ISSN 1744-1986

Technical Report N° 2005/02

Using a Distributional Thesaurus to Resolve Coordination Ambiguities

Francis Chantree
Adam Kilgarriff
Anne De Roeck
Alistair Willis

25th February 2005

Department of Computing
Faculty of Mathematics and Computing
The Open University
Walton Hall,
Milton Keynes
MK7 6AA
United Kingdom

http://computing.open.ac.uk

Using a Distributional Thesaurus to Resolve Coordination Ambiguities

Abstract

We present a novel method for resolving coordination ambiguities. This type of ambiguity is one of the most pervasive and challenging. We test the hypothesis that the most likely reading of a coordination ambiguity can be indicated by the distributional similarity of terms. Our experiments show that words or phrases in a coordination which have distributional similarity also tend to have "coordination first" characteristics.

1 Introduction

Coordination ambiguity is a structural (i.e. syntactic) ambiguity. Compared with prepositional phrase (PP) attachment ambiguity, which is also a structural ambiguity, it has received little attention in the literature. This is despite the fact that coordinations are known to be a "pernicious source of structural ambiguity in English" (Resnik, 1999).

We test the hypothesis that the preferred reading of a coordination ambiguity can be predicted by looking at distributional similarities between the head words of phrases that are coordinated. The hypothesis states that phrases with distributional similarity are likely to be coordinated before modifiers of those phrases takes scope. For example, in the phrase

> *old boots and shoes,*

the fact that *boots* and *shoes* have strong distributional similarities suggests that they are likely to be a syntactic unit. In other words the coordination would be performed first, before the modifier *old* takes effect. This idea is suggested in (Kilgarriff, 2003).

In this paper, we test whether this hypothesis is true for a set of coordination ambiguities drawn from requirements engineering - a suitable domain where misunderstandings can lead to costly mistakes.

We extract a collection of sentences from our tagged corpus of requirements specification documents, with each sentence containing one coordination ambiguity. For each of these, we identify preferred readings by means of an ambiguity survey, where we ask participants to express a judgement. In this way, we obtain a consensus judgement for each example, and this forms the evaluation dataset for our experiments.

For each of the sentences, we investigate the distributional similarity between the head words involved in the coordination by comparing rankings produced by the Sketch Engine[1] (Kilgarriff et al., 2004), which obtains distributional information from the British National Corpus[2]. Where matches are found, we predict a coordination first reading. We introduce two further variants of the base line experiment. We evaluate all results against our test set of consensus judgements.

We first describe the coordination ambiguity problem and related research. We then outline how we create our evaluation dataset and describe our experimentation methodology. This is followed by description of our experiments, evaluation of the

[1] http://www.sketchengine.co.uk

[2] http://natcorp.ox.ac.uk

results, conclusions, and some ideas for future research.

2 Coordination Ambiguity

Coordinating conjunctions are potentially a widespread problem as they are common in English. Together, *and* and *or* account for approximately 3% of the words in the British National Corpus. The former account for (87.07%) and the latter for (10.34%) of the conjunctions extracted from a corpus of biomedical abstracts (Nenadic et al., 2004). We confine our investigations to *and*, *or* and *and/or*.

In the phrase

old boots and shoes

the external modifier *old* applies either to both the boots and the shoes or to just the boots. We refer to the former case as “coordination first”, and to the latter case as “coordination last”[3]. In our experiments we will concentrate on coordinations of this type where at least two readings are possible.

There are other types of coordinations which are not of interest to us. For instance, although words of almost all types can be coordinated, and the external modifier can also be a word or phrase of almost any type, some coordinations are not syntactically ambiguous. The coordination of *yellow* and *red* in the sentence

a yellow and red flag,

is one such example. This is because the adjective *yellow* by itself cannot be modified by a determiner. We do not include syntactically unambiguous sentences in our surveys.

Phrases with dissimilar parts of speech can be coordinated, such as

Joe is athletic and a fine sportsman.

