# The Le*fff* 2 syntactic lexicon for French: architecture, acquisition, use

**Benoît Sagot[*], Lionel Clément[†], Éric Villemonte de La Clergerie[*] and Pierre Boullier[*]**

[*]INRIA, Projet Atoll
Domaine de Voluceau, Rocquencourt, B.P. 105
78153 Le Chesnay, France
{benoit.sagot, eric.de_la_clergerie, pierre.boullier}@inria.fr

[†]LaBRI, projet Signes
351, cours de la Libération
33405 Talence, France
lionel.clement@labri.fr

**Abstract**

In this paper, we introduce a new lexical resource for French which is freely available as the second version of the Le*fff* (Lexique des formes fléchies du français – Lexicon of French inflected forms). It is a wide-coverage morphosyntactic and syntactic lexicon, whose architecture relies on properties inheritance, which makes it more compact and more easily maintainable and allows to describe lexical entries independantly from the formalisms it is used for. For these two reasons, we define it as a *meta-lexicon*. We describe its architecture, several automatic or semi-automatic approaches we use to acquire, correct and/or enrich such a lexicon, as well as the way it is used both with an LFG parser and with a TAG parser based on a meta-grammar, so as to build two large-coverage parsers for French. The web site of the Le*fff* is `http://www.lefff.net/`.

## 1. Introduction

Precision and recall of a natural language processing chain is influenced not only by the grammar. Other components, including the pre-syntactic processing and the parser generator, play a major role. However, by its importance at all levels, the lexicon is particularly important.

But a large-coverage lexicon is a very rich and very large set of highly structured information. Moreover, it describes linguistic properties of linguistic items. Hence, a linguistically justified and operationally efficient structuration is required. Moreover, appropriate acquisition, supplementation and correction methods are needed, that have to be as automatic as possible.

In this paper, we introduce a new syntactic lexicon for French that satisfies these criteria. In particular, it relies on an original properties inheritance model. This lexicon, the Le*fff* 2 (*Lexique des formes fléchies du français*[1]), partly originates in the morphological lexicon of French verbs, the Le*fff* 1, whose automatic acquisition has been presented in (Clément et al., 2004).

## 2. Architecture

The Le*fff* 2 is described in an intentional way that allows for factorization of information, thanks to a hierarchical inheritance structure. The intentional lexicon is a lexicon of lemmas, whereas the extensional lexicon is a lexicon of inflected forms. In order to describe this architecture, we will describe the process of compilation from this intentional form to the extensional form used by parsers, summed up in Figure 1.

This compilation process can be divided into two steps: a morphological step and a syntactic step. Both steps start from the lemmas files that associate to each lemma a morphological class and a syntactic class. These files contain most of the lexical information that is stored in the Le*fff*.

The morphological step uses also a morphological description of French, which describes each morphological class, in order to inflect all lemmas. Moreover, a file of special inflected forms allows to add extra forms to lemmas when needed (orthographic variants, abbreviations, etc. . . ). Thus, the result of this morphological step is a set of 4-uples (the role of the morphosyntactic flag is described below): (*lemma*, *form*, *morphosyntactic tag*, *morphosyntactic flag*). Table 1 shows the information associated with lemma *boire* ("to drink") in the intensional lexicon.

| | | |
|---|---|---|
| `boire` | `v-re3` | `@verbe_standard` |

Table 1: Lexical entry for lemma *boire* ("to drink") in the intensional lexicon. Morphological class `v-re3` is the standard class for so-called third group verbs with an infinitive ending *-re*[2], and `@verbe_standard` is the syntactic class of transitive verbs with an optional direct object and possible pronominalization (although pronominalization phenomena are not yet treated with a satisfying level of detail).

The syntactic step works as follow. A first file describes a set of syntactic classes by an inheritance graph: each class is a disjunction of inherited classes or atomic properties. An example thereof, namely the syntactic class `@commencer`, is given in Table 2. Each atomic property, described in a second file, defines a part of the syntactic information represented by classes that inherit from this property. It can define a part of speech, give a lexical weight, or add some information to the syntactic structure itself (sub-categorization properties, realization properties of sub-categorized complements, phenomena like control or attributives, etc.). Morphosyntactic flags coming from the morphological step are special atomic properties. They encode the (small) part of the syntactic structure that depends from the morphosyntactic tag of a form, and not only from its lemma (e.g., the subject of a verb is mandatory, except at the infinitive form; most forms have the flag "De-

[1] Lexicon of French inflected forms

[2] We do not describe here our morphological formalism. The verb *boire* is usually considered as irregular. However, with appropriate collision rules to manage phenomena that happen at the boundary between the stem and the suffix, it is possible to fit *boire*'s inflection into the general class of *-re* third class verbs.

