Building a morphological and syntactic lexicon by merging various linguistic resources

Miguel A. Molinero
Grupo LYS
University of A Coruña
A Coruña, Spain
mmolinero@udc.es

Benoît Sagot
Project ALPAGE
INRIA
Paris, France
benoit.sagot@inria.fr

Lionel Nicolas
Laboratoire I3S (Équipe RL)
Université de Nice-Sophia Antipolis
Sophia Antipolis, France
lnicolas@i3s.unice.fr

Abstract

This paper shows how large-coverage morphological and syntactic NLP lexicons can be developed by interpreting, converting to a common format and merging existing lexical resources. Applied on Spanish, this allowed us to build a morphological and syntactic lexicon, the Le*ff*e. It relies on the Alexina framework, originally developed together with the French lexicon Le*fff*. We describe how the input resources — two morphological and two syntactic lexicons — were converted into Alexina lexicons and merged. A preliminary evaluation shows that merging different sources of lexical information is indeed a good approach to improve the development speed, the coverage and the precision of linguistic resources.

1 Introduction

In the environment of Natural Language Processing (NLP), linguistic resources, such as lexicons and grammars, are required for many high-level applications. However, the current situation for most languages is that several scattered resources exist, with different coverage levels, different linguistic backgrounds and different lexical formalisms. Nevertheless, none of these resources combines in a satisfying way the following properties:

- coverage: all words, including rare ones, in all categories should be included;
- quality: manually and automatically developed resources contain various errors;
- richness: applications such as (deep) parsing require at least morphological and syntactic information, including subcategorization frames.

However, each existing resource for a given language is a provider of valuable lexical information. Merging these resources and expanding them thanks to semi-automatic techniques is therefore a promising idea. Anyhow, this requires to be able to interpret all input resources despite partly incompatible lexical models, to convert them into a common model and format, and then to merge these converted lexicons. None of these three steps is trivial. This approach has been successfully applied on French for developing the syntactic lexicon Le*fff* (*Lexique des formes fléchies du français*), within a lexicon development framework named Alexina (Sagot et al., 2006; Sagot and Danlos, 2008; Danlos and Sagot, 2008).

In this paper, we confirm the validity of this approach by applying it to Spanish, in order to build a wide-coverage morphological and syntactic lexicon for this language, the Le*ff*e (*Léxico de formas flexionadas del español*). Such a lexicon can be directly used in advanced NLP applications, particularly in those involving deep parsing. The Le*ff*e is developed within the same framework as the Le*fff*, the Alexina framework, and distributed under the same free license, the LGPL-LR.[1] The flexibility and completeness of the Alexina format allows for a straightforward integration with deep grammatical formalisms (LFG, LTAG) which require detailed syntactic data for all forms.

The work described in this paper is one of the starting points of the recently created Victoria project, which aims at developing

[1] Lesser General Public License for Linguistic Resources

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 126–133

techniques and tools for efficient acquisition and correction of large-coverage linguistic resources with inter-language links. The first phase of the project focuses on Spanish, Galician[2] and French.

This paper is organized as follows: first, in Section 2, we introduce the Alexina model. Section 3 describes the existing Spanish resources we used. Along Section 4 we show how these resources were merged, and in Section 5 we briefly evaluate the resulting lexicon. We present our conclusions and future work in Section 6.

2 Representing lexical information: the Alexina model

A detailed description of all words belonging to a language is needed in order to perform high-level NLP tasks such as deep parsing. This information is usually compiled into a lexicon, which could be defined as a list of words associated with their corresponding morphological and syntactic information. Alexina is a framework compatible with the LMF[3] standard, whose goal is to represent lexical information in a complete, efficient and readeable way (Sagot, 2005; Danlos and Sagot, 2008). The Alexina model allows to describe rich morphological and syntactic lexical information, which can be used in NLP tools relying on various grammatical formalisms.