Here, an adjective and a noun phrase are coordinated. However, these are not so common and, usually, when words such as *and* and *or* have dissimilar phrases on either side they will be acting as “connectors” rather than coordinating conjunctions. For this reason, such coordinations are not of interest to us, and we exclude such sentences from our surveys. The full set of the criteria that we use for eliminating certain usages of coordinating conjunctions is given in table 3.

3 Related Research

Previous research has tackled coordination ambiguity in a variety of different ways.

(Agarwal and Boggess, 1992) present an algorithm that attempts to identify which phrases are coordinated by coordinating conjunctions. They achieve an accuracy rate of 81.6% for the conjunctions *and* and *or*. Their method identifies parts of speech and case labels of the head words of the phrases. This means, however, that adjectives and other modifiers are subsumed into phrases without consideration of their possible effect on the more distant of the coordinated phrases. Their method is a potentially useful, and relatively straightforward, way of matching candidate coordinated phrases, though it does not deal adequately with coordination last readings.

(Kurohashi and Nagao, 1992) present a method of analysing coordination structures in Japanese, with the aim of parsing them more successfully. They use word similarity as part of their method. This leads them to simplifying long sentences into several shorter ones. We are not seeking to alter text, but their use of similarities means that it is interesting compare their performance with our own. They achieve a precision of 81.3%.

(Goldberg, 1999) uses unsupervised learning to determine the attachment of ambiguous coordinate phrases. She simplifies the text using a chunker, and then extracts the headwords of the coordinated phrases. Her data, which is unannotated, includes a lot of noise. Also, as her method is a straightforward re-implementation of a PP-attachment method (Ratnaparkhi, 1998), it cannot model useful information that is specific to coordination ambiguity. However, she is one of very few researchers who present an actual coordination ambiguity disambiguation system. Goldberg achieves an accuracy of 72%.

(Resnik, 1999) is the contribution that is of most relevance to our research. He investigates the role of semantic similarity in resolving coordination am-

[3] Other terminology can be used, e.g. “low attachment” and “high attachment”, depending on where the second coordinated word attaches in the parse tree (Goldberg, 1999). We believe, however, that our terminology is better suited to our task.

biguities involving nominal compounds. (Note that this is not the same as the distributional similarity which we use.) An example of such a construction is

a bank and warehouse guard

where the guard guards either both places or just the warehouse. Resnik uses several heuristics to disambiguate such constructions, one of which is semantic similarity. He looks up the nouns in WordNet and determines which of the classes that subsume them both has the highest information content. Nouns that cannot be found in WordNet, are treated as instances of the most general class. Without using other back-off strategies, or further analysis, this procedure results in 71.2% precision and 66.0% recall. He does not look at other types of coordination ambiguities, so his dataset is much more focused than ours.

4 Developing an Evaluation Dataset

4.1 Human Judgements

Ambiguity is speaker- and context-dependent, and so there are no absolute criteria for judging it. Therefore, we capture human judgements about the ambiguity of the examples in our surveys in order to form our evaluation dataset. Rather than rely upon the judgement of one human reader, we take a consensus opinion from multiple readers. Such an approach is known to be a very effective method albeit an expensive one (Berry et al., 2003). We have found that people's perceptions of ambiguity can vary widely. For instance, let us say that severe disagreement occurs when 20% of participants choose a reading other than the majority one, ignoring all instances where the coordination is judged to be ambiguous. Over a quarter of the sentences in our coordination ambiguity surveys show such severe disagreement.

4.2 The Ambiguity Surveys

The sentences in our ambiguity surveys are drawn from our corpus of requirements specifications. Sentences - or non-sentential titles, bullet points etc - that contain coordinating conjunctions are identified. All such constructions are referred to as "sentences" from here on. We do not include all the sentences containing a coordination that we find. The heuristics that we used to eliminate sentences from our surveys are listed in Table 3. A breakdown of these sentences by the head word type of the coordinated phrases, with some examples, is given in Table 1. A breakdown of the sentences by the head word type of the external modifier is given in Table 2.

