A morphological and syntactic wide-coverage lexicon for Spanish: The Leffe

Miguel Molinero, Benoît Sagot, Lionel Nicolas

A morphological and syntactic wide-coverage lexicon for Spanish: The Le*ff*e

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

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

Lionel Nicolas
Équipe RL
Laboratoire I3S
Sophia Antipolis, France
lnicolas@i3s.unice.fr

Abstract

In this paper, we introduce the *Léxico de Formas Flexionadas del Español* (Le*ff*e), a wide-coverage morphological and syntactic Spanish lexicon based on the Alexina lexical framework. We explain how the Le*ff*e has been created by merging together several heterogeneous lexicons and how the Alexina lexical framework has been applied to Spanish. We also introduce a semi-automatic technique based on a tagger to detect the lexicon's deficiencies. A preliminary evaluation shows the potential of the Le*ff*e and the relevance of both creation and extension processes.

1 Introduction

High-level Natural Language Processing (NLP) tools require reliable linguistic resources, such as lexicons and grammars. Nowadays, such relevant resources exist for English, but are often absent or incomplete for other languages, even major ones. For example, some lexical resources exist for Spanish, but none of them combines satisfactorily 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.

The Le*ff*e[1] is a wide-coverage morphological and syntactic lexicon based on the Alexina framework [13, 15, 1]. This lexicon follows the linguistic criteria applied on the French lexicon Le*fff*[2] taking advantage of the linguistic proximity between Spanish and French as Romance languages.

The main contributions of this piece of research are the following:

- we present a morphological and syntactic wide coverage lexicon for Spanish;
- we describe an enhanced available lexical framework,
- we expose a simple semi-automatic PoS tagger-based approach to detect numerous missing entries in a lexicon (including homonyms).

The work described here is one of the starting points of the recently created Victoria project. This project aims at developing techniques and tools for an efficient acquisition and correction of the linguistic resources necessary to symbolic syntactic parsers. The first phase of the project focuses on Spanish, Galician[3] and French.

This paper is organized as follows: we present first a brief description of Spanish language in Section 2. In Section 3 we describe the lexical framework used to formalize the linguistic information. Then, we briefly describe how some available resources were used to develop the Le*ff*e. In Section 5 we present a semi-automatic technique to correct and extend the lexicon. Finally in Section 6 we show preliminary evaluations of the lexicon and present our conclusions in Section 7.

2 The Spanish language in brief

Spanish is a Romance Language, just like Italian, French, Portuguese, and many others. Despite being spoken as a mother tongue by more than 400 million people, this language is little formalized within the framework of NLP when compared with English.

Spanish is an inflected language, with a two-gender system, about fifty conjugated forms per verb, but limited inflections for nouns, adjectives, and determiners. It is morphologically characterized with the Latin alphabet plus the letter *ñ* and the digraphs *ch*, *ll* and *rr*. Apart from this, the acute accents are commonly used and they enable homophones to be distinguished: e.g., *te* ('you', object pronoun) and *té* ('tea').

Regarding syntax and grammar, it is right-branching, uses prepositions, and usually, though not always, places adjectives after nouns. Its syntax is generally Subject Verb Object, though variations are valid and very common. The subject is usually omitted but appears in an implicit fashion. Contrary to English, but similarly to other Romance languages, it is verb-framed, i.e., many Spanish verbs directly encode motion path, and may leave out the manner of motion or express it in a complement of manner: e.g., *entrar* (go in), *salir* (go out), *subir* (go up). It also use a noticeable range of pronominal verbs.

[1] *Léxico de formas flexionadas del español* - Lexicon of Spanish inflected forms
[2] *Lexique des formes fléchies du français* - Lexicon of French inflected forms
[3] A co-official language in north-west Spain.

A corner stone of this work relies on the fact that Spanish is similar to other Romance Languages in many ways. Indeed, a network of correspondances can easily be established between their features.

