Growing TreeLex

Anna Kupść[1] and Anne Abeillé[2]

[1] Université de Bordeaux, ERSSàB/SIGNES and IPIPAN;
Université Michel de Montaigne, Domaine Universitaire, UFRL
33607 Pessac Cedex, France
[2] Université Paris7, LLF/CNRS
UMR 7110, CNRS-Université Paris 7, Case 7031, 2, pl. Jussieu
75251 Paris Cedex 05, France
akupsc@u-bordeaux3.fr, anne.abeille@linguist.jussieu.fr

Abstract. TreeLex is a subcategorization lexicon of French, automatically extracted from a syntactically annotated corpus. The lexicon comprises 2006 verbs (25076 occurrences). The goal of the project is to obtain a list of subcategorization frames of contemporary French verbs and to estimate the number of different verb frames available in French in general. A few more frames are discovered when the corpus size changes, but the average number of frames per verb remains relatively stable (about 1.91–2.09 frames per verb).

Key words: Verb valence, subcategorization, treebank

1 Introduction

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

Information about the combinatory potential of a predicate, i.e., the number and the type of its arguments, is called a subcategorization 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, whereas the predicative adjective *fier* 'proud' selects a prepositional complement introduced by the preposition *de*. 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). This implies that subcategorization lexicons which store such syntactic information have to be developed for each language individually.[3] In addition to their importance in language learning, they play a crucial role in many NLP

[3] Work on mapping theory has revealed partial correlations between lexical semantics and subcategorization frames, see for example [10] for linking relations of verbs' arguments. We are not aware of any similar work done for other types of predicates, e.g., adjectives or adverbs.

applications related both to parsing, e.g., [4], [6], [24], and generation, e.g., [9], [17].

The (un)availability of such lexical resources is still a bottleneck for text processing. Traditionally, they have been developed manually by human experts, e.g., [21, 18] (for English) or [15, 16, 19, 25] (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. [5, 13]. Recent years have witnessed also an increased interest in obtaining such resources for French, either by applying statistical techniques, e.g., [2], [7], adapting the existing lexicons, e.g., [14, 11], or using heurisitics to extract valence information [23, 22, 8] for French verbs; a syntactic lexicon of French prepositions has been lately created by [12].

In this paper we present another effort on 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, [1], a journalistic 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 main goal of the project is to obtain a list of different subcategorization frames of French verbs as well as to enrich corpus annotations with this information. We aim also at estimating the number of verb frames in general and propose different methods to reduce the ambiguity rate.

2 Corpus Annotations

In the corpus, all main syntactic constituents are annotated but their internal structure is not indicated. For example, the boundaries of the adverb phrase *pas encore* 'not yet' in Fig. 1 are marked but its components (i.e., words *pas* 'not' and *encore* 'yet') are treated on a par, i.e., no structural relation between them is indicated.

The adopted annotation schema distinguishes a VP only for infinitive phrases. Instead, for inflected verbs, a verbal nucleus (VN) is defined and it contains the main verb, auxiliaries, negation, pronominal clitics and adverbs which follow the auxiliary. The head verb is not explicitly indicated but we assume that the last verb in VN is the head. Note that pronominal clitics, e.g., the pronominal subject *il* 'he' in Fig. 1, are not treated as syntactic NPs but are part of VN.

Syntactic functions are annotated only for verbal dependents. As shown in Fig. 1, the verb *sait* 'knows' has the subject (NP) and the object (a subordinate phrase, Ssub) indicated but no relation is specified between the noun *état-major* 'management' and its AP modifier *français* 'French'. Functions are treated as

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

Fig. 1. Example of annotation schema: *L'état-major français sait qu'il a gagné une bataille mais pas encore la guerre* 'The French management knows that they won a battle but not yet the war'

relations between constituents (e.g., VN and NP *une bataille* 'a battle' in Fig. 1) and they do not link directly the head and its dependents (i.e., V and NP).

