PrepLex: a lexicon of French prepositions for parsing

Karën Fort
Calligramme and TALARIS projects
LORIA/INRIA Lorraine / Nancy, France
Karen.Fort@loria.fr

Bruno Guillaume
Calligramme project
LORIA/INRIA Lorraine / Nancy, France
Bruno.Guillaume@loria.fr

Abstract

PrepLex is a lexicon of French prepositions which provides all the syntactic information needed for parsing. It was built by comparing and merging several authoritative lexical sources. This lexicon also includes information about the prepositions or classes of prepositions that appear in French verb subcategorization frames. This resource has been developed as a first step in making current French preposition lexicons available for effective natural language processing.

1 Introduction

When defining lexical entry classes according to categories, an obvious distinction appears between two types of classes. First, the closed classes, comprising elements which can be exhaustively enumerated, for example pronouns or determiners. Second, open classes for which it is impossible to list all the elements (for example, they may vary according to the domain). The four main open classes are nouns, verbs, adjectives and adverbs. The lexicon construction methodology has to be adapted according to the type of class that is being dealt with.

The status of the class of prepositions is difficult to determine. A priori, prepositions may seem to be a closed class, with elements which can be enumerated. In practice, however, a comparison of the different available resources shows that it is not an easy task to exhaustively list prepositions. Besides, they represent more than 14% of French lemma tokens.[1]

A complete lexicon for parsing applications should contain subcategorization information for predicative words (Briscoe and Carroll, 1993; Carroll and Fang, 2004). This subcategorization information often refers to prepositions in the description of their arguments. Arguments are commonly used with a particular preposition (for example *compter sur* [count on]) or a set of semantically linked prepositions (such as *aller* [go] *LOC*, where *LOC* can be any locative preposition).

For deep parsing, we need to distinguish between indirect complements, required by the verb, and adjuncts which do not appear in the verb valence. The following two examples (1a) and (1b) have the same surface structure, in which the two preposition uses for *avec* can only be distinguished semantically: in the first case, it introduces an oblique complement, whereas in the second case, it introduces an adjunct. This issue can be solved using finer-grained semantic information.

1a. *Jean se bat avec Paul*
[Jean fights against Paul]
1b. *Jean se bat avec courage*
[Jean fights with courage]

This distinction leads us to allow two different preposition uses and therefore causes lexical ambiguity. In order to limit this ambiguity, it is important for a lexicon to identify the prepositions which can have both functions (we will call these "argument" prepositions).

[1] see for example, on a newspaper corpus: https://www.kuleuven.be/ilt/blf/rechbaselex_kul.php\#freq (Selva et al., 2002)

Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions, pages 17–24,
Prague, Czech Republic, June 2007. ©2007 Association for Computational Linguistics

Our work aims at providing the community with a lexicon that can be directly used by a parser. We focused on syntactic aspects and extended the work to some semantic elements, like semantically linked sets of prepositions (as *LOC*). The generated lexicon is freely available and is expected to be integrated into larger resources for French, whether existing or under development.

Section 2 describes the sources and the comparative methodology we used. Section 3 details the results of the comparison. Section 4 explains how the lexicon was created from the above-mentioned results. Finally, Section 5 shows an example of use of the lexicon in a parsing application.

2 Methodology

In order to use prepositions for parsing, we need a large list, containing both garden-variety prepositions and prepositions that appear in verb subcategorization frames.

2.1 Using syntactic lexicons

Obviously, some lexicons already exist which provide interesting lists of prepositions. This is the case of Le*fff* (Sagot et al., 2006), which contains a long list of prepositions. However, the syntactic part of the lexicon is still under development and it provides only few prepositions in verb subcategorization frames. Besides, some prepositions in Le*fff* are obsolete or rare. The French-UNL dictionary (Sérasset and Boitet, 2000) also contains prepositions, but its coverage is quite limited and the quality of its entries is not homogeneous. Other sources present prepositions in verb subcategorization frames, but the lists are not quite consistent.

We thus collected, as a first step, prepositions from a certain number of resources, lexicons and dictionaries for the garden-variety list, and syntactic lexicons for the argument prepositions list. Two resources belong to both categories, Le*fff* and French-UNL dictionary:

- Le*fff* (Lexique des Formes Fléchies du Français/French inflected form lexicon (Sagot et al., 2006)) is a large coverage (more than 110,000 lemmas) French morphological and syntactic lexicon (see table 1 for an example of a Le*fff* syntactic entry).

