Putting the corpus into the dictionary

Adam Kilgarriff
Lexical Computing Ltd
adam@lexmasterclass.com

Abstract

A corpus is an arbitrary sample of language, whereas a dictionary aims to be a systematic account of the lexicon of a language. Children learn language through encountering arbitrary samples, and using them to build systematic representations.

These banal observations suggest a relationship between corpus and dictionary in which the former is a provisional and dispensable resource used to develop the latter. In this paper we use the idea to, first, review the Word Sense Disambiguation (WSD) research paradigm, and second, guide our current activity in the development of the Sketch Engine, a corpus query tool. We develop a model in which a database of mappings between collocations and meanings acts as an interface between corpus and dictionary.

1 Putting the dictionary in the corpus.

Consider SEMCOR, a hugely successful project and resource, very widely used and stimulating large amounts of WSD work. It is clearly a dynamic and important model, only exceeded in its take-up and impact by the WordNet database itself.

SEMCOR inserts dictionary sense labels into the corpus. It “puts the dictionary into the corpus”; like our title, but the other way round. Call this the PDIC model, as opposed to our preferred PCID model.

If one thinks of WSD as a task on the verge of hitting the marketplace and being widely used in applications, then PDIC is appropriate, as it represents the WSD task successfully done, and can be used as a model for what a system should do. However it is widely acknowledged that WSD is not in any such near-to-market situation (as shown by discussions at the SENSEVAL-3 workshop[1]) and we stand by our deep reservations about the nature of the WSD task (Kilgarriff 1997a, 1997b) which imply this is unlikely to change. An alternative model, closer to the observations of the opening paragraph, is that the larger task is at a further remove from applications and is better seen as lexical acquisition. We are not yet at a stage (and probably never will be) at which a general–purpose WSD module is a relevant goal, but there are many language interpretation tasks which cannot be done without richer lexical representations. In the PCID model, the corpus serves to enrich the lexicon.

1.1 Levels of abstraction

The direct approach to corpus-dictionary linkage is to put pointers to dictionary senses into the corpus (in the PDIC model, as in SEMCOR) or (in the PCID model) to put pointers to corpus instances of words into the dictionary. The direct approach has a number of drawbacks. The primary practical one is fragility. If the corpus (PCID model) or the dictionary (PDIC model) is edited or changed in any way, maintenance of the links is a headache. (This has been an ongoing issue for SEMCOR, as new versions of WordNet call for re-building SEMCOR in ways which cannot in general be fully and accurately automated; see Daudé et al (2000, 2001)). The theoretical one concerns levels of abstraction. A dictionary is an abstract representation of the language, in which we express differences of meaning but are not engaged with specifics of differences of form. The corpus is at the other end of the scale: the differences of form are immediately present but differences of meaning can only be inferred. What is needed is an intermediate level which links both to the meaning-differences in the dictionary and to the specific instances in the corpus.

Our candidate for this role is the grammatical-relation triple, comprising <*grammatical-relation, word1, word2*> (examples are <*object, drink, beer*> and <*modifier, giant, friendly*>). Triples such as

[1] Notes available at http://www.senseval.org/senseval3

these[2] have, of late, been very widely used in NLP, as focal objects for parsing and parse evaluation (*eg* Briscoe and Carroll 1998, Lin 1998), thesaurus-building (*eg* Grefenstette 1992) and for building multilingual resources from comparable corpora (*eg* Lű and Zhou 2004). Approaching from the other end, they are increasingly seen as core facts to be represented in dictionaries by lexicographers, who usually call them just collocations (COBUILD 1987, Oxford Collocations Dictionary 2003). In our own work, we compile sets of all the salient collocations for a word into 'word sketches', which then serve as a way of representing corpus evidence to the lexicographer (Kilgarriff and Tugwell 2001, Kilgarriff and Rundell 2002, Kilgarriff et al 2004).

