Improving Subcategorization Acquisition using Word Sense Disambiguation

Anna Korhonen and Judita Preiss*
University of Cambridge, Computer Laboratory
15 JJ Thomson Avenue, Cambridge CB3 0FD, UK
Anna.Korhonen@cl.cam.ac.uk, Judita.Preiss@cl.cam.ac.uk

Abstract

We investigate the change in performance of automatic subcategorization acquisition when a word sense disambiguation (WSD) system is employed to guide the acquisition process. As a subgoal, this involves creating a probabilistic WSD system, which we evaluate on the SENSEVAL-2 English all-words task data. We carry out an evaluation of the enriched subcategorization acquisition system using 29 'difficult' English verbs which shows that WSD helps to improve the acquisition performance.

1 Introduction

Subcategorization information is vital for successful parsing, however, manual development of large subcategorized lexicons has proved difficult because predicates change behaviour between sublanguages, domains and over time. Additionally, manually developed subcategorization lexicons do not provide the relative frequency of different subcategorization frames (SCFs) for a given predicate, essential in a probabilistic approach.

Over the past years, several approaches have been proposed for automatic acquisition of subcategorization from corpus data (e.g. (Briscoe and Carroll, 1997; Carroll et al., 1998; Sarkar and Zeman, 2000)). Although these approaches vary largely according to the methods used and the number of SCFs being extracted, they perform similarly. They mostly gather information about syntactic aspects of subcategorization and do not distinguish between various predicate senses. As no lexical/semantic information is typically exploited, system output is noisy and the accuracy of the resulting lexicons shows room for improvement.

Recently, Korhonen (2002) has proposed a method which makes use of the predominant sense of a verb. Like the previous methods, this method also acquires subcategorization specific to a verb *form* rather than *sense*. However, it guides the acquisition process using back-off (i.e. probability) estimates based on the predominant sense of a verb in WordNet (Miller et al., 1990) (as determined by the frequency data in the associated SemCor corpus). These estimates help to correct the acquired SCF distribution and deal with sparse data. Where the sense is assigned correctly, significant improvement is reported in acquisition performance.[1]

This result shows that for many verbs, there is some single predominating sense in corpus data which accounts for most of the verbs subcategorization behaviour and can therefore be usefully deployed to improve automatic acquisition. However, for many highly polysemous verbs, the

*This work was supported by UK EPSRC project GR/N36462/93: 'Robust Accurate Statistical Parsing (RASP)'.

[1] On a set of 45 test verbs, the F-measure (Section 4.2.1) improves by 17 against the baseline method, which assumes no sense.

distribution of senses in balanced corpus data tends to be flat rather than zipfian (Preiss et al., 2002). For this important group of medium and high frequency verbs, the predominant sense is not frequent enough for back-off estimates based on just this sense to yield maximum benefit. To improve the acquisition performance, we need to consider non-predominant senses as well.

In addition, the assumption that the predominant sense is predictable, i.e. static across balanced corpora is questionable. While the good results obtained suggest that is often the case, the ultimate goal of automatic acquisition is to be able to produce domain specific lexicons. SCF frequencies have been shown to vary across corpus type (e.g. written vs. spoken language) and genre (e.g. financial vs. balanced text) and much of this variation is reported to be due to the effects of different corpus genres on verb sense and the effect of verb sense on subcategorization (Roland et al., 2000; Roland and Jurafsky, 2001). Due to a lack of large manually sense annotated corpora for each corpus type/genre, the variable predominant sense would be better determined using automatic word sense disambiguation (WSD). Such a system could also improve acquisition for verbs which do not have a clear predominant sense.

A small-scale experiment with manually sense annotated data (i.e. 100% accurate WSD) has shown that it is possible to improve SCF acquisition of the 'difficult' highly polysemous verbs by considering their non-predominating senses as well (Preiss and Korhonen, 2002). In this paper, a similar, but larger scale experiment is reported using a real WSD system. We introduce a new probabilistic combination WSD system, which produces probability distributions on senses, and we show that the system performs comparably on the Senseval-2 English all-words task data (Palmer et al., 2002). Information from this system is incorporated in the subcategorization system using a novel method. Finally, an experiment is reported with 29 difficult verbs which shows that real WSD can be used to improve the accuracy of subcategorization acquisition.