This led us to consider the benefits of reusing resources describing related languages when building the L*effe*. Such approach presents many advantages :

- a more flexible and complete formalism could be found to develop the L*effe*,
- establishing interlingual links between resources written with a common formalism results easier,
- the data contained in resources describing other related languages can be more easily acquired.

According to these statements, we identified the L*efff*, an enhanced morphological and syntactic wide-coverage French lexicon based on the Alexina format (See section 3), as the best candidate.

3 The Alexina framework

A detailed lexical description of all words (or as many as possible) 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 lexical forms associated with morphological and syntactic information. Alexina is a framework that represents lexical information in a complete, efficient and readable way [11, 1], and is compatible with the LMF[4] [2] standard. The flexibility and completeness of the Alexina format allow a straightforward integration with deep grammatical formalisms (LFG, LTAG) which require detailed syntactic data for all forms, and allow to model lexical information for diverse languages. It is indeed the lexical framework of the L*efff*, a large-coverage morphological and syntactic lexicon for French, but also that of other lexical resources for languages such as Polish, Slovak, and soon English.

The Alexina model is based on a two-level representation, detailed below, that separates the description of a lexicon from its use:

- The intensional lexicon factorizes the lexical information by associating each lemma with a morphological class and deep 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 with a detailed structure that represents all its morphological and syntactic information; it is directly used by NLP tools such as parsers.

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 described in [11, 12]. 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.

3.1 The Intensional format

Each entry in the intensional lexicon is usually defined by a lemma and a POS. Nevertheless, it is possible to find several entries with same lemma and POS but differing in the morphological and syntactic information. This allows to split one lemma into different semantic meanings which implies different syntactic constructions.

An intensional entry details the following information:

- a *morphological class*, which defines the patterns that build all inflected forms of the lemma [12];
- a *category* (or part-of-speech, often written POS), that is taken from the chosen tagset — L*effe* uses the Multext (Parole) tagset; categories can be divided in two types: open (productive) categories (adjectives, adverbs, verbs, nouns) and closed (grammatical) categories;
- a (deep-syntax) *subcategorization frame*, that explicits how the lemma might be used in valid syntactic constructions: it lists the canonical syntactic functions of the lemma's possible arguments,[5] and the possible realizations of each of these functions;[6,7]
- additional syntactic information (control, raising, attributes...);
- possible *(re)distributions*, that define how the deep-syntax subcategorization frame is to be transformed so as to build extensional surface-syntax subcategorization frames (usual (re)distributions are *%actif*, *%passif*, *%se_moyen*).[8]

For example, here is the intensional (slightly simplified[9] for clarity reasons) entry in the L*effe* for the Spanish lemma

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

[5] The L*effe* uses the following syntactic functions: *Suj* for subjects, *Obj* for direct objects that can be cliticized into an accusative clitic pronoun, *Obja* for indirect objects introduced by the preposition *a*, *Loc* and *Dloc* for locative and delocative arguments, *Att* for (subject, object or *a*-object) attributes, and *Obl* (and *Obl2*) for other (non-cliticizable) arguments. More detailed defining criteria for their French counterparts in the L*efff* can be found in [14].

[6] Possible realizations are threefold:
– clitic pronouns: *cln* (nominative clitic), *cla* (accusative clitic), *cld* (dative clitic), *serefl* (reflexive *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., *a-sn*)

[7] Note that realizations have the same (French) names as their French counterparts in the L*efff*. This should change in the next version of the L*effe*.

[8] As for realizations, redistributions have the same (French) names as their French counterparts in the L*efff*. This should also change in the next version of the L*effe*.

[9] In particular, additional syntactic features such as control information are not shown.

diagnosticar$_1$, i.e., *diagnosticar* in the sense of the English *to diagnose*.

```
diagnosticar1
        V4
        Lemma;v;
        <arg0:Suj:cln|scompl|sinf|sn,
        arg1:Obj:(cla|scompl|sn)>;
        %actif,%passif
```

It describes a transitive entry with the following informations:

- its morphological class is `V4`, the class of the first-conjugation verbs (ending *-ar*) whose stem changes for present subjunctive(*c* changes to *que*);
- its semantic predicate can be represented by the `Lemma` as is, i.e., *diagnosticar*;
- 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. ;
- it allows for two different redistributions: active (`%actif`) and passive (`%passif`).

