A Subcategorization Acquisition System for French Verbs

Cédric Messiant
Laboratoire d'Informatique de Paris-Nord
CNRS UMR 7030 and Université Paris 13
99, avenue Jean-Baptiste Clément, F-93430 Villetaneuse France
cedric.messiant@lipn.univ-paris13.fr

Abstract

This paper presents a system capable of automatically acquiring subcategorization frames (SCFs) for French verbs from the analysis of large corpora. We applied the system to a large newspaper corpus (consisting of 10 years of the French newspaper 'Le Monde') and acquired subcategorization information for 3267 verbs. The system learned 286 SCF types for these verbs. From the analysis of 25 representative verbs, we obtained 0.82 precision, 0.59 recall and 0.69 F-measure. These results are comparable with those reported in recent related work.

1 Introduction

Many Natural Language Processing (NLP) tasks require comprehensive lexical resources. Handcrafting large lexicons is labour-intensive and error-prone. A growing body of research focuses therefore on automatic acquisition of lexical resources from text corpora.

One useful type of lexical information for NLP is the number and type of the arguments of predicates. These are typically expressed in simple syntactic frames called subcategorization frames (SCFs). SCFs can be useful for many NLP applications, such as parsing (John Carroll and Briscoe, 1998) or information extraction (Surdeanu et al., 2003). Automatic acquisition of SCFs has therfore been an active research area since the mid-90s (Manning, 1993; Brent, 1993; Briscoe and Carroll, 1997).

Comprehensive subcategorization information is currently not available for most languages. French is one of these languages: although manually built syntax dictionaries do exist (Gross, 1975; van den Eynde and Mertens, 2006; Sagot et al., 2006) none of them are ideal for computational use and none also provide frequency information important for statistical NLP.

We developed *ASSCI*, a system capable of extracting large subcategorization lexicons for French verbs from raw corpus data. Our system is based on a approach similar to that of the well-known Cambridge subcategorization acquisition system for English (Briscoe and Carroll, 1997; Preiss et al., 2007). The main difference is that unlike the Cambridge system, our system does not employ a set of predefined SCF types, but learns the latter dynamically from corpus data.

We have recently used *ASSCI* to acquire *LexSchem* – a large subcategorization lexicon for French verbs – from a raw journalistic corpus. and have made the resulting resource freely available to the community on the web (Messiant et al., 2008).

We describe our SCF acquisition system in section 2 and explain the acquisition of a large subcategorization lexicon for French and its evaluation in section 3. We finally compare our study with work previously achieved for English and French in section 4.

2 ASSCI: The Acquisition System

Our SCF acquisition system takes as input corpus data and produces a list of frames for each verb that occurred more than 200 times in the corpus. It the first system that automatically induces a large-scale SCF information from raw corpus data for French.

Proceedings of the ACL-08: HLT Student Research Workshop (Companion Volume), pages 55–60,
Columbus, June 2008.

Previous experiments focussed on a limited set of verbs (Chesley and Salmon-Alt, 2006), or were based on treebanks or on substantial manual work (Gross, 1975; Kupść, 2007).

The system works in three steps:

1. verbs and surrounding phrases are extracted from parsed corpus data;
2. tentative SCFs are built dynamically, based on morpho-syntactic information and relations between the verb and its arguments;
3. a statistical filter is used to filter out incorrect frames.

2.1 Preprocessing

When aiming to build a large lexicon for general language, the input data should be large, balanced and representative enough. Our system tags and lemmatizes input data using *TreeTagger* (Schmid, 1994) and then syntactically analyses it using *Syntex* (Bourigault et al., 2005). The *TreeTagger* is a statistical, language independent tool for the automatic annotation of part-of-speech and lemma information. *Syntex* is a shallow parser for extracting lexical dependencies (such as adjective/noun or verb/noun dependencies). *Syntex* obtained the best precision and F-measure for written French text in the recent EASY evaluation campaign[1].

