The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French

Benoît Sagot

The Le*fff*, a freely available and large-coverage morphological and syntactic lexicon for French

Benoît Sagot

Alpage, INRIA Paris-Rocquencourt & Université Paris 7
Domaine de Voluceau — Rocquencourt, BP 105
78153 Le Chesnay Cedex, France
benoit.sagot@inria.fr

Abstract

In this paper, we introduce the Le*fff*, a freely available, accurate and large-coverage morphological and syntactic lexicon for French, used in many NLP tools such as large-coverage parsers. We first describe Alexina, the lexical framework in which the Le*fff* is developed as well as the linguistic notions and formalisms it is based on. Next, we describe the various sources of lexical data we used for building the Le*fff*, in particular semi-automatic lexical development techniques and conversion and merging of existing resources. Finally, we illustrate the coverage and precision of the resource by comparing it with other resources and by assessing its impact in various NLP tools.

1. Introduction

Many Natural Language Processing (NLP) tools require or benefit from reliable linguistic resources, such as lexicons and grammars. In particular, for tasks such as parsing, a morphological and syntactic lexicon is a highly valuable source of information. However, such a lexicon needs (1) to have a large coverage, (2) to guarantee a high level of quality, (3) to be directly usable in NLP tools, and (4) to be available to all its potential users. Such resources now exist for English, but are often lacking or incomplete for other languages, even major ones. For example, for French, several lexical resources exist that contain syntactic information, such as Lexicon-Grammar tables (Gross, 1975), Dicovalence (van den Eynde and Mertens, 2006) or Les Verbes Français (Dubois and Dubois-Charlier, 1997), but none of them combines satisfactorily the four above-mentioned properties.

These properties are the basis of our lexical development work. In this paper, we introduce both our lexical formalism, named Alexina, and the most advanced lexical resource developed within this framework, the Le*fff* (Lexique des Formes Fléchies du Français — *Lexicon of French inflected forms*), now in its third version (3.0.1). The Le*fff* is a widely-used and freely available[1] large-coverage morphological and syntactic lexicon for French. Apart from the Le*fff*, other Alexina lexicons are being developed, in particular the Le*ff*e for Spanish (Molinero et al., 2009), and resources for Galician, Polish (Sagot, 2007), Slovak (Sagot, 2005), Persian (Sagot and Walther, 2010), Sorani Kurdish (Walther and Sagot, 2010) and soon English.[2]

[1]The Le*fff* is distributed under the LGPL-LR license. See http://alpage.inria.fr/~sagot/lefff.html or the web page of the Alexina project: https://gforge.inria.fr/projects/alexina.

[2]Moreover, other freely redistributable lexicons have been converted into Alexina morphological lexicons and are accessible on the web page of the Alexina project. This includes the Morph-it! lexicon for Italian (Zanchetta and Baroni, 2005) and the Dutch lexicon distributed with the Alpino parser (van Noord, 2007). Since this latter lexicon also contains syntactic information, a full conversion would lead to a morphological *and* syntactic Alexina lexicon for Dutch.

2. The Alexina framework

Alexina is a lexical modeling and acquisition framework that covers both the morphological and syntactic levels. Alexina allows to represent lexical information in a complete, efficient and readable way, that is meant to be independent of the language and of any grammatical formalism (Sagot, 2005; Sagot et al., 2006; Sagot, 2007). Moreover, it is compatible with the LMF[3] standard (Francopoulo et al., 2006). Therefore, an Alexina lexicon can be directly used in NLP tools. Taking parsers as an example, we are aware of Alexina lexicons, and most notably the Le*fff*, being used in parsers based on LTAG, including LTAGs generated from meta-grammars developed in various meta-grammar formalisms (Thomasset and de la Clergerie, 2005), LFG (Boullier and Sagot, 2005), Interaction Grammars (Guillaume and Perrier, 2010), Pre-Group Grammars (Béchet and Foret, 2009), and other less known formalisms.

An Alexina lexicon consists of lexical entries that correspond to *lexemes*, i.e., to a meaning of a lemma that exhibits consistent morphological and syntactic properties. The morphological information in an Alexina lexicon has a simple and standard structure. Each entry is associated with a *lemma*, a *category* (or part-of-speech) and an *inflection class*. The *morphological description* defines how to build all *wordforms* (*inflected forms*, or simply *forms*) from a given lemma depending on its inflection class and associate with each inflected form a *morphological tag* (e.g., ms for masculine singular) and a *morphosyntactic flag* (see below). The morphological formalism used in Alexina for defining inflection classes is described in Section 2.2..