Theoretical approaches use different representations of subcategorization frames. In some models, like LFG [3], the notation based on functional information is preferred (1), while in others, like LADL (lexicon–grammar of [15]), a categorial notation is adopted (2), yet in others, like HPSG [20], a mixed approach is used (3):

(1) <SUJ, OBJ>

(2) N0 V N1

(3) <SUJ:NP, OBJ:NP>

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 sentential, 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 (3) in order to obtain more complete information. The functional representation (1) will be used for a comparison.

The list of categories and functions used in the corpus is presented in Tab. 2. The list ignores two functions: MOD, which always corresponds to non-subcategorized elements, and COORD, which represents coordinated phrases, relatively rare in the corpus, and which does not provide the category information. For prepositional complements, P-OBJ, we retain the type of the preposition which introduces the complement. This allows us to normalize verb frames with respect to active and passive forms.

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

Fig. 2. Possible categories for every function of a verb. Functions: SUJ (subject), OBJ (direct object), DE-OBJ (indirect object introduced by *de*), A-OBJ (indirect object introduced by *à*), P-OBJ (a prepositional complement introduced by a different preposition), ATO (object's attribute), ATS (subject's attribute)

3 Frame Extraction

3.1 Experiment

For extraction of the verb valency, we used the part of the corpus which contains both constituent and functional annotations, i.e., about 20 000 phrases (500 000 words). Our experiment was divided into two steps: first, verbs in the main clauses, i.e., verbs with all functions specified, have been used, which resulted in a lexicon of 1362 verb lemma (12 353 occurrences). Then, we complemented annotations for other verbs, e.g., we added missing subjects to imperative and infinitive forms and we completed frames of verbs in relative clauses. This resulted in 2006 verb lemma (25 076 occurrences) in the final verb lexicon.

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 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 by 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'. Such clitics are nevertheless tightly dependent on the verb so we retain them in the subcategorization frames. In order to indicate clitics, we added two more functions: `refl` for reflexive clitics and `obj` for all other clitics. Moreover, 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*). Such duplicated functions had to be eliminated. Finally, there are frames which are missing the subject. It has been added to the imperative forms and infinitives in subordinate clauses. There are two lemma which always appear without a subject, *voici* and *voilà* '(t)here is'. They are considered indicative verbs which do not have a subject.

We normalized frames with respect to passive vs. active form. We used a list of 62 verbs which can be 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 *être* 'be' but its past tense form requires another auxiliary (*avoir* 'have'), the form is considered passive and it is transformed to an active form. We add OBJ to the frame (as SUJ is already present), whereas if the PP expressing the agent is present, i.e., P-OBJ introduced by the preposition *par* or *de*, this PP is deleted. If the passive form appears with an ATS complement (the subject's attribute), we rename this function to ATO (the object's attribute). All other functions in the frame (if any) remain unchanged.

In French, syntactic arguments don't have to be realized in a fixed order. For example, the order of complements is relatively free, cf. (4) and (5), and the subject can also appear postverbally (subject inversion). The order in which functions appear in the frames does not reflect their surface order but has been normalized based on obliqueness and they are listed as follows: SUJ, OBJ, A-OBJ, DE-OBJ, P-OBJ, ATS, ATO, obj, refl. For instance, the verb *parle* 'talks' in (4) and (5), has the same subcategorization frame for both sentences (SUJ, A-OBJ, DE-OBJ):

(4) Marie parle [de ce problème] [à Paul].
Mary talks of this problem to Paul
Mary is talking to Paul about this problem

(5) Marie parle [à Paul] [de ce problème].

In some cases, corpus annotations turned out to be insufficient to extract correct frames. For example, 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 dependents realized on a distance, e.g., in *Que peut faire le gouvernement?* 'What can the government do?', we extract (incorrectly) two objects for the verb *peut* 'can' (*que* 'what' and *faire* 'do') and none for the verb *faire*. Such cases are nevertheless quite rare.