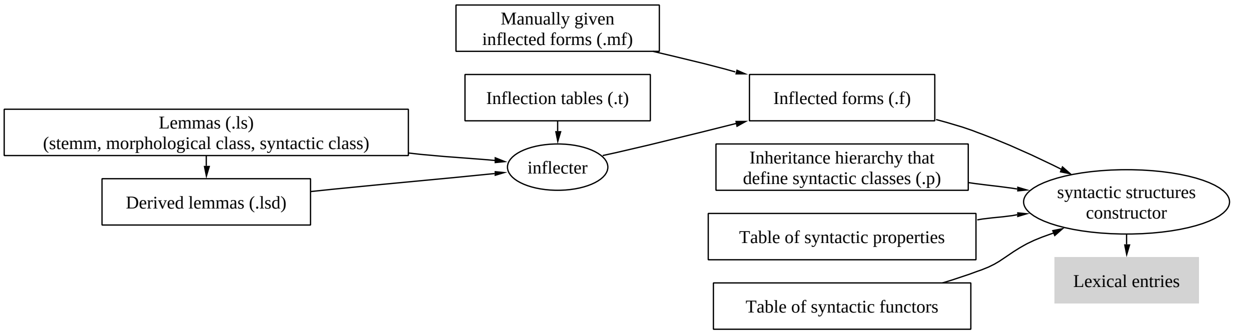
Manually given inflected forms (.mf)
Inflection tables (.t)
Inflected forms (.f)
Lemmas (.ls)
(stemm, morphological class, syntactic class)
inflecter
Derived lemmas (.lsd)
Inheritance hierarchy that define syntactic classes (.p)
syntactic structures constructor
Table of syntactic properties
Lexical entries
Table of syntactic functors

Figure 1: Overview of the compilation process that builds the extensional version of the Le*fff* from its intensional factorized version.

fault" that adds no tag-dependent atomic property). Each inflected form receives the syntactic structure obtained by combining all atomic properties it inherits from.

We recently added a mechanism to handle derivational morphology, thus taking advantage of the frequent parallelism that exists between morphology and syntax in this mechanism. Indeed, one can indicate in a lemmas file, below a given lemma, a certain amount of derivations. Each derivation can itself be followed by secondary derivations, and so on. A derivation is modeled by only one identifier, which denotes both:

- the morphologic mechanism that generates the derived lemma from the base lemma (these mechanisms are described in the morphological description that is used),
- a syntactic functor which defines the transformation that has to be applied to the syntactic structure of the base lemma in order to get the syntactic structure of the derived lemma (these syntactic functors are described in the same way than syntactic classes).

An example thereof is provided in Table 3.

## 3. Content

The Le*fff* 2 contains 404,483 inflected forms representing 625,720 entries, some of them being factorized (for example, the first and the third person of the present of *-er* verbs are grouped in one entry). This corresponds, among others, to 6,798 verbal lemmas, 37,673 nominal lemmas (excluding proper nouns), and 10,053 adjectival lemmas.

Our lexicon can be seen as a meta-lexicon, because the information it contains is stored in an inheritance graph and in a formalism-independent way. An entry consists of:

- an inflected form,
- sometimes a lexical weight, that allows to represent, for example, the fact that support verb constructions or idiosyncrasies should be preferred to normal constructions during parsing[3],
- a part of speech,
- a predicate (which can be seen as the *lemma* of several formalisms, or as the *pred* of LFG),
- a sub-categorization frame (represented in an LFG-like way that is easily convertible in other formats),
- a list of morphosyntactic and syntactic "macros", whose expansion can differ from one formalism to another, but whose semantics is formalism-independent (e.g., morphological tags, or macros such as `@CtrlSubj` or `@Impersonal`).

A few examples are shown in Table 4.

## 4. Acquisition, extension and correction

Building and maintaining a lexicon is a difficult task, both because of the number of entries needed to achieve a large coverage and because of the complexity of the information associated with each entry. Hence the need for automatic or semi-automatic techniques to acquire, extend and correct lexical information.