Head Word	% of Total	Example from Surveys
Noun	86.5	Communication and performance requirements
Verb	11.5	Proceed to enter and verify the data
Adjective	1.9	It is very common and ubiquitous

Table 1: Breakdown of Sentences by Head Word Type

Modifier	% of Total	Example from Surveys
Adject	42.3	.... define architectural components and connectors
Noun	25.0	(It) targeted the project and election managers
Prep	19.2	Facilitate the scheduling and performing of works
Relative	5.8	Assumptions and dependencies that are of importance
Adverb	5.7	(It) might be automatically rejected or flagged
Other	1.9	increased by the lack of funding and local resources

Table 2: Breakdown of Sentences by Modifier Type

We extracted 52 suitable coordination constructions and showed them to 17 participants, in 2 separate surveys. They were asked to judge whether the coordinated expression was coordination first, coordination last or "ambiguous so that it might lead to misunderstanding". In the last case, the coordinated expression is then classed as an "acknowledged ambiguity" for that participant. Clearly, there is an elusive dividing line between what would and what would not lead to genuine misunderstandings. We take the view that, by getting a sufficient numbers of participants, we obtain a fairly reliable consensus about where this line lies for each example.

4.3 Upper and Lower Bounds

We use percent agreement, as defined in (Gale et al., 1992), as an indication of the extent to which the participants are "of one voice". It is the percentage of the number of judgements that agree with the majority opinion. For our study it is 59.2%. However, this includes all the judgements of acknowledged ambiguity. If one removes these, leaving only the judgements that were deemed by the participants to be sufficiently clear-cut to be able to be judged, the figure rises to 86.7%. This effectively turns a judgement between three alternatives into a judgement be-

Reason for Exclusion	Example	Explanation
Entire sentences coordinated (conjunction used as "connector")	*I fell over and everyone laughed*	No external modifying element
1 of the coordinated phrases can't stand alone & make syntactic sense	*a yellow and red flag*	Only 1 syntactic reading: coordination first
Phrases with dissimilar head words are coordinated	*Joe is athletic and a fine sportsman*	head words with dissimilar parts of speech can't be looked up in thesaurus
Premodification, and 2nd coordinated phrase beginning with determiner	*I ate green beans and the sausages*	A modifier cannot premodify a determiner - except when it's a premodifer, e.g. *all*
Premodification, and 2nd coordinated phrase beginning with a pronoun	*I like tall women and her over there*	Modifier can't usually premodify pronoun
Coordinated phrases with the same head word	*I like green beans and red beans*	A word cannot be matched to itself in the thesaurus
Coord'd nouns have different number & followed by present tense verb	*boots and a raincoat are essential*	Only 1 syntactic reading: coordination first
Bracketings, and other types of punctuation	*I like green beans (and sausages)*	Such punctuation devices usually signify "asides", which are not affected by external modifiers. Many borderline cases though
A coordinated word was a company or proprietary name	"Bloggsystems"	Must be kept anonymous: therefore requires substitution with a dummy word that would skew the results with repeated use

Table 3: Heuristics Used to Exclude Coordinations from Our Study

tween two. We will take this figure to be the upper bound for our automated system. To our knowledge, there are no comparable statistics in the literature for coordination disambiguation, so we compare our results with those in PP-attachment disambiguation. PP-attachment research is similar as it is also concerned with attachment of syntactic units. The human disambiguation performance figures of (Ratnaparkhi et al., 1994) are often quoted as upper bounds in this field. Their figure for human performance agreement with the consensus view is 92.5%. Our human participants showed somewhat less agreement, suggesting that coordination ambiguity is a more difficult problem than PP-attachment ambiguity.

For the lower bound for our performance figures, we assign the most likely reading in all instances. In our case, the coordination first reading is the most likely one in 75% of the sentences. All reasonable systems should hopefully outperform this baseline.