  In its latest public version, 2.2.1, Le*fff* contains 48 simple prepositions and 164 multiword prepositions. It also provides information on verb subcategorization frames, which contain 14 argument prepositions.

- UNL (Universal Networking Language (Sérasset and Boitet, 2000)), is a French to disambiguated English dictionary for machine translation, which contains syntactic information in its French part (see table 1 for a UNL example entry).

  UNL has limited coverage (less than 27,000 lemmas), but it provides, in the English part, semantic information that we will consider using in the near future. UNL contains 48 simple prepositions, among which 12 appear in verb subcategorization frames.

2.2 Using reference sources

We then completed the list of prepositions using manually built resources, including lexicons, dictionaries and grammars:

- The Grevisse (Grevisse, 1997) grammar, in its paper version, allowed us to check some intuitions concerning the obsolescence or usage of some prepositions.

- The TLFi (Trésor de la langue française informatisé), that we consulted through the CNRTL[2], and that offers a slightly different list of prepositions. In particular, it contains the forms *voici* and *voilà*, that are seldom quoted in the other available resources.

- Finally, the PrepNet (Saint-Dizier, 2006) prepositions database was used to check the completeness of our list as well as the semantic information provided by other sources.

2.3 Using verb valence dictionaries

We then looked for a way to enrich the list of prepositions appearing in verb subcategorization frames in Le*fff* and UNL, using resources that focus more particularly on verbs:

[2]see: `http://www.cnrtl.fr`

Le*fff* entry for *dialoguer avec* [to talk to]

```
dialoguer: suj:sn|sinf|scompl,obja:(à-sn|avec-sn),objde:(de-sn|de-scompl|de-sinf)
```

UNL entry for *dialoguer avec* [to talk to]

```
[dialoguer] {AUX(AVOIR),CAT(CATV),GP1(AVEC),VAL1(GN)} "have_talks";
```

DICOVALENCE entry for *dialoguer avec* [to talk to]

```
VAL$      dialoguer: P0 PP<avec>
VTYPE$    predicator simple
VERB$     DIALOGUER/dialoguer
NUM$      29730
EG$       le délégué des étudiants a dialogué avec le directeur de l'école
TR$       spreken, zich onderhouden, een gesprek hebben, onderhandelen
P0$       qui, je, nous, elle, il, ils, on, celui-ci, ceux-ci
PP_PR$    avec
PP$       qui, lui_ton, eux, celui-ci, ceux-ci, l'un l'autre
LCCOMP$   nous dialoguons, je dialogue avec toi
```

SynLex entry for *adapter avec* [to adapt to]

```
adapter   '<suj:sn,obj:sn,obl:avec-sn>'
```

Table 1: Description of some entries with the preposition *avec* [with] in valence dictionaries

- DICOVALENCE, a valence dictionary of French, formerly known as PROTON (van den Eynde and Mertens, 2002), which has been based on the pronominal approach. In version 1.1, this dictionary details the subcategorization frames of more than 3,700 verbs (table 1 gives an example of a DICOVALENCE entry).

  We extracted the simple and multiword prepositions it contains (i.e. more than 40), as well as their associated semantic classes.

- We completed this argument prepositions list with information gathered from SynLex (Gardent et al., 2006), a syntactic lexicon created from the LADL lexicon-grammar tables (Gross, 1975) (see table 1 for a SynLex entry).

Using these sources, we conducted a systematic study of each preposition, checking its presence in each source, whether in verb subcategorization frames or not, as well as its associated semantic class(es). We then grouped the prepositions that appear both as lexical entries and in verb subcategorization frames.

As multiword prepositions show specific characteristics (in particular, their number) and raise particular issues (segmentation), we processed them separately, using the same methodology.

3 Source comparison results

3.1 Simple prepositions

We thus listed 85 simple prepositions, among which 24 appear in verb subcategorization frames (see table 2).

It is noticeable that the different sources use quite different representations of syntactic information as shown in table 1. Le*fff* offers a condensed vision of verbs, in which valence patterns are grouped into one single entry, whereas SynLex uses a flatter representation without disjunction on syntactic categories for argument realization or for optional arguments. To summarize, we could say that DICOVALENCE lies somewhere between Le*fff* and SynLex, since it uses disjunctive representation but has a finer description of syntactic information and hence splits many entries which are collapsed in Le*fff*.