In Section 2 we introduce the basic subcategorization acquisition system and report the modifications made to the system to enable it to use WSD. Section 3 describes our probabilistic WSD system. Results and discussion are presented in Section 4 and conclusions are drawn in Section 5.

2 Subcategorization Acquisition

2.1 Baseline System

Building on the SCF acquisition framework of Briscoe and Carroll (1997), Korhonen (2002) has proposed a system which uses knowledge of verb semantics to guide the process of subcategorization acquisition.[2]

The system exploits the knowledge that semantically similar verbs are similar in terms of subcategorization. Levin (1993) has demonstrated that verb *senses* divide into semantic classes distinctive in terms of subcategorization. Korhonen (2002) shows that many verb *forms* also divide into such classes, according to their predominant sense. For instance, the verb form specific SCF distributions for *fly* and *move* correlate quite closely because the predominant senses of these verbs (according to the WordNet frequency data) are similar. They both belong to the Levin "Motion verbs".

The system of Korhonen (2002) works by first identifying the sense, i.e. the semantic class for a predicate. The semantic classes are based on Levin classes (Levin, 1993); mostly on broad classes (e.g. 51. "Motion verbs") rather than subclasses (e.g. 51.2 "*Leave* verbs").[3] Verbs are classified according to their predominant sense in WordNet. This is done using a mapping which links WordNet synsets with Levin classes.[4]

After the semantic class is identified, the system of Briscoe and Carroll (1997) is used to acquire a putative SCF distribution from corpus data. This system employs a robust statistical

[2]This system currently only treats verbs but plans are under way to extend it to other parts of speech (nouns and adjectives).

[3]The broad classes are more useful because they allow adequate generalizations to be made, but are still distinctive enough in terms of subcategorization to provide good accuracy.

[4]See the work of Korhonen (2002) for details of the mapping.

parser (Briscoe and Carroll, 2002) and a comprehensive classifier which is capable of distinguishing 163 verbal SCFs – a superset of those found in the ANLT (Boguraev and Briscoe, 1987) and COMLEX Syntax dictionaries (Grishman et al., 1994).

The SCF distribution is smoothed using the probability (i.e. "back-off") estimates of the semantic class of the verb. The smoothing method is linear interpolation (e.g. (Manning and Schütze, 1999)). The back-off estimates are obtained using the following method:

(i) 4-5 representative verbs are chosen from a verb class.[5]

(ii) SCF distributions are built for these verbs by manually analysing c. 300 occurrences of each verb in the British National Corpus (BNC) (Leech, 1992).

(iii) The resulting SCF distributions are merged, giving equal weight to each distribution.

The back-off estimates for the "Motion verb" *fly*, for example, are constructed by merging the SCF distributions for 5 other "Motion verbs" e.g. *move*, *slide*, *arrive*, *travel*, and *sail*.

As a final step, a simple empirically determined threshold is used on the probability estimates after smoothing to filter out noisy SCFs.

2.2 Combining with WSD

Preiss and Korhonen (2002) modified the baseline system for their small-scale experiment so that it could benefit from disambiguating the first and/or second most frequent senses of verbs. Different corpus datasets were created for the (1-2) senses being disambiguated (initial senses) and for the remaining senses (which were grouped together). SCFs were acquired separately for each of these datasets. For each dataset corresponding to the initial senses, the back-off estimates of the relevant sense were used for smoothing. No smoothing was done in the case of the dataset of grouped senses. Finally, the SCF lexicons acquired for different datasets were merged, so that each lexicon received a weight corresponding to its size. This yielded a SCF distribution specific to a verb form rather than sense.

Although Preiss and Korhonen (2002) reported an improvement using this method[6] they encountered sparse data problems in subcategorization acquisition, since many datasets were simply too small to yield an accurate lexicon. Separating out the data into different datasets not only generated noise but was also unnecessary: the lexicons have to be merged in the end, to allow sensible comparison with the baseline system and the use of the extant verb form specific gold standard data.

We therefore employed a different method which does not involve separating data. Instead it involves using back-off estimates specific to the sense distribution in our data, as determined by our WSD system. Thus the method is identical to the baseline method presented in Section 2.1, but a different method is adopted for constructing back-off estimates: they are now constructed from the back-off estimates of all the senses our WSD system has detected (not just the predominant), so that the contribution of each set of estimates is weighted according to the frequency of the corresponding senses in corpus data.

