A Large Subcategorization Lexicon for Natural Language Processing Applications

Anna Korhonen △, Yuval Krymolowski *, Ted Briscoe △

△ Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue, Cambridge CB3 0GD, UK
{Anna.Korhonen, Ted.Briscoe}@cl.cam.ac.uk
*Department of Computer Science, Technion
Israel Institute of Technology, Haifa 32000, Israel
{yuvalkr}@cs.technion.ac.il

Abstract

We introduce a large computational subcategorization lexicon which includes subcategorization frame (SCF) and frequency information for 6,397 English verbs. This extensive lexicon was acquired automatically from five corpora and the Web using the current version of the comprehensive subcategorization acquisition system of Briscoe and Carroll (1997). The lexicon is provided freely for research use, along with a script which can be used to filter and build sub-lexicons suited for different natural language processing (NLP) purposes. Documentation is also provided which explains each sub-lexicon option and evaluates its accuracy.

1. Introduction

Accurate, comprehensive subcategorization lexicons are vital for the development of successful parsing technology (Carroll et al., 1998; Arun and Keller, 2005), important for various (computational) linguistic tasks (such as automatic verb classification, selectional preference acquisition, psycholinguistic experiments (Lapata et al., 2001; Schulte im Walde and Brew, 2002; McCarthy and Carroll, 2003)) and useful for any NLP application which can benefit from information related to predicate-argument structure (e.g. Information Extraction, Machine Translation (Hajič et al., 2002; Surdeanu et al., 2003)).

Several large, manually developed subcategorized lexicons are available for English, e.g. the COMLEX Syntax (Grishman et al., 1994) and the ANLT (Boguraev and Briscoe, 1987) dictionaries. However, manually built lexicons are prone to errors which are difficult to detect automatically and it is costly to extend these resources to cover information not currently encoded. One important type of information absent from most lexicons is statistical information concerning the relative frequency of different SCFs for a given predicate. This information, essential for a probabilistic approach, is almost impossible to collect by hand as it is highly domain-sensitive, i.e. it varies with predominant word senses, which change across corpora and domains.

These problems suggest that when aiming to obtain a subcategorization lexicon useful for a real-world task, automatic acquisition of SCFs and their frequencies from repositories of unannotated text (such as corpora and the web) is a more promising approach. The automatic approach is now viable and gathers statistical information as a side-effect of the acquisition process which can easily be adapted to new domains with adequate corpus data.

Over the past decade several systems have been proposed for this purpose for both English and other languages, e.g. (Brent, 1993; Briscoe and Carroll, 1997; Carroll and Rooth, 1998; Kawahara et al., 2000; Ferrer, 2004). The different systems vary according to methods used and the number of SCFs being extracted, but they perform quite similarly: they mainly deal with verbs, they do not distinguish between different predicate senses, they mostly gather information about the syntactic aspects of subcategorization (the type, number and/or relative frequency of SCFs given specific predicates) and they perform (at their best) around 80-85% token recall[1].

Further research is needed before highly accurate lexicons encoding information also about semantic aspects of subcategorization (e.g. different predicate senses, the mapping from syntactic arguments to semantic representation of argument structure, selectional preferences on argument heads, diathesis alternations, etc.) can be obtained automatically. Also, currently many argument-adjunct tests cannot yet be exploited since they rest on semantic judgments that cannot yet be made automatically.

While research into further improving the systems will continue, the state of the art has already developed to the point where the best existing systems are capable of detecting comprehensive SCF (frequency) information with accuracy high enough to benefit practical NLP tasks, e.g. (Schulte im Walde and Brew, 2002; Korhonen et al., 2003; McCarthy and Carroll, 2003).

Given this, we now provide the NLP community with a large-scale subcategorization lexicon acquired automatically from five corpora and the Web using the current version of Briscoe and Carroll's (1997) comprehensive system (Korhonen, 2002b; Korhonen and Preiss, 2003). The lexicon includes SCF (frequency) information for 6,397 (American and British) English verbs. We provide the resource freely for research use, together with a script which can be used to build sub-lexicons suitable for different NLP purposes and with documentation which explains each (sub-)lexicon option and evaluates its accuracy.

[1] Token recall is the percentage of SCF tokens in a sample of manually analysed text that were correctly acquired by the system.

We introduce the system we used for SCF acquisition in section 2. The process of constructing the large lexicon is described in section 3. Section 4. evaluates the performance of the large lexicon and a few representative sub-lexicons extracted from it. Section 5. summarises the paper.