1.2 Aside: Parsing and lemmatizing

Few would dispute that collocations which incorporate grammatical information (and are thereby triples like <*object, drink, beer*>) are a more satisfactory form of lexical object than 'raw' collocates – those words occurring within a window of 3 or 5 words to the right of, or to the left of, or on either side of the node word. Windowing approaches operate as a proxy for grammatical information and are appropriate only where there is no parser available, or it is too slow, or too often wrong. Historically, these factors have often applied and most older work uses windowing rather than grammar. As we are able to work with grammar, we do. We are repeatedly struck by how much cleaner results we get. We also find it preferable to work with lemmas rather than raw word forms, where a lemmatiser is available for a language.

1.3 Terminology

"Grammatical relation triples" being unwieldy, I shall call these objects simply "collocations", or say that the one word is the other's "collocate". Strictly, the items in the triples are lemmas which include a word class label (*noun, verb, adj* etc) as well as the base form; in examples, the word class labels will be omitted for readability. Naturally, some grammatical relations are duals (object, object-of) or symmetrical (and/or); for a full treatment see Kilgarriff and Tugwell (2001).

2 The collocation database

In the proposed model, between the dictionary and the corpus sits a database. For each dictionary headword, there is a set of records in this database comprising

1) a collocate (including grammatical relation)
2) a pointer to the dictionary sense
3) a set of pointers to corpus instances of the headword in this collocation

The database is, in the first instance, generated from the corpus, so the corpus links are immediately available. To start with, the dictionary pointers are not filled in for polysemous words. (For monosemous words, the links can immediately be inserted.) A word sketch (see Fig 1) is an example of such a database. The corpus links are present, implemented as hyperlinked URLs: for on-line readers, clicking on a number opens up a concordance window for the collocation (go to www.sketchengine.co.uk to self-register for an account).

2.1 Limitations and potential extensions

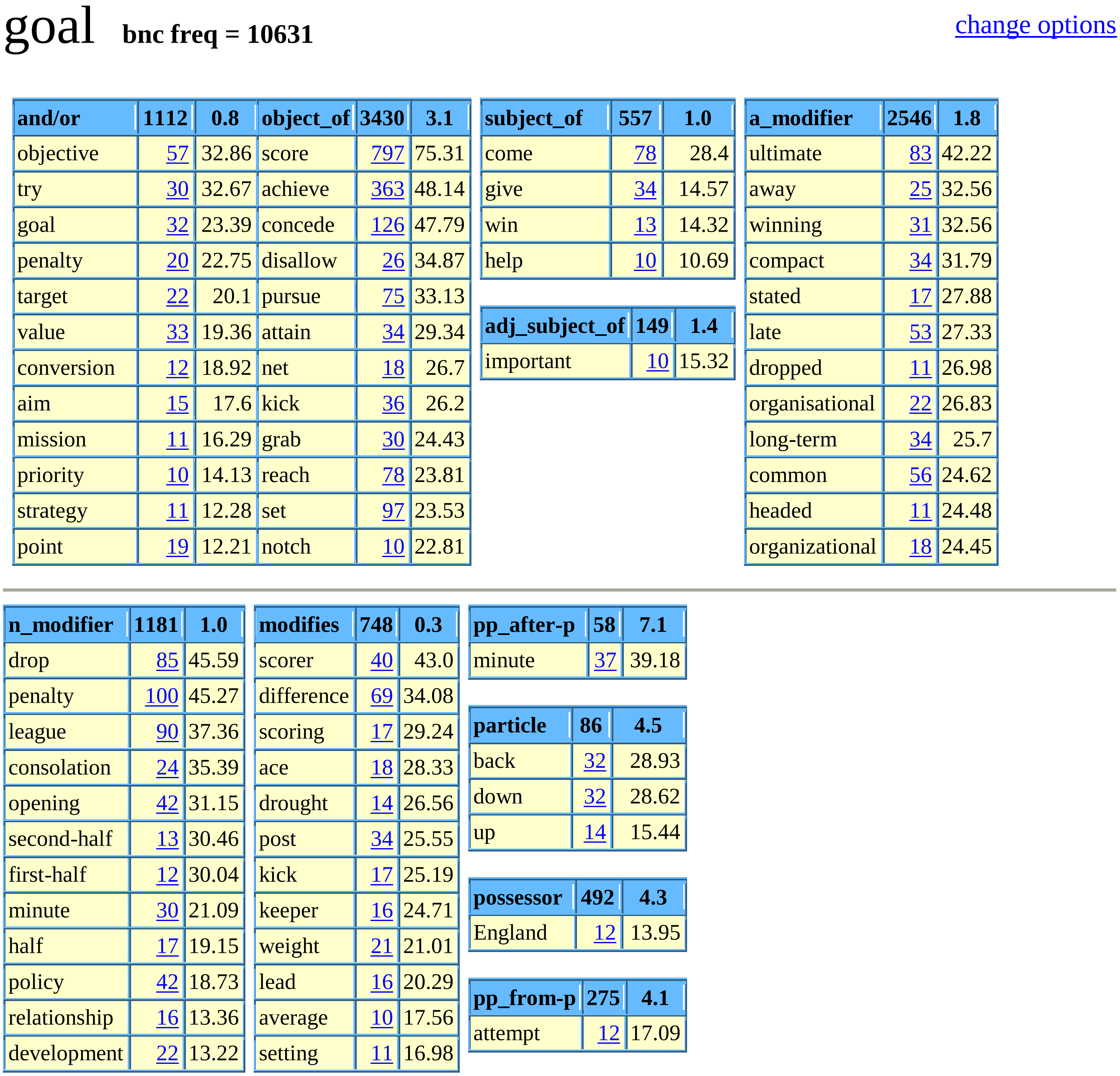
The word sketch model is dependent on Yarowsky's "One sense per collocation" (Yarowsky 1993). To the extent that this does not hold, the model will be inadequate and we will need to make the structure of the database record richer.