Alexina is based on two representation levels:

- The intensional lexicon factorizes the lexical information by associating each lemma with a morphological class and deep syntactic information (a deep subcategorization frame, a list of possible restructurations, and other syntactic features such as information on control, attributes, mood of sentencial complements, etc.);

- The extensional lexicon, which is generated automatically by *compiling* the intensional lexicon, associates each inflected form with a detailed structure that represents all its morphological and syntactic information: morphological tag, surface subcategorization frame corresponding to one particular redistribution, and other syntactic features.

The intensional representation is used for an efficent description, while the extensional is directly used by NLP tools such as parsers.

The remainder of this section briefly describes the format of the intensional and extensional lexicons and the formalism used for describing the morphological and syntactic information within the Alexina model.

The first task achieved by the compilation process, which turns an intensional lexicon (an `.ilex` file) into an extensional lexicon (a `.lex` file), is to inflect lemmas according to their morphological class. Morphological classes are defined in a formalized morphological description (Sagot, 2005; Sagot, 2007). In case a lemma inflects in a very specific way, and/or if a lemma has additional inflected forms apart from those generated by its morphological class, these forms are "manually" listed in an additional file (the corresponding `.mf` file).

As sketched above, the compilation process also maps deep syntactic information into surface syntactic information. Deep syntactic information (deep subcategorization frames and other syntactic information) is common to all redistributions, whereas each redistribution corresponds to different surface syntactic information, and therefore to different extensional entries.

For example, here is the intensional entry in the Le*fff* for the French lemma *clarifier*$_1$ (i.e., *clarifier* in the sense of English *clarify*), slightly simplified:[4]

```
clarifier_1 v-er
            Lemma;v;
            <arg0:Suj:cln|scompl|sinf|sn,
            arg1:Obj:(cla|scompl|sn)>;
            %actif,%passif,%passif_impersonnel
```

It describes a transitive entry whose morphological class is `v-er`, the class of so-called first-group verbs. Its semantic predicate can be represented by the `Lemma` as is, i.e., *clarifier*. Its category is *verb*

[2] A co-official language in north-west Spain.

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

[4] In particular, additional syntactic features such as control information are not shown, for clarity reasons.

(v). It has two arguments canonically realized by the syntactic functions Suj (subject) and Obj (direct object).[5] Each syntactic function is associated with a list of possible realizations,[6] which are between brackets if it is faculative. This entry allows for three different redistributions: active (%actif), passive (%passif), impersonnal passive (%passif_impersonnel, *il a été clarifié (par Pierre) que Marie ne viendrait pas*, in English *it has been clarified (by Pierre) that Mary wouldn't come*).

The compilation process builds one extensional entry for each inflected form and each compatible redistribution, by applying formalized definitions of these redistributions (which can be found in file constructions). For example, the only inflected forms of *clarifier* that is compatible with the passive redistribution are the past participle forms. The (simplified) extensional passive entry for *clarifiés* is the following (Kmp is the morphological tag for past participle masculine plural forms):

```
clarifiés    v
[pred='clarifier1<arg1:Suj:cln|scompl|sn,
arg0:Obl2:(par-sn)>',@passive,@pers,@Kmp];
%passif
```

As said before, merging linguistic resources requires a careful interpretation of their underlying models, followed by their conversion into a common model that is able to preserve as much (valuable) information as possible. The Alexina model has been evolved over the last 5 years, alongside with the development of the Le*fff* and resources for other languages (Polish, Slovak, and others). The Le*fff* has been mostly developed by semi-automatic acquisition techniques and by merging lexical information extracted from other freely available resources. It has been used in different NLP tools including deep parsers for French based on various formalisms (LTAG, LFG, etc.). This all has allowed to develop Alexina in order to represent a great range of lexical phenomena. This fact, besides the linguistic proximity between French and Spanish as Romance languages, explains why Alexina already covers all lexical phenomena we encountered while working on Spanish, and no changes in the format were needed.

