as heads of object NPs for *glue* in a large-enough corpus. We now generate a histogram with words arranged along the x-axis according to their similarity to *part.* All being well, this gives a bell-shaped curve with its peak at *part.* The lexicographer has ideally to convey not only the prototypical items at the peak of the curve, but also how steep it is: do a small cluster of words account for most occurrences, or is it a flat curve where many more forms are normal? (A related everyday problem for lexicographers is how to rate high-frequency words against lower frequency ones, when selecting which collocates to present in the dictionary. Typically the high-frequency word has higher absolute frequency as a collocate, but the lower-frequency word is a more striking collocate, with a more specific meaning. The lower-frequency word paints a more vivid but less general picture of the meaning of the nodeword.)

[2] Words occurring repeatedly as heads of object noun phrases of verbal *glue* in the BNC were: *edge material surface overlap bit strip* (and *eye* in the metaphorical sense.)

[3] More sophisticated metalanguage has the drawback that it is ignored: dictionary users do not read the Front Matter ('explanatory material') so do not understand metalinguistic devices that are not familiar to them from other genres (Béjoint, 1994).

For many purposes, particularly at the semantic end of the scale, these complexities are reduced to a boolean: acceptable/not acceptable. For lexicographers and others working towards the text end of the scale, they are pressing concerns.

3 Lexicons, tools and formalisms

At a first pass, Information Extraction (IE) and lexicography **tools** may seem out of place in a classification of lexicons. However, the tools use formalisms which can be looked at in two ways. They are search languages for finding patterns in corpora. But then, where a set of search patterns has been successful in finding all and only the occurrences of some lexical phenomenon, it also serves well as the textual aspect of a lexical entry for that object.

Consider *keep tabs on*. Of 174 occurrences of *tabs* in the BNC, half relate to the idiom. It occurs with the verb in the forms *keep keeps kept keeping keepin'*, with *tabs* premodified by *close careful effective*, and once in "on whom they must keep tabs". A good lexicographic tool will allow the lexicographer to specify these patterns (bar the last) in an extended regular expression language as

[lem="keep"] [pos="adj"]? [word="tabs"] [word="on"]

(assuming a part-of-speech-tagged, lemmatised corpus as input), and this is a search string the lexicographer might use to gather a set of instances of the expression prior to writing a definition.

But this string (supplemented by others for passive, "on whom ..." and other variants) also serves as the LHS for a lexical entry: it will identify all and only the occurrences of the idiom in an input text.[4] While a lexicography tool is clearly not a lexicon, the tool defines a formalism which a lexicon may ause (and which a lexicon developed with that tool will use by default). We generally think of lexicons as static resources, while tools relate to activity. The distinction is not salient when the formalisms are the focus of our attention.

4 Word Sense Disambiguation

How can WSD systems be classified alongside lexicons?

WSD research has the following pattern:

1. Use a lexical resource and/or corpus and/or human input to develop a 'profile' of the contexts associated with each of the word's senses;

2. For a word-in-context to be disambiguated, identify which profile is the best match.

For Lesk (1986) and related programs based on machine-readable dictionaries, stage 1 is null or minimal: the lexical resource is the machine-readable dictionary, either in its original form or modified, and the ingenuity lies in stage 2. For much corpus-based WSD (c-WSD) work, stage 1 is the critical phase. Most c-WSD papers report, first, on an algorithm for producing a special purpose lexicon, and then, on its use. It is c-WSD systems that are my concern in this paper.

The WSD task is a pure textual one, of finding the correct lexical entry given the text. By way of semantics, c-WSD lexicons generally have only atomic, distinct meanings (occasionally related to each other in a hierarchy or by some measure of closeness).

1 and 2 above are generally presented as a whole. The form of the output of stage 1, the profiles, is not presented as something of value in its own right. Evaluation procedures consider the whole process, rather than focusing on the efficacy of the two elements separately. And texts of the same type are used for training and for testing the system. No authors discuss the possibility of the two stages being separated. However they are quite distinct in most c-WSD algorithms and there are several arguments for separating them.

