

A Subcategorization Frames 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
firstname.lastname@lipn.univ-paris13.fr

Abstract

This paper presents a system intended to automatically acquire subcategorization frames (SCFs) of verbs from the analysis of large corpora. The system has been applied to a newspaper corpus (made of 10 years of the French newspaper *Le Monde*) and acquired subcategorization information for 3267 verbs. 286 SCFs were dynamically learnt for these verbs. From the analysis of 25 representative verbs, we obtained 0.83 precision, 0.59 recall and 0.69 F-measure. These results are comparable with those reported in recent work.

1 Introduction

Nowadays, most Natural Language Processing (NLP) tools require deep lexical resources. However, hand-crafting lexicons is labour-intensive and error-prone. There is therefore a growing body of research regarding the automatic acquisition of lexical resources, especially from electronic corpora.

A part of the required lexical information for NLP applications is the number and the types of the arguments related to predicates, i.e. the subcategorization frames (SCFs) of the predicative items. SCFs are useful in many NLP applications, such as parsing (John Carroll and Briscoe, 1998) or information extraction (Surdeanu et al., 2003). Thus, automatic acquisition of such information has become a major area of research since the early 90s (Manning, 1993; Brent, 1993; Briscoe and Carroll, 1997).

Subcategorization information is currently not available for most languages; it is the case for French, even if some partial lexical bases (mostly

manually built) exist. We developed *ASSCI*, a system capable of extracting large subcategorization lexicons for French verbs from raw corpus. Our approach is based on an adaptation of the work done in Cambridge (Briscoe and Carroll, 1997; Preiss et al., 2007), which is a well-tried system for English. Using *ASSCI*, we have induced *LexSchem*, a large subcategorization lexicon for French verbs, from a raw journalistic corpus. We do not use a fixed set of SCFs defined beforehand, but the list of SCFs is dynamically learnt from the corpus. The resulting resource is made available to the community on the web (see below).

Most of previous theoretical work about subcategorization make a distinction between arguments and adjuncts. Typically, arguments are obligatory and should be part of the SCFs whereas adjuncts should not. In sentence (1), the prepositional phrase “*sur le Sahel*” is an argument and should be included in the SCF whereas “*en 1972-1973*” is a time phrase and should not be included in the SCF.

(1) La sécheresse s' abattit sur le Sahel en 1972-1973 .

(The drought came down on Sahel in 1972-1973.)

However, there is evidence that no linguistic criterion is relevant enough to distinguish, whatever the context, between arguments and adjuncts. Depending on the theory and / or the application, a complement can be considered back and forth as argument or as adjunct. We should then consider a continuum between arguments and adjuncts, that can represent more accurately the nature of the link between a verb

and its complements. We need to “translate” this continuum in terms of SCFs. (Manning, 2003) proposes to describe subcategorization as a probability distribution over argument frames.

We first describe our SCF acquisition system in section 2; we show 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 a large corpus and produces a list of frames for each verb that occurred more than 200 times in the corpus. It is one of the first system that automatically induces large scale lists of SCFs from a large corpus for French. Previous experiments only concerned a limited set of verbs (Chesley and Salmon-Alt, 2006), or were based on treebanks or on a huge amount of manual work (Gross, 1975; Kupść, 2007).

The system is made of three modules:

1. verbs and surrounding phrases are extracted;
2. tentative SCFs are dynamically built, based on morphosyntactic information and relations between the verb and its arguments;
3. a statistical filter is used to throw out incorrect frames.

2.1 Preprocessing

The input corpus must be large enough, balanced and representative. The corpus is first tagged and lemmatized using *TreeTagger* (Schmid, 1994) and then syntactically annotated by *Syntax* (Bourigault et al., 2005). The *TreeTagger* is a language independent tool for the automatic annotation of part-of-speech and lemma information using probabilities. *Syntax* is a shallow parser specialized in the extraction of lexical dependencies (such as adjective/noun, or verb/noun associations). The parsing strategy is based on heuristics and statistics, since no lexical information is available at this stage. The dependencies extracted by the parser include arguments and adjuncts such as location or time phrases indistinctly. *Syntax* tags relations using heuristics and

statistics: the word on the left side of the verb is generally assumed to be the subject, the one on the right is assumed to be the object. A set of exceptions tags all other cases.

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

(3) is the preprocessed input of *ASSCI* for sentence (2) (after the *TreeTagger* annotation and *Syntax*'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 first module takes as input the analysis of the corpus by *Syntax* and extracts each verb which is sufficiently frequent (at least 200 occurrences) in the corpus and its dependencies in the analysis. In some cases, this module has to explore “deep” relations in the analysis. For examples, when a preposition is part of the dependencies, the pattern extractor is looking for 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 second module considers the dependencies according to their syntactic category (e.g., noun phrase) and to their relation to the verb (e.g., object), if any. The module tries to dynamically construct frames from these features (the complete

list of features is as follows: 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 dependency, the corresponding SCF is “intransitive” (INTRANS). There is no available list of SCFs for French. Therefore, contrary to most of previous work (e.g., (Preiss et al., 2007)), we do not have a predefined set of SCFs. The frames are learnt dynamically, depending on the information found in the corpus. This module counts the number of occurrences of each SCF and the total number of occurrences of each verb.

