Merging Lexicons for Higher Precision Subcategorization Frame Acquisition

Laura Rimell[*], **Thierry Poibeau**[†], **Anna Korhonen**[*]

[*] Dept of Theoretical and Applied Linguistics & Computer Laboratory, University of Cambridge, UK
[†] LaTTiCe, UMR8094, CNRS & ENS, France

Abstract

We present a new method for increasing the precision of an automatically acquired subcategorization lexicon, by merging two resources produced using different parsers. Although both lexicons on their own have about the same accuracy, using only sentences on which the two parsers agree results in a lexicon with higher precision, without too great loss of recall. This "intersective" resource merger is appropriate when both resources are automatically produced, hence noisy, or when precision is of primary importance, and may also be a useful approach for new domains where sophisticated filtering and smoothing methods are unavailable.

1. Introduction

Verb subcategorization frame (SCF) lexicons contain information about the subcategorization preferences of verbs, that is, the tendency of verbs to select the types of syntactic phrases with which they co-occur. For example, the verb *believe* can take a noun phrase complement, a clausal complement, or both together, while the verb *see* can take a noun phrase or a clausal complement, but not both together (Figure 1). SCF lexicons can serve as useful resources for applications requiring information about predicate-argument structure, including parsing (Carroll and Fang, 2004), semantic role labeling (Bharati et al., 2005), verb clustering (Schulte im Walde, 2006), information extraction (Surdeanu et al., 2003), and machine translation (Han et al., 2000).

Manually developed resources containing subcategorization information (Boguraev et al., 1987; Grishman et al., 1994) typically have high precision but suffer from a lack of coverage, making automatic acquisition desirable. The automatic acquisition of SCF information requires extraction of co-occurrence information from large amounts of unstructured text. A typical approach involves using a parser to discover the grammatical relations (GRs, i.e. dependencies) headed by each verb instance, then deciding which GR patterns constitute instances of various SCFs, either by heuristically matching a set of pre-defined patterns, or by accepting all patterns found within the data with a minimum frequency. The resulting set of SCF instances are amalgamated into an SCF lexicon, containing a probability distribution over SCFs for each verb lemma (Briscoe and Carroll, 1997; Korhonen, 2002; Preiss et al., 2007; Messiant et al., 2008; Lapesa and Lenci, 2011). Automatically acquired resources typically have higher coverage than manually developed ones, but suffer from a lack of precision.

SCF	Example
NP	Mary *believed* [$_{\text{NP}}$Susan].
CCOMP	Mary *believed* [$_{\text{CCOMP}}$that the book had been returned].
NP-CCOMP	Mary *believed* [$_{\text{NP}}$Susan] [$_{\text{CCOMP}}$that the book had been returned].
NP	Mary *saw* [$_{\text{NP}}$Susan].
CCOMP	Mary *saw* [$_{\text{CCOMP}}$that the book had been returned].
NP-CCOMP	*Mary *saw* [$_{\text{NP}}$Susan] [$_{\text{CCOMP}}$that the book had been returned].

Figure 1: Sample subcategorization frames taken by two verbs. The asterisk represents an ungrammatical sentence.

A number of filtering and smoothing techniques have been proposed in order to improve the precision of automatically acquired SCF lexicons. Filtering SCFs which are attested below a relative frequency threshold for any given verb, where the threshold is applied uniformly across the whole lexicon, has been shown to be effective (Korhonen, 2002; Messiant et al., 2008). However, this technique relies on empirical tuning of the threshold, necessitating a gold standard in the appropriate textual domain, and it is insensitive to the fact that some SCFs are inherently rare. The most successful methods of increasing accuracy in SCF lexicons rely on language- and domain-specific dictionaries to provide back-off distributions for smoothing (Korhonen, 2002).