2. System for Automatic Subcategorization Acquisition

The system of Briscoe and Carroll is one of the most comprehensive subcategorization acquisition systems available for English, capable of categorizing 163 verbal SCFs and returning relative frequencies for each SCF found for a verb. The current version of the system makes use of the RASP (Robust Accurate Statistical Parsing) toolkit (Briscoe and Carroll, 2002). Raw corpus data are first tokenized, tagged, lemmatised, and parsed with RASP using a tag-sequence grammar written in a feature-based unification formalism. This yields complete though intermediate parses. RASP has several modes of operation; Briscoe and Carroll's system currently invokes it in a mode which outputs intermediate phrase structure analyses.

Verb subcategorization patterns (local syntactic frames including the syntactic categories and head lemmas of constituents) are then extracted from parsed sentences, from subanalyses which begin/end at the boundaries of specified predicates. The patterns are classified into SCFs using a comprehensive classifier which is capable of categorizing the 163 verbal SCFs—a superset of those found in the ANLT and COMLEX dictionaries (see (Briscoe, 2000) for the detailed description). They abstract over specific lexically-governed particles and prepositions and specific predicate selectional preferences but include some derived semi-predictable bounded dependency constructions, such as particle and dative movement. Lexical entries are constructed for each verb and SCF combination, and the basic lexicon is built.

As no lexical or semantic information is typically exploited during parsing, the basic lexicon is inevitably noisy. A filtering component may therefore be applied which can be used to remove noisy SCFs from the lexicon. The same component can also be used to improve the quality of automatically acquired SCF distributions and/or to create sub-lexicons suitable for different purposes. Multiple options are provided by the filter. The basic ones include:

1. Selecting only certain verbs or verb types from the lexicon, e.g. a user can provide the list or frequency (range) of verbs which should be included in the sub-lexicon.

2. Smoothing the automatically acquired SCF distributions for individual verbs in order to deal with the sparse data problem (the fact that many relevant SCFs are either too low in frequency or altogether absent from the lexicon) and/or to improve their accuracy. Three smoothing techniques are provided[2]:

   (i) add-one smoothing (Laplace, 1995)

   (ii) Katz backing-off (Katz, 1987)

   (iii) linear interpolation (Chen and Goodman, 1996)

   Each technique deals with the sparse data problem, but Katz backing-off and linear interpolation make use of specific back-off (probability) estimates. Linear interpolation has the strongest impact because it makes a linear combination of the acquired SCF distributions and back-off estimates, affecting equally low and high frequency SCFs.

   For Katz backing-off and linear interpolation, verbs are first classified according to their most frequent sense(s) in WordNet (Miller et al., 1990), as determined by the frequency data in the associated SemCor corpus. Their automatically acquired SCF distributions are then smoothed using the back-off estimates of the respective verb class(es).

   The back-off estimates are based on lexical-semantic classes of verbs (Levin, 1993; Korhonen and Briscoe, 2004). They make use of the knowledge that semantically similar verbs are often similar also in terms of subcategorization (e.g. the SCF distributions for the 'motion' verbs such as *fly*, *run*, *travel*, and *move* correlate quite closely). They were built separately for each verb class, by choosing a number of representative verbs from the class and by merging their manually built SCF distributions.

   This method, described in detail in (Korhonen, 2002b; Korhonen and Preiss, 2003), helps to correct the acquired SCF distributions and deal with sparse data. The most frequent WordNet senses of 2685 medium-high frequency verbs were hand-classified to lexical classes so that this technique could be applied in a large scale.

3. Selecting the SCF (sub-)set from the lexicon on the basis of

   (i) empirically defined (a) uniform or (b) SCF-specific filtering thresholds based on the absolute or relative frequencies of SCFs;

   (ii) statistical confidence tests (the binomial hypothesis test, log likelihood ratio test, t-test);

   (iii) the SCFs in the manually built COMLEX and ANLT dictionaries.

The final subcategorization lexicon provides a lexical entry for each verb and SCF combination found in corpus data. A lexical entry specifies (at minimum) the verb and the SCF[3] in question, the syntax of detected arguments, the raw and relative frequencies of the SCF given the verb, the POS tags of the verb tokens, the argument heads in different argument positions, and the frequency of possible lexical rules (e.g. the passive rule) applied during parsing.

Figure 1 shows a small sample entry for the verb *send* with the NP frame (SCF number 24) (e.g. *John sent a message*)

[2] For the details of these techniques and their application to the task see (Korhonen, 2002b).

