TreeLex: A Subcategorisation Lexicon for French Verbs

Anna Kupść
Université de Bordeaux, ERSSàB/SIGNES
and Polish Academy of Sciences
UFRL, Domaine Universitaire
F33607 Pessac Cedex
akupsc@u-bordeaux3.fr

Anne Abeillé
Université Paris 7, LLF
UMR 7110, Case 7031
2, place Jussieu
75251 Paris cedex 05
abeille@linguist.jussieu.fr

Abstract

TreeLex is a subcategorization lexicon of French verbs, automatically extracted from a syntactically annotated corpus. The lexicon comprises 1362 verbs (12353 occurrences). We present not only a list of verbs with their subcategorization frames but we also estimate the number of different verb frames available in French in general. Additionally, we estimate the average number of frames per verb. After applying various factorization techniques, we obtain 58 frames for a function-based representation (on average, 1.72 frames per verb), and 160 frames for a richer representation based on function-category information (on average, 1.91 frames per verb).

1 Introduction

The paper presents TreeLex, a subcategorisation lexicon for French extracted from a syntactically annotated corpus.

Information about a combinatory potential of a predicate, i.e., the number and the type of its arguments, is called a subcategorisation frame or valence. For example, the verb *embrasser* 'kiss' requires two arguments (the subject and an object), both of them realized as a noun phrase. This kind of syntactic properties is individually associated with every predicate, both within a single language and cross-linguistically. For example, the English verb *miss* has two NP arguments but the second argument of its French equivalent *manquer* is a PP (and semantic roles of the two arguments are reversed).[1] This implies that subcategorisation lexicons which store this kind of syntactic information have to be developed for each language individually. In addition to their importance in language learning, they play a crucial role in many NLP applications related both to parsing, e.g., (Briscoe and Carroll, 1993), (Carroll and Fang, 2004), (Surdeanu et al., 2003), and generation, e.g., (Danlos, 1985), (Han et al., 2000).

The (un)availability of such lexicons is still a bottleneck for text processing. Traditionally, they have been developed manually by human experts, e.g., (Procter, 1978; Hornby, 1989) (for English) or (Gross, 1975; Guillet and Leclère, 1992; Mel'cuk et al., 1984 1988 1992 1999; van den Eynde and Mertens, 2003) (for French), which guarantees their high quality, but they cannot be directly used in NLP applications. With the development of corpora and adaptation of statistical techniques for NLP, more efficient methods became available, which allowed for an automatic construction of syntactic lexicons for many languages (English, Spanish, German, Chinese), cf. (Cahill et al., 2002; Frank et al., 2002). Recent years have witnessed also an increased interest in obtaining such resources for French, either by applying statistical techniques, e.g., (Bourigault and Frérot, 2005), (Chesley and Salmon-Alt, 2005), adapting the existing lexicons, e.g., (Gardent et al., 2006; Falk et al., 2007), or using heuristics to extract

[1] Theoretical work in mapping theory has revealed partial correlations between lexical semantics and subcategorization frames (see for example (Davis and Koenig, 2000) for linking relations of nominal arguments).

valence information (Sagot et al., 2006; Sagot and Danlos, 2007; Danlos and Sagot, 2007) for French verbs; a syntactic lexicon of French prepositions has been created by (Fort and Guillaume, 2007). Finally, we mention two European research and development initiatives concerning French (among others): EAGLES (GENLEX, (Menon and Modiano, 1993)) and LE-PAROLE (Ruimy et al., 1998). The projects aimed at providing a general multilingual architecture for and creating multilingual resources, including syntactic verb lexicons.

In this paper we present yet another effort for automatic extraction of a syntactic lexicon for French verbs. The approach we have adopted differs form those mentioned above as it relies on syntactic and functional corpus annotations. We use the treebank of Paris7, (Abeillé et al., 2003), a newspaper corpus based on articles from *Le monde* (1989–1993), a French daily newspaper. The corpus contains morphological, syntactic and functional annotations for major constituents. The annotations have been manually validated, which makes the corpus a valuable resource for linguistic research but also for NLP applications.