This paper presents a different approach to acquiring a higher precision SCF resource, namely the merging

of two automatically acquired resources by retaining only the information that the two resources agree on. Previous work on language resource merging has generally focused on increasing coverage by adding information from one resource to another, e.g. (Crouch and King, 2005; Molinero et al., 2009), which focus on merging multiple levels of information from disparate resources. More closely related to our work, (Necsulescu et al., 2011; Bel et al., 2011; Padró et al., 2011) merge two manually built SCF lexicons, unifying SCFs when possible but with the goal of retaining information from both lexicons. Treating language resource merger as (roughly) a union operation is appropriate for manually developed resources, or when coverage is a priority. However, when working with automatically acquired resources it may be worthwhile to adopt the approach of merger by intersection.

We focus here on the fact that the tagger and parser are one source of noise in automatic SCF acquisition, and combine two lexicons built with different parsers. This approach is similar in spirit to parser ensembles, which have been used successfully to improve parsing accuracy (Sagae and Lavie, 2006; Sagae and Tsujii, 2007). We build two SCF lexicons using the framework of (Korhonen, 2002; Preiss et al., 2007), which was designed to classify the output of the RASP parser (Briscoe et al., 2006), and which we extend to classify the output of the unlexicalized Stanford parser (Klein and Manning, 2003). We then build a combined lexicon that includes only SCFs that are agreed on by both parsers. Using this simple combination approach, we obtain a lexicon with higher precision than the lexicon built with either parser alone.

2. Previous Work

Manually developed resources containing subcategorization information include ANLT (Boguraev et al., 1987) and COMLEX (Grishman et al., 1994). Automatically aquired SCF resources for English include (Briscoe and Carroll, 1997; Korhonen, 2002; Korhonen et al., 2006a; Preiss et al., 2007), and for other languages such resources as (Messiant et al., 2008) for French, and (Lapesa and Lenci, 2011) for Italian. The state of the art system for SCF acquisition in English is that of (Preiss et al., 2007), which we adopt and extend here. It uses manually defined rules to identify SCFs based on the output of the RASP parser.

The only previous work we are aware of on combining SCF lexicons is (Necsulescu et al., 2011; Bel et al., 2011; Padró et al., 2011). However, they combine manually developed lexicons. To our knowledge there is no previous work on combining automatically acquired SCF lexicons.

Parser ensembles have previously been used to improve parsing accuracy (Sagae and Lavie, 2006; Sagae and Tsujii, 2007), as well as for applications such as extraction of protein-protein interactions (Miyao et al., 2009).

3. System Description

We adapted the SCF acquisition system of (Preiss et al., 2007). First, corpus data is parsed to obtain GRs for each verb instance. We use the RASP parser and the unlexicalized Stanford parser. Second, a rule-based classifier matches the GRs for each verb instance with a corresponding SCF. The classifier of (Preiss et al., 2007) is based on the GR scheme of (Briscoe et al., 2006), used by the RASP parser. Since the Stanford parser produces output in the Stanford Dependencies (SD) scheme (de Marneffe et al., 2006), we developed a new version of the classifier for the Stanford output. We also made some minor modifications to the RASP classifier. At this stage we added a parser combination step, creating a new set of classified verb instances by retaining only instances for which the two classifiers agreed on the SCF. A lexicon builder then extracts relative frequencies from the classified data and builds lexical entries, and the resulting lexicons are filtered.

3.1. Parsing

SCF acquisition requires an unlexicalized parser, i.e. a parser that does not already have a notion of SCF probabilities conditioned on particular verb lemmas, so as not to bias the outcome towards the parser's existing knowledge. RASP is a modular statistical parsing system which includes a tokenizer, tagger, lemmatizer, and a wide-coverage unification-based tag-sequence parser, and has been used in a number of previous SCF acquisition experiments. The Stanford system includes a tokenizer, tagger, lemmatizer, and an unlexicalized[1] stochastic context-free grammar parser. We are unaware of any previous SCF acquisition using the Stanford parser.

