Automatic extraction of subcategorization frames for French

Paula Chesley
Linguistics Department, University at Buffalo, USA & ATILF-CNRS, France
pchesley@buffalo.edu

Susanne Salmon-Alt
ATILF-CNRS, France
salt@atilf.fr

This paper describes the integration of corpus-based syntactic subcategorization frames into a large-scale, theory-neutral lexical resource for French (Romary et al. (2004)). This database is the first to implement the Lexical Markup Framework (LMF), an international initiative towards ISO standards for lexical databases (ISO TC 37/SC 4). The subcategorization frames have been acquired via a dependency-based parser (Bick (2003)), whose verb lexicon is currently incomplete with respect to subcategorization frames. Therefore, we have implemented probabilistic filtering as a post-parsing step. Similar methods have proven efficient for English (Brent (1992), Manning (1993), Briscoe and Carroll (1997)) and other languages (Sarkar and Zeman (2000)). On a test set of 104 frequent verbs, most of which are in the Test Suites for Natural Language Processing for French (Lehmann et al. (1996)), we achieve lower bounds on type precision at 86.8% and on token recall at 54.3%. Additionally, our work demonstrates that using a language resource, such as the VISL parser, with a currently unevaluated (and potentially high) error rate can yield robust results in conjunction with probabilistic filtering of the resource output. For languages currently lacking robust parsers and large-scale treebanks, such findings are favorable indeed.

We created a multi-genre corpus of 200 random occurrences of the 104 verbs using the Frantext online literary database\textsuperscript{1}. We then had this corpus parsed with the VISL parser. As opposed to many works on subcategorization frame extraction, we do not assume a set of a priori frames to be extracted before filtering. Rather, the frames are determined throughout the filtering stage. However, we limit the syntactic categories treated to the following:

- direct objects;
- prepositional phrases (PPs) headed by a particular preposition;
- subordinate clauses and small clauses with various markers;
- raising and control verbs;
- predicative adjectival phrases;
- reflexive verbs.

The resulting subcategorization frames can consist of any combination of the above categories.

\textsuperscript{1}This resource is available at \url{http://www.frantext.fr/categ.htm}.
We implemented the binomial distribution to filter the noisy parser output. Specifically, this output conflates syntactic complements, syntactic modifiers, and incorrect PP and complement attachments. A sentence with a direct object, a subcategorized PP, and three additional modifying PPs is treated in the same fashion as a true subcategorization frame. As a result of this approach, many true subcategorization frames will be embedded in other erroneous frames proposed by the parser. In our implementation of the binomial distribution, these frames neither count as evidence for nor against observing the true frame.

The key element in using the binomial distribution for such purposes consists in finding the proper error rate for each frame; in so doing we adapted the method used in Brent (1992) to better fit our data and methodology. Brent’s procedure consists in examining every occurrence of a frame with every verb above a certain number of occurrences. From these occurrences we construct a histogram based on the number of co-occurrences of cues and the verbs with a sufficient amount of corpus attestations. We look for a binomial distribution toward the lower end of the histogram that signals the false cues. The average in this distribution is a proper estimation of the rate of false cues. This method requires a large number of examples in order to form regularities in frame distribution. In our data, the putative frames output by the parser are relatively rare. However, each subcategorized constituent in the frame does appear frequently. We thus decided to find the error rate for each subcategorized constituent, and to sum the error rates of all subcategorized constituents in a frame to find the error rate for the entire frame. This implementation allows for all realizations of a given syntactic constituent to contribute to the global error rate of the frame, and not only to the error rate of the specific, potentially rare putative frame. It also yields robust results in spite of the noisy parser output and sparse subcategorization frame data.

Our results from this experiment are 24 unique subcategorization frames and 176 verb-frame combinations. To evaluate this work, we submitted our results to two native speakers of French also researching subcategorization frames. This method of evaluation was chosen since currently there is no electronic resource explicitly encoding subcategorization frames for French. Inter-rater agreement was judged reliable, at $K = .82$. We take the lower bound for precision to be the intersection of subcategorization frames both raters judged correct; this figure is 86.8%. Conversely, the upper bound for precision, 96.4%, is the union of the frames at least one rater judged felicitous. These results are consistent with those of the works mentioned above. Token recall was found to be 54.3%. This figure was found by examining four random occurrences of each verb from a corpus of online French newspaper articles. Crucially, this recall rate is probably closer to a lower bound, since the 104 verbs in our sample are extremely frequent and polysemous, and thus exhibit many different subcategorization frames. It is natural to assume that our recall will be higher for verbs exhibiting fewer subcategorization frames. Our relatively low recall rate is also due to our conception of the subcategorization frames, since we did not initially wish to conflate reflexive subcategorization frames with their non-reflexive counterparts.

Our plans for future work include determining how to deal with reflexive verbs with respect to subcategorization frames and how to better exploit noisy parser data. Nevertheless, as the first work undertaken on extracting subcategorization frames from corpora for French, we find these results encouraging, given the subcategorization challenges specific to French, such as

---

2The criteria they were given for this categorical judgment task are available at [http://www-student.cse.buffalo.edu/~pchesley/subcats/criteria.pdf](http://www-student.cse.buffalo.edu/~pchesley/subcats/criteria.pdf).

3It is worth noting that the results of Manning (1993) have a lower ratio of learned frames to verbs than us. However, Manning’s recall rate, calculated similarly to ours, is 82%. This difference is probably best explained in that his recall is for random verb occurrences.
clitics, argument composition in general, and post-verbal constituent order.

References

[Bick2003] E. Bick. 2003. A CG & PSG Hybrid Approach to Automatic Corpus Annotation. pages 1–12. Corpus Linguistics, Lancaster.

[Brent1992] M. Brent. 1992. Robust Acquisition of Subcategorization Frames from Unrestricted Text: Unsupervised Learning with Syntactic Knowledge. Master’s thesis, Johns Hopkins University, Baltimore, MD.

[Briscoe and Carroll1997] T. Briscoe and J. Carroll. 1997. Automatic Extraction of a Subcategorization from Corpora. pages 356–363. Proceedings of the 5th ACL Conference on Applied Natural Language Processing.

[Lehmann et al.1996] Sabine Lehmann, Dominique Estival, and Stephan Oepen. 1996. TSNLP — Des jeux de phrases-test pour l’évaluation d’applications dans le domaine du TALN. In Proceedings of the 3rd Traitement Automatique du Langage Naturel (TALN) Conference, Marseille, May.

[Manning1993] C. Manning. 1993. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. pages 235–242. Proceedings of the 31st ACL.

[Romary et al.2004] L. Romary, G. Francopooulo, and S. Salmon-Alt. 2004. Standards going concrete: from LMF to Morphalou. COLING workshop.

[Sarkar and Zeman2000] Anoop Sarkar and Daniel Zeman. 2000. Automatic extraction of subcategorization frames for czech. In Proceedings of the 18th conference on Computational linguistics, pages 691–697, Morristown, NJ, USA. Association for Computational Linguistics.