The goal of the project is to obtain a list of different subcategorisation frames of French verbs as well as to enrich corpus annotations with this information for each verbal occurrence. We aim also at estimating the ambiguity rate of verb frames and propose different methods to reduce it.

2 Frame Extraction

2.1 Representation

Theoretical approaches use different forms of subcategorization frames. Some theoretical models, like LFG (Bresnan, 1982), prefer the notation based on functional information (1a), while others, like LADL (lexicon–grammar of (Gross, 1975)) adopt category-based notation (1b), yet others, like HPSG (Pollard and Sag, 1994), use a mixed approach (1c):

(1) *laver*:

a. <SUJ, OBJ>

b. N0 V N1

c. <SUJ:NP, OBJ:NP>

SUJ	NP, VPinf, Ssub, VN
OBJ	NP, AP, VPinf, VN, Sint, Ssub
DE-OBJ	VPinf, PP, Ssub, VN
A-OBJ	VPinf, PP, VN
P-OBJ	PP, AdP, VN, NP
ATO	Srel, PP, AP, NP, VPpart, VPinf, Ssub
ATS	NP, PP, AP, AdP, VPinf, Ssub, VPpart, Sint, VN

Figure 1: Possible categories for every function. Functions: SUJ (subject), OBJ (direct object), DE-OBJ (indirect object introduced by *de*), A-OBJ (indirect object introduced by *à*), P-OBJ (a complement with a different preposition), ATO (object's attribute), ATS (subject's attribute)

The first two approaches are not fully informative as both functions and categories can have multiple realizations. For example, a subject can be either nominal or phrasal whereas a postverbal NP can be considered either a direct object or an attribute. Since the corpus we are using contains both kinds of information, we adopt a mixed representation in order to obtain more complete information. The list of categories and functions used in the corpus is presented in Tab. 1. We are ignoring MOD, which always corresponds to non-subcategorized elements, and COORD as repeated coordination is relatively rare in the corpus. For prepositional complements, P-OBJ, we retain the type of the preposition which introduces the complement. This allows us to normalize frames with respect to active and passive forms. The adopted annotation schema does not distinguish a VP. Instead, a verbal nucleus (VN) is defined and it contains the main verb, auxiliaries, negation and pronominal clitics. The head verb is not explicitly indicated but we assume that the last verb in VN is the head.

2.2 Experiment

The automatic extraction of subcategorization frames is more difficult for French than for English. The higher complexity is due to pronominal clitics, which can express grammatical functions on a par with phrases, and a more flexible word order (e.g., a postverbal NP can be an inverted subject rather than an object). In the corpus, syntactic functions are treated as attributes of constituents rather than as relations between the head and its dependents. An example of the annotation schema is given in Fig. 2.

For extraction of the verb valency, we used the

```
<SENT>
<NP fct="SUJ">L’état-major
<AP>français</AP> </NP>
<VN>sait</VN>
<Ssub fct="OBJ">qu’
<VN fct="SUJ">il a gagné</VN>
<NP fct="OBJ">une bataille</NP>,
<COORD>mais
<AdP>pas encore</AdP>
<NP>la guerre</NP>
</COORD>
</Ssub>.
</SENT>
```

Figure 2: Example of annotation schema

part of the corpus which contains functional annotations, i.e., 15 000 phrases (about 300 000 words).[2] In the experiment presented here, only verbs in the main clauses, i.e., verbs with all functions specified, have been used, which resulted in 1362 verb lemmas (12 353 occurrences).