[3] The SCFs are indicated by number codes from (Briscoe, 2000).

```
#S(EPATTERN
  :TARGET  |send|
  :SUBCAT  (VSUBCAT NP)
  :CLASSES (24)
  :FREQCNT 44
  :RELFREQ 0.53
  :TLTL    (VVD VVD VVZ VVD VV0 VV0 VV0 VVN
            VV0 VVG VVG VV0 VVN VV0 VVD VVD
            VVD VVD VVG VVN VV0 VVD VVD VVD
            VVD VVD VV0 VVD VVN VVD VV0 VVD
            VVN VVD VVN VVD VVD VVD VV0 VV0
            VV0 VV0 VV0 VVD)
  :SLTL    (((|Edward| NP1)) ((|he| PPHS1))
            ((|Rick| NP1)) ((|He| PPHS1))
            ((|she| PPHS1)) ((|she| PPHS1))
            ((|She| PPHS1)) ((|Teacher| NN2))
            ((|who| PNQS)) ((|language| NN1))
            ((|it| PPH1)) ((|Minton| NP1))
            ((|He| PPHS1)) ((|Service| NN1))
            ((|he| PPHS1)) ((|you| PPY))
            ((|he| PPHS1)) ((|They| PPHS2))
            ((|They| PPHS2)) ((|Someone| PN1))
            ((|friend| NN2)) ((|prince| NN1))
            ((|She| PPHS1)) ((|Romania| NP1))
            ((|American| NN2)) ((|He| PPHS1))
            ((|Klaus| NP1)) ((|Britain| NP1))
            ((|We| PPIS2)) ((|Renwick| NP1))
            ((|She| PPHS1)) ((|you| PPY))
            ((I PPIS1)) ((I PPIS1))
            ((|military| NN1)) ((|she| PPHS1))
            ((|Who| PNQS)) ((|Claudel| NP1))
            ((|king| NN1)) ((|Clement| NP1))
            ((II MC)) ((|he| PPHS1))
            ((|Edward| NP1)) ((|he| PPHS1)))
  :OLT1L   (((|delegate| NN2)) ((|envoy| NN2))
            ((|instruction| NN2))
            ((|Elsa| NP1)) ((|spray| NN1))
            ((|signal| NN1)) ((|Ace| NN1))
            ((|thankyou| NN1)) ((|it| PPH1))
            ((|essay| NN2)) ((|message| NN1))
            ((|information| NN1))
            ((|information| NN1))
            ((|story| NN1)) ((|top| NN2))
            ((|wire| NN1)) ((|Android| NN2))
            ((|he| PPHO1)) ((|pound| NN2))
            ((|message| NN1)) ((|it| PPH1))
            ((|we| PPIO2)) ((I PPIO1))
            ((|bill| NN1)) ((|Body| NN2))
            ((|she| PPHO1)) ((|it| PPH1))
            ((|he| PPHO1)) ((|he| PPHO1))
            ((|minister| NN1))
            ((|team| NN1)) ((|division| NN2))
            ((|force| NN2)) ((|message| NN1))
            ((|you| PPY)) ((|report| NN1))
            ((|Claudel| NP1)) ((|they| PPHO2))
            ((|surf| NN1)) ((|report| NN1))
            ((|it| PPH1)) ((|message| NN1))
            ((|Sophia| NP1)) ((|unit| NN2)))
  :OLT2L   NIL
  :OLT3L   NIL
  :LRL     4)
```

Figure 1: Sample lexical entry for *send* with the SCF NP

```
#S(EPATTERN
  :TARGET  |verb|
  :SUBCAT  (syntax of arguments for SCF)
  :CLASSES ((SCF number code(s)) frequency
           of SCF in ANLT)
  :FREQCNT frequency of the SCF
           with the verb
  :RELFREQ the relative frequency of
           the SCF with the verb
  :TLTL    (verbs and their POS tags)
  :SLTL    (argument heads and their
           POS tags in subject position)
  :OLT1L   (argument heads and their
           POS tags in the 1st argument position)
  :OLT2L   (argument heads and their
           POS tags in the 2nd argument position)
  :OLT3L   (argument heads and their
           POS tags in the 3rd argument position)
  :LRL     number of lexical rules applied
           during parsing)
```

Figure 2: Legend for a lexical entry

which gathers information from 44 subcategorization patterns found in corpus data. The different fields of the entry are explained in the legend provided in figure 2.