3.2. Classifying Verb Instances

The classifier attempts to match the set of GRs produced for each verb instance against its inventory of SCFs, using a set of rules which were manually developed by examining parser output on development sentences. The classifier is implemented in Lisp and

[1]Though the Stanford parser is unlexicalized, the rules provided with the parser to generate GRs from a constituent parse are mildly lexicalized; for example, they can distinguish some raising verbs. This affects only a small number of SCFs. We made use of the information when it was available.

SCF: EXTRAP-TO-NP-S
It matters to them that she left.
RASP
`ncsubj(matter it _)`
`iobj(matter to)`
`ccomp(that matter leave)`
`ncsubj(leave she _)`
Stanford
`nsubj(matter it)`
`prep(matter to)`
`ccomp(matter leave)`
`nsubj(leave she)`

Figure 2: Example sentence with RASP and SD GRs (incidental formatting has been normalized). The classifier rules identify this SCF only when the word *it* is in subject position and the preposition is *to*.

examines the graph of GRs headed by the verb, finding the SCF which matches the greatest number of GRs. For example, if the verb has a direct object (NP) and an indirect object (PP), then the classifier will find SCF NP-PP, not NP. (Note that we do not include subjects in the SCF name, since they are obligatory in English.) For the RASP parser, we used a re-implementation in Lisp of the rule set in (Preiss et al., 2007). We made some minor modifications to the rules based on examination of development data.

Despite commonalities between the GR scheme of (Briscoe et al., 2006) and the SD scheme, the realization of a particular SCF can nevertheless exhibit a number of differences across schemes. Rather than converting the SD output to (Briscoe et al., 2006) format, a complex many-to-many mapping that would likely lose information, we chose to develop a new version of the classifier, based on examination of development data parsed by the Stanford parser. Figure 2 shows an example of parser output in the two schemes.

3.3. Merging Classifier Output

For the combined lexicon, we merged the classifier output on a sentence-by-sentence basis. A sentence was considered to exemplify an SCF for a verb only if both classifiers, RASP and Stanford, agreed on that SCF based on the parser output. Note that we did not merge the results of the lexicon building step (Section 3.4.), which would mean accepting an SCF on a verb-by-verb basis, if the two lexicons agreed that the verb takes that SCF. We chose not to use this strategy since we believed it would allow more errors of both parsers to pass through the pipeline.[2]

In some cases, differences in the two GR schemes allowed the parsers to take different views on the data. For example, RASP cannot distinguish the SCFs ADVP (*He meant well*) and PARTICLE (*She gave up*), since it analyzes both *well* and *up* as particle-type non-clausal modifiers. However, Stanford distinguishes the two as adverbial modifier and particle, respectively. In such cases we used the more fine-grained analysis in the resulting lexicon.

3.4. Lexicon Building and Filtering

The lexicon builder amalgamates the SCFs hypothesized by the classifier for each verb lemma. SCFs left underspecified by the classifier are also treated here. As the gold standard SCF inventory is very fine-grained, there are a number of distinctions which cannot be made based on parser output. For example, the gold standard distinguishes between transitive frame NP with a direct object interpretation (*She saw a fool*) and NP-PRED-RS with a raising interpretation (*She seemed a fool*), but parsers in general are unable to make this distinction. We used two different strategies at lexicon building time: weighting the underspecified SCFs by their frequency in general language, or choosing the single SCF which is most frequent in general language. For example, we either assign most of the weight to SCF NP with a small amount to NP-PRED-RS, or we assign all the weight to NP.

The goal of the parser combination method is to increase the precision of the acquired lexicon, which is also the goal of the various filtering methods for removing noise from SCF lexicons. In order to investigate the role of filtering in the context of parser combination, we filtered all the acquired lexicons using uniform relative frequency thresholds of 0.01 and 0.02.

4. Experiments

4.1. Gold Standard

We used the gold standard of (Korhonen et al., 2006b), consisting of SCFs and relative frequencies for 183 general-language verbs, based on approximately 250 manually annotated sentences per verb. The verbs were selected randomly, subject to the restriction that they take multiple SCFs. The gold standard includes 116 SCFs. Because of the Zipfian nature of SCF distributions – a few SCFs are taken by most verbs, while a large number are taken by few verbs – only 36 of these SCFs are taken by more than ten verbs in the gold standard.