The triples formalism does not readily express some kinds of information which are known to be relevant to WSD. An intermediate database to link dictionary to corpus should have a place for all relevant facts. Two kinds of fact which do not obviously fit the triples model are grammatical constructions, and domain preferences.

Many, possibly all, grammatical constructions can be viewed as grammatical relations (with the "other word" field null). Thus a verb like *found*, when in the passive, means "set up" ("the company was founded in 1787"). In this case we associate the triple <*passive, found,* _> with the "set up" meaning. We have already implemented a range of "unary relations" within the Sketch Engine, and believe this approach will support the description of all grammatical constructions.

As much recent work makes clear, domains are central to sense identification (*eg* Agirre *et* al 2001, Buitelaar and Sacaleanu 2001, McCarthy *et al* 2004). However it is far from clear how domain

[2] Naturally, details vary between authors. Briscoe and Carroll do not in fact use triples, but tuples with further slots for particles and further grammatical specification.



goal **bnc freq = 10631** change options

and/or	1112	0.8
objective	57	32.86
try	30	32.67
goal	32	23.39
penalty	20	22.75
target	22	20.1
value	33	19.36
conversion	12	18.92
aim	15	17.6
mission	11	16.29
priority	10	14.13
strategy	11	12.28
point	19	12.21

object_of	3430	3.1
score	797	75.31
achieve	363	48.14
concede	126	47.79
disallow	26	34.87
pursue	75	33.13
attain	34	29.34
net	18	26.7
kick	36	26.2
grab	30	24.43
reach	78	23.81
set	97	23.53
notch	10	22.81

subject_of	557	1.0
come	78	28.4
give	34	14.57
win	13	14.32
help	10	10.69

adj_subject_of	149	1.4
important	10	15.32

a_modifier	2546	1.8
ultimate	83	42.22
away	25	32.56
winning	31	32.56
compact	34	31.79
stated	17	27.88
late	53	27.33
dropped	11	26.98
organisational	22	26.83
long-term	34	25.7
common	56	24.62
headed	11	24.48
organizational	18	24.45

n_modifier	1181	1.0
drop	85	45.59
penalty	100	45.27
league	90	37.36
consolation	24	35.39
opening	42	31.15
second-half	13	30.46
first-half	12	30.04
minute	30	21.09
half	17	19.15
policy	42	18.73
relationship	16	13.36
development	22	13.22