The syntactic level of the Alexina model deserves a more detailed description. Each phrase or pronoun whose existence, type, morphosyntactic properties and distribution is controlled by a given form in a given sentence is considered as the *realization* of a *(syntactic)*

[3]Lexical Markup Framework, the ISO/TC37 standard for NLP lexicons.

argument of this form. In languages such as French or English, such a *predicative* form may be among others a verb, an adjective or a noun.
In most cases, a syntactic argument corresponds to a *semantic argument* of the form, i.e., a participant to the process expressed by this form. In this case the syntactic argument can also be called an *actant* of the form. However, a syntactic argument may have no semantic counterpart, and is then called a *pseudo-argument*. This is for example the case of the *se/s'* in French pronominal verbs such as *s'évanouir* (*to faint*). Conversely, it may be impossible to provide a syntactic counterpart to a semantic argument. This if the case in French in so-called *se-moyen* constructions (Abeillé, 2002, p. 193).
The set of syntactic arguments of a given form and the associated constraints is modeled by the means of a *sub-categorization frame*. It encodes the set of arguments of the form, as well as additional syntactic properties (e.g., control, various constraints on the arguments, etc.).
A sub-categorization frame associated with a predicative form is defined as a list of syntactic arguments of this form; each of them is assigned a *syntactic function* with respect to the predicative form, i.e., a consistent set of morphological and syntactic constraints, as well as the set of its possible *realizations*; pseudo-arguments are also included in the sub-categorization frame, with no associated syntactic function. The notion of syntactic function (or grammatical function) is widespread across formalisms and approaches (Tesnière, 1959; Kaplan and Bresnan, 1982; Perlmutter and Postal, 1983). We define them on a per-language basis by the mean of several syntactic criteria that can be sketched as follows:

- the *commutation principle*, taking into account both pronouns and phrases, contrarily to (van den Eynde and Mertens, 2003): if a pronoun or a phrase can be replaced at the same position or another by another pronoun or phrase (both pronouns or phrases being mutually exclusive), without changing the dependency structure underlying the sentence, then they occupy the same syntactic function;

- the *unique realization principle*: for a given predicate, a syntactic function is realized at most once.[4]

With such criteria, the *linking* between semantic arguments and syntactic functions is not necessarily unique. Several sub-categorization frames may be found among the various forms of a given lexeme, for at least two reasons. First, there are form-dependant specificities in the syntactic behavior (e.g., in French, the infinitive form of a verb may have a non-realized subject). Second, and more importantly, a same inflected form of a given lexeme may have its semantic arguments linked to (final) syntactic functions in many ways: this is the well-studied phenomenon of regular syntactic alternations (e.g., active, passive or impersonal for French or English verbs). Therefore, inspired by previous work (Perlmutter and Postal, 1983; Candito, 1999), we define *initial syntactic functions* as follows:[5]

- for most lexemes, there exists a non-marked mapping between their semantic arguments and syntactic functions that leads to syntactically non-marked constructions;

- initial syntactic functions are defined as identical to final syntactic functions for the non-marked case;

- therefore, the set of initial syntactic functions is the same as the set of final syntactic functions;

- marked mappings (e.g., passive for French or English verbs) or mappings that lead to syntactically marked constructions (e.g., impersonal constructions for French or English verbs) are defined as *redistributions* of the set of initial syntactic functions;

- each redistribution must be defined formally on a per-language basis; a redistribution may assign a syntactic argument to a different syntactic function than its initial one, affect the list of realizations and/or change some of its properties (optionality, control, etc.).

Given the correspondence between the lemma and the inflected forms of a lexeme, as well as the correspondence between an initial sub-categorization frame of a lexeme and the various final sub-categorization frames its inflected form may receive, the Alexina model is based on a two-level representation that separates the description of a lexicon from its use:

- The lexicon proper, or *intensional lexicon*, factorizes the lexical information: each entry corresponds to a lexeme and provides its lemma, morphological class and initial syntactic information; it is used for lexical resource development;

- The *extensional lexicon*, which is generated automatically by *compiling* the intensional lexicon, associates each inflected form of a given entry with a detailed structure that represents its morphological information

[4]Note that with such a principle, the clitic subject inversion in French needs an appropriate treatment. Indeed, in sentences such as *Ainsi Pierre viendra-t-il demain* (*Thus Pierre will come tomorrow*), one could argue that both *Pierre* and *il* should receive the syntactic function *subject*, which is apparently incompatible with the unique realization principle.