[2]We also did not combine parsers by voting on individual GRs, to generate a new *parse* with higher accuracy than the individual parser output; this would have been difficult due to the differences between the GR schemes.

4.2. Corpus Data

The input corpus consisted of up to 10,000 sentences for each of the 183 verbs, from the British National Corpus (BNC) (Leech, 1993), the North American News Text Corpus (NANT) (Graff, 1995), the Guardian corpus, the Reuters corpus (Rose et al., 2002), and TREC-4 and TREC-5 data. Data was taken preferentially from the BNC, using the other corpora when the BNC had insufficient examples.

4.3. Evaluation Measures

We used type precision, recall, and F-measure for lexicon evaluation, as well as the number of SCFs present in the gold standard but missing from the unfiltered lexicon (i.e. not acquired, rather than filtered out). We also measured the distributional similarity between the acquired lexicons and the gold standard using various measures.

5. Results and Discussion

Tables 1 and 2 show the overall results for each parser alone as well as the combination, using the two different methods of resolving underspecified SCFs. We note first that the single-parser systems show similar accuracy across the different filtering thresholds. In Table 1, both systems achieve an F-score of about 18 for the unfiltered lexicon, and between 45 and 50 for the uniform frequency thresholds of 0.01 and 0.02. In Table 2, the accuracy is slightly higher overall, with both systems achieveing F-scores of about 21-22 for the unfiltered lexicon, and between 51-57 for the uniform frequency thresholds. The RASP-based system achieves higher accuracy than the Stanford-based system across the board, due to higher precision. We attribute this difference to the fact that the RASP classifier rules have been through several generations of development, while the Stanford rule set was first developed for this paper and has had the benefit of less fine-tuning, rather than to any difference in suitability of the two parsers for the task.

The merged lexicon shows a notable increase in precision at each filtering threshold compared to the single-parser lexicons, with, in most cases, a corresponding increase in F-score. In Table 1, the unfiltered lexicon achieves an F-score of 26.7, the lexicon with a uniform frequency threshold of 0.01 an F-score of 53.6, and with a uniform frequency threshold of 0.02 an F-score of 51.1. In Table 2, the unfiltered lexicon achieves an F-score of 35.7, the lexicon with a uniform frequency threshold of 0.01 an F-score of 59.4, and with a uniform frequency threshold of 0.02 an F-score of 56.8. Depending on the settings, the increase in precision over the higher of the single-parser lexicons ranges from about four points (Table 1, bottom row) to over 11 points (Table 2, middle row). This increase is achieved without developing any new classifier rules.

Filtering Method		RASP	Stanford	Comb.
Unfiltered	P	9.6	10.0	15.7
	R	95.8	95.4	90.3
	F	17.5	18.2	26.7
Uniform 0.01	P	42.7	38.6	50.8
	R	59.0	59.8	56.7
	F	49.6	46.9	53.6
Uniform 0.02	P	52.6	43.9	56.7
	R	48.8	47.2	46.6
	F	50.6	45.5	51.1

Table 1: Type precision, recall, and F-measure for 183 verbs. Underspecified SCFs weighted by frequency in general language.

Filtering Method		RASP	Stanford	Comb.
Unfiltered	P	12.1	12.9	22.8
	R	83.6	86.8	82.4
	F	21.2	22.5	35.7
Uniform 0.01	P	48.6	42.8	59.9
	R	62.5	62.7	58.9
	F	54.7	50.9	59.4
Uniform 0.02	P	61.5	51.4	68.3
	R	52.8	51.3	48.6
	F	56.8	51.3	56.8

Table 2: Type precision, recall, and F-measure for 183 verbs. Underspecified SCFs by taking the single most frequent SCF from the set.