The dependencies extracted by the parser include both arguments and adjuncts (such as location or time phrases). The parsing strategy is based on heuristics and statistics only. This is ideal for us since no lexical information should be used when the task is to acquire it. *Syntex* works on the general assumption that the word on the left side of the verb is the subject, where as the word on the right is the object. Exceptions to this assumption are dealt with a set of rules.

```
(2) Ces propriétaires exploitants
achètent ferme le carburant  la
compagnie .
(These owners buy fast the fuel to
the company.)
```

(3) is the preprocessed *ASSCI* input for sentence (2) (after the *TreeTagger* annotation and *Syntex*'s analysis).

```
(3) DetMP|ce|Ces|1|DET;3|
AdjMP|propriétaire|propriétaires|2|ADJ;3|
NomMP|exploitant|exploitants|3||DET;1,ADJ;2
VCONJP|acheter|achètent|4||ADV;5,OBJ;7,PREP;8
Adv|ferme|ferme|5|ADV;11|
DetMS|le|le|6|DET;7|
NomMS|carburant|carburant|7|OBJ;4|DET;6
Prep|à|à|8|PREP;4|NOMPREP;10
DetFS|le|la|9|DET;10|
NomFS|compagnie|compagnie|10|NOMPREP;8|DET;9
Typo|.|.|11||
```

2.2 Pattern Extractor

The pattern extraction module takes as input the syntactic analysis of *Syntex* and extracts each verb which is sufficiently frequent (the minimum of 200 corpus occurrences) in the syntactically analysed corpus data, along with surrounding phrases. In some cases, this module makes deeper use of the dependency relations in the analysis. For example, when a preposition is part of the dependencies, the pattern extractor examines whether this preposition is followed by a noun phrase or an infinitive clause.

(4) is the output of the pattern extractor for (3).

```
(4) VCONJP|acheter
NomMS|carburant|OBJ_Prep|à+SN|PREP
```

Note that +SN marks that the *"à"* preposition is followed by a noun phrase.

2.3 SCF Builder

The next module examines the dependencies according to their syntactic category (e.g., noun phrase) and their relation to the verb (e.g., object), if any. It constructs frames dynamically from the following features: a nominal phrase; infinitive clause; prepositional phrase followed by a noun phrase; prepositional phrase followed by an infinitive clause; subordinate clause and adjectival phrase. If the verb has no dependencies, its SCF is "intransitive" (INTRANS). The number of occurrences for each

[1] http://www.limsi.fr/Recherche/CORVAL/easy
The scores and ranks of *Syntex* at this evaluation campaign are available at http://w3.univ-tlse2.fr/erss/textes/pagespersos/bourigault/syntex.html#easy

SCF and the total number of occurrences with each verb are recorded.

This dynamic approach to SCF learning was adopted because no sufficiently comprehensive list of SCFs was available for French (most previous work on English (e.g., (Preiss et al., 2007)) employs a set of predefined SCFs because a relatively comprehensive lists are available for English).

The SCF candidate built for sentence (2) is shown in (5)[2].

(5) SN_SP[à+SN]

2.4 SCF Filter

The final module filters the SCF candidates. A filter is necessary since the output of the second module is noisy, mainly because of tagging and parsing errors but also because of the inherent difficulty of argument-adjunct distinction which ideally requires access to the lexical information we aim to acquire, along with other information and criteria which current NLP systems (and even humans) find it difficult to identify. Several previous works (e.g., (Briscoe and Carroll, 1997; Chesley and Salmon-Alt, 2006)) have used binomial hypothesis testing for filtering. Korhonen et al. (2000) proposes to use the maximum likelihood estimate and shows that this method gives better results than the filter based on binomial hypothesis testing. This method employs on a simple threshold over the relative frequencies of SCFs candidates. (The maximum likehood estimate is still an option in the current Cambridge system but an improved version calculates it specific to different SCFs - a method which we left for future work).

The relative frequency of the SCF *i* with the verb *j* is calculated as follows:

$$rel\_freq(scf_i, verb_j) = \frac{|scf_i, verb_j|}{|verb_j|}$$

$|scf_i, verb_j|$ is the number of occurrences of the SCF *i* with the verb *j* and $|verb_j|$ is the total number of occurrences of the verb *j* in the corpus.

These estimates are compared with the threshold value to filter out low probability frames for each verb. The effect of the choice of the threshold on the results is discussed in section 3.