Note that large lexical entries in big lexicons gather information from hundreds or thousands of subcategorization patterns found in corpus data. The information stored in such entries has proved useful for a number of NLP tasks, including parsing (Carroll et al., 1998), lexical classification (Korhonen et al., 2003) and the acquisition of selectional preferences (McCarthy and Carroll, 2003) and diathesis alternations (McCarthy, 2000).

3. Lexicon

6,433 verbs were first selected for inclusion in the large lexicon from both British and American English: all the 5,583 verbs listed in the American COMLEX dictionary and the 850 most frequent non-COMLEX verbs in the British National Corpus (Leech, 1992). To obtain as comprehensive subcategorization (frequency) information as possible up to 10,000 sentences containing an occurrence of each of these verbs were included in the input data to subcategorization acquisition. The sentences where extracted from 5 different corpora:

1. The British National Corpus (BNC)
2. The North American News Text Corpus (NANT) (Graff, 1995)
3. The Guardian corpus
4. The Reuters corpus (Rose et al., 2002)
5. The data used for two Text Retrieval Evaluation Conferences[4]: TREC-4 and TREC-5

Where the BNC and NANT provided the required 10,000 sentences, the other corpora were not used. Where the other

[4] http://trec.nist.gov/data/docs_eng.html

Resource	Total size (in words)	No. of extracted sentences	Percentage of data
BNC	100M	4859668	31%
NANT	350M	8033566	51%
Reuters	185M	1159034	7%
Guardian	29M	338881	2%
TREC	240M	1243312	7%
The Web		247150	2%
Total	904M	15881611	100%

Table 1: The data resources

corpora were used, this was done in the order of preference indicated above. According to our previous experiments (Korhonen, 2002b) around 250 input sentences per verb are required, on average, when aiming to acquire a relatively comprehensive SCF distribution[5]. Therefore, for any verb with less than 250 sentences in these corpora (2,049 out of the 6,433), additional sentences were extracted from the Web using the Google Web APIs[6]. In the end, only 36 verbs had to be excluded from the lexicon because no corpus or Web data was found for them.

The resulting data for 6,397 verbs includes 15.9M sentences in total. Table 1 shows each corpus / data resource used, its total size in words (except for the Web[7]), and the number of sentences and the percentage of the data extracted from it. The data was finally processed using the subcategorization SCF acquisition system described in section 2. which produced a large lexicon containing 212,741 SCFs in total, 33 SCFs per verb on average.

4. Evaluation

This large unfiltered lexicon constitutes the basic lexicon which we have made available at http://www.cl.cam.ac.uk/users/alk23/subcat/lexicon.html. Together with the lexicon we provide filtering software which users may use to remove noisy SCFs from the lexicon, to improve the acquired SCF distributions and/or to create sub-lexicon(s) suitable for the specific (NLP) use in mind. A short description of the basic filtering options was given in section 2. A more comprehensive description is given in the documentation which we provide at the website. The documentation explains each filtering option and evaluates the accuracy of the resulting sub-lexicon. Because the filtering options and the resulting sub-lexicons are numerous, we concentrate on describing and evaluating the accuracy of five representative lexicons here — the basic lexicon and 4 sub-lexicons extracted from it :

Lexicon 1: The basic unfiltered lexicon

Lexicon 2: A sub-lexicon created by selecting from the basic lexicon only those SCFs whose relative frequency is higher than a SCF-specific threshold

Lexicon 3: A sub-lexicon created by first smoothing the automatically acquired SCF distributions with the back-off estimates using linear interpolation (with the weight of 0.5) and then setting a uniform threshold of 0.01 on the probability estimates from smoothing

Lexicon 4: A sub-lexicon created by selecting from the basic lexicon all the SCFs which are also listed in the ANLT and/or COMLEX dictionaries plus the ones whose relative frequency is higher than a SCF-specific threshold.