An interesting effect of merging can be observed in the unfiltered case. The unfiltered lexicons all have an extreme bias towards recall over precision. Because of noise in the parser and classifier output, most SCFs are hypothesized for each verb. However, the merged lexicon shows higher precision even in the unfiltered case: effectively, the merger acts as a kind of filter.

The combined lexicon does show somewhat lower recall than the single-parser lexicons. This is probably due to the fact that the intersection of the two classifier outputs resulted in a much smaller number of sentences in the input to the lexicon builder. Recall that the original dataset contained up to 10,000 sentences per verb. Not all of these sentences were classified in each pipeline, either due to parser errors or to the GRs failing to match the rules for any SCF. On average, the RASP classifier classified 6,500 sentences per verb, the Stanford classifier 5,594, and the combined

classifier only 1,922. It should be noted that classifying more sentences does not necessarily mean better accuracy, since the classifications are noisy; in some cases it is preferential not to match on any SCF. In fact, the Stanford-based lexicon was based on fewer sentences than the RASP-based lexicon without loss of recall. However, the input corpus for the combined lexicon was effectively much smaller than the input corpus for the other two lexicons, which probably contributed to the loss of recall.

We found that the best results for the individual parsers were obtained with the higher threshold (0.02), and for the combination with the lower threshold (0.01). Again, this is probably due to the smaller effective number of sentences classified; rare SCFs were more likely to fall below the threshold. As the threshold value increases, the precision and F-score for the single-parser lexicons approach that of the combined lexicon, because increasing the threshold always has the effect of increasing precision at the expense of recall. Using a parser combination achieves the same effect without the need to tune the threshold.

The next measure we look at is the number of SCFs that were present in the gold standard but missing from the unfiltered lexicons, i.e. never identified at all by the SCF aquisition system (rather than filtered out). For this measure we use the weighting method of treating underspecified SCFs (as in Table 1); otherwise the assignment of probability mass to the most frequent SCF in the underspecified cases means that many more SCFs are missed. The results are shown in Table 3. The merged lexicon clearly suffers on this measure, as there were seven SCFs that it did not identify at all; however, these SCFs are all rare, so they are presumably not the most important ones for downstream applications. For example, the merged lexicon does not identify the frame PP-WHAT-TO-INF, e.g. *They deduced from Kim what to do*, or TO-INF-SUBJ, e.g. *To see them hurts*, both of which are rare in general language according to the ANLT dictionary. Sometimes SCFs were missed because each parser/classifier identified the SCF, but never both on the same sentence, and in other cases neither individual parser/classifier identified a true positive.

The one missing SCF for the unfiltered Stanford lexicon was POSS-ING, e.g. *She dismissed their writing novels*. The Stanford tagger consistently tags the gerund as NN rather than VVG, which makes the SCF impossible to identify.

On the other hand, the merged lexicon shows a clear increase in the number of SCFs it can identify accurately. Table 4 shows the SCFs identified with at least 50% accuracy (F-score) in the unfiltered lexicon; the combined system was able to do this for 17 SCFs, compared to 8 and 7 for the RASP- and Stanford-based systems, respectively. This includes the very important PP frame, e.g. *They apologized to him*, which is very frequent in general language and relies for its identification on accurate argument-adjunct discrimination. Several frames with *wh*-elements were also identified with greater than 50% accuracy in the combined lexicon but not the single-parser lexicons, such as WH-TO-INF, e.g. *He asked whether to clean the house*.

We next compare the acquired lexicons to the gold standard using various measures of distributional similarity: Kullback-Leibler divergence (KL), Jensen-Shannon divergence (JS), cross entropy (CE), skew divergence (SD), and rank correlation (RC). These measures all compare the SCF probability distributions learned by the SCF acquisition system for each verb lemma. Such measures are a useful complement to the type precision, recall, and F-score evaluation, because unlike the type-based measures, the distributional similarity measures compare the *frequencies* learned by the SCF acquisition system. We use several measures since they exhibit different sensitivity to noise in the data; see (Korhonen and Krymolowski, 2002) for a discussion of the application of the various distributional similarity measures to SCF acquisition.