3.2 The Extensional format

The compilation process builds one extensional entry for each inflected form and each compatible redistribution, by applying formalized definitions of these redistributions. For example, the only inflected forms of *diagnosticar* that are compatible with the passive redistribution are the past participle forms. The (simplified) extensional passive entry for *diagnosticados* (*diagnosed*) is the following (`MP00SM` is the morphological tag for past participle masculine plural forms):

```
diagnosticados    v
[pred='diagnosticar1<arg1:Suj:cln|scompl|sn,
arg0:Obl2:(por-sn)>',@passive,@pers,@MP00PM];
%passif
```

As can be seen 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 (*por-sn*) was added.

4 Reusing other lexical resources

In order to create a first version of the L*eff*e, the first step was of course to reuse available Spanish lexical resources.

Reusing available linguistic resources is a handy way to start developing new ones. However, it requires to interpret all input resources even though their lexical models are partially incompatible, convert them into a common model and format, and finally merge the converted lexicons. None of these three steps is trivial.

Indeed, available resources might describe a given language from different points of view and/or using different linguistic criteria. This can be used to acquire information covering different aspects of a language. When considering whether a resource was worth using or not for this task, we payed more attention to quality or richness than coverage. After all, combining several resources shall lead to a good coverage that will generally be wider than the largest of them. Thus, we ensured as more important the reliability of the information put into the new resource. The application of the technique described in section 5 allowed us later to regain more coverage.

As stated in the introduction, several resources are available for Spanish, but none of them fulfilled our requirements:

- wide coverage, good precision and satisfying richness,
- complete separation between lexical and grammatical information, i.e., independence from the grammatical formalism it is going to be used with,
- clear and compact format easily readable by humans,
- free availability in terms of access, modification and distribution;
- easily linkable with resources in other languages.

Nevertheless, in order to create a first version of the L*eff*e, we reused the following resources:

Multext is an international project [6] which aims, among other things, at developing standards and specifications for the encoding and processing of linguistic corpora.

The USC lexicon is a large morphological Spanish lexicon [16], created for PoS tagging tasks in the research group *Gramática del Español* of the University of Santiago de Compostela (Spain).

ADESSE is a database of Spanish verbs developed at the University of Vigo (Spain) [3] with syntactic and some semantic information. It is a high quality work which includes subcategorization frames for more than 4,000 verbs. However, it is restricted to verbs and does not include morphological information;

The Spanish Resource Grammar (SRG) is an open-source multi-purpose large-coverage and precise grammar for Spanish [7] grounded in the theoretical framework of Head-driven Phrase Structure Grammar (HPSG). It includes a lexicon describing syntactic information for Spanish in a well organized hierarchy of syntactic classes. However, it is not easily readable, and specific to the HPSG formalism.

In order to merge these resources, we followed a process described in details in [9]. We briefly remind it here.

As mentioned in Section 2, Multext and USC lexicons only include morphological information, whereas the SRG Lexicon and ADESSE include syntactic information. Therefore, we proceeded in the following way:

1. we built a morphological baseline lexicon by converting the Multext lexicon into the Alexina format and added some Alexina-specific entries (prefixes, suffixes, named entities, punctuation signs);
2. we converted the USC Lexicon into the Alexina format and merged it with the baseline lexicon extracted from Multext, so as to obtain the morphological base of the L*eff*e;
3. we converted the syntactic information from ADESSE and the SRG lexicon into the Alexina format;

4. we merged the morphological L*eff*e from step 2 and both verbal syntactic lexicons built during step 3; the result was the L*eff*e beta.