3 Existing lexical resources for Spanish

Several resources are available for Spanish. However, none of them fulfills all our requirements:

- Large coverage, good precision and satisfying richness (as explained in the introduction);
- Complete separation between lexical and grammatical information;
- Clear and compact format easily readable by humans;
- Freely available in terms of access, modification and distribution;
- Easily linkable with resources describing other languages;

Nevertheless, many valuable information can be found in these existing resources. The following ones were used at some point in the development of the Le*ff*e:

Multext is an international project (Ide and Véronis, 1994) whose goals are to develop standards and specifications for the encoding and processing of linguistic corpora, and to develop tools, corpora and linguistic resources embodying these standards. It includes morphological (but not syntactic) lexicons for several languages, including Spanish, that rely on a widely-used tagset;

The USC lexicon is a large morphological lexicon (Álvarez et al., 1998), created for PoS tagging tasks in the research group *Gramática del Español* of the University of Santiago de Compostela (Spain).

[5] The complete set of syntactic functions used in the Le*fff* and in the Le*ff*e is the following: Suj (subject), Obj (direct object that can be cliticized into an accusative clitic), Objde (indirect object canonically introduced by preposition *de* that can be cliticized into a genitive clitic), Objà or Obja (indirect object canonically introduced by *à* in French or *a* in Spanish), Loc (locative), Dloc (delocative), Att (attribute), Obl and Obl2 (oblique non-cliticizable arguments).

[6] Clitic realizations in French are cln, cla, cld, en and y for the nominative, accusative, dative, *en* (genitive) and *y* clitic pronouns. Direct realizations are sn, sinf, scompl, qcompl and sa for nominal, infinitive, phrasal, indirect interrogative and adjectival phrases. Prepositional realization are of the form *prep-real*, where *prep* is a preposition and *real* a direct realization.

ADESSE is a database for Spanish verbs developed at the University of Vigo (Spain) (García-Miguel and Albertuz, 2005) with syntactic and some semantic information. It is a high quality work which includes subcategorizarion frames for more than 4,000 verbs. However, it is restricted to verbs and includes no morphological information;

The Spanish Resource Grammar (SRG) is an open-source multi-purpose large-coverage and precise grammar for Spanish (Marimon et al., 2007). It is grounded in the theoretical framework of Head-driven Phrase Structure Grammar (HPSG) and includes a lexicon describing syntactic information for Spanish in a well organized hierarchy of syntactic classes. However, its is not easily readable, and specific to the HPSG formalism.

4 Converting and merging existing resources for building the Le*ff*e

The construction of the Le*ff*e has been sucessfuly achieved by interpreting all input resources mentioned above (despite their partially incompatible lexical models), converting them into the Alexina format, and finally merging the converted lexicons. As said in the previous section, the Multext and the USC lexicons only include morphological information, whereas the SRG and the ADESSE lexicons include syntactic information. Therefore, we decided to proceed in the following way:

1. Build a morphological baseline lexicon by converting the Multext lexicon into the Alexina format and adding some Alexina-specific entries (prefixes, suffixes, named entities, punctuation signs);

2. Converting the USC Lexicon into the Alexina format and merging it with the baseline lexicon extracted from Multext, so as to get the morphological basis of the Le*ff*e;

3. Converting the ADESSE and the SRG lexicon, which are syntactic-only, into the Alexina format;

4. Merging the morphological Le*ff*e from step 2 and both verbal syntactic lexicons built during step 3; the result is the current Le*ff*e, i.e., the Le*ff*e beta.

We shall now describe sucessively the four following tasks: converting a morphogical lexicon into the Alexina format (steps 1 and 2), converting the ADESSE and SRG syntactic lexicons into the Alexina format (step 3), merging morphologial lexicons (step 4) and merging syntactic lexicons (step 4).

4.1 Converting a morphological lexicon into the Alexina format

A morphological lexicon can be seen as a set of triples of the form *(form,lemma,tag)*. However, in an architecture such as Alexina, which aims at representing also syntactic information, each (intensional) entry corresponds to one lemma. As explained in Section 2, each lemma is associated with a morphological class, which is formally defined in a morphological description of the language. Therefore, in order to convert a morphological lexicon into the Alexina format, such a morphological description has to be extracted automatically from a set of *(form,lemma,tag)* triples.