modifies	748	0.3
scorer	40	43.0
difference	69	34.08
scoring	17	29.24
ace	18	28.33
drought	14	26.56
post	34	25.55
kick	17	25.19
keeper	16	24.71
weight	21	21.01
lead	16	20.29
average	10	17.56
setting	11	16.98

pp_after-p	58	7.1
minute	37	39.18

particle	86	4.5
back	32	28.93
down	32	28.62
up	14	15.44

possessor	492	4.3
England	12	13.95

pp_from-p	275	4.1
attempt	12	17.09

Fig 1. Word sketch for *goal* (reduced to fit in article)

information should be expressed: hand-developed inventories of domains have many shortcomings, but data-driven approaches to domain induction are not yet mature and suffer from the arbitrariness of the corpora they use. The incorporation of domain information into the database model remains further work.

Whereas a collocate tends to be associated with one and only one sense, so can be used to generate a Boolean rule of the form “collocation X implies sense Y”, both grammatical constructions and domains provide preferences. *Royalty* (singular) usually means kings and queens, whereas *royalties* (plural) usually means payments to authors. However a rule “singular implies kings-and-queens” should not be Boolean: we often talk about, eg, "royalty payments" which are payments to authors, not to (or from) kings and queens. The facts are preferences, or probabilistic, rather than categorical. Our current model does not incorporate preferences or probabilities, and they raise theoretical problems: are the probabilities not as arbitrary as the corpora they were drawn from? This, again, is further work.

The formalism will allow Boolean combinations of triples and of senses, so it is possible to say, eg, "triple X AND triple Y imply NOT sense Z". We envisage that unary relations (eg, grammatical constructions) will often be used to rule out senses, or in conjunction with collocates.

Once solutions to the domains issue are found, we will be able to view the database connecting corpus to dictionary as a database of collocations, constructions and domains: a CoCoDo database.

2.2 Linking collocations to senses

There are a number of ways in which the pointers to dictionary senses might be added. Over the last forty years the WSD community has developed a host of strategies for assigning collocates to dictionary senses (Ide and Véronis 1998, Kilgarriff and Palmer 2000, SENSEVAL-2 2001, SENSEVAL-3 2004). Many of them can be applied (depending, obviously, on the nature of the dictionary and the information it provides).

We have specified the problem as the disambiguation of collocates rather than corpus instances. In practice, collocates (more widely or narrowly construed) are the workhorse of almost all WSD. The core is of identifying a large set of collocates (or, more broadly, sentence patterns or text features) which are associated with just one of the word's senses, which then may be found in a sentence or text to be disambiguated. The task of assigning collocates is a large part of the task of assigning instances.

Other differences between the task as specified here and the traditional WSD task are as follows.

1) **Dictionary structure:** We can link to any substructure of the dictionary entry; if the entry has subsenses, or multiwords embedded within senses or *vice versa*, we can link to the appropriate element, so need not make invidious choices about whether to use 'fine-grained' or 'coarse-grained' senses.
2) **Other dictionary information:** Since the larger goal is enrichment of lexical resources, where a resource is already rich, the information it contains is given. It can be used in WSD, and does not need to be duplicated. One resource we have looked closely at, the database version of Oxford Dictionary of English (McCracken 2003), contains particularly full information on domain, taxonomy, multiwords, grammatical and phonological patterning etc., all sense-specific. This is all immediately available, both for disambiguation and, obviously, in the output resource.
3) **Precision-recall tradeoff:** There is no commitment to disambiguating all corpus instances (or all collocates). Like many NLP tasks, WSD exhibits a precision-recall tradeoff. If a system need not commit itself when the evidence is not clear, it can achieve high accuracy for those cases where it does present a verdict. WSD has usually been conceptualised as a task where a choice must be made for every instance (so precision=recall) and in the PDIC model this seems appropriate. But in the PCID model it is not necessary. What we would like is **some** corpus-based information about all dictionary senses, and it is immaterial if there are some corpus instances which do not contribute to any lexical entry. Once we view the WSD task in this light, we welcome high-precision, low-recall strategies (for example Magnini et al 2001, which achieved precision 5% higher than the next highest-precision system in the SENSEVAL-2 English all-words task, with 35% recall). We can do WSD without the shadow of an apparent 60% precision ceiling (SENSEVAL-3 2004) hanging over us.
4) **Mixed-initiative methods** Once WSD is seen as a step towards the enrichment of lexical resources, it becomes valid to ask how humans may be involved in the process. Kilgarriff and Tugwell (2001), and Kilgarriff, Koeling, Tugwell and Evans (2003) present a system in which a lexicographer assigns collocates to senses, and this then feeds Yarowsky (1995)'s