5 Our Experiments

5.1 The Sketch Engine

The Sketch Engine thesaurus is a distributional thesaurus in the tradition of (Sparck-Jones, 1986) and (Grefenstette, 1994); it measures similarity between any pair of words according to the number of corpus contexts they share. The corpus is parsed and all triples comprising a grammatical relation and two collocates, (eg ⟨*object, drink, wine*⟩ or ⟨*modifier, wine, red*⟩) are identified. Contexts are shared where the relation and one collocate remain the same, so ⟨*object, drink, wine*⟩ and ⟨*object, drink, beer*⟩ count towards the similarity between wine and beer. Shared collocates are weighted according to the product of their mutual information, and the similarity score is the sum of these weights across all shared collocates, as in (Lin, 1998). Distributional thesauruses circumvent a host of difficult questions about the nature of meaning (including pairs of opposites black/white, old/young as near neighbours): as argued in (Kilgarriff, 2003), this is altogether helpful for NLP tasks such as ours. Initial evidence that distributional thesauruses may outperform manual ones (such as WordNet or Roget) for NLP tasks is provided by (McLauchlan, 2004) and (Calvo et al., 2005).

5.2 Method

The head words in each coordinated phrase are looked up in the Sketch Engine's thesaurus. Lemmatised verbs, nouns and adjectives can be entered. The thesaurus gives the words with the most distributional similarity, up to a cutoff limit which one can specify. We choose cutoff limits to get even distribution on a log scale. We use only the rankings of matches given by the similarity measure, and not the actual similarity values. Research has shown that the ranking of the matches in a distributional thesaurus is a more useful metric than the actual similarity measures (McLauchlan, 2004). One reason for this is that the thesaurus tends to weight less similar words too highly (McLauchlan, 2004). Also, common words generate high similarity measures for more matches than less common words tend to. Therefore, the strength of matches cannot easily be compared using the similarity measure alone.

5.3 Refinements to the Method

We investigate two alternative ways of using our data.

Firstly, instead of entering lemmas, we match the underlying verbs of the coordinated words, where underlying verbs exist. This is to avoid problems of sparseness, in cases where lemmas are too rare to give reliable lists of matches. This procedure has been possible in 69% of the sentences. Where no underlying verbs can be found, we use a back-off procedure of reverting back to the original words. We also use the back-off procedure when either of the underlying verbs is archaic or unrelated to the word under examination. This is done to avoid getting irrelevant data about words that are not in common usage with the meaning that is intended.

The second approach excludes from our evaluation dataset those sentences that are strongly shown to contain acknowledged ambiguities. Sentences for which the number of acknowledged ambiguity judgements exceeds both the number of coordination first judgements and the number of coordination last judgements fall into this category. The sentences that remain should contain less perceived ambiguity, and therefore the thesaurus matching method should give a higher performance.

5.4 Performance

True positives in this study are sentences that yield thesaurus matches and are judged to be coordination first coordinations, more often than either of the other two alternatives, by our participants. We calculate precision as the number of true positives divided by the total number of sentences where a thesaurus match was found. We calculate recall as the number of true positives divided by the total number of coordinations that were judged to be coordination first coordinations by our participants. Graphs showing our results for precision, recall and f-measure are shown in Figures 1, 2 and 3 respectively.

When lemmas were entered into the thesaurus, very good precision was achieved for the topmost matches, but poor recall. The lines for "lemma, including ambiguities" in Figures 1, 2 and 3 show this. Unsurprisingly, precision declined and recall increased when the number of matches considered was increased. Precision is much more important to us than recall: we wish our thesaurus matching method to be simply an indicator of how coordinations should be read rather than a catch-all method. We envisage using this method as one of a toolbox of heuristics which will disambiguate many coordinations with good precision. To ensure that we do not obtain many true positives at the expense of also obtaining a lot of false positives, we choose to use an f-measure with a weighting of α = 0.9, strongly in favour of precision:

$$FMeasure = \frac{1}{\alpha\frac{1}{Precision} + (1-\alpha)\frac{1}{Recall}}$$

When underlying verbs were entered into the thesaurus (using the back-off where necessary), precision improved with larger numbers of matches, though it was worse when few matches were considered. The lines for "u/l verb (including ambiguities)" in Figures 1, 2 and 3 are of relevance here. Recall was approximately the same, giving a combined f-measure that was higher except when considering a very few matches.