We developed a fully-automatic technique for extracting morphological classes from such a set of triples. For each lemma, it extracts the longest prefix that is common to all its inflected forms, which is considered as the stem, and builds an ordered list of *(suffix,tag)* pairs.[7] If at least 3 lemmas lead to the same list of *(suffix,tag)* pairs, this list is turned into the definition of a morphological class, and all corresponding lemmas are associated with this class. Moreover, the stems of all these lemmas are analyzed, so as to build the most specific (reasonable) regular pattern that matches them all. This allows to prevent further lemmas to be added with an incompatible morphological class, but also to use the morphological description as an ambiguous lemmatizer with limited overgeneration. For example, while converting the Spanish Multext lexicon, a morphological class is built from a list of *(suffix,tag)* pairs that include the ending *-ar* for the infinitive, *-a* for the third person singular of the indicative present, and *-ué* for the first person singular

[7] At this point, the process discards all entries that do not have their lemma as one of their inflected forms.

of the indicative past. An example of such a verb is *halagar* (*to flatter*), which has the inflected forms *halaga* (*he flatters*) and *halagué* (*I flattered*). Because the stems of all lemmas in this class end in *-g*, the regular pattern `.*g` is associated to this morphological class.

Morphological classes that include only one or two lemmas are not built. Instead, the inflected forms of the corresponding lemmas are listed in the corresponding `.mf` file (see Section 2).

We applied this technique to build our baseline lexicon by converting the Spanish Multext lexicon into an Alexina lexicon, including a morphological description of Spanish. The same technique has also been applied to convert the USC lexicon into the Alexina format, which created a different morphological description, since the set of lemmas, the tagsets and sometimes the set of inflected forms for a given lemma are different from one lexicon to another. Section 4.3 explains how we merged these two morphological lexicons.

4.2 Converting the ADESSE and SRG lexicons into the Alexina format

Our most important source of syntactic information is the ADESSE lexicon, a database containing syntantic information for Spanish verbs. ADESSE is a carefully developed resource that includes much valuable information. We parsed and transformed it into the Alexina format as follows. Each verb in the ADESSE lexicon was transformed into one or more Le*ff*e entries with dummy morphological information, by converting ADESSE argument structures into Alexina subcategorization frames. The result is a lexicon with complete and reliable syntactic information for a significant number of Spanish verbs (3,427 unique verb lemmas).

Since some verb lemmas included in Multext or in the USC lexicon are not covered by the ADESSE lexicon and because cross validation is generally useful, we also extracted information from the SRG lexicon. However, we shall see that the technique we used is not fully reliable, and the SRG lexicon itself has a lower precision than the ADESSE lexicon. Thus, we gave a lower level of confidence to syntactic information extracted from SRG, as explained in Section 4.4.

The SRG classifies lemmas according to a hierarchy of syntactic classes. Mapping one class into the Le*ff*e format allows to extract as many entries as there are lemmas belonging to this class. We used the Le*fff* as bridge in order to establish a mapping between SRG syntactic classes and Alexina syntactic descriptions. The syntactic proximity between Spanish and French allows to retain Le*ffff* syntactic descriptions in the Spanish lexicon with very few modifications (almost only translating prepositions). The technique can be described as follows: [8]

1. First, a list of the most common verb classes in SRG were extracted;
2. A representative lemma of each of these classes was taken from SRG; this lemma must belong only to a single class in SRG and its translation into French should have the same syntactic behaviour than the Spanish one (something easy to fulfill thanks to the linguistic proximity between French and Spanish).
3. We look into the Le*fff* for the translation of these lemmas and extracted their associated syntactic information;
4. A link was created between the SRG class and the extracted Le*fff* syntactic description, manually adapted for becoming a Le*ff*e syntactic description[9];
5. Finally, we assigned to each SRG entry the corresponding Le*ff*e syntactic description.