As a starting point, we used the frames extracted directly from the corpus, without any modification and then, we experimented with several methods to compact the frames. First, we separated the function tags indicating clitic arguments. If there are several clitics attached to a verb, e.g., in *Il l’a vue* ‘He has seen her’, the subject *Il* ‘He’ and the direct object *l’* ‘her/it’, the two functions are indicated with a single tag `SUJ/OBJ` and they have to be separated. Clitics are not always associated with grammatical functions, e.g., *y* in the idiomatic expression *il y a* ‘there is/are’ or the reflexive clitic *se* in inherently reflexive verbs such as *s’evanouir* ‘to faint’. We keep such clitics as frame elements. Additionally, a clitic and a constituent can have the same function. For example, in *Paul en mange-t-il beaucoup?* ‘Has Paul eaten lots of them?’ there are two subjects (*Paul* and *il*) and two objects (*en* and *beaucoup*). Hence, the repeated functions have to be eliminated. Finally, there are frames which are missing the subject. In the current experiment, it has been added to the imperative forms in the main clauses. There are two verbs which always appear without a subject, *voici* and *voilà* ‘(t)here is’. They are considered indicative verb forms which do not have a subject.

We normalized frames with respect to passive vs. active form. We used a list of 62 verbs which are inflected with the auxiliary *être* ‘be’ in order to distinguish past tense (fr. *passé composé*) and passive forms. If a verb appears with the auxiliary ‘be’ but its past tense form requires another auxiliary (*avoir*), the form is considered passive and it is transformed to an active form. We add OBJ whereas if the PP expressing the agent is present, i.e., P-OBJ introduced by the preposition *par* or *de*, this element is deleted. If the passive verb has ATS complement (subject’s attribute), we rename this function to ATO (object’s attribute).

Because of the relatively free order between French complements and subject inversion, we normalize the surface order of functions. Hence, for the two sentences in (2) and (3), we extract the same subcategorisation frame (SUJ, A-OBJ, DE-OBJ):

(2) Marie parle de ce problème à Paul.
Mary talks of this problem to Paul

Mary is talking to Paul about this problem

(3) Marie parle à Paul de ce problème.
Mary talks to Paul of this problem

We traced a few problems related to corpus annotations. For example, according to the annotation schema, only adverbial phrases but not adverbs alone have a grammatical function assigned. Therefore, the adverb *bien* ‘well’ is not recognized as a complement in *Elle va bien* ‘She is doing well’. Then, only locally realized arguments of a verb are annotated so we do not capture long-distance dependents, e.g., in *Que peut faire le gouvernement?* ‘What can the government do?’, we extract two objects (*que* ‘what’ and *faire* ‘make’) for the verb *peut* ‘can’ and none for the verb *faire*. Such cases are nevertheless quite rare.

2.3 Results

In the experiment described here we used only explicit information in the corpus, i.e., 1362 verbs (12353 tokens) present in the main clause.[3] In the

[2] The corpus has been recently enlarged and it contains currently about 20 000 phrases (500 000 words).

[3] All in all, the functionally annotated part of the corpus contains 2187 different verbs (30916 tokens). The verbs we did not use in the current experiment comprise infinitives or verbs in relative clauses. Their frames have to be (automatically) completed, e.g., by adding the missing subject to an infinitive, before they are extracted.

two following subsections, we compare results obtained for the two representations mentioned in (1): the functional (1a) and the mixed approach (1c).

2.3.1 Functional Representation

As indicated in Fig. 3, after neutralization of passive and active forms, we obtain 142 different subcategorisation frames, with an average of 1.9 frames per verbal lemma. Unsurprisingly the verb with the highest number of frames is *être* 'be' with 26 frames, whereas more than half of the verbs (849 lemmas) have exactly one subcategorisation frame.

Then we perform several operations in order to eliminate superfluous clitic arguments. We clean the frames so that double functions are removed. After these modifications, we reduced the number of frames almost three times and we obtained 58 frames, with an average of 1.8 frames per verb lemma. If we additionally compact frames where a complement is realized either as an NP or a reflexive clitic, the ambiguity rate drops to 1.72 per verb, although the number of frames remains the same. The verb *être* still has the most frames (16) but the number of verbs with a single frame increases to 886.

Only 6 verbs have 10 frames or more and they are the most ambiguous French verbs: *être* 'be', *avoir* 'have', *faire* 'make', *rendre* 'return', *passer* 'pass', *laisser* 'allow'. Their frames with frequency counts are shown in Fig. 4.