We combined the different back-off estimates using linear interpolation (Chen and Goodman, 1996). Let $p_j(scf_i)$, $j = 1 \ldots n_{\mathrm{bo}}$ (where n_{bo} is the number of back-off estimates) be the probabilities of SCFs in different back-off distributions The estimated probability of the SCF in the resulting combined back-off distribution is calculated as follows:

$$P(scf_i) = \sum_{j=1}^{n_{\mathrm{bo}}} \lambda_j \cdot p_j(scf_i)$$

[5] The verb for which subcategorization is being acquired is always excluded.

[6] When 100% accurate WSD was used to simply separate the first sense from any other sense (for 7 verbs) there was an increase in the F-measure from 74.3 to 77.6. In the case of 3 verbs where three sense groups were distinguished there was an increase in the F-measure from 75.0 to 78.8. See Section 4 for calculation of F-measure.

where the λ_j denote weights for the different distributions and sum to 1. The values for λ_j are determined specific to a verb and are obtained from the probabilistic WSD system.

3 Probabilistic WSD

WSD systems can either choose a single sense for a word, or they can produce a probability distribution on the word's senses. We created a system which produces a probability distribution for each noun, verb, adjective and adverb in a text. This makes the system particularly suitable for the subcategorization acquisition application: we extract the probability distributions for our chosen verbs and combine them by computing an average. This yields a probability distribution on senses for each verb[7] which is integrated into the SCF acquisition system.

3.1 System Description

Our system is designed along the lines of Stevenson and Wilks (2001), who use voting to combine a number of knowledge sources to produce a WSD system. Each of our component modules produces a probability distribution on senses. These probability distributions are combined, using the independence assumption, by multiplication to yield probability distribution on senses for each word. Our modules are based on those described in Yarowsky (2000), Mihalcea (2002) and Pedersen (2002), and can be found in Table 1. We trained all modules (except **tagger** and **frequency** which are not trained) on SemCor, the English all-words SENSEVAL-2 task test data and all data for the English lexical sample SENSEVAL-2 task. A part of this training corpus is held out to create a development corpus, which we use to obtain a confidence in each module for each of our desired verb.[8] The confidence value of each module is used to decide the level of smoothing for the module: the probability distribution for a word from a module with low confidence is smoothed extensively to more approximate a uniform distribution.

We decided on the optimal combination of modules based on the accuracy (F-measure) on the English all-words task (for this evaluation, the system was trained on all corpora apart from the English all-words task). When the system is run in a forced choice mode (the sense with the highest probability is chosen), its precision is 63.83% and recall 62.55% on the English all-words task. This would place the system in the third place (F-measure) in the English all-words task (initial results).

4 Experiment

4.1 Test Data

Preiss et al. (2002) showed that high frequency polysemous verbs whose predominant sense is not very frequent are likely to benefit most from WSD. As these verbs are particularly important for practical NLP applications, we focused on them – despite the fact that it made our task harder: being exceptionally difficult for both WSD and subcategorization acquisition, these verbs are frequently used to examine the true potential and limits of automatic acquisition.

We chose 29 of these verbs for investigation. The verbs were chosen at random, subject to the constraint that they occur in the SemCor data in at least two broad Levin-style senses. The WordNet senses of these verbs were mapped to Levin senses, using as a starting point the mappings provided by Korhonen (2002) and Bonnie Dorr (the 'LCS database').[9] Those WordNet senses not covered in these mappings were mapped to Levin senses (either original ones or Dorr's 26 additional Levin-style senses) manually. Senses very low in frequency and those which could not be mapped to any extant Levin-style senses were left out of consideration. The maximum number of Levin senses considered per verb was 4. These typically map to several WordNet senses, as Levin assumes more coarse-grained sense distinctions than WordNet.

[7] Note that this probability distribution is on WordNet 1.7.1 senses plus an extra 'not available' sense. The senses are mapped to Levin using the mapping described in Korhonen (2002).

[8] If a verb does not appear in the development corpus, the confidence of the module for that verb is taken to be an average of all the module's confidences.