Lexicon 5: A lexicon created as lexicon 4, but before filtering the SCF distributions are smoothed using linear interpolation (with the weight of 0.5)

4.1. Method

Automatically acquired SCF lexicons are usually evaluated against a gold standard obtained either through manual analysis of corpus data, or from SCF entries in a large dictionary. Manual analysis is usually the more reliable method and it has the benefit that it can be used to also evaluate the frequencies of SCFs, but our corpus data was too large for exhaustive manual analysis. Meanwhile, obtaining a gold standard from a dictionary is quick and can be applied to a large number of verbs, but the resulting standard may miss SCFs present in the corpus data or contain SCFs absent from the corpus data (particularly for low frequency verbs). As neither method was fully ideal for evaluation of our large lexicon(s), we used them both:

- 183 test verbs[8] were evaluated against manual analysis of some of the corpus data (at least 300 corpus occurrences per test verb[9])
- 5,659 verbs occurring in ANLT and/or COMLEX where evaluated against the SCFs in these dictionaries

The results were calculated using 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:

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

For the 183 test verbs with manually analysed SCF frequency data, we could also calculate the similarity between the acquired unfiltered[10] 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 evaluation of subcategorization acquisition can be found in Korhonen and Krymolowski (2002).

[5] This particularly applies to verbs taking multiple SCFs.

[6] http://www.google.com/apis/

[7] From the Web we extracted 963M words, but only a small part of this proved relevant because many retrieved documents included only one occurrence of the verb we were looking for.

[8] The verbs were selected in random but subject to the constraint that they take multiple SCFs.

[9] Given SCF distributions are Zipfian, this was sufficient to yield a realistic / reasonable distribution for most medium to high frequency SCFs and some low frequency ones.

[10] Note that no threshold was applied to remove the noisy SCFs from the distributions.

4.2. Results

Table 2 shows the average results for the 183 verbs in the 5 lexicons as evaluated against the manual analysis of corpus data. The basic unfiltered lexicon is, as expected, very noisy, yielding only 21.9 F-measure. Lexicon 2, created using the simple thresholding technique is substantially more accurate (58.6 F-measure) but it has low recall (46.1%). The smoothing technique employed when creating lexicon 3 addresses the sparse data problem, resulting in much better recall (63.3%) and better overall results (69.2 F-measure). The most accurate lexicons are, expectedly, those obtained by supplementing high frequency SCFs with lower frequency ANLT and COMLEX SCFs (lexicons 4 and 5). This method gives the best results when combined with smoothing (87.3 F-measure), yielding both better precision (93.1% vs. 90.4%) and better recall (82.2% vs. 78.0%).

The results in table 3 illustrate the impact of smoothing on the accuracy of SCF distributions. For example, the ranking of SCFs is clearly more accurate in lexicon 3 than in lexicons 1 and 2 (RC 0.81 vs. 0.59). Also the SCF distributions are notably more similar with gold standard distributions (e.g. KL 0.36 vs. 1.16).

	Lexicon				
Measures	1	2	3	4	5
Precision (%)	13.0	80.7	76.2	90.4	93.1
Recall (%)	69.4	46.1	63.3	78.0	82.2
F-measure	21.9	58.6	69.2	83.7	87.3

Table 2: Average precision, recall and F-measure for 183 verbs evaluated against the manual analysis of corpus data

	Lexicon	
Measures	1 & 2	3
KL	1.16	0.36
JS	0.10	0.04
CE	2.45	1.65
SD	0.65	0.21
RC	0.59	0.81
IS	0.84	0.95

Table 3: Average distributional similarity results for 183 verbs evaluated against the manual analysis of corpus data

	Lexicon		
Measures	1	2	3
Precision (%)	8.4	65.1	58.6
Recall (%)	80.4	40.7	56.0
F-measure	15.2	50.1	57.3

Table 4: Average precision, recall and F-measure for 5,659 verbs evaluated against ANLT and COMLEX

Table 4 shows the average results for the 5,659 verbs in lexicons 1-3 when evaluated against ANLT and COMLEX SCFs. These results are significantly lower than those in Table 2, particularly for lexicon 3 which yields only 57.3 F-measure. This is partly because the benefit of smoothing is less visible (the smoothing technique is only applicable to medium to high frequency verbs) but mostly because the dictionary-based gold standard is inadequate for the reasons discussed earlier. The fact that also the results in table 2 are slightly lower than usual[11] indicates some discrepancy between the large test data and the manually obtained gold standard. However the problem is smaller than with the dictionary-based gold standard.

The results reported here give nevertheless a general idea of the accuracy of each (sub-)lexicon. However, the optimal filtering and/or (sub-)lexicon option depends entirely on the intended use of the lexicon. For example, if the aim is to use SCF frequencies to aid parsing, a user may want to maximise the accuracy (rather than the coverage) of the lexicon. As shown in table 2, the most accurate lexicon for general language texts can be obtained by extracting ANLT, COMLEX and high frequency SCFs from the basic lexicon (see lexicons 4 and 5). However, for highly domain-specific (e.g. biomedical, astronomy, law) texts (for which ANLT and COMLEX SCFs may not be fully valid) the most accurate lexicon may well be obtained by extracting only high frequency SCFs from the basic lexicon (see lexicon 2). On the other hand, some NLP tasks may benefit from a lexicon which provides good coverage at the expense of accuracy. For example, Korhonen et al. (2003) obtained the best results with automatic verb classification when using an unfiltered SCF lexicon as input data (similar to lexicon 1). In this case noisy SCFs contained information useful for the task.