	RASP	Stanford	Comb
Missing	0	1	7

Table 3: Missing SCFs in unfiltered lexicon.

	RASP	Stanford	Comb
INTRANSITIVE	•	•	•
TRANSITIVE	•	•	•
NP-PP	•	•	•
PARTICLE	•	•	•
PARTICLE-NP	•	•	•
PARTICLE-NP-PP	•	•	•
PARTICLE-PP	•		•
WH-TO-INF	•	•	•
ADVP			•
EXTRAP-TO-NP-S			•
HOW-S			•
HOW-TO-INF			•
PP			•
PP-HOW-TO-INF			•
WH-S			•
WHAT-S			•
FIN-CLAUSE-SUBJ			•

Table 4: SCFs identified with F-score of at least 50 in unfiltered lexicon.

Measure	RASP	Stanf	Comb
KL distance	0.376	0.376	0.337
JS divergence	0.072	0.083	0.059
cross entropy	1.683	1.680	1.619
skew divergence	0.345	0.358	0.297
rank correlation	0.627	0.599	0.666

Table 5: Distributional similarity measures comparing unfiltered lexicons to the gold standard, on SCFs common to both gold and acquired lexicon. Lower value means greater correlation: KL, JS, CE, SD. Higher value means greater correlation: RC.

	RASP	Stanford	Comb
SCFs proposed	94.5	90.2	54.7

Table 6: Average number of SCFs proposed per verb in the unfiltered lexicons. Average over 183 verbs in gold standard.

Table 5 shows the results of the distributional similarity comparisons on the unfiltered acquired lexicons. In each case the merged lexicon shows greater similarity to the gold standard than either of the single-parser lexicons.

Finally, an indication of how the parser combination acts as a kind of filter is given in Table 6, which shows the number of SCFs proposed for each verb lemma. The single-parser classifiers posit a higher number of SCFs: some genuine higher frequency SCFs, followed by a long noisy tail of false positives. The parser combination proposes only half the number of SCFs per verb lemma in the unfiltered lexicon.

6. Conclusion

We have combined the SCF classifier output for two parsers to produce a higher precision verb subcategorization lexicon than those resulting from the single-parser classifiers. This higher precision is achieved without the need for dictionaries or other external resources. Although there is a significant initial investment in defining the parser-specific SCF classifier rules for a particular unlexicalized parser to form part of the merged system, the resulting SCF acquisition system can subsequently be used across a variety of domains without additional effort. The improved precision is particularly interesting in the case of the unfiltered SCF lexicons, since the merger effectively acts as a kind of filter on incorrect SCFs. The unfiltered, merged lexicon is not accurate enough for downstream applications, but the filtered, merged lexicon also exhibits higher precision than the filtered single-parser lexicons. The interaction between parser combination and various filtering methods should be further investigated.

Future work should attempt to overcome the fact that the number of sentences successfully classified decreased dramatically with the parser combination, resulting in loss of recall. Using a larger input corpus would be a natural first step. Another natural extension which we leave for future work is to use a more nuanced version of the "intersective" merger; for example, increasing the likelihood of an SCF when the parsers/classifiers agree, but still retaining the sentences where they do not agree. It may also be possible to identify and leverage the particular strengths of each parser to aid in SCF identification.

Acknowledgements

This work was funded by the EU FP7 project 'PANACEA' and the Royal Society (UK). Thierry Poibeau is supported by the laboratoire d'excellence (labex) Empirical Foundation of Linguistics.

7. References

Núria Bel, Muntsa Padró, and Silvia Necsulescu. 2011. A method towards the fully automatic merging of lexical resources. In *Proceedings of the Workshop on Language Resources, Technology and Services in the Sharing Paradigm, IJCNLP-11*, Chiang Mai, Thailand.