[9] The database is available from `http://www.umiacs.umd.edu/~bonnie/verbs-English.lcs`

Module	Description
Tagger	We use the Acquilex tagger (Elworthy, 1994), which produces a probability distribution on CLAWS-II tags. We combine these to produce a distribution on noun, verb, adjective and adverb.
Frequency	The frequency information is taken from WordNet, and converted into a probability distribution.
PoS	Part of speech (PoS) of surrounding words (one word before, two words before, etc.) produces a probability distribution on senses.
GR	The grammatical role (subject, direct object and indirect object) is taken into account for nouns, along with the corresponding verb to produce a probability distribution.
Head	Information about the word being a head of a (noun, verb, etc.) phrase is taken into account to produce a probability distribution.
Trigram	PoS trigrams produce a probability distribution on senses of certain words. We used the NSP software for this module.[a]

[a]This is available from `http://www.d.umn.edu/~tpederse/nsp.html`

Table 1: Probabilistic Modules

The test verbs and their senses are shown in Table 2. The senses, indicated by number codes from Levin's and Dorr's classifications are listed in the order of their frequency in SemCor, starting from the predominant sense (marked as 1st).[10]

4.2 Evaluation

4.2.1 Method

We took a sample of 20 million words of the BNC corpus and extracted all sentences containing any of the test verbs. After the extraction process, we retained on average 1000 sentences per verb. These sentences were disambiguated using the probabilistic WSD system described in Section 3 and then processed by the modified subcategorization system outlined in Section 2.2. The latter constructs and uses for each test verb an individual set of back-off estimates, built by taking into account the different (2-4) senses of test verbs and the frequency of these senses in the corpus data (as detected by the WSD system).

[10]This table displays senses as Levin subclasses where such are available (e.g. 13.5), regardless of whether we assumed a narrow or broad class (e.g. 13) in our method for subcategorization acquisition. Note also that one additional sense was used which does not appear in Levin or Dorr: 017.

The results were evaluated against a manual analysis of the corpus data. This was obtained by analysing c. 300 occurrences for each test verb in our BNC test data. 5-21 gold standard SCFs were found for each verb (16 SCFs per verb on average).

We calculated type precision (the percentage of SCF types that the system proposes which are correct), type recall (the percentage of SCF types in the gold standard that the system proposes) and F-measure:

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall} \qquad (1)$$

We also compared the similarity between the acquired unfiltered[11] and gold standard SCF distributions using various measures of distributional similarity: the Spearman rank correlation (RC), Kullback-Leibler distance (KL), Jensen-Shannon divergence (JS), cross entropy (CE), skew divergence (SD), and intersection (IS). The details of these measures and their application to subcategorization acquisition can be found in Korhonen and Krymolowski (2002).

Finally, we recorded the total number of SCFs missing in the distributions, i.e. the type of false

[11]No threshold was applied to remove the noisy SCFs from the distributions.

Verb	Senses			
	1st	2nd	3th	4th
absorb	45.4	14	22.1	
bear	058	31.2	014	26.4
choose	13.5	017		
compose	47.8	26.4		
conceive	26.4	29.1		
concentrate	45.4	31.1		
continue	55.1	37.7		
count	027	31.4	29.1	
descend	51.1	55.1		
distinguish	23.1	30.2	29.1	
embrace	47.8	36.2		
establish	45.4	29.4	9.1	
find	13.5	29.4	30.1	
force	002	12		
grasp	018	15.1		
induce	26.4	002		
keep	15.2	55.1	023	
mark	25.1	29.1	23.1	
offer	13.3	005		
proclaim	29.3	37.7		
provide	13.4	29.2		
roar	43.2	37.3	51.3	
seek	35.6	005	51.3	017
settle	9.6	017	36	
strike	18.1	31.1	43.2	
submit	11.1	37.7	13.2	
wait	47.1	53.1	29.5	
watch	30.2	35		
write	25.2	37.1	36.1	

Table 2: Test verbs and their senses

Measures	Method		
	No Sense	Predominant	WSD
Precision (%)	72.9	72.3	74.6
Recall (%)	31.3	38.9	42.2
F-measure	43.8	50.6	53.9
RC	0.59	0.57	0.61
KL	1.20	0.93	0.56
JS	0.10	0.10	0.09
CE	2.72	2.44	2.30
IS	0.72	0.80	0.97
Unseen SCFs	175	129	22

Table 3: Average results for 29 verbs

negatives which did not even occur in the unfiltered distributions. This was to investigate how well a method deals with sparse data, i.e. how accurate the back-off estimates are.

For comparison, we also reported results for the baseline system described in Section 2.1 which backs-off to the predominant sense, and for another version of this system which assumes no sense at all (i.e. no back-off estimates are employed and no smoothing is done).

4.2.2 Results

Table 3 shows average results for the 29 verbs with the two versions of the baseline system and for the modified system which employs WSD.