Such a way to process could lead to some incomplete or erroneous entries. To restrict their impact, we decided to ignore extracted information in case of doubt.

Despite our efforts, it is possible that no syntactic information is found at all for some lemmas of our baseline lexicon. The opposite situation is very rare, that is, not to find morphological information, since it is

[8] Steps 1 and 5 were automatically acomplished, while steps 2, 3 and 4 were manually done for the 40 most frequent SRG classes, which covered more than 3,000 verbal lemmas.

[9] In practice, we needed only to translate prepositions.

much more commonly available and easier to acquire. So the very basic condition to acquire a word is to find its morphological information.

4.3 Merging morphological resources

Once in the Alexina format, a morphological lexicon can be seen as a set of *(lemma,class)* pairs, where *class* denotes the inflection class of the entry. Therefore, merging a main morphological lexicon L with an additional morphological lexicon L' consists in converting morphological classes of L' into morphological classes of L. This merging process is applied PoS by PoS, to avoid problems related to cross-PoS homonymy.

In order to achieve this mapping, we rely on lemmas that are common to both lexicons. Given a class from L', we extract from L' all corresponding lemmas that are also in L. Then we look for the classes of these lemmas in L. Usually, the large majority of the lemmas involved have the same class in L, but exceptions do occur. These exceptions correspond to mismatches between L and L', and therefore to errors in L and/or L'. They can be solved automatically by giving the priority to L (or L'), or checking them manually.

We applied this technique with L being the baseline lexicon extracted from Multext (so as to preserve the Multext tagset) and L' being the result of the conversion of the USC lexicon into the Alexina format. The result of this merging process is the morphological part of the Le*ff*e. Section 5 gives quantitative figures about it and compares it to other morphological lexicons.

4.4 Merging syntactic resources

Once the morphological part of the Le*ff*e is obtained, we must complete it with syntactic information. For verbs, this information is obtained by merging the Alexina version of the ADESSE and SRG lexicons, i.e., two intensional lexicons. For other categories, not covered by the ADESSE lexicon, we used the syntactic information extracted from the Alexina version of the SRG lexicon. Finally, some entries (prepositions, auxiliaries, a few very specific verbs) have been written or completed manually.

Contrarily to (Danlos and Sagot, 2008), our two input lexicons did not use the same criteria to distinguish between different entries of a same lemma. Therefore, we were not able to merge intensional entries. Rather, the merging process we used relies on the notion of *expanded intensional lexicon*. As seen above, an intensional entry includes a subcategorization frame in which each syntactic function may be facultatively realized and may have a list of realization alternatives. Such an intensional entry can be converted into a set of *expanded intensional entries*: each of these entries has a subcategorization frame that is *fully-specified* (no alternatives, no facultative argument), in such a way that all these entries, taken together, cover all cases covered by the original intensional entry. For example, an intensional entry with the subcategorization frame `<Suj:cln|sn,Obj:(sn)>` corresponds to 4 expanded intensional entries with the following subcategorization frames: `<Suj:sn>`, `<Suj:cln>`, `<Suj:sn,Obj:sn>` and `<Suj:cln,Obj:sn>`.

The idea is the following: we first expand both our input intensional lexicons (the Alexina versions of the ADESSE and SRG lexicons); then we merge these expanded intensional lexicons; finally, we re-factorize the merging result into an intensional lexicon. The expansion and merging steps are straightforward (here, merging is simply computing the union of all expanded entries). The refactorization step computes the optimal factorization of a list of (possibly expanded) intensional entries, and involve no particular linguistic knowledge.

The result is a syntactic-only lexicon, which is trivially merged with the morphological lexicon. For those morphological entries that were not covered by the syntactic-only lexicon, we decided to give them the syntactic features that were the most common among entries of the same PoS. This is obviously a baseline. For example, all verbal lemmas that are not covered by ADESSE and by SRG received the following subcategorization frame: `<Suj:sn|cln,Obj:(sn|cla)>` (transitive verb with facultative direct object). However, we rely on existing semi-automatic techniques for extending and correcting our lexicon in the near future (Nicolas et al., 2008).