The SCF candidate built for sentence (2) is (5)¹.

(5) SN_SP [à+SN]

2.4 SCF Filter

The third stage aims at filtering the results. It is necessary to filter them since the output of the second module is noisy, mainly because of tagging and parsing errors. Several authors, in previous experiments (e.g., (Briscoe and Carroll, 1997; Chesley and Salmon-Alt, 2006)) have used binomial hypothesis testing at this stage. Anna Korhonen proposes to use the maximum likelihood estimate and shows that this method gives better results than binomial hypothesis testing (Korhonen et al., 2000). This method consists on a simple threshold over the relative frequencies of SCFs’ candidates. The maximum likelihood estimate is still used at Cambridge but with a specific threshold for each SCF.

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.

Then, these estimates are compared to a threshold to filter out the set of low probability frames for each verb. The effect of the choice of the threshold on the results is discussed in section 3.

¹SN stands for noun phrase and SP for 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, “clean” enough to be easily and efficiently parsed. The systems needs enough occurrences for each verb in order to acquire relevant information. Only verbs with more than 200 occurrences are analyzed by the system.

3.2 LexSchem: The Acquired Lexicon

3267 verbs can be found with more than 200 occurrences in the corpus. From these verbs, we induced 286 distinct SCFs. The extracted lexicon is 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 to consult the SCFs acquired for each verb and which verbs are taking a SCF is also available at the same address. For more details on the lexicon and its format, see (Messiant et al., 2008).

3.3 Gold Standard

We have shown in previous work that the comparison with a *gold standard* is problematic (Poibeau and Messiant, 2008). This method is still the easiest and fastest way to evaluate this kind of resource. We manually built a *gold standard* for 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 various frequency in the corpus (from 200 to 100.000 occurrences). To build this reference, we used the *Trésor de la Langue Française Informatisé (TFLI)*, a large French dictionary available on the web², including information about the argument structure (and therefore the SCFs).

3.4 Evaluation Measures

We calculated type precision, token recall and F-measure for these 25 verbs. We obtain the best results (0.822 precision, 0.587 recall and 0.685 f-measure) for a threshold around 0.032 (see figure

²<http://atilf.atilf.fr/>

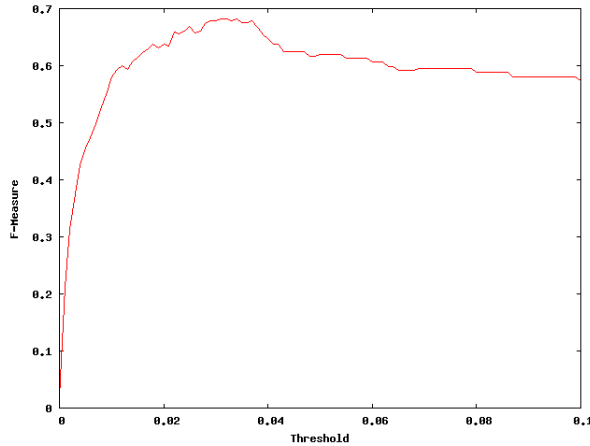


Figure 1: The effect of the threshold on the F-Measure

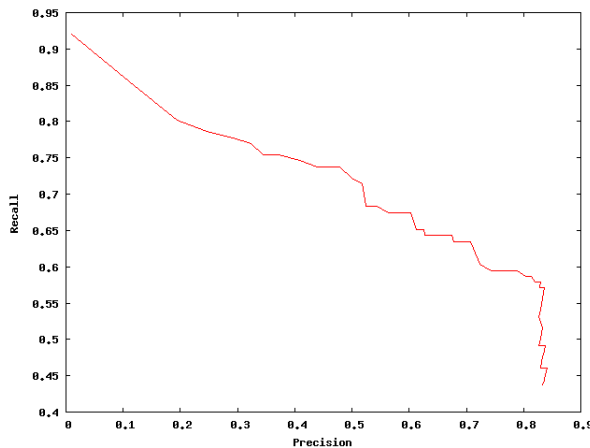


Figure 2: The relation between precision and recall

1). Figure 2 shows that, from this point, even an important loss in recall can not improve the precision which is always lower than 0.85.

Comparison of different versions of *ASSCI* is given in table 1. The different versions of the system are:

- Baseline: 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.

System	Precision	Recall	F-Measure
Baseline	0.010	0.921	0.020
<i>ASSCI-1</i>	0.789	0.595	0.679
<i>ASSCI-2</i>	0.822	0.587	0.685

Table 1: Comparison of different versions of *ASSCI*

The unfiltered output is very noisy (0.01 precision) but a simple threshold on the relative frequencies is really improving the results (*ASSCI-1*).

Every step of the acquisition can generate 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 it binds dependencies: in some cases, the analysis fails to identify relevant dependencies; in some other cases incorrect dependencies are generated. The SCF builder is another important source of errors because of ambiguity or lack of information to build some frames (e.g. pronouns). Finally, the filtering module rejects some correct SCFs and accept some incorrect ones despite the threshold. We may correct these errors by improving the filtering method or refining the thresholds.