3.2 Multiword prepositions

We obtained a list of 222 multiword prepositions, among which 18 appear in verb subcategorization frames (see table 3). It is to be noticed that only DICOVALENCE and SynLex contain multiword prepositions in verb subcategorization frames. As for Le*fff*, it provides an impressive list of multiword

	Lexicons					Subcategorization frames			
	Le*fff*	TLFi	Grevisse	PrepNet	UNL	Le*fff*	DV[a]	SynLex	UNL
à	X	X	X	loc		319	895 (18 loc)	887 (70 loc)	246
après	X	X	X	loc	X	2	12	1	
aussi					X				
avec	X	X	X	X	X	35	193 (1 loc)	611 (1 loc)	49
chez	X	X	X	loc	X		9 (5 loc)		1
comme	X				X	14	11	10	3
de	X	X	X	deloc	X	310	888 (117 deloc)	1980 (69 deloc)	282
depuis	X	X	X	deloc	X		2	1	
derrière	X	X	X	loc	X		3		
devers	X	X	X						
dixit	X								
emmi		X							
entre	X	X	X	loc	X		19 (3 loc)	4	
hormis	X	X	X	X	X				
jusque	X	X	X		X		7 (7 loc)		
lès	X	X	X						
moyennant	X	X	X	X	X				
par	X	X	X	loc	X	3	38 (4 loc)	73	8
parmi	X	X	X	loc	X		7 (3 loc)	7	
passé		X			X				
selon	X	X	X	X	X		1	1	
voici		X			X				

Table 2: Some simple prepositions in different sources

[a]DICOVALENCE

prepositions (more than 150) which represents an excellent basis for our work.

4 Lexicon construction

The first selection criterion we applied to build the lexicon is that a preposition should appear in at least one source among the above-mentioned ones. Also, we consider a preposition to be an argument preposition if it appears in at least one verb subcategorization frame.

4.1 Manual filtering

We then filtered the prepositions according to very simple criteria. In particular, we identified some prepositions to be removed as they were:

- erroneous, this is the case, for example, of *aussi* (adverb rather than preposition), which is present in the UNL dictionary as a preposition,
- obsolete or very rare, like *emmi* (from TLFi), *devers* (from Le*fff*, TLFi, Grevisse) or *comme de* (from DICOVALENCE).

We also checked the semantic features given in the sources and removed erroneous ones, like *avec* as locative in SynLex and DICOVALENCE.

4.2 Some remarks

Some sources include as prepositions forms that are not universally considered to be prepositions in linguistics. This is the case, in particular, for:

- *comme*, which is not present in the three reference sources (Grevisse, TLFi and PrepNet) as it is ambiguous and can also be used as a conjunction,

	Lexicons					Subcategorization frames			
	Le*fff*	TLFi	Grevisse	PrepNet	UNL	Le*fff*	DV[a]	SynLex	UNL
à cause de	X		X	X					
à la faveur de			X	X					
à partir de	X		X	deloc				1	
afin de	X	X	X	X					
au nord de				loc					
au vu de	X								
auprès de	X	X	X	loc			27 (1 loc)	35	
comme de							1		
conformément à	X			X					
d'avec			X				1	6	
d'entre	X								
en faveur de	X		X	X			13		
face à	X		X				2		
il y a	X								
jusqu'à	X			loc	X		10 (2 loc)		
jusqu'en	X								
jusqu'où	X								
loin de	X		X	loc					
par suite de			X						
pour comble de	X								
près de	X		X	loc					
quant à	X	X	X						
tout au long de	X			X					
vis-à-vis de	X		X	X				1	

Table 3: Some multiword prepositions in different sources

[a]DICOVALENCE

- *il y a* or *y compris*, which only appear in Le*fff*,
- *d'avec*, which only appears in Grevisse and verb subcategorization frames in DICOVALENCE and SynLex.

We decided to keep those forms in the lexicon for practical reasons, keeping the parsing application in mind.

Moreover, even if its coverage is quite large, the created lexicon is obviously not exhaustive. In this respect, some missing entries should be added, namely:

- prepositions from the DAFLES (Selva et al., 2002), like, for example, *au détriment de*,
- prepositions appearing in reference grammars, like *question*, in Grammaire méthodique du français (Riegel et al., 1997),
- some locative prepositions (and, through metonymy, time prepositions) that are prefixed by *jusqu'*, for example *jusqu'auprès de*. This elided form of *jusque* should probably be treated separately, as a preposition modifier. The same goes for *dès*, followed by a time preposition (or a locative one, through metonymy).

However, it is to be noticed that none of these missing prepositions appear in verb subcategorization frames.

