Towards efficient production of linguistic resources: the *Victoria* Project

Lionel Nicolas[⋆], Miguel A. Molinero[◇], Benoît Sagot[⊕],
Elena Sánchez Trigo[•], Éric de La Clergerie[⊕], Miguel Alonso Pardo[◇],
Jacques Farré[⋆], Joan Miquel Vergés[•].

[⋆] Équipe RL, Laboratoire I3S, UNSA+CNRS, France {lnicolas,jf}@i3s.unice.fr
[◇] Grupo LYS, Univ. de A Coruña, España mmolinero@udc.es
[⊕] Projet ALPAGE, INRIA Rocquencourt + Paris 7, France {benoit.sagot, Eric.De_La_Clergerie}@inria.fr
[•] Grupo Cole, Univ. de Vigo, España {etrigo,jmv}@uvigo.es, alonso@udc.es

Abstract

In order to produce efficient Natural Language Processing (NLP) tools, reliable linguistic resources are a preliminary requirement. When available for a given language, the resources are generally far below the expectations in terms of quality, coverage or usability. This paper presents a project whose ambition is to enhance the production capacities of linguistic resources through the creation and intensive use of interconnected acquisition and correction tools, inter-lingual transfer processes and a collaborative online development framework.

1 Introduction

The efficiency and linguistic relevance of most NLP tools depends directly or indirectly on the quality and coverage of the resources they rely on. For major languages such as Spanish and French, many well known and widely used resources are still in a precarious state of development. For languages with a smaller speech community, such as Galician[1], they are generally non-existent.

Such an absence is a direct consequence of the cost induced by their development: their complexity and/or size makes their manual improvement a labor-intensive, complex and error prone task requiring massive expert work.

Such an important effort could obviously be balanced by sharing it among several people or groups interested in those resources. Nevertheless, long-term collaboration can be problematic for license, management, distance, time or financial reasons. Thus, linguistic resources are generally developed in a somewhat isolated way.

Owing to these issues, building linguistic resources takes years of constant effort which often fails to achieve visible or useful results. Therefore, quick and efficient acquisition and correction of linguistic resources is an unsolved problem of considerable interest to the NLP community.

In order to face it, the *Victoria* project aims at:

- First and foremost, developing a chain of tools automatizing the acquisition and correction processes in order to reduce labor work and enhance the quality, homogeneity, connectivity and coverage of the linguistic resources produced.
- Exploring inter-lingual transfer processes of linguistic knowledge in order to build or upgrade resources for a given language taking advantage of similar resources formalizing other linguistically related languages.
- Allowing people to combine their efforts through a shared web development framework.

These three objectives are dedicated to the more general goal of producing and providing freely available high-quality linguistic resources.

In its current state of development, the project focuses on the resources necessary to build symbolic syntactic parsers[2] for French, Spanish and Galician. As a long term goal, the project will extend to other kinds of resource and other Romance languages.

The main contributions of this paper are the following:

1. It exposes a strategy with several guidelines that may be reused for other projects with similar objectives.
2. It reports the viability of transferring some formalized linguistic knowledge between two related languages.
3. It presents theoretical and generic techniques to sequentially and incrementally detect and correct shortcomings in linguistic resources.
4. It lists the tools, techniques and resources that have already been produced.

This paper is organized as follows. In section 2, we shortly introduce the project itself and briefly recall the past and existing projects with common points to ours. In section 3 and section 4, we explain our strategy for enhancing the production capacities. We then detail in section 5 what has been achieved and what is still ongoing. Finally, in section 6, we detail our future orientations and developments and conclude in section 7.

2 Overview

This project brings together researchers from the computer science field and researchers from the human translation field of four different French and Spanish teams. The

[1] A co-official language spoken in the north-west of Spain.

[2] Morphological rules, morpho-syntactic lexicons and lexicalised grammar.

International Conference RANLP 2009 - Borovets, Bulgaria, pages 318–323

COLE team (Grupo COLE) from the Univ. of Vigo, the LYS team (Grupo Lys) from the Univ. of A Coruña, the Alpage project (Projet Alpage) from the Univ. of Paris 7 and the INRIA Institute of Rocquencourt and the RL team (Équipe RL) from the I3S laboratory of the Univ. of Nice Sophia Antipolis and the CNRS Institute.