[5]The notion of initial syntactic function is not standard. It is an alternative to the more widespread notion of *lexical rule* (see for example (Kaplan and Bresnan, 1982)). Lexical rules can be applied iteratively, starting with a base entry, and successively generating derived entries. The difference with the initial vs. final syntactic functions approach is that lexical rules can be applied on a derived entry, the result being itself a derived entry on which another lexical rule may be applied, and so on. On the opposite, a redistribution is a mapping between initial and final syntactic functions that applies to an initial sub-categorization frame, and the resulting final sub-categorization frame cannot undergo another redistribution. This allows to avoid termination issues, and does not prevent from defining a redistribution as, e.g., the sequence of two redistributions (see Section 2.3.).

and (some of) its possible syntactic behaviours; it is directly used by NLP tools such as parsers.[6]

The set of syntactic functions, the set of possible realizations and the set of redistributions defined for French and used in the L*efff* are described in Section 2.1.. Sections 2.2. and 2.3. respectively describe of the formalisms used in Alexina for defining inflection classes and redistributions. Finally, Section 2.4. and 2.5. define and illustrate respectively the format of the intensional and extensional lexicons.

2.1. Syntactic functions, realizations and redistributions in the L*efff*

For verbs, the L*efff* uses the following syntactic functions (defined here in a simplified way):

- Suj for subjects: cliticization with the nominative clitic;
- Obj for direct objects: cliticization with the accusative clitic, commutable with *ceci/cela* (*this/that*), impacted by passivization when it is possible;
- Objà for indirect objects canonically introduced by the preposition *à*: commutable with *à+non-clitic pronoun* (in the sense of (van den Eynde and Mertens, 2006)) but not with *ici* (*here*) or *là(-bas)* (*there*), may be cliticizable into the dative clitic or *y*;
- Objde for indirect objects introduced by the preposition *de*: cliticization with *en*, not commutable with *d'ici* (*from here*) or *de là* (*from there*),
- Loc for locative arguments: commutable with *ici* (*here*) or *là(-bas)* (*there*), cliticizable with *y*;
- Dloc for delocative arguments: commutable with *d'ici* (*from here*) or *de là* (*from there*), cliticizable with *en*;
- Att for (subject, object or *à*-object) attributes and pseudo-objects (e.g., *3 euros* in *j'ai acheté ceci 3 euros — I bought this 3 euros*),
- Obl and Obl2 for other (non-cliticizable) arguments;Obl2 is used for verbs with two oblique arguments, such as *plaider auprès de quelqu'un en faveur de quelqu'un d'autre* (*to plead in front of somebody for somebody else*).

For predicative adjectives and nouns, that can be headed respectively by a copula or a support verb, the same set of functions are used. The argument of a preposition is considered as an Obj. Adverbs may have arguments with the syntactic function Objà (*contrairement*) or Objde (*indépendamment*).
Possible realizations are threefold:

- clitic pronouns: cln (nominative clitic), cla (accusative clitic), cld (dative clitic), y, en, seréfl (reflexive *se*), seréc (reciprocal *se*);
- direct phrases: sn (noun phrase), sa (adjectival phrase), sinf (infinitive clause), scompl (completive clause), qcompl (interrogative clause);
- prepositional phrases: a direct phrase introduced by a preposition (e.g., à-sn, de-scompl,[7] pour-sinf).

For verbs, the inventory of possible redistributions is the following:

- %actif, a dummy "redistribution" that has almost no effect on the initial sub-categorization information;[8]
- %passif for the standard passive in *par*;
- %passif_de for the passive in *de* (*"Pierre est aimé de Marie"*/*"Pierre is loved by Mary"*);
- %impersonnel for (active) impersonal constructions with inverted subject, if any;
- %passif_impersonnel for passive impersonal constructions with inverted subject, if any;
- %se_moyen for modelling constructions such as *"ce livre se vend bien"*/*"this book sells good"* on the basis of the underlying transitive construction for the same verb;[9]
- %se_moyen_impersonnel, the impersonal counterpart of the previous redistribution (see *il se vend beaucoup de livres ici*/*there are many books sold here*);

For adjectives, we have defined two redistributions:

- %adj_impersonnel when an adjective is the lexical head of an impersonal construction (see *il est difficile de travailler*/*it is hard to work*);
- %adj_personnel for other cases.

For now, all other categories only use the %default construction that builds a final sub-categorization frame which is identical to the initial one.

[6]Note that in our approach, redistributions are computed during the compilation process of the lexicon. Therefore, taking parsing as an example, there is no need for on-the-fly transformations (e.g., raising), but at the same time the same lexeme will generate both an active and passive past participle (see below).

[7]de-scompl and à-scompl are exceptions, in so far that they do not correspond to *à* or *de* followed by a clause, but to *à ce* or *de ce* followed by a clause ("*Pierre se souvient de *(ce) que Marie est belle*"/"Pierre remembers that Marie is beautiful").