3.2 Results

Below we present an analysis how different representations and parametrization techniques influence the number of extracted frames and their ambiguity rate. These results are provided for the initial data set, i.e., frames of verbs in main clauses. The impact of the size of the data set used is discussed in sec. 3.3.

Functional Representation As indicated in Fig. 3, after neutralization of passive and active forms, we obtain 142 different subcategorization frames, with an average of 1.9 frames per verb 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 subcategorization frame. Then we perform several operations in order to eliminate superfluous clitic arguments. We clean the frames so that duplicated functions are removed. After these modifications,

	# frames	average	max. nr of 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

Fig. 3. Functional representation

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 appears with 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 the functional approach is that we have lost categorial information available in the corpus. For example, 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.

Mixed Representation A mixed representation (with categories and functions), after depassivization, gives a gross total of 783 different subcategorization frames, with an average of 2.47 frames per lemma, and almost 58% of the lemmas which have only one frame. With the clitic factorization described in section 3.1, 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 the lexical value of a prepositional complement (indirect complements introduced by prepositions other than *à* or *de*). The average number of subcategorization frames drops slightly (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 neutralize different realizations of the attribute (ATS and ATO) and types of a subordinate clause (interrogative, Sint, vs. subordinate, Ssub). 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 all other arguments being the same,

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)

Fig. 4. Subcategorization 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%)

Fig. 5. 10 most frequent frames (functional representation)

the two realizations are merged 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. It is clear that the mixed approach is more precise than the functional

	# frames	average	max. nr of frames	1 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
attribute & 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

Fig. 6. Mixed representation

```
être (27), avoir (22), faire (17), passer (12), rendre (12), rester (12),
porter (12), laisser (11), aller (10), dire (10), tenir (10), trouver
(10)
```

Fig. 7. 12 Most ambiguous verbs (10 frames or more); mixed representation

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 further reduced if we compact frames with optional complements.

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 complements). We observe as well that verbs with a sentential complement are more frequent than with an infinitival one (both for verb types and tokens).

3.3 More Data

As indicated in Fig. 9, the size of the data set influences the results. If we consider all verbs (in main and subordinate clauses), the number of all frames and the ambiguity rate increase, for both representations. Although these changes are noticeable (8 new frames discovered for the functional and 20 for the mixed approach), they are not dramatic given that the number of verbs considered raises by almost 70%. Moreover, the frequency of the new frames is very low, e.g., all new functional frames appear only once in the corpus, see Fig. 10, whereas only 6 of the 20 new mixed frames occur more than once, cf. Fig. 11. In particular, it should be verified if these frames are attested on a different data set

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%)

Fig. 8. 10 most frequent frames (mixed representation)

representation	# lemmas	# frames	average	max. nr of frames	verbs with 1 frame	verbs with ≥ 10 frames
function	1362	58	1.72	16 (être)	859 (63%)	6 (0.4%)
	2006	66	1.93	21 (être)	1183 (59%)	12 (0.6%)
mixed	1362	160	1.91	27 (être)	833 (61.1%)	13 (0.9%)
	2006	180	2.09	29 (être)	1168 (58.2%)	29 (1.4%)

Fig. 9. Comparison of results: verbs in main clauses (1362 types) vs. all verbs (2006 types)

or whether additional factorization techniques should incorporate them to the existing frames. For example, the reflexive clitic in Fig. 10 might be a result of insufficient factorization. Similarly, the frames with an apparently impersonal subject (`SUJ:il`) might be due to insufficient data: *il* 'it' is either an impersonal or a personal (3sg. masc) clitic. In the latter case, it can be replaced by an NP, hence the subject realization should be specified as SUJ:NP.