Re-usable resource The output of stage 1, like other lexical information, is expensive and difficult to gather but, once gathered, can be re-used by different clients. Separating the two stages will make it possible to produce NLP applications which perform WSD without needing to generate profiles, the hard part of WSD.

Semi-automatic methods
Fully automatic methods are essential only for stage 2. For stage 1, human input may well be appropriate.

Evaluation Stages 1 and 2 can be evaluated separately, making comparisons between systems more precise.

[4] I gloss over difficult questions of idiom identity. The BNC also contained

> On election day they will keep a running tab on who has voted in order to ...

Questions for the reader: (1) Is it the same meaning? (2) Is it the same idiom?

Text varieties/ sense varieties The main argument *against* separating the stages in corpus-based WSD is that it is appropriate to **train** and **test** a system on the same kinds of text. Then, the discriminators between senses which are found in the training data will tend to be those that apply in the test data. Only one corpus is required, and the paradigm is familiar: experiments can be repeated through partitioning the corpus differently between training set and test set, and so forth.

This paradigm misses a crucial point, a point which is central to the theme of the workshop. The prior question is not whether the training corpus matches the test corpus (or text the WSD program is to be applied to) but whether the senses are appropriate for the test corpus/application text (and task). The usual scenario in work to date has been that the word senses are taken from a general purpose dictionary, so are for general English, whereas the material to be disambiguated is, say, Wall Street Journal text. So, the profiles the program develops will be for general English senses according to the WSJ: i.e., a severe but widely overlooked mismatch (see also (Basili, Della Rocca, and Pazienza, 1997)).

Re-use of the c-WSD lexicon will be more appealing where it is human-readable.

c-WSD plays a role in the view presented in this paper both directly and indirectly. Directly, because data to support WSD is important lexical information. Indirectly, through expertise in data-driven methods. c-WSD has been an active research area for several years now, and serves as a testing ground for bottom-up techniques for lexicon development. In contrast to Brent-style automatic acquisition of grammatical information, WSD clearly requires lexical specifications that are sensitive to all the specificities and oddities of the contexts that a particular word-sense occurs in: the lexicon needs to stay right at the text end of the scale. It seems likely that much information about collocates, adjuncts and arguments would be more accurate and more useful if it stayed closer to the text, so techniques currently under investigation for WSD are likely to become salient for a wide range of kinds of lexical data.

5 Spiral-bound regular expression formalisms

Here I note that three tools have independently developed very similar formalisms for corpus searching (so potentially also for textual specifications in lexicons). They are XKWIC/CQP (Schulze and Christ, 1994), from IMS, Stuttgart ('Xkwic'); the ALEMBIC workbench (Day et al., 1997) from MITRE Corp. ('AWB'), and Mother of Perl (Doran et al., 1997) from U Penn ('MOP').[5]

First, a disclaimer noting the differences. Xkwic was produced for linguists and lexicographers, whereas AWB and MOP were produced for rapid development of Information Extraction systems. MOP assumes its users are programmers, so makes less concessions to user-friendliness than Xkwic. MOP is described as a programming language, and permits embedding of arbitrary Perl code, whereas Xkwic only has one kind of action associated with a pattern-match, viz., "return the match".

The commonalities are these. In each, words are first tokenised, providing the spiral in the spiral-bound notebook metaphor used to describe MOP. Then, assorted other programs add information about tokens or spans of tokens. Each adds a page to the notebook, with information on different pages related via the spiral. These programs typically include part-of-speech taggers, lemmatisers, sentence-identifiers, noun-phrase identifiers, taggers for names and places, and so forth. Patterns can then be constructed by reference to the raw form of the token, or any of the other added information. Thus, the pattern for *keep tabs on*, shown above and repeated here,

[lem="keep"] [pos="adj"]? [word="tabs"] [word="on"]

assumes that the input has passed through a part-of-speech tagger (which fills the `pos` field for each word) and a lemmatiser (filling the `lem` field). This and other examples are in the Xkwic formalism but can straightforwardly be translated into AWB or MOP.