When we remove from the data all the lines that were judged to be acknowledged ambiguities more frequently than either of the other options, the results remain similar. The lines for "lemma (excluding ambiguities)" in Figures 1, 2 and 3 are of relevance here. The precision is only better than the lemma results when ambiguities were included when the number of matches considered is 100, and the recall is roughly the same. The f-measure line for excluding ambiguities with lemmas gives average performance, except when large numbers of matches are considered where it shows the poorest performance of all.

Precision for underlying verbs excluding ambiguities is uniformly worse than when underlying verbs are used and ambiguities are not excluded. However, recall for the former is uniformly better than for the latter. The lines for "underlying verb (excluding ambiguities)" in Figures 1, 2 and 3 are of relevance here. The f-measure for underlying verbs excluding ambiguities is the best indicator of all.

5.5 Evaluation

The approximately logarithmically linear improvement of all the recall statistics show that the the-

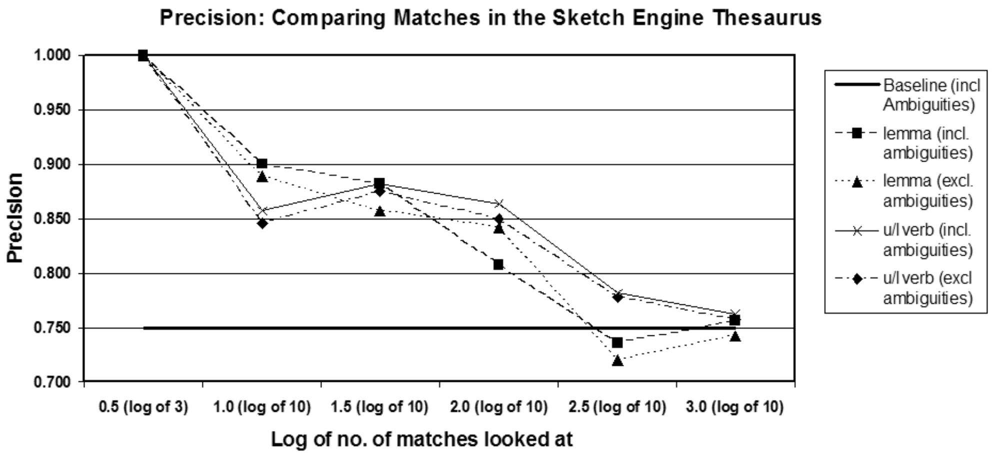

Figure 1: Precision

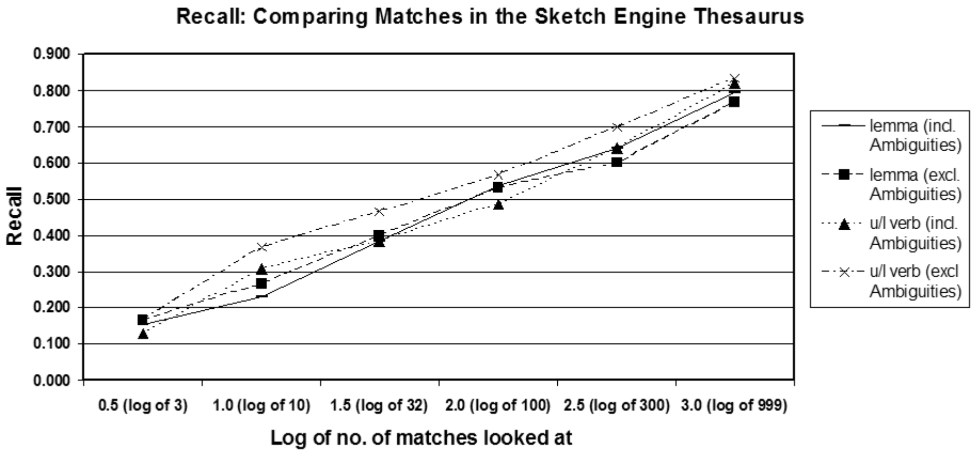

Figure 2: Recall

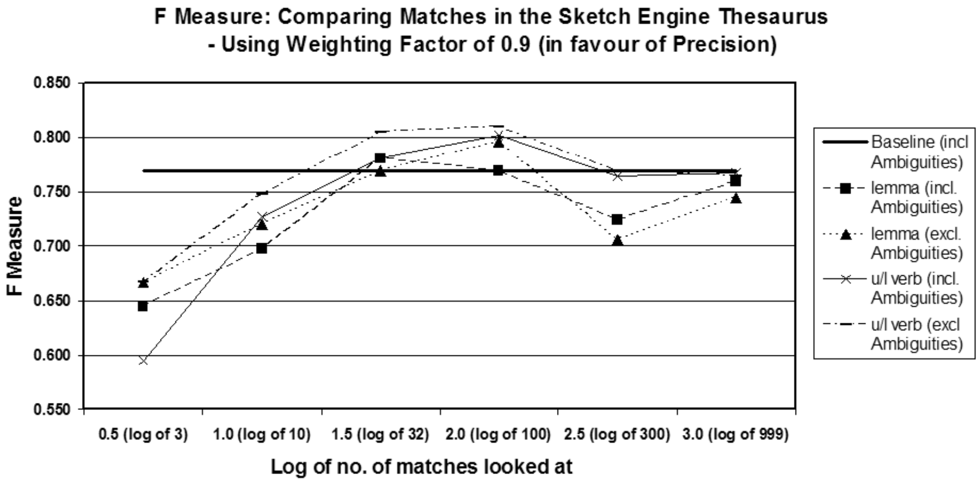

Figure 3: FMeasure

Where presented	Recall (%)	Precision (%)	F-Measure α = 0.9 (%)
This Paper - lemmas, incl ambiguities (32 matches)	38.5	88.2	78.1
This Paper - lemmas, incl ambiguities (100 matches)	53.8	80.8	76.9
(Agarwal and Boggess, 1992)	n/a	81.6	n/a
(Kurohashi and Nagao, 1992)	n/a	81.3	n/a
(Goldberg, 1999)	n/a	72	n/a
(Resnik, 1999) (unweighted)	66.0	71.2	70.6
(Resnik, 1999) (weighted)	69.7	77.4	76.6

Table 4: Comparison of Performances

saurus exhibits consistent distributional similarity characteristics. It is predictable that finding matches becomes rapidly more difficult, and that precision generally decreases the more matches are considered. However, a "shelf", when up to 32 and 100 matches are considered, can be observed in the precision and the f-measure data. The refinements to the matching method give only limited improvement over the plain lemma matching with no ambiguities excluded.

