LexSchem: A Large Subcategorization Lexicon for French Verbs

Cédric Messiant, Anna Korhonen, Thierry Poibeau

LexSchem: A Large Subcategorization Lexicon for French Verbs

Cédric Messiant *, Anna Korhonen **, Thierry Poibeau *

* 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
** Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue, Cambridge CB3 0FD, UK
Anna.Korhonen@cl.cam.ac.uk

Abstract

This paper presents *LexSchem* – the first large, fully automatically acquired subcategorization lexicon for French verbs. The lexicon includes subcategorization frame and frequency information for 3268 French verbs. When evaluated on a set of 20 test verbs against a gold standard dictionary, it shows 0.79 precision, 0.55 recall and 0.65 F-measure. We have made this resource freely available to the research community on the web.

1. Introduction

A lexicon is a key component of many current Natural Language Processing (NLP) systems. Hand-crafting lexical resources is difficult and extremely labour-intensive - particularly as NLP systems require statistical information about the behaviour of lexical items in data, and the statistical information changes from dataset to another. For this reason automatic acquisition of lexical resources from corpora has become increasingly popular.

One of the most useful lexical information for NLP is that related to the predicate-argument structure. Subcategorization frames (SCFs) of a predicate capture at the level of syntax the different combinations of arguments that each predicate can take. For example, in French, the verb "acheter" (to buy) subcategorizes for a single nominal phrase as well as for a nominal phrase followed by a prepositional phrase governed by the preposition "à".

Subcategorization lexicons can benefit many NLP applications. For example, they can be used to enhance tasks such as parsing (John Carroll and Guido Minnen and Ted Briscoe, 1998; Abhishek Arun and Frank Keller, 2005) and semantic classification (Sabine Schulte im Walde and Chris Brew, 2002) as well as applications such as information extraction (Surdeanu et al., 2003) and machine translation.

Several subcategorization lexicons are available for many languages, but most of them have been built manually. For French these include e.g. the large French dictionnary *"Le Lexique Grammaire"* (Maurice Gross, 1975) and the more recent *Lefff* (Benoît Sagot and Lionel Clément and Eric de La Clergerie and Pierre Boullier, 2006) and *Dicovalence* (`http://bach.arts.kuleuven.be/dicovalence/`) lexicons.

Some work has been conducted on automatic subcategorization acquisition, mostly on English (Michael R. Brent, 1993; Christopher D. Manning, 1993; Ted Briscoe and John Carroll, 1997; Anna Korhonen and Yuval Krymolowski and Ted Briscoe, 2006) but increasingly also on other languages, from which German is just one example (Sabine Schulte im Walde, 2002). This work has shown that although automatically built lexicons are not as accurate and detailed as manually built ones, they can be useful for real-world tasks. This is mostly because they provide what manually built resources don't generally provide: statistical information about the likelihood of SCFs for individual verbs.

We have recently developed a system for automatic subcategorization acquisition for French which is capable of acquiring large scale lexicons from un-annotated corpus data (Cédric Messiant, 2008). To our knowledge, only one previously published system exists for SCF acquisition for French SCFs (Paula Chesley and Susanne Salmon-Alt, 2006). However, no further work has been published since the initial experiment with this system, and the lexicon resulting from the initial experiment (which is limited to 104 verbs) is not publicly available.

Our new system is similar to the system developed in Cambridge (Ted Briscoe and John Carroll, 1997; Judita Preiss and Ted Briscoe and Anna Korhonen, 2007) in that it extracts SCFs from data parsed using a shallow dependency parser (Didier Bourigault and Marie-Paule Jacques and Cécile Fabre and Cécile Frérot and Sylwia Ozdowska, 2005) and is capable of identifying a large number of SCFs. However, unlike the Cambridge system (and most other systems which accept raw corpus data as input), it does not assume a list of predefined SCFs. Rather it learns the SCF types from data. This approach was adopted because at the time of development no comprehensive manually built inventory of French SCFs was available to us.