[8]It corresponds to the non-marked case, but is not assigned to all verbs. For example, some meteorological verbs have only the impersonal redistribution.

[9]Neutral constructions, with or without *se*, are considered as corresponding to a different lexeme, and therefore a different although semantically related lexical entry (Danlos and Sagot, 2008).

2.2. The morphological formalism

In the Alexina formalism, inflection is modelled as the affixation of a prefix and a suffix around a stem, while *sandhi* phenomena may occur at morpheme boundaries (see below), sometimes conditioned by stem properties. The formalism, which shares some widespread ideas with the DATR formalism (Evans and Gazdar, 1990) , relies on the following scheme:

- The core of a morphological description is a set of inflection tables that define corresponding inflection classes; inflection tables can (partly or completely) inherit from one another,
- Each inflection table defines a set of forms, each one of them being defined by a morphological tag and by a prefix and a suffix that, together with the stem, constitute the sequence of morpheme-like units *prefix_stem_suffix*;
- *Sandhi* phenomena allow to link the inflected form to the underlying *prefix_stem* and *stem_suffix* sequences by applying regular transformations; such rules may use classes of characters (e.g., [:aou:] can be defined as denoting one of the characters *a*, *o* or *u* with or without diacritics, as illustrated in Table 1);
- Forms can be controlled by tests over the stem (e.g., a given rule can apply only if a given regular expression matches the stem and/or if another one does not match the stem);
- Forms can be controlled by "variants" of the inflection classes (e.g., forms can be selected by one or more flags which complement the name of the class).

Tables 1 and 2 illustrate this model by showing respectively a few *sandhi* rules and an excerpt of a verbal inflection class.
Within the Alexina architecture, a morphological description using this formalism can generate two tools:

- an inflection tool that generates all inflected forms of a given lemma according to its morphological class;
- an ambiguous lemmatization tool, that computes for a given form (associated or not with a category) all possible candidate lemmas (existing or not) that are consistent with the morphological description and have this form among their inflected forms.

```
<letterclass name="aou" letters="a à â ä o ô ö u û ü ù"/>
<letterclass name="ou" letters="o ô ö u û ü ù"/>

<sandhi source="g_[:aou:]" target="ge_[:aou:]"/>
<sandhi source="[:ou:]y_es$" target="[:ou:]i_es$"/>
<sandhi source="et_2e$" target="ett_e$"/>
```

Table 1: A letter class definition and three *sandhi* rules from our Alexina description of French morphology (the "_" models a morpheme boundary; the first *sandhi* associates for example *mangeons* with *mang_ons*, the second associates for example *broient* with *broy_ent*, the third associates for example *jette* with *jet_2e*)

```
<table name="v-er" canonical_tag="W" stems="..*">
    <form suffix="er" tag="W" synt="Infinitive"/>
    <alt>
        <form suffix="2e" tag="PS13s"
              var="dbl" synt="ThirdSing"/>
        <form suffix="e" tag="PS13s" rads="..*ay"
              var="std" synt="ThirdSing"/>
        <form suffix="e" tag="PS13s"
              var="std" synt="ThirdSing"/>
    </alt>
...
    <form suffix="a" tag="J3s" synt="ThirdSing"/>
    <form suffix="ai" tag="J1s"/>
...
```

Table 2: Excerpts of the inflection class for French regular first group verbs in the Alexina morphological description used by the Le*fff*. The attributes **tag** and **synt** respectively define the morphological tag and the morphosyntactic flag. The attribute **var**, if present, indicates that the form must be generated only if the inflection class of the input lemma has the corresponding variant (e.g., **v-er:std**). The attributes **rads** (resp. **except**) indicates that the form must be generated only if the stem matches (resp. does not match) the corresponding pattern. Alternatives are represented with **alt** tags; within an alternative, at least one form must be generated, otherwise an error occurs.

2.3. The redistribution formalism

For a given language, redistributions are defined formally as a sequence of *elementary transformations* to be applied on the initial sub-categorization frame in order to produce the final sub-categorization frame. More precisely, each inflected form generated for an intensional entry tries to combine itself with each redistribution associated with this intensional entry. In some cases, it may lead to an incompatibility (e.g., in the Le*fff*, non-third person singular for the %impersonal redistribution). One of the ways to control this is by the means of the morphosyntactic flag associated with each inflected form (e.g., all past participle forms in the Le*fff* receive the morphosyntactic flag PastParticiple, which is required by the %passive redistribution in order to apply).
Each elementary transformation belongs to one of the following categories

- it may control the compatibility of the inflected form with the redistribution ({Only *morphosyntacticFlag*} or {Skip *morphosyntacticFlag*});
- it may add or remove a realization from the realization list of a syntactic function[10] (e.g., {Obj -cla} removes the cla realization — accusative clitic — from the list of realizations of the object syntactic function);

[10] Apart from syntactic functions, the special 0 may be used as a syntactic function, in order to add elements that do not realize any syntactic function (e.g., in French, the impersonal pronoun *il* for the %impersonal redistribution.