We see that the performance improves with the number of senses considered. The WSD yields 3.3 better F-measure than the predominant sense, which in turn yields 6.8 better F-measure than 'no sense'. The improvement can be observed on all measures (the only exceptions are precision and RC, which are slightly worse for the predominant sense than 'no sense'), but particularly on those which evaluate the capability of the system to deal with sparse data. From the total of 175 gold standard SCFs unseen in the unsmoothed lexicon, 107 are unseen after using the predominant sense method, and only 22 remain unseen after WSD is employed.

The effect of WSD is particularly clear on the more sensitive measures of distributional similarity which consider unfiltered (noisy) SCF distributions and (unlike precision and recall) evaluate the actual frequencies/ranks of SCFs. IS indicates that there is a large intersection between the acquired and gold standard SCFs when WSD is used (0.97, as opposed to 0.80 with the predominant sense). The improvement on RC is smaller (0.04), demonstrating that WSD improves the ranking of SCFs slightly.

From the entropy-based similarity measures (KL, CE and JS), KL improves the most with WSD (0.37 from the predominant and 0.56 from 'no sense'). JS, which is considered the most robust of these measures, shows smaller but nevertheless noticeable improvement.

Table 4 lists F-measure and JS results for each of the individual test verbs. We see that, generally, WSD benefits the most those verbs which are highly polysemous with 3-4 senses (e.g. *bear, count, distinguish, roar, wait*) or verbs whose various senses differ substantially in terms of subcategorization (e.g. *conceive, continue, embrace, grasp*).

For example, a clear improvement is seen with many of the verbs whose one sense involves mainly NP/PP SCFs (e.g. *He* **grasped**

Verb	F-measure		JS	
	Pred.	WSD	Pred.	WSD
absorb	40.0	40.0	0.08	0.07
bear	47.6	54.6	0.12	0.10
choose	62.5	62.5	0.06	0.06
compose	50.0	50.0	0.10	0.09
conceive	38.1	52.2	0.11	0.10
concentrate	50.0	50.0	0.21	0.15
continue	48.3	53.3	0.06	0.06
count	59.3	64.3	0.08	0.06
descend	61.5	61.5	0.03	0.03
distinguish	37.5	47.1	0.03	0.03
embrace	54.6	61.5	0.09	0.08
establish	23.5	33.3	0.04	0.04
find	48.0	48.0	0.15	0.14
force	66.7	66.7	0.17	0.16
grasp	45.5	54.6	0.07	0.05
induce	61.5	61.5	0.05	0.03
keep	50.0	50.0	0.14	0.13
mark	38.1	38.1	0.08	0.08
offer	47.6	47.6	0.06	0.06
proclaim	53.9	56.0	0.13	0.10
provide	42.9	42.9	0.06	0.06
roar	69.2	74.1	0.11	0.09
seek	66.7	60.0	0.16	0.12
settle	40.0	46.2	0.16	0.15
strike	61.5	64.0	0.16	0.14
submit	54.6	54.6	0.03	0.02
wait	31.6	47.6	0.10	0.09
watch	48.5	48.5	0.19	0.17
write	56.3	60.6	0.16	0.12

Table 4: F-measure and JS for test verbs

the door's handle, and he entered the chamber of secrets) and another one involves sentential SCFs (e.g. *Does anyone* **grasp** *that this was done in 2000, and is old news?*).

Due to diathesis alternations, an occurrence of one SCF is likely to give rise to another, related SCF. Thus SCFs tend to occur in data as 'families'. Detection of a verb sense can therefore result in detection of a whole family of new (gold standard) SCFs.

One verb shows worse performance when WSD is used: *seek*. Surprisingly, this verb is highly polysemous and its senses differ substantially in terms of subcategorization. In theory, it is possible that if senses differ a lot in terms of subcategorization and one of them is clearly predominating in the data, then the detection of any of the other senses may result in noise. Our results show, however, that this is not usually the case.

The verbs which do not show (clear) improvement with WSD (e.g. *choose, compose, induce, watch*) are not as highly polysemous (in our coarse grained gold standard), although some of their senses do differ substantially in terms of subcategorization. It is possible that these verbs occurred in our data mostly in their predominating sense and therefore WSD made little (or no) difference. This is difficult to evaluate without sense disambiguated data.