The final result is a morphological and syntactic lexicon with an important coverage in terms of morphological information but a more restricted one in terms of syntactic information. Indeed, for morphological entries[10] for which no syntactic information could be found, we added default syntactic features corresponding to the most common ones among entries with the same PoS. For example, all verbal lemmas that were not covered by ADESSE or SRG received the following subcategorization frame: `<Suj:sn|cln,Obj:(sn|cla)>` (transitive verb with optional direct object). However, the application of semi-automatic techniques to extend and correct a lexicon, as described in [10], should help us fixing this aspect.

5 Tagger-based identification of missing entries

The next step after obtaining a first version of the L*eff*e was to continue upgrading it by adding missing entries. Usually, this task is manually performed and thus, is a time-costly process. We now present a simple but effective semi-automatic technique which greatly eases the process by identifying possible missing entries.

We distinguish two types of missing entries:

1. totally non referenced forms,
2. missing homonyms of forms referenced in the lexicon, i.e., forms non associated to a different Part-of-Speech (PoS).

In order to detect missing entries, a PoS tagger [5, 8] might be used to discover new PoS tags thanks to its ability to guess PoS tags for unknown words. The tagger we use is trained with a Spanish training corpus of approx. 500,000 words extracted from the Ancora[11] corpora and L*eff*e as an external lexicon.

According to the kind of missing entries we are trying to identify, the tagger is used in two different ways.

When looking for non referenced forms, we simply rely on the tagger's ability to guess tags for unknown words.

When looking for missing homonyms, we allow the tagger to assign new tags to known forms that are different from those included in its lexicon by forcing it to consider known forms as unknown. Indeed, the default strategy for most taggers when facing a form included in their internal lexicon is to consider as candidate tags only the ones associated there. Thus, when facing a missing homonym of a form, the tagger will never consider as a potential candidate the correct missing tag. In order to obtain such behavior, we simply bypass the internal lexicon. Thus, the tagger guesses new tags basing itself on the morphology of the form and its local context.

Obviously, such a process introduces ambiguity on purpose. In order to keep it beyond limits, we only force one form in a sentence at a time to be considered as unknown. Thus, to guess PoS tags for all words in the sentence, the sentence is entirely tagged several times.

Since forms belonging to closed categories[12] are generally well described (and their homonyms correctly included too), only forms belonging to open categories[13] are forced as unknown.

Of course, taggers make mistakes, particularly when dealing with unknown (forced or not) forms. A well-known situation for a tagger is to consider an unknown proper noun as a common noun. However, the scope of the process span an entire corpora and not only one sentence. Thus, considering a large amount of text allows us to compute a statistical ranking of the suspected missing forms which balance the false positives produced by tagging errors. This ranking takes into account the precision rate $prec_t$ for a tag t, as evaluated relatively to the training corpus, and n_{wt} and the number of occurrences of the form w tagged as t. More precisely, we assign to each couple form w and tag t a *score* $S_{\text{sc}}(w,t)$ defined as follows:

$$S_{\text{sc}}(w,t) = prec_t \cdot \log(n_{w/t}) \quad (1)$$

Thanks to this *ranking*, we are able to generate an ordered list of candidate pairs *(form,PoS)* which minimizes the appearance of false positives. As we will see in section 6, this list was good enough to be manually reviewed in a short amount of time.

6 Preliminary Evaluation

In order to evaluate the quality of L*eff*e, currently in beta version, we performed the following tests: on the one hand, we compared L*eff*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 L*eff*e 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). We computed the following properties for the other lexicons:

- SRG: 76,000 unique *(lemma,PoS)* pairs[14] (53.9% fewer than L*eff*e), some of them associated with syntactic information;
- Multext: 510,710 unique *(form,PoS)* pairs[15] (24.9% fewer than L*eff*e), and no syntactic information is provided;
- Spanish gilcUB-M Dictionary: 70,000 lemmas[15](57.6% fewer than L*eff*e), and no syntactic information is provided;
- USC Lexicon: 490,000 unique *(form,PoS)* pairs (27.95% fewer than L*eff*e), and no syntactic information is provided.

We also tested the morphological coverage of our lexicon in the context of a real application: a morphological

[10] The condition to add an entry to the L*eff*e was to acquire at least its morphological information.