The majority of frames detected on the smaller sample are confirmed, i.e., their frequency increases or remains the same: 93% of functional and 83% of mixed frames found in both data sets. For the remaining shared frames, their frequency drops on the bigger data set. The main reason for this apparent paradox is that some frames get 'corrected' by supplementary data: for example, the 'impersonal' *il* subject turns out to be a personal pronoun if the verb appears with an NP subject as well (e.g., the frequency of `SUJ:il, OBJ:NP, A-OBJ:PP` drops from 6 to 3 in the final data set), or the frame has been reclassified with a different frame realized by the same verb (for instance, the initial `SUJ:NP, obj:en` is regrouped with `SUJ:NP, DE-OBJ:PP` which was found for the verb *faire* 'make/do' in the larger sample; this makes the overall frequency of the former frame drop from 7 to 6).

As frequencies change, the final ranking of frames is slightly different as well. For the ten top frames more regrouping occurs among the function-based

```
SUJ, DE-OBJ, P-OBJ, refl (1); SUJ, DE-OBJ, P-OBJ, ATS (1);
SUJ, A-OBJ, DE-OBJ, refl (1); SUJ, A-OBJ, DE-OBJ, ATS (1);
SUJ, A-OBJ, ATS, refl (1); SUJ, OBJ, DE-OBJ, ATS (1);
SUJ, A-OBJ, ATO (1); SUJ, obj:te (1)
```

Fig. 10. 8 additional functional frames with their frequencies

```
SUJ:NP, OBJ:VPinf, DE-OBJ:PP (11); SUJ:il, OBJ:AdP, obj:y (3);
SUJ:NP, OBJ:NP, P-OBJ:PP, refl:CL (3);
SUJ:NP, OBJ:NP, A-OBJ:VPinf, refl:CL (2);
SUJ:il, A-OBJ:PP, DE-OBJ:VPinf (2); SUJ:NP, OBJ:NP, A-OBJ:AP (2)
```

Fig. 11. 6 of the additional mixed frames which occur more than once

frames (5 frames get promoted), whereas only strictly intransitive verbs appear more frequently if the mixed representation is considered. Table 12 presents a comparison of 10 top frames for both representations (numbers correspond to frequency counts in the bigger and smaller, in square brackets, data sets). Frames which get a higher rank in the final evaluation are boldfaced.

functional representation			mixed representation		
frame	v.types	v.tokens	frame	v.types	v.tokens
SUJ,OBJ	1431 [913]	13461 [6407]	SUJ:NP,OBJ:NP	1387 [854]	10257 [4157]
SUJ	730 [351]	3166 [1035]	**SUJ:NP**	717 [339]	3137 [1011]
SUJ, ATS	17 [16]	2582 [1951]	SUJ:NP,ATS:XP	17 [16]	2561 [1932]
SUJ,OBJ,A-OBJ	224 [162]	1083 [517]	SUJ:NP,OBJ:Ssub	115 [95]	1987 [1186]
SUJ,DE-OBJ	202 [129]	1028 [558]	SUJ:NP,OBJ:VPinf	40 [40]	1138 [839]
SUJ,A-OBJ	155 [103]	733 [359]	SUJ:NP,DE-OBJ:PP	162 [91]	843 [380]
SUJ,P-OBJ	150 [85]	494 [233]	SUJ:NP, OBJ:NP,A-OBJ:PP	183 [120]	770 [348]
SUJ,OBJ,DE-OBJ	186 [75]	468 [160]	SUJ:NP,A-OBJ:PP	128 [79]	518 [223]
SUJ,OBJ,P-OBJ	154 [81]	399 [197]	SUJ:NP,P-OBJ:PP	145 [80]	471 [218]
SUJ,OBJ,ATO	45 [32]	248 [114]	SUJ:NP, OBJ:NP,P-OBJ:PP	149 [75]	387 [185]

Fig. 12. Comparison of 10 most frequent frames

Finally, Fig. 9 shows that the number of most ambiguous verbs is doubled, for both representations. This indicates that frame ambiguity is in fact more common than predicted by our initial sample. The 29 verbs which belong to the class of more than 10 frames (mixed representation) are indicated in Fig. 13; the verbs which entered this class are written in boldface. The numbers in brackets indicate the number of frames associated with these verbs in the bigger and smaller samples.