As indicated in Fig. 5, the most frequent frames are SUJ-OBJ (more than half of the lemma, i.e., the verb types), SUJ (about a quarter of the lemmas), then SUJ-A-OBJ and SUJ-DE-OBJ and ditransitive verbs. Very few lemmas have a predicative complement but they are frequently used.

The drawback of this approach is that we have lost categorial information available in the corpus. For example the distinction between verbs with a sentential complement and verbs with a nominal complement are indistinguishable. Therefore, we turn to a mixed approach in order to obtain more complete information.

2.3.2 Mixed Representation

A mixed representation (with categories and functions), after depassivization, gives a gross total of 783 different subcategorisation frames, with an average of 2.47 frames per lemma, and almost 58% of

	# frames	avge	max. frames	1 frame	
				%	#
passive	142	1.9	26 (*être*)	62.3%	849
clitics	58	1.8	16 (*être*)	63.1%	859
reflexive	58	1.72	16 (*être*)	65.1%	886

Figure 3: Functional representation

être (16 frames | 3842 tokens): SUJ, ATS (1632); SUJ (112); SUJ, OBJ, ATS (66); SUJ, OBJ (46); SUJ, P-OBJ (27); SUJ, DE-OBJ (21); SUJ, DE-OBJ, ATS (14); SUJ, P-OBJ, ATS (9); SUJ, A-OBJ (6); SUJ, A-OBJ, ATS (5); SUJ, OBJ, DE-OBJ (2); SUJ, OBJ, A-OBJ (2); SUJ, OBJ, A-OBJ, ATS (1); SUJ, A-OBJ, obj:en (1); SUJ, OBJ, P-OBJ (1); SUJ, P-OBJ, obj:en (1)

avoir (16 frames | 607 tokens): SUJ, OBJ (211); SUJ, OBJ, P-OBJ (65); SUJ, OBJ, ATO (11); SUJ (7); SUJ, A-OBJ (5); SUJ, OBJ, DE-OBJ (5); SUJ, OBJ, obj:y (4); SUJ, OBJ, A-OBJ (4); SUJ, obj:y (3); SUJ, P-OBJ (2); SUJ, A-OBJ, obj:y (1); SUJ, OBJ, P-OBJ, obj:y (1); SUJ, A-OBJ, DE-OBJ (1); SUJ, obj:y_en (1); SUJ, DE-OBJ (1); SUJ, DE-OBJ, P-OBJ (1)

faire (12 frames | 205 tokens): SUJ, OBJ (103); SUJ (19); SUJ, OBJ, A-OBJ (11); SUJ, OBJ, DE-OBJ (9); SUJ, ATS, refl (3); SUJ, obj:en (3); SUJ, P-OBJ, refl (2); SUJ, OBJ, P-OBJ (2); SUJ, OBJ, refl (2); SUJ, OBJ, obj:y (1); SUJ, DE-OBJ, ATO (1); SUJ, A-OBJ, refl (1)

rendre (12 frames | 34 tokens): SUJ, OBJ, ATO (15); SUJ, ATS (4); SUJ, A-OBJ, refl (3); SUJ, P-OBJ, ATS (2); SUJ, OBJ, A-OBJ (2); SUJ, OBJ (2); SUJ, P-OBJ, refl (1); SUJ, OBJ, DE-OBJ, refl (1); SUJ, OBJ, DE-OBJ, ATO (1); SUJ, OBJ, refl (1); SUJ, OBJ, A-OBJ, DE-OBJ (1); SUJ, obj:me (1)

passer (11 frames | 89 tokens): SUJ, P-OBJ (17); SUJ, DE-OBJ (16); SUJ (9); SUJ, OBJ (9); SUJ, A-OBJ (8); SUJ, A-OBJ, DE-OBJ (6); SUJ, OBJ, P-OBJ (2); SUJ, OBJ, refl (2); SUJ, OBJ, A-OBJ (2); SUJ, DE-OBJ, refl (1); SUJ, ATS (1)

laisser (10 frames | 43 tokens): SUJ, OBJ (23); SUJ, OBJ, A-OBJ (3); SUJ, OBJ, ATO (2); SUJ, A-OBJ (1); SUJ, OBJ, P-OBJ (1); SUJ (1); SUJ, OBJ, DE-OBJ (1); SUJ, OBJ, refl (1); SUJ, ATO (1); SUJ, OBJ, P-OBJ, refl (1)