In this paper, we report work where we used this recent system to automatically acquire the first large subcategorization lexicon for French verbs. The resulting lexicon, *LexSchem*, is made freely available to the community under LGPL-LR (Lesser General Public License For Linguistic Resources) license.

We describe *ASSCI*, our SCF acquisition system, in section 2. *LexSchem* (the automatically acquired lexicon) is introduced and evaluated in section 3. We compare our work against previous work in section 4.

2. ASSCI : the subcategorization acquisition system

ASSCI takes raw corpus data as input. The data is first tagged and syntactically analysed. Then, our system produces a list of SCFs for each verb that occurred frequently enough in data (we have initially set the minimum limit to 200 corpus occurrences). *ASSCI* consists of three modules: a pattern extractor which extracts patterns for each target verb; a SCF builder which builds a list of candidate SCFs for the verb, and a SCF filter which filters out SCFs deemed incorrect. We introduce these modules briefly in the subsequent sections. For a more detailed description of *ASSCI*, see (Cédric Messiant, 2008).

2.1. Preprocessing : Morphosyntactic tagging and syntactic analysis

Our system first tags and lemmatizes corpus data using the *Tree-Tagger* and then parses it using *Syntex* (Didier Bourigault and Marie-Paule Jacques and Cécile Fabre and Cécile Frérot and Sylwia Ozdowska, 2005). *Syntex* is a shallow parser for French. It uses a combination of heuristics and statistics to find dependency relations between tokens in a sentence. It is a relatively accurate parser, e.g. it obtained the best precision and F-measure for written French text in the recent EASY evaluation campaign[1].
Our below example illustrates the dependency relations detected by *Syntex* (2) for the input sentence in (1):

(1) La sécheresse s' abattit sur le Sahel en 1972-1973 .
(The drought came down on Sahel in 1972-1973.)

(2) DetFS|le|La|1|DET;2|
NomFS|sécheresse|sécheresse|2|SUJ;4|DET;1
Pro|se|s'|3|REF;4|
VCONJS|abattre|abattit|4|SUJ;2,REF;3,PREP;5,PREP;8
Prep|sur|sur|5|PREP;4|NOMPREP;7
DetMS|le|le|6|DET;7|
NomMS|sahel|Sahel|7|NOMPREP;5|DET;6
Prep|en|en|8|PREP;4|NOMPREP;9
NomXXDate|1972-1973|1972-1973|9|NOMPREP;8|
Typo|.|.|10||

Syntex does not make a distinction between arguments and adjuncts - rather, each dependency of a verb is attached to the verb.

2.2. Pattern extractor

The pattern extractor collects the dependencies found by the parser for each occurrence of a target verb. Some cases receive special treatment in this module. For example, if the reflexive pronoun *"se"* is one of the dependencies of a verb, the system considers this verb like a new one. In (1), the pattern will correspond to *"s'abattre"* and not to *"abattre"*. If a preposition is the head of one of the dependencies, the module explores the syntactic analysis to find if it is followed by a noun phrase (+SN]) or an infinitive verb (+SINF]).
(3) shows the output of the pattern extractor for the input in (1).

(3) VCONJS|s'abattre :
Prep+SN|sur|PREP_Prep+SN|en|PREP

2.3. SCF builder

The SCF builder extracts SCF candidates for each verb from the output of the pattern extractor and calculates the number of corpus occurrences for each SCF and verb combination. The syntactic constituents used for building the SCFs are the following:

1. SN for nominal phrases;
2. SINF for infinitive clauses;
3. SP[*prep*+SN] for prepositional phrases where the preposition is followed by a noun phrase. *prep* is the head preposition;
4. SP[*prep*+SINF] for prepositional phrases where the preposition is followed by an infinitive verb. *prep* is the head preposition;
5. SA for adjectival phrases;
6. COMPL for subordinate clauses.

When a verb has no dependency, its SCF is considered as INTRANS.
(4) shows the output of the SCF builder for (1).