This filtering process also allowed us to identify some issues, in particular elisions in multiword forms, like *afin de*, *afin d'*, or contractions like *face à*, *face au* or *à partir de*, *à partir du*, which will be processed in the segmentation step.

Others, like *lès*, which is only used in toponyms in dashed forms (e.g. Bathelémont-lès-Bauzemont), will be processed during named entity segmentation.

4.3 Results

We obtained a list of 49 simple prepositions, of which 23 appear in verb subcategorization frames in at least one source and are therefore considered to be argument prepositions (see table 4).

We also obtain a list of more than 200 multiword prepositions, among which 15 appear in verb subcategorization frames in at least one source and are therefore considered to be argument prepositions (see table 5).

For the time being, we limited the semantic information in the lexicon to *loc* (locative) and *deloc* (source), but we intend to extend those categories to those used in DICOVALENCE (time, quantity, manner). We have already added those to the prepositions database that is being populated.

We also referred to the sources to add the categories of the arguments introduced by argument prepositions.

PrepLex is currently distributed in a text format suitable both for hand-editing and for integration in a parser or other natural language processing tools. In the format we propose, syntactic information is described via feature structures. These feature structures are always recursive structures of depth 2. The external level describes the structure in terms of "arguments" whereas the internal level gives a finer syntactic description of either the `head` or of each argument. This format aims at being modular and at defining some "classes" that share redundant information. In the case of prepositions, the skeleton of the feature structure used by all entries is:

```
Prep : [
head [cat=prep, prep=#, funct=#]
comp [cat=#, cpl=@]
]
```

When instantiated for a particular preposition, 3 feature values are to be provided (written with "#" in the above description) and the last parametrized feature (written with `@`) is optional. When they are in the head sub-structure, features are referred to by their names whereas, in other cases, a prefix notation is used.

```
à [prep=a|LOC; funct=aobj|loc|adj;
   comp.cat=np|sinf; comp.cpl=void|ceque]
après [prep=apres|LOC; funct=obl|loc|adj;
       comp.cat=np]
avec   [prep=avec; funct=obl|adj;
        comp.cat=np]
à_travers [prep=a_travers; funct=obl|adj;
           comp.cat=np]
```

Technically, the only difficult part is to decide how to represent semantic classes of prepositions like *LOC*. Here, we chose to define the whole set of argument prepositions as well as all the semantic classes (noted in uppercase) as possible atomic values for the *prep* feature. We then used the disjunction `a|LOC` to indicate that the preposition *à* can be used, either as a specific preposition or as a locative preposition.

Additionally, we decided to add to the lexicon information about the sources in which the preposition appears, in order to allow filtering for some specific applications. In the case of argument prepositions, we also added information about the preposition's frequency in the source, as well as a relevant example.

We also decided to add corpus-based frequencies to the lexicon. Thus, for each preposition, we provide its frequency per 1000 words, either as found in the DAFLES (Selva et al., 2002), from a newspaper corpus composed of *Le Monde* and *Le Soir* (1998), or as extracted directly from *Le Monde* (1998) with a simple *grep* command, without tagging.

5 Using the lexicon in a NLP system

We briefly expose some parsing problems related to prepositions.

5.1 Segmentation issues

The first issue that appears when integrating prepositions in a parsing system is that of segmentation. In particular, contractions have to be processed specifically so that *au* is identified as the equivalent of *à le*. The same goes for *de*, which can appear in some multiword prepositions and can be elided as *d'*. However, these phenomena are not specific to prepositions. They can be addressed either in the lexicon (for example Le*fff* explicitly contains both

Lexicons						Subcategorization frames				
Le*fff*	TLFi	Grevisse	PrepNet	UNL	**PrepLex**	Le*fff*	DV	SynLex	UNL	**PrepLex**
44	69	55	36	46	**49**	14	24	18	11	**23**

Table 4: Total number of simple prepositions by source

Lexicons						Subcategorization frames				
Le*fff*	TLFi	Grevisse	PrepNet	UNL	**PrepLex**	Le*fff*	DV	SynLex	UNL	**PrepLex**
166	11	77	89	2	**206**	0	16	4	0	**15**

Table 5: Total number of multiword prepositions by source

au cours de and *au cours d'*), or during the segmentation step.

We decided on the second solution as it improves lexicon maintainability.