- it may change the optionality status of the surface realization of a syntactic function (e.g., {Obj ()} makes optional the surface realization of the object);
- it may build a list of realizations for a syntactic function (replacing the previous one if any) from an existing list, namely that of the same syntactic function or another one (e.g., the realizations of the subject for the %passive redistribution are built by the elementary transformation {Suj <Obj[cla>cln,de-sinf>sinf,seréfl>,seréc>]});
- it may add a new piece of information in the additional information section (e.g., {Macros @être} adds a macro which means the auxiliary must be *être*, which is the case for example for the passive past participle);
- it may simply replace a regular pattern by another in the additional information section (e.g., {@CtrlObjObjà @CtrlSujObjà} turns a macro expressing the fact that the control of the object on the *à*-object must be transformed into a control of the subject on the *à*-object, which is necessary for the %passive redistibution);

Each of these elementary transformations can be controlled by a morphosyntactic flag (e.g., Infinitive:{Suj ()} makes optional the subject only for the inflected form that has the Infinitive flag). Moreover, each elementary transformation may be mandatory (if it is not applicable, it means that the inflected form and the redistribution are incompatible) or optional (it is then prefixed by ?). Finally, any redistribution can be used as an elementary transformation (e.g., %passive_impersonal is simply defined as %passif + %impersonnel).
An example of redistribution definition in the L*efff* is shown in Table 3.

```
%se_moyen =
    %active
    + ?{@avoir } # removes the @avoir macro
    + {Macros @être}
    + {Macros @se_moyen}
    + {0 se}
    + {Suj <Obj[cla>,de-sinf>sinf,seréfl>,seréc>]}
    + {Suj )(} # if the Obj was optional, the Suj is as well,
                # and must therefore be made mandatory
    + ?{@AttSuj } + ?{@AttObj @AttSuj}
    + ?{@Ctrl.* } + ?{@Comp.* }
```

Table 3: The formal definition of the %se_moyen redistribution in the L*efff*. This redistribution models sentences such as *"ce livre se vend bien"* (*"this book sells good"*).

2.4. The Intensional format

Each entry in the intensional lexicon corresponds to a unique meaning of the corresponding lemma. It contains the following information:

- a *morphological class*, which defines the patterns that build its inflected forms (see Section 2.2.);
- a *category* (or part-of-speech); categories can be divided in two types: open (productive) categories (adjectives, adverbs, verbs, nouns) and closed (grammatical) categories;
- the initial sub-categorization frame;
- additional syntactic information (e.g., control, raising, attributes) represented by *macros* (e.g., @CtrlSujObj indicates that if it is realized as an infinitive phrase, the object is controlled by the subject);
- the list of possible redistributions.

For example, the intensional entry (slightly simplified for clarity reasons) in the L*efff* for the French lemma *diagnostiquer*$_1$*/to diagnose* is as follows:

diagnostiquer$_1$
v-er:std Lemma;v;
<arg$_0$:Suj:cln|sn,arg$_1$:Obj:(cla|sn)>;
%actif,%passif,%se_moyen

It describes a transitive entry with the following information:

- its morphological class is v-er:std, the class of standard first-conjugation verbs (ending *-er*);
- its semantic predicate can be represented by the Lemma as is, i.e., *diagnostiquer*;
- its category is *verb* (v);
- it has two arguments canonically realized by the syntactic functions Suj (subject) and Obj (direct object); each syntactic function is associated with a list of possible realizations, but the Obj is optional as shown by the brackets;
- it allows for three different redistributions: %active, %passive, and %se_moyen.

2.5. The Extensional format

The compilation process builds one extensional entry for each inflected form and each compatible redistribution, by inflecting the lemma according to the definition of its morphological class and by applying the formalized definitions of these redistributions. For example, the only inflected forms of *diagnostiquer* that are compatible with the passive redistribution are the past participle forms. The (simplified) extensional passive entry for *diagnostiqués/diagnosed* is the following (Kms is the morphological tag for past participle masculine plural forms):

diagnostiqués v
[pred='diagnostiquer$_1$<arg$_1$:Suj:cln|sn,
arg$_0$:Obl2:(par-sn)>',@passive,@pers,@Kms];
%passive

The original direct object (Obj) has been transformed into the passive Subject and an optional Agent (Obl2) realized by a noun phrase preceded by a preposition (*par-sn*) was added.