(4) S'ABATTRE+s'abattre ;;;
SP[sur+SN]_SP[en+SN]

2.4. SCF filter

Each step of the process is fully automatic, so the output of the SCF builder is noisy due to tagging, parsing or other processing errors. It is also noisy because of the difficulty of the argument-adjunct distinction. The latter is difficult even for humans. Many criteria that exist for it are not usable for us because they either depend on lexical information which the parser cannot make use of (since our task is to acquire this information) or on semantic information which even the best parsers cannot yet learn reliably. Our approach is based on the assumption that true arguments tend to occur in argument positions more frequently than adjuncts. Thus many frequent SCFs in the system output are correct.
We therefore filter low frequency entries from the SCF builder output. We currently do this using the maximum likehood estimates (Anna Korhonen and Genevieve Gorrell and Diana McCarthy, 2000). This simple method involves calculating the relative frequency of each SCF (for a verb) and comparing it to an empirically determined threshold. The relative frequency of the SCF *i* with the verb *j* is calculated as follows:

[1]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

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

$|scf_i, verb_j|$ 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.
If, for example, the frequency of the SCF `SP[sur+SN]_SP[en+SN]` is less than the empirically defined threshold, the SCF is rejected by the filter. The MLE filter is not perfect because it is based on rejecting low frequency SCFs. Although relatively more low than high frequency SCFs are incorrect, sometimes rejected frames are correct. Our filter incorporates special heuristics for cases where this assumption tends to generate too many errors. With prepositional SCFs involving one PP or more, the filter determines which one is the less frequent PP. It then re-assigns the associated frequency to the same SCF without this PP.
For example, `SP[sur+SN]_SP[en+SN]` could be split to 2 SCFs : `SP[sur+SN]` and `SP[en+SN]`. In our example, `SP[en+SN]` is the less frequent prepositional phrase and the final SCF for the sentence `(1)` is `(5)`.

`(5) SP[sur+SN]`

Note that `SP[en+SN]` is here an adjunct.

3. LexSchem

We used *ASSCI* to acquire *LexSchem*, the first fully automatically built large subcategorization lexicon for French verbs. We describe this work and the outcome in the subsequent sections.

3.1. Corpus

The automatic approach benefits from a large corpus. In addition, as we want our lexicon to be suitable for general use (not only for a particular domain use), the corpus needs to be heterogeneous enough to cover many domains and text types. We thus used ten years of the French newspaper *Le Monde* (two hundred millions words in total). *Le Monde* is one of the largest corpora for French and "clean" enough to be parsed easily and efficiently.

3.2. Description of the lexicon

Running *ASSCI* on this corpus data, we extracted 11,149 lexical entries in total for different verb and SCF combinations. The lexicon covers 3268 verb types (a verb and its reflexive form are counted as 2 different verbs) and 336 distinct SCFs.
Each entry has 7 fields :

- **NUM**: the number of the entry in the lexicon;
- **SUBCAT**: a summary of the target verb and SCF;
- **VERB**: the verb;
- **SCF**: the subcategorization frame;
- **COUNT**: the number of corpus occurences found for the verb and SCF combination;
- **RELFREQ**: the relative frequency of the SCF with the verb;
- **EXAMPLES**: 5 corpus occurrences exemplifying this entry (the examples are provided in a separate file).

The following shows the *LexSchem* entry for the verb *"s'abattre"* with the SCF `SP[sur+SN]`.

```
:NUM:       05204
:SUBCAT:    s'abattre : SP[sur+SN]
:VERB:      S'ABATTRE+s'abattre
:SCF:       SP[sur+SN]
:COUNT:     420
:RELFREQ:   0.882
:EXAMPLE:   25458;25459;25460;25461;25462
```