The default structure for the search is a sequence, where each square-bracketed item corresponds to a token. There can be no constraints on a token (empty square brackets) or multiple constraints:

[5] The list is not exhaustive: as was evident from presentations at the 1997 Summer School on Information Extraction in Frascati, systems at New York University and Sheffield University have similar characteristics, and there are also similarities with corpus search engines such as SARA and CorpusBench. There has also been a related proposal for a TIPSTER standard (Onyshkevych, 1996).

```
[lem= "keep" & pos="N.*"]
```

specifies the nominal lemma *keep* (as in the castle's keep). Note also the regular expression matching at two different levels: both over strings of characters on the RHS of the equals sign, so `[pos="N.*"]` matches `NN1` for singular noun or `NN2` for plural noun, and over strings of tokens, so an optional adjective is shown by `[pos="adj"]?`. Full regular-expression matching is supported at both levels in all three formalisms.

All three also have comparable mechanisms for referring to spans, or bracketings, so that, given S or NP markup in the input, tokens can be constrained to be in the same S or NP. They also all have mechanisms for specifying "within N words of". The write-ups all include similar arguments in favour of formalisms that are modular with respect to programs which add linguistic annotation, so that one can always add another variety of annotation when a new NLP tool becomes available.

6 WSD algorithms and spiral-bound features

In much c-WSD, a lexical entry is simply a list of collocates with weightings (which may be probabilities) attached for each sense. Then, the runtime algorithm takes each sense of the nodeword (i.e. the word to be disambiguated), and, for all the words in the context, combines the weightings. The output is the word with the highest weighting. (The artifice and ingenuity goes into the compile-time lexicon generation, not the runtime disambiguation.) Such algorithms have generally treated the context as a set or bag of words, irrespective of position with respect to the nodeword. They tend to use information from lower-frequency, content words, rather than higher-frequency grammar words, where positional information is critical. Leading work of this genre includes (Yarowsky, 1992; Gale, Church, and Yarowsky, 1993; Schütze, 1998).

The approach fares well when the different senses of the word occur in different domains, which tends to occur when they are at the 'homograph' end of the homograph/polysemy spectrum. To discriminate finer-grained senses, or to reach beyond an accuracy threshold, a richer feature set is required (Leacock, Towell, and Vorhees, 1993).

The research focuses on algorithms: most could be used with any feature set. One could in principle use <word, position> pairs as features (with position defined relative to the nodeword and ranging from, say, +3 to -3). Parts of speech, lemmas, bracketings are all sometimes optimal ways to express salient features for WSD. Sometimes it is a word sequence, such as preceding *of the* which is critical to disambiguation. In short, all features of the spiral-bound regular expression formalisms are of potential use for c-WSD. Of course, each extension comes with a cost, in terms of escalating numbers of features and correspondingly sparse data. It is only viable to extend the repertoire of features if one also introduces methods for determining which are salient for each word. Papers exploring this route in different ways are (Hearst, 1991; Leacock, Towell, and Vorhees, 1993; Yarowsky, 1995; Pedersen, Bruce, and Wiebe, 1997).

Note that if one sees the lexicon generation phase of c-WSD as a one-off, resource development activity, it becomes viable to spend substantially longer on it than if it is seen as a regularly-repeated compile-time activity.

7 Relation to theme of workshop, and way ahead

If lexical resources are to be customised, automatically or semi-automatically, then there is only one plausible source for the genre that the lexicon is to be customised to: the corpus. In this paper, **all** lexicography is understood as a process of generalising away from corpus data. Lexical tuning, then, sits comfortably within the whole spread of approaches to lexicon production. (A discussion of how this would integrate with existing lexical resources would be another paper.)

There is a gap between text-oriented tools for lexicon-generation, and sophisticated data-driven methods for producing profiles of the contexts of a word, phrase or word sense as in c-WSD research. Tools that have been widely used for corpus exploration rapid manual customisation of IE systems are converging on a 'spiral-bound regular expression formalism'.