3. Lexical data

3.1. Sources of lexical information

Lexical information included in the Le*fff* originate in different works:

- automatic acquisition (with manual validation) thanks to statistical techniques applied on raw corpora (Clément et al., 2004; Sagot, 2005);
- automatic acquisition (with manual validation) of specific syntactic information (Sagot, 2006, ch. 7);
- manual correction and extension guided by automatic techniques, such as simple statistics on tagged corpora (Molinero et al., 2009) or error mining in parsing results (Sagot and de La Clergerie, 2006);
- careful linguistic study of some phenomena and their representation in other resources, conversion of (part of) these resources in the Alexina format, and manually validated automatic merging with the Le*fff*; we mainly used Lexicon-Grammar Tables (Gross, 1975), Dicovalence (van den Eynde and Mertens, 2006) and the Lexique des Verbes Français (Dubois and Dubois-Charlier, 1997). This was applied among other to impersonal constructions (Sagot and Danlos, 2008), pronominal constructions (Danlos and Sagot, 2008), adverbs in *-ment* (Sagot and Fort, 2007), several classes of frozen verbal expressions (Danlos et al., 2006), verbs in *-iser* and *-ifier* (Sagot and Fort, 2009)
- finally, a certain amount of nominal and adjectival entries have their origin in the Multext morphological lexicon for French (Veronis, 1998).

3.2. Quantitative data

At the extensional level, the current version of the Le*fff* (3.0.1) contains 536,375 entries corresponding to 110,477 distinct lemmas covering all categories. Detailed figures are given in Table 4.

Category	intensional entries	distinct lemmas	extensional entries
verbs	7,107	6,825	361,817
verbal idioms	1,868	1,850	3,295
nouns	37,755	37,530	78,338
adjectives	10,504	10,483	34,096
adverbs	4,019	3,584	4,062
prepositions	226	225	655
proper nouns	52,482	52,185	52,552
other[11]	833	632	1,342

Table 4: Quantitative data about the Le*fff*

[11]This includes all kinds of conjunctions, determiners, interjections, punctuation marks, pronouns, prefixes and suffixes, as well as special entries for named entities and unknown words.

4. Evaluation

The evaluation of a lexical resource in itself is not easy, as no gold standard can be used. However, we performed three types of evaluation: (1) quantitative comparison with other resources, (2) comparative evaluation of NLP tools based on a lexicon, depending on whether they use the Le*fff* or no lexical resource, and (3) comparative evaluation of NLP tools based on a lexicon, depending on whether they use the Le*fff* or another lexical resource. The two latter types of evaluation are illustrated respectively on a part-of-speech tagger and on a deep parser.

4.1. Quantitative comparison with other resources

We provide in Table 5 direct comparison with other lexical resources in terms of morphological coverage (number of distinct lemmas). However, including more and more rare or archaic words in a lexicon may prove inappropriate, for it increases the size of the lexicon and the lexical ambiguity, with a very low improvement in the coverage of real texts.

Category	Le*fff*	Morphalou	Multext	Dicovalence
verbs	6,825	8,789	4,782	3,729
nouns	37,530	59,002[12]	18,495	0
adjectives	10,483	22,739	5,934	0
adverbs	3,584	1,579	1,044	0
preps	225	(51)	117	0

Table 5: Quantitative comparison of the amount of unique lemmas in various resources

4.2. Evaluating a POS tagger using the Le*fff* vs. no lexicon

In (Denis and Sagot, 2009), the authors compared the performance of a maximum-entropy-based part-of-speech tagger for French depending on the amount of lexical information extracted from the Le*fff* it relies on. This tagger, trained solely on the French TreeBank (Abeillé et al., 2003), exhibits a 97.0% accuracy (86.1% for words unknown to the training corpus). An additional coupling with the Le*fff* by adding Le*fff*-based features to the model increases this figure up to 97.7% (90.1% for words unknown to the training corpus), which is state-of-the-art for French.

The explanation for this significant improvements is that the Le*fff*-based features reduce data sparseness and provide useful information on the right context: first, fewer errors on words unknown to the training corpus (a direct result of the use of a morphosyntactic lexicon) necessarily leads to fewer erroneous contexts for other words, and therefore to better tagging; second, the possible categories of tokens that are on the right of the current tokens are valuable pieces of information, and they are available only from the lexicon.