Two of the five corpus sentences exemplifying this entry are shown as follows (the syntactic analysis of *Syntex* is also available):

```
25458===Il montre la salle :  On a
fait croire aux gens que des hordes s'
abattraient sur Paris .

25459===Dans ces conditions , sa réponse au
problème politique corse est avant tout
policière :  avant 1981 , comme entre
1986 et 1988 , la répression s' abat sur
les terroristes , souvent assimilés à des
délinquants de droit commun , et le pouvoir
rejette toute idée de dialogue avec les "
séparatistes " .
```

3.3. Evaluation

We evaluated *LexSchem* against a gold standard from a dictionary. Although this approach is not ideal (e.g. a dictionary may include SCFs not included in our data, and vice versa – see e.g. (Thierry Poibeau and Cédric Messiant, 2008) for discussion), it can provide a useful starting point. We chose a set of 20 verbs listed in Appendix to evaluate this resource. These verbs were chosen for their heterogeneity in terms of semantic and syntactic features, but also because of their varied frequency (200 to 100,000) in the corpus. We compared our lexicon against the *Trésor de la Langue Française Informatisé (TLFI)* - a freely available French lexicon containing verbal SCF information from a dictionary. We had to restrict our scope to 20 verbs because of problems in turning this resource into a gold standard[2].
We calculated type precision, type recall and F-measure against the gold standard, and obtained 0.79 precision, 0.55 recall and 0.65 F-measure. These results are shown in table 1, along with: 1) the results obtained with the only previously published work on automatic subcategorization acquisition (from raw corpus data) for French verbs (Paula Chesley and Susanne Salmon-Alt, 2006), and 2) those reported with the previous Cambridge system when the system was used to acquire a large SCF lexicon for English with a baseline filtering technique comparable to the one employed in our work (VALEX sub-lexicon 2) (Anna Korhonen and Yuval Krymolowski and Ted Briscoe, 2006). Due to the differences in the data, SCFs, and experimental setup, direct comparison of these results is unmeaning-

[2]See (Thierry Poibeau and Cédric Messiant, 2008) for details.

	Our work	Chesley & Salmon-Alt (2006)	Korhonen *& al.* (2006)
# test verbs	20	104	183
Precision	0.79	0.87	0.81
Recall	0.55	0.54	0.46
F-Measure	0.65	0.67	0.58

Table 1: Comparison with recent work in French and English

Verb	# SCFs	Precision	Recall
aimer	5	0.80	0.80
apprendre	5	0.60	0.50
chercher	2	1.00	0.67
comprendre	3	0.33	0.33
compter	5	0.80	0.50
concevoir	5	0.60	0.75
continuer	4	1.00	0.80
croire	6	0.83	0.50
donner	3	1.00	0.30
exister	4	0.50	0.50
jouer	7	0.86	1.00
montrer	3	0.67	0.40
obtenir	2	1.00	0.50
offrir	4	0.75	0.75
ouvrir	2	1.00	0.22
posséder	2	0.50	1.00
proposer	5	0.80	0.44
refuser	2	1.00	0.40
rendre	4	1.00	1.00
s'abattre	2	1.00	1.00

Table 2: The number of SCFs detected and the performance figures per each test verb

ful. However, their relative similarity seems to suggest that *LexSchem* is a state-of-the-art lexicon.

The type precision and recall scores for each test verb are given in table 2.

3.4. The web distribution of LexSchem

LexSchem is freely available to the research community under the LGPL-LR (Lesser General Public License For Linguistic Resources) license [3]: `http://www-lipn.univ-paris13.fr/~messiant/lexschem.html`. A web interface is provided at the same address which enables viewing lexical entries for each verb along with practical examples.

4. Related work

This section describes other existing syntax dictionaries and lexicons for French (most of the ones we are aware of). For comparison, it also includes a description of *VALEX* – the first large subcategorization lexicon acquired automatically for English. Table 3 summarizes the key information included in these different lexical resources.