être (29|27), avoir (22|22), faire (19|17), rester (17|12), passer (16|12), tenir (15|10), porter (15|12), trouver (14|10), venir (14|9), présenter (13|5), rendre (13|12), attendre (12|8), dire (12|10), vendre (11|6), pouvoir (11|6), voir (11|8), aller (11|10), estimer (10|9), revenir (10|8), engager (10|6), laisser (10|11), demander (10|9), montrer (10|8), devoir (10|8), appeler (10|6), déclarer (10|9), permettre (10|8), assurer (10|4), mettre (10|8)

Fig. 13. 29 Verbs with more than 10 (mixed) frames in the bigger sample

4 Conclusion

We presented results of an automatic frame extraction from a French treebank. We have succeeded in considerably reducing the number of verb 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 fact speaks in favor of the mixed approach as more informative. Moreover, these numbers do not drastically change with the size of the data set which indicates that the number and types of frames has stabilized. On the contrary, the repertoire of frames for individual verbs is still growing.

We plan different extensions to the work presented here. 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. 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:
http://erssab.u-bordeaux3.fr/article.php3?id_article=150.

References

1. Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In *Treebanks*. Kluwer.
2. 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*.
3. Joan Bresnan, editor. 1982. *The Mental Representation of Grammatical Relations*. MIT Press Series on Cognitive Theory and Mental Representation. The MIT Press, Cambridge, MA.
4. T. Briscoe and J. Carroll. 1993. Generalised probabilistic LR parsing for unification-based grammars. *Computational linguistics*.
5. 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.

6. 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.
7. Paula Chesley and Susanne Salmon-Alt. 2005. Le filtrage probabiliste dans l'extraction automatique de cadres de sous-catgorisation. In *Journé ATALA sur l'interface lexique-grammaire*, Paris.
8. Laurence Danlos and Benoît Sagot. 2007. Comparaison du Lexique-Grammaire et de Dicovalence: vers une intégration dans le Lefff. In *Proceedings TALN'07*.
9. Laurence Danlos. 1985. *La génération automatique de textes*. Masson.
10. A. Davis and J-P. Koenig. 2000. Linking as constraints on word classes in a hierachical lexicon. *Language*, 76(1):56–91.
11. Ingrid Falk, Gil Francopoulo, and Claire Gardent. 2007. Evaluer SynLex. In *Proceedings of TALN'07*.
12. Karën Fort and Bruno Guillaume. 2007. Preplex: a lexicon of French prepositions for parsing. In *ACL SIGSEM07*.
13. Anette Frank, Luisa Sadler, Josef van Genabith, and Andy Way. 2002. From treebank resources to LFG f-structures. In *Treebanks*. Kluwer.
14. Claire Gardent, Bruno Guillaume, Guy Perrier, and Ingrid Falk. 2006. Extraction d'information de sous-catégorisation à partir du lexique-grammaire de Maurice Gross. In *TALN 2006*.
15. Maurice Gross. 1975. *Méthodes en syntaxe*. Hermann.
16. Alain Guillet and Christian Leclère. 1992. *La structure des phrases simples en français*. Droz, Genève.
17. Chung-hye Han, Juntae Yoon, Nari Kim, and Martha Palmer. 2000. A feature-based lexicalized tree adjoining grammar for Korean. Technical report, IRCS.
18. A. S. Hornby. 1989. *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, Oxford, 4th edition.
19. 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.
20. Carl Pollard and Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. Chicago University Press / CSLI Publications, Chicago, IL.
21. Paul Procter, editor. 1978. *Longman Dictionary of Contemporary English*. Longman, Burnt Mill, Harlow.
22. 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*.
23. 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, Gênes, Italie*.
24. M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth. 2003. Using predicate-argument structures for information extraction.
25. 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.