In their study, (Denis and Sagot, 2009) also compare the effect of the use of increasingly large sub-parts of

[12]Note that in Morphalou, masculine and feminine variants of a noun are considered as two different lemmas, whereas in the Le*fff* they are forms of the same lemma (ex.: *fermier*/*fermière* — *farmer*.

the L*efff* in a way that approximately simulates the development of a morphological lexicon, by retaining only the most frequent lemmas (frequency figures come from a large journalistic corpus). The results show that using a morphological lexicon drastically improves the tagging accuracy on unknown words, whatever the development stage. Moreover, for fixed performance levels, the availability of the full lexicon consistently reduces the need for training data by at least one half (and up to two thirds).

4.3. Evaluating a deep parser using the L*efff* vs. another lexicon

Comparative experiments with the FRMG parser for French (Thomasset and de la Clergerie, 2005) have been described, depending on the lexicon used by FRMG. FRMG is based on Tree-Adjoining Grammars, and normally relies on the L*efff*. However, the lexical information included in Lexicon-Grammar verb tables, a high-quality and large-coverage resource (Gross, 1975), were converted into the Alexina framework (Tolone and Sagot, 2009). This allowed for integrating it within FRMG, by replacing verb entries in the L*efff*. Results show that FRMG performs slightly better with the original L*efff*: according to the metrics used by the EASy French parsing evaluation campaign (Paroubek et al., 2006), f-measures on "relations" (approx. dependencies between lexical words) drop from 59.9% to 56.6% when replacing L*efff* verb entries with entries extracted from Lexicon-Grammar tables.[13]

Several explanations can be given to explain these results. First, despite their fine-grainedness, Lexicon-Grammar tables lack some information, for instance on subject and object attributes and on pronominal redistributions. Second, the limit between arguments and modifiers tends to be different from that used in the L*efff*; the higher number of verb arguments listed in Lexicon-Grammar tables have some negative effects on the disambiguation heuristics used by FRMG. Moreover, the higher number of entries in Lexicon-Grammar leads to higher ambiguity levels, which in turn increases parsing times (and therefore the amount of sentences that can not be parsed before a fixed timeout) and affects the precision of disambiguation heuristics. Finally, the L*efff* has been developed from the very beginning for and in parallel with NLP applications, which is not the case for Lexicon-Grammar tables.

5. Conclusion and perspectives

Since its first versions (Sagot et al., 2006), the L*efff* has turned into a widely used morphological and syntactic lexical resource for French. Its lexical framework, Alexina, is used for developing L*efff*-like resources for several other languages. Moreover, the lexical data in the L*efff* is under continuous improvement thanks to various semi-automatic techniques.

The next step of the L*efff*'s development shall be twofold. First, we intend to carry on the improvement of the precision and coverage through linguistic studies and various semi-automatic techniques, with a strong manual validation effort. Second, we aim at extending the L*efff* to the semantic level, by coupling it with semantic resources such as the WOLF (Wordnet Libre du Français — *Free French Wordnet*) (Sagot and Fišer, 2008).

[13] Note that these figures cannot be directly compared with classical f-measures based on evalb metrics. FRMG is one of the best performing parsers for French, as proved during the EASy/Passage evaluation campaigns.

6. References

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

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

Denis Béchet and Annie Foret. 2009. PPQ : a pregroup parser using majority composition. In Timothy Fowler and Gerald Penn, editors, *Proceedings of the ESSLLI 2009 Workshop on Parsing with Categorial Grammars, 20-24 July 2009, Bordeaux, France the ESSLLI 2009 Workshop on Parsing with Categorial Grammars*, pages 33–37, Bordeaux, France.

Pierre Boullier and Benoît Sagot. 2005. Efficient and robust LFG parsing: SXLFG. In *Proceedings of IWPT 2005*, pages 1–10, Vancouver, Canada.

Marie-Hélène Candito. 1999. *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées*. Ph.D. thesis, Université Paris 7.

Lionel Clément, Benoît Sagot, and Bernard Lang. 2004. Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of LREC 2004*, pages 1841–1844, Lisbon, Portugal.

Laurence Danlos and Benoît Sagot. 2008. Constructions pronominales dans dicovalence et le lexique-grammaire – intégration dans le L*efff*. In *Proceedings of the 27th Lexis and Grammar Conference*, L'Aquila, Italy.

Laurence Danlos, Benoît Sagot, and Susanne Salmon-Alt. 2006. French frozen verbal expressions: from lexicon-grammar to NLP applications. In *Proceedings of the 25th Lexis and Grammar Conference*, Palermo, Italy.

Pascal Denis and Benoît Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of PACLIC 2009*, Hong Kong.

Jean Dubois and Françoise Dubois-Charlier. 1997. *Les verbes français*. Larousse-Bordas, Paris, France.