4.1. Dictionaries and lexicons for French

The **Lexicon-Grammar (LG)** is the earliest resource for subcategorization information for French. (Maurice Gross, 1975; Maurice Gross, 1994) – a manually built dictionary including subcategorization information for verbs, adjectives and nouns. It is not ideally suited for computational use but work currently in progress is aimed at addressing this problem (Claire Gardent and Bruno Guillaume and Guy Perrier and Ingrid Falk, 2005). Only part of this resource is publicly available.

As mentioned earlier, the **Trésor de la Langue Française Informatisé (TLFI)** is derived from a syntax dictionary and (like we noticed with evaluation of 3.), requires substantial manual work for NLP use.

The **Lefff** is an automatically acquired morphological lexicon for 6798 verb lemmas (Benoît Sagot and Lionel Clément and Eric de La Clergerie and Pierre Boullier, 2006) which has been manually supplemented with partial syntactic information.

DicoValence is a manually built resource which contains valency frames for more than 3700 French verbs (van den Eynde and Mertens, 2006). It relies on the pronominal paradigm approach of (Karel van den Eynde and Claire Blanche-Benveniste, 1978).

Note that the information provided by *LG, the TLFI, the Lefff and DicoValence* is type-based, i.e. no statistical information about the likelihood of SCF for words is available.

TreeLex (`http://erssab.u-bordeaux3.fr/article.php3?id\_article=150`) is a subcategorization lexicon automatically extracted from the French TreeBank (Anna Kupść, 2007). It covers about 2000 verbs. 160 SCFs have been identified (1.91 SCF per verb on average). To our knowledge, this lexicon has yet not been evaluated in terms of accuracy.

Like other resources mentioned in this section, *TreeLex* relies on manual effort. Resources built in this matter are not easily adapted to different tasks and domains.

As far as we know, the only published work on subcategorization acquisition for French is (Paula Chesley and Susanne Salmon-Alt, 2006) which proposes a method to acquire SCFs from a French cross-domain corpus. The work relies on the VISL parser which has an "unevaluated (and potentially high) error rate" while our system relies on *Syntex* which has been evaluated and discovered accurate by *EASY evaluation campaign*. We acquired and made publicly available a large subcategorization lexicon for 3268 verbs (336 SCFs) whereas Paula Chesley and Susanne Salmon-Alt (2006) only reported an experiment with 104 verbs (27 SCFs).

4.2. The first automatically acquired large scale lexicon for English : VALEX

An interesting comparison point for us is **VALEX** – a large verb subcategorization lexicon created for English (Anna Korhonen and Yuval Krymolowski and Ted Briscoe, 2006).

[3] `http://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/lgpllr.html`

This lexicon was acquired automatically using the system developed at Cambridge (Ted Briscoe and John Carroll, 1997) which identifies 163 SCF types (these abstract over lexically-governed particles and prepositions). The input data used for building *VALEX* consisted of 904 million words in total. It was extracted from five large corpora and the web. The resulting lexicon provides SCF (frequency) information for 6,397 English verbs. It includes 212,741 SCF entries, 33 per verb on average.

Because *VALEX* builds on over a decade of subcategorization acquisition research for English, the release is fairly comprehensive and offers also some ideas for further development of *LexSchem*. First, five different versions of the lexicon are provided in the web release at `http://www.cl.cam.ac.uk/~alk23/subcat/lexicon.html`. The idea is to provide different lexicons for the needs of different NLP tasks which vary in terms how accurate lexicons they require. For example, if the aim is to use SCF frequencies to aid parsing, it may be better to maximise the accuracy (rather than the coverage) of the lexicon. On the other hand, an NLP task such as lexical classification tends to benefit from a lexicon which provides good coverage at the expense of accuracy.

The accuracy is controlled by using different SCF filtering options to build the different lexicons:

Lexicon 1: Unfiltered, noisy SCF lexicon.

Lexicon 2: High frequency SCFs selected only.

Lexicon 3: High frequency SCFs supplemented with additional ones from manually built dictionaries.

Lexicon 4: High frequency SCFs after smoothing with semantic back-off estimates.