Table 4 compares some of our results with those of other researchers. Our f-measure results are good overall. At the end of the scale with few matches, our method shows very high precision, though we cannot match Resnik's recall. However, Resnik's problem is more narrowly defined than ours and may yield more matches. The methods of Agarwal and Boggess and of Kurohashi and Nagao are applied to problems which are somewhat different to ours. But the fact that their precision results are only a little above 80% indicates that our method is not any less useful. All the results discussed here fall short of performance in predicting PP-attachment ambiguity, where results above 90% have been achieved for some time (Stetina and Nagao, 1997).

The baseline for all the researchers listed in Table 4, where one is given, is the accuracy achieved if all the ambiguities are said to be coordination first. Ours is somewhat higher than those of the most comparable experiments in the literature: 75% to Resnik's 66% and Goldberg's 64%. This may be due to the wide range of ambiguities that we cover, to our consensus approach to decision making, or simply because our corpus is smaller. Our results fall within our upper and lower bounds, except at the ends of the scale.

6 Conclusions

We conclude from our research that coordinated words which have distributional similarity also tend to be of the coordination first type. The most highly ranked matches are especially reliable indicators of this. We also conclude that there may be an optimum number of matches which gives suitable precision and recall results for a system which indicates "most likely" readings of coordination ambiguities. The precision would be unacceptable, however, if our method was used to capture an optimum number of such readings. Our method compares favourably with other studies that aim to predict readings of coordination ambiguities. That all such research has lower performance than comparable research in the field of PP-attachment disambiguation, indicates the level of difficulty presented by coordination ambiguity.