Figure 4: Subcategorisation frames (functional representation) for 6 most ambiguous verbs (10 frames or more)

frame	# verb types	tokens
SUJ, OBJ	913 (67.0%)	6407 (51.9%)
SUJ, ATS	16 (1.2%)	1951 (15.8%)
SUJ	351 (25.8%)	1035 (8.4%)
SUJ, DE-OBJ	129 (9.5%)	558 (4.5%)
SUJ, OBJ, A-OBJ	162 (11.9%)	517 (4.2%)
SUJ, A-OBJ	103 (7.5%)	359 (2.9%)
SUJ, P-OBJ	85 (6.2%)	233 (1.9%)
SUJ, OBJ, P-OBJ	81 (5.9%)	197 (1.6%)
SUJ, OBJ, DE-OBJ	75 (5.5%)	160 (1.3%)
SUJ, A-OBJ, refl	55 (4.0%)	132 (1.1%)

Figure 5: 10 most frequent frames (functional representation)

the lemmas which have only one frame. With the clitic factorization described in section 2.2, we obtain 300 different frames, with an average of 2.32 frames per lemma. The number of unambiguous verbs (with only one frame) does not raise much: 803 lemmas, that is almost 59% of the verbs.

We further factorize the subcategorization frames by the neutralization of lexical value of a prepositional complement (indirect complements introduced by prepositions other than *à* or *de*). The average number of subcategorization frames drops (2.27 frames per lemma) and so does the total number of frames (222). The number of unambiguous verbs (with only one subcategorization frame) remains the same (803). We then neutralise different realizations of the attribute and types of a subordinate clause (indicative vs. subjunctive). The number of different frames drops to 173, whereas the ambiguity rate achieves 2.21. Next, we regroup frames which differ only in subject realization. For example, if the subject of a verb can be expressed either as a nominal or a clitic argument with the same frame, the two realizations are collapsed to form a single frame. This leads to 160 verb frames with 2 frames per verb on average. The final modification, concerning the neutralization of a complement as either a reflexive clitic or an NP, results in 1.91 frames per verb, or 858 unambiguous verbs.

As shown in Fig. 7, there are 12 verbs with more than 10 frames, with a maximum of 27 frames for *être* ‘to be’. The general results are presented in Fig. 6.

	# frames	avge	max.	1 frame	
			frame	%	#
passive	453	2.47	100 (*être*)	57.9%	783
clitics	300	2.32	86 (*être*)	58.9%	803
prepositions	222	2.27	72 (*être*)	58.9%	803
attribut & subordinate	173	2.21	43 (*être*)	59.0%	804
subject	160	1.99	27 (*être*)	61.2%	833
reflexive	160	1.91	27 (*être*)	62.9%	858

Figure 6: Mixed representation

être (27), avoir (24), faire (17),
passer (12), rendre (12), rester (12),
porter (12), laisser (11), aller (10),
dire (10), tenir (10), trouver (10)

Figure 7: 12 Most ambiguous verbs (10 frames or more)

frame	# verb types	tokens
SUJ:NP, OBJ:NP	854 (62.7%)	4157 (33.6%)
SUJ:NP, ATS:XP	16 (1.2%)	1932 (15.6%)
SUJ:NP, OBJ:Ssub	95 (7.0%)	1186 (9.6%)
SUJ:NP	339 (24.9%)	1011 (8.2%)
SUJ:NP, OBJ:VPinf	40 (2.9%)	839 (6.8%)
SUJ:NP, DE-OBJ:PP	91 (6.7%)	380 (3.1%)
SUJ:NP, OBJ:NP, A-OBJ:PP	120 (8.8%)	348 (2.8%)
SUJ:NP, A-OBJ:PP	79 (5.8%)	223 (1.8%)
SUJ:NP, P-OBJ:PP	80 (5.9%)	218 (1.7%)
SUJ:NP, OBJ:NP, P-OBJ:PP	75 (5.5%)	185 (1.5%)

Figure 8: 10 most frequent frames (mixed representation)

It is clear that the mixed approach is more precise than the functional one, since it comprises ca. 3 times more frames. But the average number of frames and the ambiguity rate are comparable. The number of frames may be reduced if we further compact frames where complements are optional.