decision-list learning algorithm. In general, in the proposed architecture, both people and systems can identify collocate-to-sense mappings, with each potentially learning from evidence provided by the other and correcting the other's errors or omissions. (There will be a set of issues around permissions: which agents (human or computer) can add or edit which mappings.) Ideally, the process of identifying the mappings for a word is a mixed-initiative dialogue in which the lexicographer refines their analysis of the word's space of meaning in tandem with the system refining, in real time, the WSD program which allocates instances to senses and thereby provides the lexicographer with evidence.

2.3 Dictionary-free methods

While most WSD work to date has been based on a sense inventory from an existing resource, some (eg Schűtze 1998) has used unsupervised methods to arrive, bottom-up, at its own senses inventory.

If the PCID model is being used to create a brand new dictionary, or if a fresh analysis of a word's meaning into senses is required, or if some dictionary-independent processing is required as a preliminary or complement to a dictionary-specific process, then dictionary-free methods are suitable. Methods such as Schűtze's are based on clustering instances of words. Our strategy will be to cluster collocates. One method we have already implemented uses the thesaurus we have created from the same parsed corpus as was used to create the word sketches. Looking at the verbs that *goal* is object of, in Figure 1, we see a number of verbs with closely related meanings, and we would ideally like to form them into two clusters, one for sporting *goal*s and one for life *goal*s (these being the two main meanings of *goal).* In the thesaurus entry for *disallow*, we find, within the top ten items, *concede* and *net,* thus providing evidence that these three items cluster together.

Another method we shall be implementing shortly depends on the observation that a single instance of a word may exemplify more than one collocation. The instance "score a drop goal" exemplifies both <*object, score, goal*> and <*modifier, goal, drop*> so provides evidence that these two collocations should be mapped to the same sense.

Collocate-clustering is best seen as a partial process, marking collocates as sharing the same sense only when there is strong evidence to do so and remaining silent elsewhere. It then provides good evidence for other processes, dictionary-based or manual, to build on.

3 The dispensable corpus

As mentioned in the opening paragraph, a corpus is an arbitrary sample. A person's mental lexicon, while developed from a set of language samples, has learnt from them and moved on.[3] The corpus is dispensable. In a PDIC approach, this clearly does not apply: if the corpus is thrown away, all the evidence linking dictionary to corpus is lost too. Likewise for a PCID approach with direct corpus-dictionary linking. But in the model presented here, if the corpus is thrown away, the collocate-to-sense mappings are rich, free-standing lexical data in their own right (and could readily be used to find new corpus examples for each collocate or sense).

4 WordNet proposal

The paper is largely programmatic. We have, as indicated above, starting exploring a number of the ideas, using the Sketch Engine (http://www.sketchengine.co.uk) platform and its predecessor, the WASPbench. We now want to develop it further, and are considering which dictionary (if any) to develop it with. (The Sketch Engine identifies all items –collocations, triples, word instances- as URLs, thereby supporting distributed development, open access, and connectivity with other resources.)

Dictionary-free development is attractive and under discussion, but, to develop a rich and accurate resource, a large investment will be required. It is unlikely the resulting resource would be in the public domain.

Collaborations with dictionary publishers, to enrich their existing dictionaries, are under discussion. They too would not give rise to public-domain resources.

[3] The success of the memory-based learning paradigm, in both NLP and psycholinguistics, may be seen as casting doubt on this claim.