[11] http://clic.ub.edu/ancora/index.php, July 2009.

[12] Such as prepositions, pronouns and determiners.

[13] Adverbs, common nouns, proper nouns, verbs, adjectives.

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

[15] According to ELRA webpage `http://catalog.elra.info`, April 2009.

pre-processor [4] developed by the COLE[16] and LYS[17] groups. We first performed a test with our baseline lexicon and another one with the Le*ff*e.

The corpus of raw text we used as input for these tests was obtained from Wikipedia Sources[18]. 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 pre-processor and thus remained unknown. It is worth noting that unknown words are the main 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 has noticeable benefits. The Le*ff*e has beaten other large lexicons in the morphological preprocessing task. Even if the difference is slight, this demonstrates the advantage 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 used the notion of *expanded intensional entry* [9] which is just a defactorized Alexina subcategorization frame. Thus, each *expanded intensional entry* describes one fully-specified syntactic behaviour.

The expanded intensional lexicon acquired from SRG contains 42,689 unique entries, i.e., fully-specified subcategorization frames, while the one obtained from the ADESSE contains 39,040 entries. After merging these lexicons, the number of such unique entries jumps to 66,028. Finally, the Le*ff*e, which associates default syntactic information with all verbs not covered by the result of this merging, contains 91,507 unique expanded entries. After factorization, the Le*ff*e contains 16,311 verbal entries.

Once the first version of Le*ff*e was built, we used the technique described in section 5 to upgrade its coverage. We used a corpus built from a subset of the Spanish part of the Europarl[19] containing approx. 6 million words. The only restriction applied to this corpus was to avoid the inclusion of sentences containing foreign words, since they would lead to false positives.

A ranking of suspected missing pairs *(form, tag)* was obtained. The quality of this list was not exceptional since it included many false positives, but, thanks to this list, we did include in the Le*ff*e at a very small cost (it was manually done by one person in two days) nothing less than 1,800 lemmas. We must note that the original coverage of the Le*ff*e was very high and thus it is reasonable to think that the proportion of false positive would have been reduced when dealing with lexicons with a smaller coverage.

Table 2 shows the number of lemmas added to the Le*ff*e classified by categories. The great majority were proper nouns, since they were very incomplete in Le*ff*e up to this point. The approx. 1,800 intensional entries added to the Le*ff*e correspond to more than 3,700 inflected forms in the extensional lexicon. For example, we added the verbs it abstraer (to abstract) and *documentar* (to document), the adjective francoespañol (Franco-Spanish), the common noun *biocarburante* (biofuel), the adverb *precipitadamente* (hastily) and the proper noun *Niza* (Nice).

Apart from the correct entries, the list allowed us to detect some systematic deficiencies, such as diminutives/augmentatives and adverbs ending in *-mente*. In a near future, they will be automatically generated after updating the morphological rules used to obtain the extensional lexicon from the intensional one (see sect.3).

[16] `http://www.grupocole.org`, April 2009
[17] `http://www.grupolys.org`, April 2009
[18] `http://download.wikimedia.org`, January 2009
[19] A parallel corpus from the European Parliament proceedings

7 Conclusion

In this work we have presented a morphological and syntactic wide-coverage lexicon for Spanish built by taking advantage of existing lexical resources in Spanish and French. Nowadays, for many languages, several scattered linguistic resources exist, but usually none of them is satisfying in terms of coverage, richness or precision. Nevertheless, the amount of work invested in their development should not be ignored. In fact, we believe reusing already formalized knowledge is a handy and productive way to build and/or upgrade other linguistic resources and it will be the usual strategy in the near future.

We also described a tagger-based approach to detect missing entries in a lexicon. Even when applied to a quite exhaustive lexicon, such as Le*ff*e, this simple approach has allowed us to add more than 3,700 lexical forms in a very short amount of time.[20]

The resulting lexicon, the Le*ff*e, is currently in beta version and will soon be distributed under a LGPL-LR license[21]. Although it is still far from perfect, we have shown that the Le*ff*e 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 by comparing the coverage and precision of different deep parsers that rely on the same grammar but using different morphological and syntactic lexicons such as the Le*ff*e.