In the case of the Le*fff* 2, we used several different techniques:

- Automatic acquisition and extension of the morphological lexicon, according to the method described in (Clément et al., 2004; Sagot, 2005). This method has been especially used to automatically acquire the verbal part of the lexicon, including the previously cited Le*fff* 1, but also in order to include derivational information about deverbal derivatives.
- Automatic detection of unknown words in large corpora. This has been done thanks to the spelling error corrector SxSPELL described in (Sagot and Boullier, 2005), which helps to distinguish between unknown words and spelling errors.
- Automatic acquisition of multi-word units, according to techniques similar to those described in (Dias et al., 2001).
- Automatic detection of entries with erroneous or incomplete syntactic description, thanks to error mining in the results provided by Le*fff*-based deep parsers on

[3] As for now, these lexical weights are set manually, or by rule-based methods, e.g., for multi-word units that get a weight which is higher than the sum of their components'. We intend, in a near future, to extract automatically these weights from (manually or automatically) annotated corpora.

a large corpora ((Sagot and de La Clergerie, 2006), see also (van Noord, 2004)). The basic idea underlying this work is to analyze a large corpus (several million words) and study with statistical tools what differentiates sentences for which parsing succeeded from sentences for which it failed, and in particular which forms lead significantly more often than others to a parsing failure. This allows to identify automatically which forms are erroneously or only partially described in the lexicon.

- Automatic acquisition of atomic syntactic information, in particular about support verbs and prepositional phrases sub-categorized by verbs. This is performed thanks to a statistical analysis of form and tag patterns in very large tagged corpora.

## 5. Use

Our lexicon is used by at least two very different parsing systems for French. The first one is FRMG (Villemonte de la Clergerie, 2005), a TAG parser based on a meta-grammar that generates a factorized TAG. Le*fff* entries are used as hypertags to anchor quasi-trees.

The second one is SxLFG (Boullier and Sagot, 2005), an efficient LFG parser, which uses Le*fff* entries as LFG lexical entries. Both systems have been used recently with Le*fff* in several experiments, e.g., during the French parsers evaluation campaign EASy and for large-scale deep parsing experiments on multi-million word corpora (Sagot and Boullier, 2006).

It is very difficult to give an idea of the precision and the coverage of the Le*fff*, since the influence of the grammar is also extremely important in parsing precision and parsing coverage. However, we are developing a rule-based parser[4] that relies on the Le*fff*. Applied on the constituents boundaries and kind detection task of the EASy parsing evaluation campaign for French (Paroubek et al., 2005), it leads to an f-measure as high as 78.5%, which is very good (only one participant got a higher mark during the campaign).

## 6. Conclusion

We have introduced a new large-coverage syntactic lexicon for French, the Le*fff* 2, which is freely available on `http://www.lefff.net/`. It contains more than 500,000 entries, and has been successfully used in large-scale parsers using various linguistic formalisms. It is represented in a compact way, thanks to an graph of inheritance of atomic properties. Moreover, automatic and semi-automatic methods have been used to acquire, supplement and correct this lexicon. Some of these methods, as well as the overall architecture, could be used to develop similar lexicon for other languages, including languages for which no large-coverage lexicon is available.

## 7. References

Boullier, Pierre and Benoît Sagot, 2005. Efficient and robust LFG parsing: SxLFG. In *Proceedings of IWPT'05*. Vancouver, Canada.

Clément, Lionel, Benoît Sagot, and Bernard Lang, 2004. Morphology Based Automatic Acquisition of Large-coverage Lexica. In *Proceedings of LREC'04*. Lisbon, Portugal.

Dias, G., S. Guilloré, J.C. Bassano, and J.G.P. Lopes, 2001. Extraction automatique d'unités lexicales complexes: Un enjeu fondamental pour la recherche documentaire. *Traitement Automatique des Langues (T.A.L.)*, 41(2).

Paroubek, Patrick, Louis-Gabriel Pouillot, Isabelle Robba, and Anne Vilnat, 2005. EASy : campagne d'évaluation des analyseurs syntaxiques. In *Proceedings of the EASy workshop of TALN 2005*. Dourdan, France.

Sagot, Benoît, 2005. Automatic acquisition of a Slovak lexicon from a raw corpus. In *Lecture Notes in Artificial Intelligence 3658 (© Springer-Verlag), Proceedings of TSD'05*. Karlovy Vary, Czech Republic.

Sagot, Benoît and Pierre Boullier, 2005. From raw corpus to word lattices: robust pre-parsing processing. In *Proceedings of L&TC 2005*. Poznań, Poland.