A lot of errors are related to the intransitive SCF. We tried to address this problem with an INTRANS-specific higher threshold (*ASSCI-2*) which improves the precision of the system but there is still intransitive false negatives. The intransitive form of verbs is found very frequently in the corpus but it doesn’t appear in the *gold standard*. Most of the time, it is due to the domain and the corpus: undercurrent object (e.g. for “*acheter*” (*to buy*)) or imperative form. A better evaluation (e.g., manual annotation of the corpus) should not yield these errors anymore. In other cases (e.g. interpolated clauses), the parser can not find the dependencies. We will soon use a new version of *Syntex* that deals with this problem.

Our results (*ASSCI-2*) are roughly similar to those obtained by the only directly comparable work for French (Chesley and Salmon-Alt, 2006) (0.87 precision and 0.54 recall).

Even if our results are satisfactory compared with recent similar work, there is still a lot of errors, especially in recall. The next step would be to evaluate whether this resource is useful for NLP applications.

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

Subcategorization lexicons was first built manually. For example, Maurice Gross built a large French dictionary called “*Les Tables du LADL*” (Gross, 1975). This dictionary is not directly useable for NLP application but work currently 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). *Dicovalence* is a manually built valency dictionary based on the pronominal approach (van den Eynde and Blanche-Benveniste, 1978; van den Eynde and Mertens, 2006). There is also semi-automatic approaches e.g., acquisition of subcategorization information from treebanks, manually annotated corpus (O’Donovan et al., 2005; Kupść, 2007).

These approaches are not comparable to ours for several reasons. Firstly, manual and semi-automatic work is time-consuming, error-prone and not reproducible. Secondly, most of the time, the acquired resources are “binary” and do not contain probabilistic information about the SCFs’ distribution. Therefore, our interest is more specifically focused towards fully automatic methods.

4.2 Automatic Work

Different experiments have been made for the automatic acquisition of subcategorization frames since the 1990s (Brent, 1993; Briscoe and Carroll, 1997). These experiments were initially done for English but the approach has successfully been applied to various other languages since the beginning of the 2000s. For example, (Schulte im Walde, 2002) has induced a subcategorization lexicon for German verbs from a lexicalized PCFG. Our approach is related to the work done at Cambridge since it fully corresponds to our need. Their system has been regularly improved and evaluated; it currently achieves among the better results on the task (Briscoe and Carroll, 1997; Korhonen et al., 2000; Preiss et al., 2007). In this last 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). Differences between these works and ours are due to 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).

As far as we know, the only directly comparable work about subcategorization acquisition for French is (Chesley and Salmon-Alt, 2006) who propose a method to acquire 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 *Syntax* which is, according to the *EASY evaluation campaign*³, the best parser for French (at least on newspaper corpora). Additionally, we acquired a large subcategorization lexicon which is available on the web (286 distinct SCFs for 3267 verbs) whereas they only have 27 SCFs for 104 verbs.

5 Conclusion

We have presented a system developed to acquire a large subcategorization lexicon for French verbs. On a large French newspaper corpus, the system produces a lexicon of 286 SCFs corresponding to 3267 verbs. This lexicon has been evaluated by comparing SCFs acquired for an heterogeneous set of 25 verbs to a manually built resource. The resource acquired by our system is freely available on the web under the LGPL-LR license and through a web interface.

Future work will include improvements of the filtering module with SCF-specific thresholds (e.g. for prepositional phrases) or binomial hypothesis testing, exploration of new ways of evaluating through the integration of the results in practical NLP applications and acquisition of semantic information from the SCFs (e.g., semantic classes (Levin, 1993)).

³<http://www.limsi.fr/Recherche/CORVAL/easy/>.

The scores and ranks of *Syntax* at this evaluation campaign are available at <http://w3.univ-tlse2.fr/erss/textes/pagespersos/bourigault/syntax.html#easy>

The main asset of our system is its ability to produce proposals than can be validated by linguists.

Acknowledgements

This research was done as part of the ANR MDCO 'CroTal' project. It was supported by the British Council and the French Ministry of Foreign Affairs -funded 'Alliance' grant, by the EPSRC project 'ACLEX', and the Royal Society, UK.

Cédric Messiant's PhD is funded by a DGA/CNRS 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.
- Christopher D. Manning, 2003. *Probabilistic syntax*, pages 289–341. R. Bod, J. Hay, S. Jannedy.
- Cédric Messiant, Anna Korhonen, and Thierry Poibeau. 2008. LexSchem : A Large Subcategorization Lexicon for French Verbs. In *Language Resources and Evaluation Conference (LREC)*, Marrakech.
- Ruth O'Donovan, Michael Burke, Aoife Cahill, Josef van Genabith, and Andy Way. 2005. Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks. *Computational Linguistics*, 31(3):329–366.
- 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	posséder	comprendre
concevoir	proposer	montrer
rendre	s'abattre	jouer
offrir	continuer	ouvrir
aimer	croire	exister
obtenir	refuser	programmer
acheter	rester	s'ouvrir
venir		