For the development of the idea within the academic community, a public domain resource is wanted. The obvious candidate is WordNet. The proposal is then to develop a collocations database with links to WordNet senses, on the one hand, and collocates found statistically in a large corpus on the other. The WordNet links would be identified using the whole range of disambiguation strategies which have been developed for WordNet (including, potentially, the multilingual and web-based ones). We believe this could be a resource that takes forward our understanding of words and language and which supports a wide range of NLP applications.

References

Agirre E., Ansa O., Martínez D., Hovy E. Enriching WordNet concepts with topic signatures. In *Proceedings of the SIGLEX workshop on "WordNet and Other Lexical Resources: Applications, Extensions and Customizations". NAACL, 2001*. Pittsburgh

Briscoe E. J. and J.Carroll 2002. Robust accurate statistical annotation of general text. In Proc LREC 2002.

Buitelaar P. and B. Sacaleanu. 2001. Ranking and selecting synsets by domain relevance. In *Proceedings of the SIGLEX workshop on "WordNet and Other Lexical Resources: Applications, Extensions and Customizations", NAACL 200*, Pittsburgh.

COBUILD 1987. Collins COBUILD English Dictionary. Ed. J. Sinclair.

Daudé J., Padró L. and Rigau G. 2000. Mapping WordNets Using Structural Information Proc ACL. Hong Kong.

Daudé J., Padró L. and Rigau G. 2001. A Complete WN1.5 to WN1.6 Mapping, Proc NAACL Workshop "WordNet and Other Lexical Resources: Applications, Extensions and Customizations". Pittsburg, PA.

Grefenstette, G. 1992. "Sextant: exploring unexplored contexts for semantic extraction from syntactic analysis" Proc ACL, Newark, Delaware: 324--326.

Ide, N. and J. Véronis, Editors. 1998. Special issue on word sense disambiguation: The state of the art. Computational Linguistics, 24(1).

Kilgarriff, A. 1997a "I don't believe in word senses." Computers and the Humanities 31: 91-113.

Kilgarriff, A. 1997b. "What is Word Sense Disambiguation Good For?" Proc. NLPRS: Phuket, Thailand.

Kilgarriff, A., R. Koeling, D. Tugwell, R. Evans 2003. "An evaluation of a lexicographer's workbench: Building lexicons for machine translation." Workshop on MT tools, European ACL, Budapest.

Kilgarriff, A., P. Rychly, P. Smrz and D. Tugwell 2004. "The Sketch Engine" Proc. Euralex. Lorient, France, July: 105-116.

Kilgarriff A. and M. Rundell 2002. "Lexical profiling software and its lexicographic applications - a case study." Proc EURALEX, Copenhagen, August: 807-818.

Kilgarriff A. and D. Tugwell 2001. "WASP-Bench: an MT Lexicographers' Workstation Supporting State-of-the-art Lexical Disambiguation". Proc MT Summit VIII, Santiago de Compostela, Spain: 187-190.

Kilgarriff, A. and M. Palmer 2000. Editors, Special Issue on SENSEVAL. Computers and the Humanities 34 (1-2).

Lin, D. 1998. A Dependency-based Method for Evaluating Broad-Coverage Parsers. Journal of Natural Language Engineering.

Lü, Y. and Zhou, M. 2004. Collocation Translation Acquisition Using Monolingual Corpora Proc ACL 2004, Barcelona: 167-174.

McCarthy, D., Koeling, R., Weeds, J. and Carroll, J. (2004) Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics.* Barclona, Spain. pp 280-287

McCracken J. and A. Kilgarriff. 2003. Oxford Dictionary of English - current developments. Proc. EACL.

Magnini, B., Strapparava, C., Pezzulo, G. and Gliozzo, A. 2001. Using Domain Information for Word Sense Disambiguation. In Proc. SENSEVAL-2: 111-114.

Oxford Collocations Dictionary for Students of English. 2003. Ed. Lea. OUP.

Schűtze, H. 1998. Automatic Word Sense Discrimination in Ide and Véronis 1998, op cit.

SENSEVAL-2 (2001) See http://ww.senseval.org

SENSEVAL-3 (2004) See http://ww.senseval.org

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

Yarowsky, D. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. Proc. ACL: 189-196.