5 Conclusion

Our results showed that a state-of-the-art WSD system can improve the accuracy of SCF acquisition for difficult verbs. Interestingly, they also showed that it is not only the *degree* of polysemy which determines the need of WSD (Preiss et al., 2002) but also how much the senses differ in terms of subcategorization.

Further research is warranted to improve the results further. We intend to investigate better ways of integrating the acquired sense (frequency) information into the SCF system, and continue refining our method for subcategorization acquisition. Time will also be invested in automatically acquiring a large training corpus for the probabilistic WSD system (e.g. (Mihalcea and Moldovan, 1999)), which should increase the system's performance.

Acknowledgements

We would like to thank Ted Briscoe for his help in the early stages of this work. Our thanks also go to Rada Mihalcea for her encouragement and Yuval Krymolowski for his technical help.

References

B. K. Boguraev and E. J. Briscoe. 1987. Large lexicons for natural language processing utilising the grammar coding system of the *Longman Dictionary of Contemporary English. Computational Linguistics*, 13(4):219–240.

E. J. Briscoe and J. Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of ACL ANLP97*, pages 356–363.

E. J. Briscoe and J. Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceed-*

ings of the 3rd International Conference on Language Resources and Evaluation, pages 1499–1504.

J. Carroll, E. Briscoe, and A. Sanfilippo. 1998. Parser evaluation: A survey and a new proposal. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 447–454.

S. F. Chen and J. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318.

D. Elworthy. 1994. Does Baum-Welch re-estimation help taggers? In *Proceedings of the 4th Conference on Applied NLP*, pages 53–58.

R. Grishman, C. Macleod, and A. Meyers. 1994. Comlex syntax: building a computational lexicon. In *International Conference on Computational Linguistics, COLING-94*, pages 268–272.

A. Korhonen and Y. Krymolowski. 2002. On the robustness of entropy-based similarity measures in evaluation of subcategorization acquisition systems. In *Proceedings of the 6th Conference on Natural Language Learning*, pages 91–97.

A. Korhonen. 2002. *Subcategorization Acquisition*. Ph.D. thesis, University of Cambridge.

G. Leech. 1992. 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

B. Levin. 1993. *English Verb Classes and Alternations*. Chicago University Press.

C. D. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

R. Mihalcea and D. I. Moldovan. 1999. An automatic method for generating sense tagged corpora. In *Proceedings of AAAI-99*, pages 461–466.

R. Mihalcea. 2002. Word sense disambiguation using pattern learning and automatic feature selection. *Journal of Natural Language and Engineering*, pages 343–358.

G. Miller, R. Beckwith, C. Felbaum, D. Gross, and K. Miller. 1990. Introduction to WordNet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244.

M. Palmer, C. Fellbaum, S. Cotton, L. Delfs, and H. T. Dang. 2002. English tasks: All-words and verb lexical sample. In Preiss and Yarowsky (Preiss and Yarowsky, 2002), pages 21–24.

T. Pedersen. 2002. Machine learning with lexical features: The duluth approach to Senseval-2. In Preiss and Yarowsky (Preiss and Yarowsky, 2002), pages 139–142.

J. Preiss and A. Korhonen. 2002. Improving subcategorization acquisition with WSD. In *Proceedings of the Word Sense Disambiguation Workshop*, pages 102–108.

J. Preiss and D. Yarowsky, editors. 2002. *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguating Systems*.

J. Preiss, A. Korhonen, and E. J. Briscoe. 2002. Subcategorization acquisition as an evaluation method for WSD. In *Proceedings of LREC*, pages 1551–1556.

D. Roland and D. Jurafsky. 2001. Verb sense and verb subcategorization probabilities. In S. Stevenson and P. Merlo, editors, *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issue*. Cambridge University Press, Jon Benjamins, Amsterdam. To appear.

D. Roland, D. Jurafsky, L. Menn, S. Gahl, E. Elder, and C. Riddoch. 2000. Verb subcatecorization frequency differences between business-news and balanced corpora. In *ACL Workshop on Comparing Corpora*, pages 28–34.

A. Sarkar and D. Zeman. 2000. Automatic extraction of subcategorization frames for Czech. In *19th International Conference on Computational Linguistics*, pages 691–697.

M. Stevenson and Y. Wilks. 2001. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3):321–350.

D. Yarowsky. 2000. Hierarchical decision lists for word sense disambiguation. *Computers and the Humanities*, 34(1/2):179–186.