If we consider the most frequent subcategorization frames, we see that, as in the previous approach, most verbs have the direct transitive frame, followed by the strict intransitive one (SUJ, without any complement). We observe as well that verbs with a sentential complement are more frequent than with an infinitival one (both for verb lemmas and occurrences).

3 Comparison with Other Approaches

The LADL tables comprise 38 main frames for simple verbs. They are based on a category of arguments, rather than on their functions, and they include lexical values of certain prepositions. Therefore, the tables distinguish frames where only an infinitive complement is possible (table 1) or a prepositional complement introduced by the preposition *à* (table 33). Our results differ not only in the representation schema but also in number of the obtained frames. For example, we have frames with attributive complements (subject and object attributes) or subjectless verbs which are not present in LADL. On the other hand, as we retain inherent clitics and distinguish different types of subject realization (e.g., impersonal *il* 'it', NP or phrasal), we obtain supplementary frames.

(Candito, 1999) and (Abeillé, 2002) describe families of trees for the FTAG grammar. However, they provide abstract subcategorisation patterns which are not associated with a big lexicon. They distinguish 45 families with a verbal head; 15 of them have nominal arguments, 24 have phrasal arguments and 6 contain an adverbial complement. The grammar contains, as here, a frame for subjectless verbs, a few verbs with an inherent clitic (e.g., *s'évanouir* 'to faint' or *s'appeler* N 'one's name is N'). It is clear, however, that our description is more fine grained.

Finally, we contrast our resource with another lexical database for French verbs: DicoValence, (van den Eynde and Mertens, 2003). The database is dictionary-based, and not corpus-based, as it comprises all the verbs from *Le petit Robert*. It is not theory neutral since it is based on the pronominal approach of (Blanche-Benveniste et al., 1984). The lexicon was developed manually (over more than 6 years) rather than obtained automatically. It is bigger than our lexicon (3700 verbs) but the average ambiguity rate is comparable: 2.4 subcategorization frames per verb. There are 93 different subcategorisation frames defined in DicoValence. They are based either on pronominal form or on semantics (temporal or manner complements for instance), with little categorial information: nominal or prepositional complements are indicated but there is no further distinction for nominal or sentential complements. Finally, grammatical functions are not explicitly indicated either.

4 Conclusion

The presented results of automatic frame extraction from a French treebank are encouraging. We have succeeded in considerably reducing the number of frames by applying different factorization techniques. Despite the important difference in number of frames for the two kinds of representations we adopted, the average number of frames per verb is very similar. This result speaks in favour of the mixed approach as more informative.

We plan different extensions to the work presented here. First, we want to include other verbs from the corpus and not only the verbs in main phrases. Second, we envisage extraction of subcategorization frames for other predicates (adjectives, nouns or adverbs). The frames need also to be validated and evaluated as we plan to use them to complete the syntactic annotations in the treebank. Finally, the lexicon can be easily integrated with other resources so it can be incorporated into syntactic parsers or NLP applications processing French.

The lexicon is freely available from the authors' web page at `http://erssab.u-bordeaux3.fr/article.php3?id_article=150`.

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In *Treebanks*. Kluwer.

Anne Abeillé. 2002. *Une grammaire électronique du français*. CNRS Editions.

C. Blanche-Benveniste, J. Delofeu, J. Stefanini, and K. van den Eynde. 1984. *Pronom et syntaxe. L'approche pronominale et son application au français*. SELAF, Paris.

Didier Bourigault and Cécile Frérot. 2005. Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*.