Lexicon 5: High frequency SCFs after smoothing with semantic back-off estimates and supplemented with additional SCFs from manually built dictionaries.

LexSchem was released with a comparable filtering method and similar accuracy than Lexicon 2 of *VALEX* (see the comparison of results in the previous section). Future work could release other, more or less accurate versions of the lexicon after the filtering component of the system undergoes first further development.

Another idea for future work concerns lexical entries. As seen above in Section 3, the lexical entries of *LexSchem* provide various information. They could be further improved by gathering in them argument head and associated frequency data in different syntactic slots. In the case of *VALEX*, such information has proved useful for a number of NLP tasks.

5. Conclusion

This paper introduced *LexSchem* – the first fully automatically acquired large scale SCF lexicon for French verbs. It includes 11,149 lexical entries for 3268 French verbs. The lexicon is provided with a graphical interface and is made freely available to the community via a web page. Our evaluation with 20 verbs showed that the lexicon has state-of-the-art accuracy when compared with recent work using similar technology: 0.79 precision, 0.55 recall and 0.65 F-measure.

Future work will include improvement of the filtering module (e.g. experimenting with SCF-specific thresholds or smoothing using semantic back-off estimates), automatic acquisition of SCFs for other French word classes (e.g. nouns), and automatic classification of verbs using the SCFs as features (Beth Levin, 1993; Sabine Schulte im Walde and Chris Brew, 2002). Like mentioned above, we also plan to enhance the lexical entries of the lexicon. It would be useful to include in them information about noun and preposition classes and morpho-syntactic properties of the words included in SCFs. Finally, as mentioned earlier, given different NLP applications have different requirements, it is worth building and releasing other versions of *LexSchem*.

Lexicon	Acquisition	#verbs	#SCFs	#entries
LS	LM10 (200M)	3268	336	11149
C&S06	created	104	27	176
VALEX	5 corpora (904M)	6397	213m	?
TreeLex	FrTB	2000	160	?
Lefff	mixed	6798	?	?
DV	manual	3700	?	8000
LG	manual	5208	?	13335

Table 3: Comparison of dictionaries and lexicons
'?' stands for unknown; LS: LexSchem; C&S06: Chesley & Salmon-Alt (2006); DV: DicoValence; LG: Lexicon-Grammar; LM10: Le Monde 10 years; FrTB: French TreeBank

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.

6. References

Abhishek Arun and Frank Keller. 2005. Lexicalization in crosslinguistic probabilistic parsing: The case of French. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 306–313, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Anna Korhonen and 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 Korhonen and Yuval Krymolowski and Ted Briscoe. 2006. A Large Subcategorization Lexicon for Natural Language Processing Applications. In *Proceedings of the 5th international conference on Language Resources and Evaluation*, Genova, Italy.

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.

Benoît Sagot and Lionel Clément and 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).

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

Cédric Messiant. 2008. ASSCI : A Subcategorization Frames Acquisition System For French. In *Proceedings of the Association for Computational Linguistics (ACL) Student Research Workshop*, Colombus, Ohio. Association for Computational Linguistics.

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.

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

Didier Bourigault and Marie-Paule Jacques and Cécile Fabre and Cécile Frérot and Sylwia Ozdowska. 2005. Syntex, analyseur syntaxique de corpus. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, Dourdan.

John Carroll and Guido Minnen 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).

Judita Preiss and 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.

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

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

Maurice Gross. 1994. Constructing Lexicon-Grammars. In *Computational Approaches to the Lexicon*, pages 213–263, Oxford. Oxford University Press.

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

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).

Sabine Schulte im Walde and Chris Brew. 2002. Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 223–230, Philadelphia, PA.

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.

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.

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.

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

Appendix — List of test verbs

aimer	apprendre	chercher
comprendre	compter	concevoir
continuer	croire	donner
exister	jouer	montrer
obtenir	offrir	ouvrir
posséder	proposer	refuser
rendre	s'abattre	