Akshar Bharati, Sriram Venkatapathy, and Prashanth Reddy. 2005. Inferring semantic roles using subcategorization frames and maximum entropy model. In *Proceedings of CoNLL*, pages 165–168, Ann Arbor.

B. Boguraev, J. Carroll, E.J. Briscoe, D. Carter, and C. Grover. 1987. The derivation of a grammatically-indexed lexicon from the Longman Dictionary of Contemporary English. In *Proceedings of the 25th Annual Meeting of ACL*, pages 193–200, Stanford, CA.

E.J. Briscoe and J. Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC.

E.J. Briscoe, J. Carroll, and R. Watson. 2006. The second release of the RASP system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*.

J. Carroll and A. Fang. 2004. The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *Proceedings of the 1st International Joint Conference*

on Natural Language Processing (IJCNLP), pages 107–114, Sanya City, China.

D. Crouch and T.H. King. 2005. Unifying lexical resources. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Saarbruecken, Germany.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*.

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

R. Grishman, C. Macleod, and A. Meyers. 1994. COMLEX syntax: Building a computational lexicon. In *Proceedings of COLING*, Kyoto.

C. Han, Benoit Lavoie, Martha Palmer, Owen Rambow, Richard Kittredge, Tanya Korelsky, and Myunghee Kim. 2000. Handling structural divergences and recovering dropped arguments in a Korean/English machine translation system. In *In Proceedings of the AMTA*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL*, pages 423–430.

A. Korhonen and Y. Krymolowski. 2002. On the robustness of entropy-based similarity measures in evaluation of subcategorization acquisition systems. In *Proceedings of the Sixth CoNLL*, pages 91–97, Taipei, Taiwan.

Anna Korhonen, Yuval Krymolowski, and Ted Briscoe. 2006a. A large subcategorization lexicon for natural language processing applications. In *In Proceedings of LREC*.

Anna Korhonen, Yuval Krymolowski, and Ted Briscoe. 2006b. A large subcategorization lexicon for natural language processing applications. In *Proceedings of LREC*.

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

Gabriella Lapesa and Alessandro Lenci. 2011. Modeling subcategorization through co-ocurrence. Presented at Explorations in Syntactic Government and Subcategorization, Cambridge, UK, September 2011.

G. Leech. 1993. 100 million words of English. *English Today*, 9(1):9–15.

Cédric Messiant, Anna Korhonen, and Thierry Poibeau. 2008. LexSchem: A large subcategorization lexicon for French verbs. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marrakech.

Yusuke Miyao, Kenji Sagae, Rune Saetre, Takuya Matsuzaki, and Jun'ichi Tsujii. 2009. Evaluating contributions of natural language parsers to protein-protein interaction extraction. *Bioinformatics*, 25:394–400.

Miguel A. Molinero, Benoît Sagot, and Lionel Nicolas. 2009. A morphological and syntactic wide-coverage lexicon for Spanish: The Leffe. In *Proceedings of RANLP*, Borovets, Bulgaria.

Silvia Necsulescu, Núria Bel, Muntsa Padró, Montserrat Marimon, and Eva Revilla. 2011. Towards the automatic merging of language resources. In *Proceedings of the International Workshop on Lexical Resources (WoLeR)*, Ljubljana, Slovenia.

Muntsa Padró, Núria Bel, and Silvia Necsulescu. 2011. Towards the automatic merging of lexical resources: Automatic mapping. In *Proceedings of RANLP*, Hissar, Bulgaria.

Judita Preiss, Ted Briscoe, and Anna Korhonen. 2007. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of ACL*.

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*, pages 29–31.

Kenji Sagae and Alon Lavie. 2006. Parser combination by reparsing. In *Proceedings of NAACL HLT: Short Papers*, pages 129–132.

Kenji Sagae and Jun'ichi Tsujii. 2007. Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proceedings of the CoNLL Shared Task*, pages 1044–1050.

Sabine Schulte im Walde. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194.

M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of ACL*, Sapporo.