Joan Bresnan, editor. 1982. *The Mental Representation of Grammatical Relations*. MIT Press Series on Cognitive Theory and Mental Representation. The MIT Press, Cambridge, MA.

T. Briscoe and J. Carroll. 1993. Generalised probabilistic LR parsing for unification-based grammars. *Computational linguistics*.

Aoife Cahill, Mairéad McCarthy, Josef van Genabith, and Andy Way. 2002. Parsing with PCFGs and automatic f-structure annotation. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG02 Conference*. CSLI Publications.

Marie-Hélene Candito. 1999. *Répresentation modulaire et parametrable de grammaires électroniques lexicalisées. Application au français et à l'italien.* Ph.D. thesis, Université Paris7.

John Carroll and A. Fang. 2004. The automatique acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *Proceedings of the 1st International Conference on Natural Language Processing*, Sanya City, China.

Paula Chesley and Susanne Salmon-Alt. 2005. Le filtrage probabiliste dans l'extraction automatique de cadres de sous-catǵorisation. In *Journé ATALA sur l'interface lexique-grammaire*, Paris.

Laurence Danlos and Benoît Sagot. 2007. Comparaison du Lexique-Grammaire et de Dicovalence: vers une intégration dans le Lefff. In *Proceedings TALN'07*.

Laurence Danlos. 1985. *La génération automatique de textes*. Masson.

A. Davis and J-P. Koenig. 2000. Linking as constraints on word classes in a hierachical lexicon. *Language*, 76(1):56–91.

Ingrid Falk, Gil Francopoulo, and Claire Gardent. 2007. Evaluer SynLex. In *Proceedings of TALN'07*.

Karën Fort and Bruno Guillaume. 2007. Preplex: a lexicon of French prepositions for parsing. In *ACL SIGSEM07*.

Anette Frank, Luisa Sadler, Josef van Genabith, and Andy Way. 2002. From treebank resources to LFG f-structures. In *Treebanks*. Kluwer.

Claire Gardent, Bruno Guillaume, Guy Perrier, and Ingrid Falk. 2006. Extraction d'information de souscatégorisation à partir du lexique-grammaire de Maurice Gross. In *TALN 2006*.

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

Alain Guillet and Christian Leclère. 1992. *La structure des phrases simples en français*. Droz, Genève.

Chung-hye Han, Juntae Yoon, Nari Kim, and Martha Palmer. 2000. A feature-based lexicalized tree adjoining grammar for korean. Technical report, IRCS.

A. S. Hornby. 1989. *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, Oxford, 4th edition.

Igor Mel'cuk, Nadia Arbatchewsky-Jumarie, and André Clas. 1984, 1988, 1992, 1999. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques, vol. I, II, III, IV*. Les Presses de l'Université de Montréal.

Bruno Menon and Nicole Modiano. 1993. Eagles: Lexicon Architecture. Technical Report EAG-CLWG-LEXARCH/B, EAGLES.

Carl Pollard and Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. Chicago University Press / CSLI Publications, Chicago, IL.

Paul Procter, editor. 1978. *Longman Dictionary of Contemporary English*. Longman, Burnt Mill, Harlow.

Nilda Ruimy, Ornella Corazzari, Gola Elisabetta, Antonietta Spanu, Nicoletta Calzolari, and Antonio Zampolli. 1998. The European LE-PAROLE Project and the Italian Lexical Instantiation. In *Proceedings of ALLC/ACH, 1998, Lajos Kossuth University, Debrecen, Hungary, July, 5-10 1998*, pages 149–153.

Benoît Sagot and Laurence Danlos. 2007. Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire. constructions impersonnelles. In *Proceedings of TALN'07*.

Benoît Sagot, Lionel Clément, Éric Villemonte de La Clergerie, and Pierre Boullier. 2006. The lefff 2 syntactic lexicon for french: architecture, acquisition, use. In *Actes de LREC 06, Gnes, Italie*.

M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth. 2003. Using predicate-argument structures for information extraction.

Karel van den Eynde and Piet Mertens. 2003. La valence: l'approche pronominale et son application au lexique verbal. *French Language Studies*, 13:63–104.