[2]SN stands for a noun phrase and SP for a prepositional phrase

3 Experimental Evaluation

3.1 Corpus

In order to evaluate our system on a large corpus, we gathered ten years of the French newspaper *Le Monde* (two hundred millions words). It is one of the largest corpus for French and "clean" enough to be easily and efficiently parsed. Because our aim was to acquire a large general lexicon, we require the minimum of 200 occurrences per each verb we analysed using this system.

3.2 LexSchem: The Acquired Lexicon

3267 verbs were found with more than 200 occurrences in the corpus. From the data for these verbs, we induced 286 distinct SCF types. We have made the extracted lexicon freely available on the web (http://www-lipn.univ-paris13.fr/~messiant/lexschem.html) under the LGPL-LR (Lesser General Public License For Linguistic Resources) license. An interface which enables viewing the SCFs acquired for each verb and the verbs taking different SCFs is also available at the same address. For more details of the lexicon and its format, see (Messiant et al., 2008).

3.3 Gold Standard

Direct evaluation of subcategorization acquisition performance against a *gold standard* based on a manmade dictionary is not ideal (see e.g. (Poibeau and Messiant, 2008)). However, this method is still the easiest and fastest way to get an idea of the performance of the system. We built a *gold standard* using the SCFs found in the *Trésor de la Langue Française Informatisé (TFLI)*, a large French dictionary available on the web[3]. We evaluated 25 verbs listed in Appendix to evaluate our system. These verbs were chosen for their heterogeneity in terms of semantic and syntactic features, but also because of their varied frequency in the corpus (from 200 to 100.000 occurences).

3.4 Evaluation Measures

We calculated type precision, type recall and F-measure for these 25 verbs. We obtain the best results (0.822 precision, 0.587 recall and 0.685 f-measure) with the MLE threshold of 0.032 (see fig-

[3]http://atilf.atilf.fr/

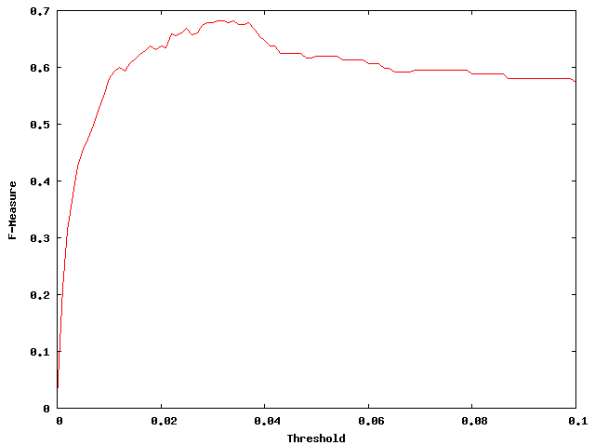

Figure 1: The relation of the threshold on the F-Measure

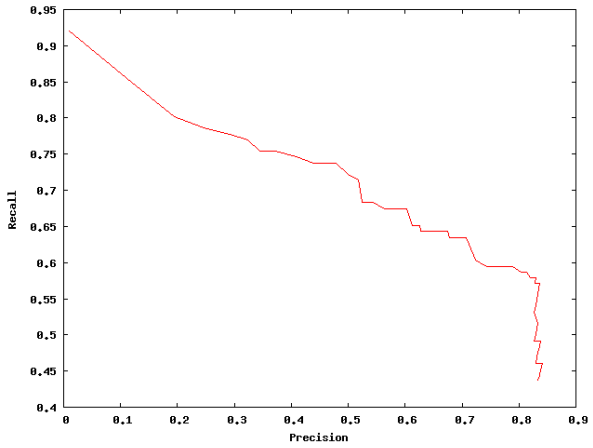

Figure 2: The relation between precision and recall

ure 1). Figure 2 shows that even by substantially lowering recall we cannot raise precision over 0.85.

Table 1 shows a comparison of three versions of *ASSCI* for our 25 verbs:

- Unfiltered: the unfiltered output of *ASSCI*;
- ASSCI-1: one single threshold fixed to 0.0325;
- ASSCI-2: one INTRANS-specific threshold (0.08) and the 0.0325-threshold for all other cases.

These results reveal that the unfiltered version of the lexicon is very noisy indeed (0.01 precision).

System	Precision	Recall	F-Measure
Unfiltered	0.010	0.921	0.020
ASSCI-1	0.789	0.595	0.679
ASSCI-2	0.822	0.587	0.685

Table 1: Comparison of different versions of *ASSCI*

A simple threshold on the relative frequencies improves the results dramatically (*ASSCI-1*).