5 Preliminary Evaluation

In order to evaluate the quality of Le*ff*e, currently in beta version, we performed the following tests: on the one hand, we have compared Le*ff*e with other known Spanish lexicons in terms of coverage; on the other hand, we measured the improvement achieved on the baseline lexicon after adding the information extracted from all other sources.

Regarding coverage, the Le*ff*e beta contains more than 165,000 unique *(lemma,PoS)* pairs, which correspond to approx. 1,590,000 extensional entries that associate a form with both morphological and syntactic information (approx. 680,000 unique *(form,PoS)* pairs). Other lexicons have the following properties:

- SRG: 76,000 unique *(lemma,PoS)* pairs[10] (53.9% less than Le*ff*e), but syntactic information is provided only for some of them;
- Multext: 510,710 unique *(form,PoS)* pairs[11] (24.9% less than Le*ff*e), and no syntactic information is provided;
- Spanish gilcUB-M Dictionary: 70,000 lemmas[11](57.6% less than Le*ff*e), and no syntactic information is provided;
- USC Lexicon: 490,000 unique *(form,PoS)* pairs (27.95% less than Le*ff*e), and no syntactic information is provided.

We have also tested the morphological coverage of our lexicon in the context of a real application: a morphological preprocessor (Graña et al., 2002; Barcala et al., 2007) developed by group COLE.[12] We performed a first test with our baseline lexicon, and a second one with the Le*ff*e beta.

We have used a corpus of raw text obtained from Wikipedia Sources[13] as an input for this test. It includes more than 4,322,000 words after clearing Wikipedia references and foreign expressions. The evaluation took into account how many words were not tagged by the preprocessor and thus remained unknown. It is worth noting that unknown words are an important cause of PoS-tagging errors. Such problems can be tackled by relying on (very) large coverage lexicons.

As can be observed in Table 1, the process allows noticeable benefits. The Le*ff*e beta has beaten other large lexicons in the morphological preprocessing task[14]. Even if the difference is slight, this demonstrates the interest of merging existing resources to create an enhanced one.

In order to measure the syntactic coverage of the lexicons at all stages of the merging process, we have used the notion of *expanded intensional entry* which describes one fully-specified syntactic behaviour (see Section 4.4). The expanded intensional lexicon acquired from SRG contains 42,689 unique entries, i.e., fully-specified subcategorization frames, while the one from ADESSE contains 39,040. After merging these lexicons, the number of such unique entries jumps to 66,028. Finally, the Le*ff*e beta, which associates default syntactic information with all verbs not covered by the result of this merge, contains 91,507 unique expanded entries. After factorization, the Le*ff*e contains 16,311 verbal entries.

6 Conclusion and future work

For many languages, several lexical resources exist, but usually none of them is satisfying in terms of coverage, richness (morphological and syntactic information is required) or precision.

In this work we have described a process to merge existing Spanish lexical resources into an enhanced one. From our point of view, this approach is nowadays the best way to produce quickly high-quality lexical resources. The theoretical and practical context described here can be used for a similar task in other languages. The resulting lexicon is a large-coverage morphological and syntactic lexicon, the Le*ff*e. This lexicon, currently in beta version, will be distributed under a LGPL-LR license[15] in the near future. Although it is still

[10] As provided by Freeling (`http://garraf.epsevg.upc.es/freeling/`) in a version from April 2008.

[11] According ELRA webpage `http://catalog.elra.info`, December 2008.

[12] http://www.grupocole.org

[13] `http://download.wikimedia.org`, January 2009

[14] It is worth noting that the distribution of entries in Multext seems not so natural, since despite being the largest in terms of number of entries, is the worse on this task. Indeed we checked that many common lemmas are missing in Multext.