The project officially started in November 2008 thanks to the financiation of the Galician Goverment (INCITE08PXIB302179PR, INCITE08E1R104022ES).

2.1 Related projects

There have been various projects aiming at building lexical resources for a large spectrum of languages. The most famous are probably the MULTEXT project[3] and its follow-up MULTEXT-East.[4] Other projects focused on specification, standardization and/or development of lexical resources, such as GENELEX , EAGLES and PAROLE.

As for the syntactic level, the DELPHIN project[5] based on the LKB framework as well as the AGFL project[6] have permitted the creation of various formalized grammars. Some other existing projects, such as LinGO Grammar Matrix[7], explore the possibility of sharing formalized linguistic knowledge among several resources in different languages.

The ongoing CLARIN[8] and FLARENET[9] initiatives aim at managing and bringing under a common framework many existing resources.

Obviously, describing each project is a complex task that would fall beyond the scope of this paper. We shall just highlight that, in most cases, resources have been built with little (or no) computer aid. So far, we are not aware of any large-scale project regarding automatic acquisition of linguistic information from plain corpora. This causes a common situation where the resources are developed until a (more or less) advanced state of development where it becomes difficult to find errors/deficiencies manually. Then, they usually get stuck and do not evolve much.

Furthermore, manual development is one of the main reasons for the poor (free) availability of the resources. Indeed, manual development greatly increases the cost of development, which sometimes prevents resources from being freely distributed.

2.2 Guidelines of the project

By studying the weaknesses and strengths of related projects, we established several guidelines to achieve our objectives. Those guidelines, detailed in section 3 and 4, can be resumed as follows:

- In order to allow collaborative work, easily accessible online consultation and edition interfaces should be available for every kind of resource produced.
- So as to maximize feedback, the resources shall always be under non-restrictive public open-source distribution license in order to avoid restricting their availability.
- In order not to limit the scope of the project to a particular set of languages or tools, the formalisms used to describe the resources shall be as general as possible.
- Existing available resources should always be considered when upgrading a particular resource, including those describing another language.
- Tools automatizing the processes of detection and correction of linguistic resources are an absolute necessity when aiming toward for the construction of high quality linguistic resources.
- The tools developed shall use as input plain text which is daily produced for most languages.

3 Enhancing collaborative work, availability and usability

To our knowledge, there are three reasons that may limit collaborative work.

First, it is unusual to find resources with dedicated consultation or edition interfaces. Manual edition is often error-prone since humans can make typing errors or introduce incoherent data. Thus, collaborative work is generally restricted to a smaller number of skilled persons.

Second, collaborative work can be limited if it cannot be achieved from anywhere.

Third, collaborative work can be technically restrained by some operating systems or platforms.

In order to prevent edition errors and allow users to focus on the data themselves without worrying about the underlying formalism, we are willing to develop a dedicated query and management interface for every resource and technique output. In order to overcome distance troubles, every dedicated interface shall be accessible online. So as to avoid technical problems that could restrain access, all interfaces shall be developed with stable Web technologies handled by most web browsers without additional installations.

In order to maximize feedback and federate people with linguistic skills around the common beneficial goal of providing high-quality resources for everybody, all formalized linguistic knowledge should be available to anybody willing to consult, use or collaborate in its development. The availability of the produced techniques, formalisms and resources by the *Victoria* project in terms of access, modification and distribution is guaranteed by a non-restrictive public open-source distribution. Such an objective is fulfilled thanks to non-restrictive public licenses like LGPL-LR[10] and CeCILL-C.[11]

In order to maximize the usability[12] of the resources produced, the choice of the most suitable formalisms has been made according to the following principles:

- Since the combined use of resources is usually a requirement when designing advanced NLP tools, all the formalisms designed and used should be general and extensible enough to permit combined uses.