Roger Evans and Gerald Gazdar. 1990. The DATR Papers: February 1990. Technical Report CSRP 139, University of Sussex, Brighton, UK.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In *Proceedings of LREC 2006*, Genoa, Italy.

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

Bruno Guillaume and Guy Perrier. 2010. Interaction grammars. *Research on Language and Computation*. To appear.

Ronald Kaplan and Joan Bresnan. 1982. Lexical-functional grammar: a formal system for grammatical representation. In Joan Bresnan, editor, *The Mental*

Representation of Grammatical Relations, pages 173–281. MIT Press, Cambridge, MA, USA.

Miguel Ángel Molinero, Benoît Sagot, and Lionel Nicolas. 2009. A morphological and syntactic wide-coverage lexicon for spanish: The Le*ff*e. In *Proceedings of RANLP 2009*, Borovets, Bulgaria.

Patrick Paroubek, Isabelle Robba, Anne Vilnat, and Christelle Ayache. 2006. Data, Annotations and Measures in EASy, the Evaluation Campaign for Parsers of French. In *Proceedings of the 5th Language Resources and Evaluation Conference (LREC 2006)*, Genoa, Italy.

David Perlmutter and Paul Postal. 1983. *Studies in Relational Grammar 1*. University of Chicago Press, Chicago, IL, USA.

Benoît Sagot and Laurence Danlos. 2008. Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire – Constructions impersonnelles. *Cahiers du Cental*, 5:107–126.

Benoît Sagot and Éric de La Clergerie. 2006. Error mining in parsing results. In *Proceedings of ACL/COLING 2006*, pages 329–336, Sydney, Australia.

Benoît Sagot and Darja Fišer. 2008. Building a free french wordnet from multilingual resources. In *Proceedings of Ontolex 2008*, Marrakech, Morocco.

Benoît Sagot and Karën Fort. 2007. Améliorer un lexique syntaxique á l'aide des tables du lexique-grammaire – adverbes en *-ment*. In *Proceedings of the 26th Lexis and Grammar Conference*, Bonifacio, France.

Benoît Sagot and Karën Fort. 2009. Description et analyse des verbes désadjectivaux et dénominaux en *-ifier* et *-iser*. In *Proceedings of the 28th Lexis and Grammar Conference*, Bergen, Norway.

Benoît Sagot and Géraldine Walther. 2010. A morphological lexicon for the persian language. In *Proceedings of LREC 2010*, Valetta, Malta.

Benoît Sagot, Lionel Clément, Éric de La Clergerie, and Pierre Boullier. 2006. The Le*fff* 2 syntactic lexicon for French: architecture, acquisition, use. In *Proceedings of the 5th Language Resource and Evaluation Conference*, Lisbon, Portugal.

Benoît Sagot. 2005. Automatic acquisition of a Slovak lexicon from a raw corpus. In *LNAI 3658, Proceedings of TSD 2005*, pages 156–163, Karlovy Vary, Czech Republic.

Benoît Sagot. 2006. *Analyse automatique du français: lexiques, formalismes, analyseurs*. Ph.D. thesis, Université Paris 7.

Benoît Sagot. 2007. Building a morphosyntactic lexicon and a pre-syntactic processing chain for Polish. In *Proceedings of LTC 2005*, pages 423–427, Poznań, Poland.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris, France.

François Thomasset and Éric de la Clergerie. 2005. Comment obtenir plus des méta-grammaires. In *Proceedings of TALN 2005*, Dourdan, France.

Elsa Tolone and Benoît Sagot. 2009. Using lexicon-grammar tables for french verbs in a large-coverage parser. In *Proceedings of LTC 2009*, Poznań, Poland.

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

Karel van den Eynde and Piet Mertens. 2006. Le dictionnaire de valence DICOVALENCE : manuel d'utilisation. http://bach.arts.kuleuven.be/dicovalence/manuel_061117.pdf.

Gertjan van Noord. 2007. Using self-trained bilexical preferences to improve disambiguation accuracy. In *Proceedings of the Tenth International Conference on Parsing Technologies (IWPT 2007)*, pages 1–10, Prague, Czech Republic.

Jean Veronis. 1998. Multext-lexicons, a set of electronic lexicons for european languages. CD-ROM distributed by ELRA/ELDA.

Géraldine Walther and Benoît Sagot. 2010. Developing a large-scale lexicon for a less-resourced language: general methodology and preliminary experiments on sorani kurdish. In *Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop)*, Valetta, Malta.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! a free corpus-based morphological resource for the italian language. In *Proceedings of Corpus Linguistics 2005*, Birmingham, UK.