Sagot, Benoît and Pierre Boullier, 2006. Deep non-probabilistic parsing of large corpora. In *Proceedings of LREC 06*. Genova, Italy. To be published.

Sagot, Benoît and Éric de La Clergerie, 2006. Trouver le coupable : fouille d'erreurs sur des sorties d'analyseurs syntaxiques. In *Proceedings of TALN 2006*. Louvain, Belgium. To appear.

van Noord, Gertjan, 2004. Error mining for wide-coverage grammar engineering. In *Proc. of ACL 2004*. Barcelona, Spain.

Villemonte de la Clergerie, Éric, 2005. From metagrammars to factorized TAG/TIG parsers. In *Proceedings of IWPT'05 (poster)*. Vancouver, Canada.

[4] This parser is based on a context-free grammar and non-probabilistic disambiguation heuristics.

| | |
|---|---|
| @commencer | |
| { | |
| < @verbe_standard_svp | *il commence son travail* |
| \| | |
| < @verbe_à_contrôle_sujet_à | *il commence à travailler* |
| < sujet_verbal_possible | |
| \| | |
| < @verbe_à_contrôle_sujet_de | *?il commence de travailler* |
| < sujet_verbal_possible | |
| \| | |
| < @verbe_impersonnel | *il commence à faire beau* |
| < @verbe_transitif_indirect_à | |
| < à-objet_infinitif_à_possible | |
| < à-objet_nominal_impossible | |
| } | |
| ; | |

Table 2: Description of the syntactic class `@commencer`, which is associated with lemmas *commencer* (“to begin”), *continuer* (“to go on”) and *recommencer* (“to start anew”) in the intensional lexicon.

| | |
|---|---|
| blanc adj-c @adj_couleur | *[base lemma]* |
| > adjectif_nominalisé | *(un) blanc* |
| > péjoratif-âtre | *blanchâtre* |
| > causatif_déadjectival-ir | *blanchir* |
| >> agentif-eur | *(le/la) blanchisseur/euse* |
| >> nom_d_action-age | *(le) blanchissage* |
| >> participe_présent_adjectivé | *blanchissant(e)(s)* |
| > nom_déadjectival-eur | *(la) blancheur* |

Table 3: Lexical entry for the adjectival lemma *blanc* (“white”) in the intensional lexicon, with some of its morphological derivatives.

| | | | |
|---|---|---|---|
| bois | | v | [pred=’boire___1<subj,(obj)>’,cat=v,@P12s] |
| bu | | v | [pred=’boire___1<subj,(obj)>’,cat=v,@active,@Kms] |
| bu | | v | [pred=’boire___1<(par-obj),subj>’,cat=v,@passive,@être,@Kms] |
| ... | | | |
| souhaite | | v | [pred=’souhaiter___1<subj,(obj\|scomp\|de-vcomp),(à-obj)>’, cat=v,@SCompSubj,@CtrlAObjDe,@PS13s] |
| souhaite | | v | [pred=’souhaiter___1<subj,(obj\|scomp\|vcomp)>’, cat=v, @SCompSubj,@CtrlSubj,@PS13s] |
| ... | | | |
| passer | | v | [pred="passer<(subj\|ssubj\|vsubj),(obj),(à-obj)>", cat=v, @W]; |
| passer | | v | [pred="passer<(subj\|ssubj\|vsubj),acomp>", cat=v, @W, @AASubj]; |
| passer | | v | [pred="passer<(subj\|ssubj\|vsubj),pour-acomp>", cat=v, @W, @AAPourSubj]; |
| passer | | v | [pred="passerSe<(subj),(de-obj)>obj", cat=v,@pron, @W]; |
| passer | | v | [pred="passerSe<(subj),de-vcomp>obj", cat=v,@pron, @W, @CtrlSubjDe]; |
| ... | | | |
| petit_à_petit | 500 | adv | [pred="petit-à-petit", cat=adv]; |

Table 4: A few lexical entries for inflected forms of the lemmas *boire* (“to drink”) *souhaiter* (“to wish”) in the extensional lexicon. As can be seen, the active and the passive past participle are distinguished, since they differ, among others, by their subcategorization frames. The `@SCompSubj` macro tells that the `scomp`, if present, must be at the subjunctive mood. The `@CtrlSubj` indicates a subject control verb, whereas `@CtrlAObjDe` indicates that the `à-obj` is the subject of the `de-vcomp`, if present. Default lexical weight (when not indicated) is 100.