Acknowledgements

This work was supported in part by the 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 the group *Gramática del Español* from USC, and especially to Guillermo Rojo, M. Paula Santalla and Susana Sotelo, for granting us access to their lexicon.

References

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

[20] The use of a tagger with a lower error rate, in particular on words unknown to the tagger's training corpus, should allow to scale up the efficiency of this approach. Such a tagger is under development in the team of one of the authors.

[21] As explained in this paper, the construction of the Le*ff*e involved the Spanish morphological lexicon developed within the Multext project, which is freely 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 it appropriate to publish the Le*ff*e beta under the LGPL-LR.

	TOTAL UNKNOWN WORDS	UNIQUE UNKNOWN WORDS
USC Lexicon	70,026	25,888
Baseline	86,521	27,234
Leffe	69,756	24,703

Table 1: *Results obtained by applying the morphological preprocessor with different lexicons.*

	LEMMAS (INTENSIONAL)	INFLECTED FORMS (EXTENSIONAL)
Adjectives	88	298
Adverbs	54	54
Verbs	26	1,693
Common nouns	117	231
Proper nouns	1,518	1,518
Total	1,803	3,740

Table 2: *Lemmas acquired using the tagger-based technique.*

[2] G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, mandy Pet, and C. Soria. Lexical Markup Framework (LMF). In *Proceedings of LREC'06*, Genoa, Italy, 2006.

[3] J. M. García-Miguel and F. J. Albertuz. 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*, Saarbrücken, Germany, 2005.

[4] J. Graña, F. M. Barcala, and J. Vilares. Formal methods of tokenization for part-of-speech tagging. *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science*, 2002.

[5] J. Graña. *Técnicas de Análisis Sintáctico Robusto para la Etiquetación del Lenguaje Natural (robust syntactic analysis methods for natural language tagging)*. Doctoral thesis, Universidad de La Coruña, Spain, 2000.

[6] N. Ide and J. Véronis. Multext: Multilingual text tools and corpora. In *Proceedings of COLING'94*, Kyoto, Japan, 1994.

[7] M. Marimon, N. Seghezzi, and N. Bel. An open-source lexicon for Spanish. In *Sociedad Española para el Procesamiento del Lenguaje Natural, n. 39*, 2007.

[8] M. A. Molinero, F. M. Barcala, J. Otero, and J. Graña. Practical application of one-pass viterbi algorithm in tokenization and pos tagging. *Recent Advances in Natural Language Processing (RANLP). Proceedings, pp. 35-40*, 2007.

[9] M. A. Molinero, B. Sagot, and L. Nicolas. Building a morphological and syntactic lexicon by merging various linguistic resources. In *Proceedings of NODALIDA'09*, Odense, Denmark, 2009.

[10] L. Nicolas, B. Sagot, M. A. Molinero, J. Farré, and É. Villemonte de La Clergerie. Computer aided correction and extension of a syntactic wide-coverage lexicon. In *Proceedings of COLING'08*, Manchester, United Kingdom, 2008.

[11] B. Sagot. 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, 2005.

[12] B. Sagot. 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, 2007.

[13] B. Sagot, L. Clément, E. Villemonte de La Clergerie, and P. Boullier. The Lefff 2 syntactic lexicon for French: architecture, acquisition, use. In *Proceedings of LREC'06*, 2006.

[14] B. Sagot and L. Danlos. Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire – Constructions impersonnelles. *Cahiers du Cental*, 2007.

[15] B. Sagot and L. Danlos. 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, 2008.

[16] C. Álvarez, P. Alvariño, A. Gil, T. Romero, M. P. Santalla, and S. Sotelo. Avalon, una gramática formal basada en corpus. In *Procesamiento del Lenguaje Natural (Actas del XIV Congreso de la SEPLN)*, pages 132–139, Alicante, Spain, 1998.