Changing the way that we use our evaluation dataset by removing ambiguities and using underlying verbs in the matching process shows an improvement that may increase if we refine these strategies. Alternatively, the limited extent of the improvement may mean that other strategies would be more beneficial.

7 Further Work

This paper is part of wider research into looking at notifying users of ambiguities in text and informing them of how likely they are to be misunderstood by readers of the text.

We are investigating the effect of the lengths of coordinated phrases and of the types of external modifiers on prediction of the most likely readings of coordination ambiguities. We are also looking at combining our distributional similarity method with a semantic similarity method, as initial results show that these may be complementary.

We hope to investigate whether our technique can scale up and successfully disambiguate chained conjunctions, such as *X and Y and Z*. These result in “explosive” ambiguity (Church and Patil, 1982), and they have a rapidly increasing syntactic complexity.

Acknowledgements

The authors wish to thank Professor Bashar Nuseibeh for his advice relating to the requirements engineering aspect of the corpus building task.

References

Rajeev Agarwal and Lois Boggess. 1992. A simple but useful approach to conjunct identification. In *Proceedings of the 30th conference on Association for Computational Linguistics*, pages 15–21. Association for Computational Linguistics.

Daniel M. Berry, Erik Kamsties, and Michael M. Krieger. 2003. From contract drafting to software specification: Linguistic sources of ambiguity. A Handbook.

Hiram Calvo, Alexander Gelbukh, and Adam Kilgarriff. 2005. Distributional thesaurus vs. wordnet: A comparison of backoff techniques for unsupervised pp attachment. In *Proceedings of Sixth International Conference on Intelligent Text Processing and Computational Linguistics: CICLING05*, pages 172–182.

Kenneth Church and Ramesh Patil. 1982. Coping with syntactic ambiguity or how to put the block in the box on the table. *Comput. Linguist.*, 8(3-4):139–149.

William Gale, Kenneth Ward Church, and David Yarowsky. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th conference on Association for Computational Linguistics*, pages 249–256. Association for Computational Linguistics.

Miriam Goldberg. 1999. An unsupervised model for statistically determining coordinate phrase attachment. In *Proceedings of the 37th conference on Association for Computational Linguistics*, pages 610–614. Association for Computational Linguistics.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers.

A. Kilgarriff, P. Rychly, P. Smrz, and D. Tugwell. 2004. The sketch engine. In *Proceedings of EURALEX 2004*.

Adam Kilgarriff. 2003. Thesauruses for natural language processing. Technical Report ITRI-03-15, Information Technology Research Institute, University of Brighton, Brighton, U.K. Also published in Proceedings of NLP-KE.

Sadao Kurohashi and Makoto Nagao. 1992. Dynamic programming method for analyzing conjunctive structures in japanese. In *Proceedings of the 14th conference on Computational linguistics*, pages 170–176. Association for Computational Linguistics.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, pages 768–774. Association for Computational Linguistics.

Mark McLauchlan. 2004. Thesauruses for prepositional phrase attachment. In *Proceedings of CoNLL-2004*, pages 73–80. Boston, MA, USA.

G. Nenadic, I. Spasic, and S. Ananiadou. 2004. Mining biomedical abstracts: What is in a term? In *Proceedings of IJCNLP (International Joint Conference on Natural Language Processing)*.

A. Ratnaparkhi, J. Reynar, and S. Roukos. 1994. A maximum entropy model for prepositional phrase attachment. In *Proceedings of the ARPA Workshop on Human Language Technology*. Morgan Kaufmann.

Adwait Ratnaparkhi. 1998. Unsupervised statistical models for prepositional phrase attachment. In *Proceedings of the Seventeenth International Conference on Computational Linguistics*.

Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130.

Karen Sparck-Jones. 1986. *Synonymy and semantic classification*. Edinburgh University Press.

Jiri Stetina and Makoto Nagao. 1997. Corpus based pp-attachment ambiguity resolution with a semantic dictionary. In *Proceedings of the 5th Workshop on Very Large Corpora*.