The rich feature-set implicit in the formalism defines a search space. Within that search space lie all the features which should appear in the textual perspectives of the lexical entries for the language, for WSD or any other purposes. As lexicons are best understood as hierarchies, critical research questions include the discovery of hierarchical relations, such as $Verb > IntransVerb > \text{“}SLEEP''$ or $\text{“}ANIMAL'' > \text{“}DOG'' > \text{“}ALSATIAN''$ in the data. AI techniques, from machine learning, statistics and data-mining, are all relevant. Ingenious methods (possibly semi-automatic) will be required for identifying which of the vast number of features are salient for which words. c-WSD has started the exploration, but there is still a long way to go before the gap is bridged and lexical specifications can flow direct from corpus into lexicon.

References

Basili, Roberto, Michelangelo Della Rocca, and Maria Teresa Pazienza. 1997. Towards a bootstrapping framework for corpus semantic tagging. In *Proc. ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How?*, pages 66–73, Washington DC, April. ACL.

Béjoint, Henri. 1994. *Tradition and Innovation in Modern English Dictionaries.* OUP, Oxford.

Brent, Michael R. 1993. From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2):243–262.

Day, David, John Aberdeen, Lynette Hirschman, Robyn Kozierok, Patricia Robinson, and Marc Vilain. 1997. Mixed initiative development of language processing systems. In *Proc. Fifth Conference on Applied Natural Language Processing*, pages 348–355, Washington DC, April. ACL.

Doran, Christine, Michael Niv, Breck Baldwin, Jeffrey Reynar, and B. Srinivas. 1997. Mother of PERL: a multi-tier pattern description language. In *Proc. Workshop on Lexicon Driven Information Extraction*, pages 13–22, Frascati, Italy, July.

Gale, William, Kenneth Church, and David Yarowsky. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(1–2):415–439.

Grefenstette, Gregory. 1994. *Explorations in Automatic thesaurus discovery.* Kluwer, Dordrecht.

Hearst, Marti A. 1991. Noun homograph disambiguation using local context in large text corpora. In *Using Corpora: Proc. Seventh Ann. Conf. of the UW Centre for the New OED*, pages 1–22, Waterloo, Canada.

Heid, Ulrich and Katja Krüger. 1996. A multilingual lexicon based on frame semantics. In Lynne Cahill and Roger Evans, editors, *Proc. AISB Workshop on Multilinguality in the Lexicon*, pages 1–13, Brighton, England, April.

Kilgarriff, Adam. 1992. *Polysemy.* Ph.D. thesis, University of Sussex, CSRP 261, School of Cognitive and Computing Sciences.

Leacock, Claudia, Geoffrey Towell, and Ellen Vorhees. 1993. Towards building contextual representations of word senses using statistical models. In Branimir Boguraev and James Pustejovsky, editors, *Acquisition of Lexical Knowledge From Text: Workshop Proceedings*, pages 10–21, Ohio. ACL Special Interest Group on the Lexicon.

Lesk, Michael E. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proc. 1986 SIGDOC Conference*, Toronto, Canada.

Onyshkevych, Boyan. 1996. Pattern specification language overview. In TIPSTER *Phase 3 kickoff Meeting: presentation*, October.

Pedersen, Ted, Rebecca Bruce, and Janyce Wiebe. 1997. Sequential model selection for word sense disambiguation. In *Proc. Fifth Conference on Applied Natural Language Processing*, pages 388–395, Washington DC, April. ACL.

Schulze, Bruno and Oliver Christ, 1994. *The IMS Corpus Workbench.* Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Schütze, Hinrich. 1998. Word sense discrimination. *Computational Linguistics.* in press.

Viegas, Evelyne and Sergei Nirenburg. 1995. The semantic recovery of event ellipsis: its computational treatment. In *Proc. IJCAI Workshop on Context and Natural Language*, Montreal, August.

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

Yarowsky, David. 1995. Unsupervised word sense disambiguation rivalling supervised methods. In *ACL 95*, pages 189–196, MIT.