[15] As explained in this paper, the construction of the Le*ff*e beta involved the Spanish morphological lexicon developed within the Multext project, which is freely

	Total unkown words	Unique unknown words
Multext	228,815	49,673
USC Lexicon	70,026	25,888
Baseline	86,521	27,234
Le*ff*e beta	69,756	24,703

Table 1: Results of applying the morphological preprocessor using different lexicons.

far from perfect, we have shown that the Le*ff*e beta has already overtaken other well known Spanish lexicons in terms of morphological and syntactic coverage.

In the near future, we plan to further evaluate the Le*ff*e as follows: we shall compare the coverage and precision of different deep parsers that rely on the same grammar but on different morphological and syntactic lexicons such as the Le*ff*e. Besides, we will continue improving Le*ff*e using techniques described here with other linguistic resources, and by applying automatic acquisition techniques as additional sources of lexical knowledge.

Acknowledgement Partially supported by Ministerio de Educación y Ciencia of Spain and FEDER (HUM2007-66607-C04-02), the Xunta de Galicia (INCITE08PXIB302179PR, INCITE08E1R104022ES, PGIDIT07SIN005206PR) and the "Galician Network for Language Processing and Information Retrieval" 2006-2009).

We would like also to thank group *Gramática del Español* from USC, and especially to Guillermo Rojo, M.ª Paula Santalla and Susana Sotelo, for granting us access to their lexicon.

References

Fco. Mario Barcala, Miguel A. Molinero, and Eva Domínguez. 2007. Xml rules for enclitic segmentation. *Lecture Notes in Computer Science: Computer Aided Systems Theory - EUROCAST 2007, Revised selected papers, pp. 273-281.*

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

José M. García-Miguel and Francisco J. Albertuz. 2005. Verbs, semantic classes and semantic roles in the adesse project. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes.*

Jorge Graña, Fco. Mario Barcala, and Jesús Vilares. 2002. Formal methods of tokenization for part-of-speech tagging. *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science.*

Nancy Ide and Jean Véronis. 1994. Multext: Multilingual text tools and corpora. In *Proceedings of COLING'94.*

Montserrat Marimon, Natalia Seghezzi, and Núria Bel. 2007. An open-source lexicon for spanish. In *Sociedad Española para el Procesamiento del Lenguaje Natural, n. 39.*

Lionel Nicolas, Benoît Sagot, Miguel A. Molinero, Jacques Farré, and Éric Villemonte de La Clergerie. 2008. Computer aided correction and extension of a syntactic wide-coverage lexicon. In *Proceedings of COLING'08.*

Benoît Sagot and Laurence Danlos. 2008. Méthodologie lexicographique de constitution d'un lexique syntaxique de référence pour le français. In *Proceedings of the workshop "Lexicographie et informatique : bilan et perspectives"*, Nancy, France.

Benoît Sagot, Lionel Clément, Éric Villemonte de La Clergerie, and Pierre Boullier. 2006. The Le*fff* 2 syntactic lexicon for French: architecture, acquisition, use. In *Proceedings of LREC'06.*

Benoît Sagot. 2005. Automatic acquisition of a Slovak lexicon from a raw corpus. In *Lecture Notes in Artificial Intelligence 3658 (© Springer-Verlag), Proceedings of TSD'05*, pages 156–163, Karlovy Vary, Czech Republic.

Benoît Sagot. 2007. Building a morphosyntactic lexicon and a pre-syntactic processing chain for Polish. In *Proceedings of the 3rd Language & Technology Conference (LTC'05)*, pages 423–427, Poznań, Poland, October.

Concepción Álvarez, Pilar Alvariño, Adelaida Gil, Teresa Romero, María Paula Santalla, and Susana Sotelo. 1998. Avalon, una gramática formal basada en corpus. In *Procesamiento del Lenguaje Natural (Actas del XIV CONGRESO de la SEPLN)*, pages 132–139, Alicante, Spain.

available for research. The Le*ff*e beta is the result of the research work described here. It merges lexical information coming from various resources, most of them with a coverage that is larger than the Spanish Multext lexicon. For this reason, we consider as appropriate to publish the Le*ff*e beta under the LGPL.

ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206