5. Summary

This paper has introduced a large computational subcategorization lexicon, acquired automatically from several corpora and the Web, which includes SCF frequency data for 6,397 English verbs. The lexicon is provided freely for research use, along with software which can be used to filter and build sub-lexicons suited for various NLP purposes. The filtering options and the accuracy of the resulting sub-lexicons is described in the documentation provided with the software. These resources potentiate the wider use of SCF lexicons in various (statistical) NLP tasks. This, in turn, should result in task-based evaluations which can provide valuable feedback for further development of automatic subcategorization acquisition.

Acknowledgments

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

[11] Compare e.g. with the results reported in (Korhonen, 2002a). Keeping in mind that the results are not fully comparable to the ones reported here (a smaller test corpus was used containing only 91 verbs), F-measure was 5-10 better, even though an older and a less accurate version of the same system was used. The main reason for this is the better fit between the smaller test data and the gold standard.

6. References

A. Arun and F. Keller. 2005. Lexicalization in crosslinguistic probabilistic parsing: The case of French. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 306–313, Ann Arbor, Michigan.

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.

M. Brent. 1993. From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics*, 19(3):243–262.

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

E. J. Briscoe. 2000. *Dictionary and System Subcategorisation Code Mappings*. Unpublished manuscript, University of Cambridge Computer Laboratory.

G. Carroll and M. Rooth. 1998. Valence induction with a head-lexicalized PCFG. In *3rd Conference on Empirical Methods in Natural Language Processing*, Granada, Spain.

J. Carroll, E. J. 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.

E. Ferrer. 2004. Towards a semantic classification of spanish verbs based on subcategorisation information. In *ACL Student Research Workshop*, Barcelona, Spain.

D. Graff. 1995. *North American News Text Corpus*. Linguistic Data Consortium. LDC95T21.

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.

J. Hajič, M. Čmejrek, B. Dorr, Y. Ding, J. Eisner, D. Gildea, T. Koo, K. Parton, G. Penn, D. Radev, and O. Rambow. 2002. Natural language generation in the context of machine translation. Technical report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore. Summer Workshop Final Report.

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

D. Kawahara, N. Kaji, and S. Kurohashi. 2000. Japanese Case Structure Analysis by Unsupervised Construction of a Case Frame Dictionary. In *Proc. of the 18th COLING*.

A. Korhonen and E. J. Briscoe. 2004. Extended Lexical-Semantic Classification of English Verbs. In *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics*, Boston, MA.

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 and J. Preiss. 2003. Improving Subcategorization Acquisition using Word Sense Disambiguation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 48–55, Sapporo, Japan.

A. Korhonen, Y. Krymolowski, and Z. Marx. 2003. Clustering Polysemic Subcategorization Frame Distributions Semantically. In *Proceedings of the 41st Annual Meeting of ACL*, pages 64–71, Sapporo, Japan.

A. Korhonen. 2002a. Semantically motivated subcategorization acquisition. In *ACL Workshop on Unsupervised Lexical Acquisition*, pages 51–58, Philadelphia.

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

M. Lapata, F. Keller, and S. Schulte im Walde. 2001. Verb frame frequency as a predictor of verb bias. *Journal of Psycholinguistic Research*, 30(4):419–435.

P. Laplace. 1995. *Philosophical Essays on Probabilities*. Springer-Verlag.

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.

D. McCarthy and J. Carroll. 2003. Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

D. McCarthy. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the 1st Conference of the North American Chapter of the ACL*, pages 229–277, Seattle, WA.

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.

T.G. Rose, M. Stevenson, , and M. Whitehead. 2002. The Reuters Corpus Volume 1 – from yesterday's news to tomorrow's language resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain.

S. Schulte im Walde and C. Brew. 2002. Inducing german semantic verb classes from purely syntactic subcategorisation information. In *Proc. of the 40th Annual Meeting of ACL*, Philadephia, USA.

M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proc. of the 41st Annual Meeting of ACL*, Sapporo.