Each step of the acquisition process generates errors. For example, some nouns are tagged as a `verb` by *TreeTagger* (e.g., in the phrase *"Le programme d'armement (weapons program)"*, "*programme*" is tagged `verb`). *Syntex* generates errors when identifying dependencies: in some cases, it fails to identify relevant dependencies; in other cases incorrect dependencies are generated. The SCF builder is another source of error because of the ambiguity or the lack of sufficient information to build some frames (e.g. those involving pronouns). Finally, the filtering module rejects some correct SCFs and accept some incorrect ones. We could reduce these errors by improving the filtering method or refining the thresholds.

Many of the errors involve intransitive SCFs. We tried to address this problem with an INTRANS-specific threshold which is higher than others (see the results for *ASSCI-2*). This improves the precision of the system slightly but does not substantially reduce the number of false negatives. The intransitive form of verbs is very frequent in corpus data but it doesn't appear in the *gold standard*. A better evaluation (e.g., a gold standard based on manual analysis of the corpus data and annotation for SCFs) should not yield these errors. In other cases (e.g. interpolated clauses), the parser is incapable of finding the dependencies. In subsequent work we plan to use an improved version of *Syntex* which deals with this problem.

Our results (*ASSCI-2*) are similar with those obtained by the only directly comparable work for French (Chesley and Salmon-Alt, 2006) (0.87 precision and 0.54 recall). However, the lexicons show still room for improvement, especially with recall. In addition to the improvements in the method and evaluation suggested above, we plan to evaluate whether lexicons resulting from our system are use-

ful for NLP tasks and applications. For example, John Carroll *& al.* shows that a parser can be significantly improved by using a SCF lexicon despite a high error rate (John Carroll and Briscoe, 1998).

4 Related Work

4.1 Manual or Semi-Automatic Work

Most previous subcategorization lexicons for French were built manually. For example, Maurice Gross built a large French dictionnary called "*Les Tables du LADL*" (Gross, 1975). This dictionary is not easy to employ for NLP use but work in progress is aimed at addressing this problem (Gardent et al., 2005). The *Lefff* is a morphological and syntactic lexicon that contains partial subcategorization information (Sagot et al., 2006), while *Dicovalence* is a manually built valency dictionnary based on the pronominal approach (van den Eynde and Blanche-Benveniste, 1978; van den Eynde and Mertens, 2006). There are also lexicons built using semi-automatic approaches e.g., the acquisition of subcategorization information from treebanks (Kupść, 2007).

4.2 Automatic Work

Experiments have been made on the automatic acquisition of subcategorization frames since mid 1990s (Brent, 1993; Briscoe and Carroll, 1997). The first experiments were performed on English but since the beginning of 2000s the approach has been successfully applied to various other languages. For example, (Schulte im Walde, 2002) has induced a subcategorization lexicon for German verbs from a lexicalized PCFG. Our approach is quite similar to the work done in Cambridge. The Cambridge system has been regularly improved and evaluated; and it represents the state-of-the-art perfomance on the task (Briscoe and Carroll, 1997; Korhonen et al., 2000; Preiss et al., 2007). In the latest paper, the authors show that the method can be successfully applied to acquire SCFs not only for verbs but also for nouns and adjectives (Preiss et al., 2007). A major difference between these related works and ours is the fact that we do not use a predefined set of SCFs. Of course, the number of frames depends on the language, the corpus, the domain and the information taken into account (for example, (Preiss et al., 2007) used a list of 168 predefined frames for English which abstract over lexically-governed prepositions).

As far as we know, the only directly comparable work on subcategorization acquisition for French is (Chesley and Salmon-Alt, 2006) who propose a method for acquiring SCFs from a multi-genre corpus in French. Their work relies on the VISL parser which have an "unevaluated (and potentially high) error rate" while our system relies on *Syntex* which is, according to the *EASY evaluation campaign*, the best parser for French (as evaluated on general newspaper corpora). Additionally, we acquired a large subcategorization lexicon (available on the web) (286 distinct SCFs for 3267 verbs) whereas (Chesley and Salmon-Alt, 2006) produced only 27 SCFs for 104 verbs and didn't produce any lexicon for public release.