An issue that is more directly linked to multiword prepositions is that of segmentation ambiguities. For example, in the following two sentences (2a) and (2b) the group of words *au cours de* is a multiword preposition in the first case, but it has to be decomposed in the second one. Other multiword prepositions can never be decomposed, for example *y compris*.

This highlights the fact that segmentation is ambiguous and that it is necessary to be able to keep the segmentation ambiguity through the whole parsing process.

2a. *Il a beaucoup travaillé au cours de cette année*
[He worked hard during the year]

2b. *Il a beaucoup travaillé au cours de M. Durand*
[He worked hard in Mr Durand's course]

5.2 *Adjunct* prepositions vs *argument* prepositions

In deep parsing we have to distinguish between prepositions introducing a verb argument and prepositions introducing adjuncts. However, we have seen that this distinction often relies on semantics and that parsing should leave the two possibilities open. Precise information about argument prepositions and verb subcategorizations eliminates many of these ambiguities.

6 Conclusion

We created a list of French prepositions for parsing applications by comparing various lexicons and dictionaries. We hence focused on syntactic aspects. Manual filtering was used to eliminate obsolete or rare prepositions, as well as a number of errors. The resulting lexicon contains more than 250 French prepositions, among which 49 are simple prepositions.

In syntactic lexicons, subcategorization frames describe prepositions introducing arguments. Prepositions appearing in verbal valence frames are called "argument prepositions". We identified 40 of them.

The produced lexicon is freely available. [3] It will be developed further. In particular, some other information sources will be incorporated. This is the case for the verbs *constructions* fields from the TFLi which contain prepositions, that can be considered as argument prepositions. We plan to use this information to improve the lexicon.

We are also populating a database with this lexical information. [3] This will help us ensure a better maintenance of the lexicon and will allow enrichment of the entries, in particular with examples and associated verbs. We are adding corpus-based frequencies to this database.

A more ambitious task would be to enrich the lexicon with fine-grained semantic information (more detailed than the general classes *loc, deloc,* ...). Many interesting linguistic studies have been conducted on prepositions, including cross-lingual approaches. However, most of them are limited to detailing the semantics of a small number of prepositions; with the exceptions of PrepNet (Saint-Dizier, 2006) for French prepositions and TPP (Litkowski and Hargraves, 2005) (The Preposition Project) for English. It is now necessary to transform those resources in order to make them directly usable by natural language processing systems.

[3] `http://loriatal.loria.fr/Resources.html`

References

Ted Briscoe and John A. Carroll. 1993. Generalized probabilistic LR parsing of natural language (corpora) with unification-based grammars. *Computational Linguistics*, 19(1):25–59.

John A. Carroll and Alex C. Fang. 2004. The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)*, pages 107–114, Sanya City, China.

Claire Gardent, Bruno Guillaume, Guy Perrier, and Ingrid Falk. 2006. Extraction d'information de sous-catégorisation à partir des tables du LADL. In *Proceedings of TALN 06*, pages 139–148, Leuven.

Maurice Grevisse. 1997. *Le Bon Usage – Grammaire française, édition refondue par André Goosse*. DeBoeck-Duculot, Paris – Leuven, 13^{th} edition.

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

Ken Litkowski and Orin Hargraves. 2005. The preposition project. In *Proc. of the ACL Workshop on Prepositions*.

Martin Riegel, Jean-Christophe Pellat, and René Rioul. 1997. *Grammaire méthodique du français*. PUF, 3^{rd} edition.

Benoit Sagot, Lionel Clément, Éric Villemonte de la Clergerie, and Pierre Boullier. 2006. The Lefff 2 syntactic lexicon for French: architecture, acquisition, use. In *Proc. of LREC 06, Genoa, Italy*.

Patrick Saint-Dizier. 2006. PrepNet: a Multilingual Lexical Description of Prepositions. In *Proc. of LREC 06, Genoa, Italy*, pages 877–885. European Language Resources Association (ELRA).

Thierry Selva, Serge Verlinde, and Jean Binon. 2002. Le DAFLES, un nouveau dictionnaire pour apprenants du français. In *Proc. of EURALEX'2002 (European Association for Lexicography), Copenhagen*.

Gilles Sérasset and Christian Boitet. 2000. On UNL as the future "html of the linguistic content" and the reuse of existing NLP components in UNL-related applications with the example of a UNL-French deconverter. In *Proceedings of COLING 2000, Saarbrücken*.

Karel van den Eynde and Piet Mertens, 2002. *La valence: l'approche pronominale et son application au lexique verbal*, pages 63–104. Cambridge University Press, 13^{th} edition.