5 Conclusion

We have introduced a system which we have developed for acquiring large subcategorization lexicons for French verbs. When the system was applied to a large French newspaper corpus, it produced a lexicon of 286 SCFs corresponding to 3267 verbs. We evaluated this lexicon by comparing the SCFs it produced for 25 test verbs to those included in a manually built dictionary and obtained promising results. We made the automatically acquired lexicon freely available on the web under the LGPL-LR license (and through a web interface).

Future work will include improvements of the filtering module (using e.g. SCF-specific thresholds or statistical hypothesis testing) and exploration of task-based evaluation in the context of practical NLP applications and tasks such as the acquisition of semantic classes from the SCFs (Levin, 1993).

Acknowledgements

Cédric Messiant's PhD is funded by a DGA/CNRS Grant. The research presented in this paper was also supported by the ANR MDCO 'CroTal' project and the British Council and the French Ministry of Foreign Affairs -funded 'Alliance' grant.

References

Didier Bourigault, Marie-Paule Jacques, Cécile Fabre, Cécile Frérot, and Sylwia Ozdowska. 2005. Syntex,

analyseur syntaxique de corpus. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, Dourdan.

Michael R. Brent. 1993. From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. *Computational Linguistics*, 19:203–222.

Ted Briscoe and John 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.

Paula Chesley and Susanne Salmon-Alt. 2006. Automatic extraction of subcategorization frames for French. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Genua (Italy).

Claire Gardent, Bruno Guillaume, Guy Perrier, and Ingrid Falk. 2005. Maurice Gross' Grammar Lexicon and Natural Language Processing. In *2nd Language and Technology Conference*, Poznan.

Maurice Gross. 1975. *Méthodes en syntaxe*. Hermann, Paris.

Guido Minnen John Carroll and Ted Briscoe. 1998. Can subcategorisation probabilities help a statistical parser? In *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*, Montreal (Canada).

Anna Korhonen, Genevieve Gorrell, and Diana McCarthy. 2000. Statistical filtering and subcategorization frame acquisition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong.

Anna Kupść. 2007. Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré. In *Actes des 14èmes journées sur le Traitement Automatique des Langues Naturelles*, Toulouse, June.

Beth Levin. 1993. *English Verb Classes and Alternations: a preliminary investigation*. University of Chicago Press, Chicago and London.

Christopher D. Manning. 1993. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In *Proceedings of the Meeting of the Association for Computational Linguistics*, pages 235–242.

Cédric Messiant, Anna Korhonen, and Thierry Poibeau. 2008. LexSchem : A Large Subcategorization Lexicon for French Verbs. In *Language Resources and Evaluation Conference (LREC)*, Marrakech.

Thierry Poibeau and Cédric Messiant. 2008. Do We Still Need Gold Standard For Evaluation ? In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marrakech.

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 the Meeting of the Association for Computational Linguistics*, pages 912–918, Prague.

Benoît Sagot, Lionel Clément, Eric de La Clergerie, and Pierre Boullier. 2006. The Lefff 2 syntactic lexicon for French: architecture, acquisition, use. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Genua (Italy).

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, Manchester, UK. unknown.

Sabine Schulte im Walde. 2002. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*, volume IV, pages 1351–1357, Las Palmas de Gran Canaria, Spain.

Mihai Surdeanu, Sanda M. Harabagiu, John Williams, and Paul Aarseth. 2003. Using Predicate-Argument Structures for Information Extraction. In *Proceedings of the Association of Computational Linguistics (ACL)*, pages 8–15.

Karel van den Eynde and Claire Blanche-Benveniste. 1978. Syntaxe et mécanismes descriptifs : présentation de l'approche pronominale. *Cahiers de Lexicologie*, 32:3–27.

Karel van den Eynde and Piet Mertens. 2006. *Le dictionnaire de valence Dicovalence : manuel d'utilisation*. Manuscript, Leuven.

Appendix — List of test verbs

compter	donner	apprendre
chercher	possder	comprendre
concevoir	proposer	montrer
rendre	s'abattre	jouer
offrir	continuer	ouvrir
aimer	croire	exister
obtenir	refuser	programmer
acheter	rester	s'ouvrir
venir		