[3] http://aune.lpl.univ-aix.fr/projects/MULTEXT/
[4] http://nl.ijs.si/ME/
[5] http://wiki.delph-in.net/moin/LkbTop
[6] http://www.agfl.cs.ru.nl/
[7] http://www.delph-in.net/matrix/
[8] http://www.clarin.eu
[9] http://www.flarenet.eu/

[10] Lesser General Public License for Linguistic Resources
[11] LGPL-compatible, http://www.cecill.info/.
[12] By usability, we mean the capacity of the resource to be integrated in NLP tools or applied to a particular language

- Foreseeing the exhaustive list of those uses is simply impossible. Therefore, it is essential for the formalisms to be compared to various kind of languages and practical tools in order to adapt and extend them.

- The formalisms need to be regularly maintained so as to guarantee their extension to uncovered phenomena.

In order to develop our morphological and lexical resources, we chose to use the Alexina framework [7, 8, 2]. This framework, which is compatible with the LMF standard [3] represents morphological and syntactic information in a complete, efficient and readable fashion. The Alexina model is based on a two-level representation distinguishing the description of a lexicon from its use. The intensional level, used for an efficient description, 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 level, used in practice by tools, is generated automatically by *compiling* the intensional lexicon thanks to the morphological rules. It 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. Alexina has already been used to develop morpho-syntactic wide-coverage lexicons for French, Spanish, Slovak and Polish and has been combined with syntactic parsers based on commonly used grammatical formalisms (TAG and LFG).

Regarding grammatical knowledge, our resources rely on a meta-grammar formalism [1] which represents the syntactic rules of a language in a hierarchical structure of classes. The classes on top of the hierarchy define general concepts as Part-of-Speech (noun, verb, etc.) and their possible attributes. Classes are then refined while descending towards the bottom of the hierarchy, adding constraints, allowed/forbidden constructions, etc. This meta-grammar formalism is theoretically compilable in most commonly used grammar formalisms. In practice, we compile our grammars into a hybrid TAG/TIG parser. Such a generic formalism is extremely useful since it permits an easy adaptation of an existing grammar to a linguistically related language. For example, Romance languages, which include major languages (Spanish, French, Italian, Portuguese, etc.) and many others with smaller speech communities (Galician, Catalan, Occitan, Sardinian, etc.), are very similar in terms of syntactic behaviors. Hence, many definitions, constraints and rules can be reused when building a new grammar for another related language. It is worth noting that the outputs produced by our parser are dependency trees.

4 Enhancing extension/correction

4.1 Using existing resources

Existing resources are generally valuable sources of data when building new resources or extending others. Ignoring the great efforts invested in order to build existing resources does not seem reasonable or productive. Such an approach depends on the resource and the kind of data one is trying to adapt. Nevertheless, various practical experiments (see sect. 5.3.2 and [2]), have shown that existing resource usually share common points. Adapting a large part of the available existing resources is often a reasonable objective.

4.1.1 Interlingual transfer processes

Since related languages share large parts of their linguistic knowledge, we do not restrict the scope of this approach to a single language and consider the existing resources describing other related languages. Such an approach is especially beneficial when working on languages with smaller speech communities and limited digital resources. It also facilitates the establishment of interlingual links required for multilingual tasks. Informally, we could say that the proximity between linguistically related languages can be used to "transfer" formalized knowledge from one resource to another.

In order to achieve such a task, one should consider separately the formalisms and the formalized knowledge.

Extending/adapting the formalisms used to describe a given language to a related one is generally fast. This statement has been verified in practice when building new resources for Spanish from French ones (see sect. 5.3)

Transferring linguistic knowledge depends on the kind of knowledge we are dealing with.

Transferring morphological knowledge seems improbable. Applying the morphological rules of a language to a related one seems risky; we have not considered it so far.

Regarding lexical knowledge, the following idea seems promising: whoever has learned two common-rooted languages must have realized that many "direct" translations are effective, i.e., it seems possible to apply a basic morphological alignment to translate some words. This concept, similar to cognates, is known as very delicate. Nevertheless, when studied more closely, this statement seems to apply mainly to less frequent words since they are generally the ones that have evolved the least from the root language (Latin in our case). For example, an infrequent word ending with *-tion* in French can often be translated by a word ending with *-ción* in Spanish.

As regards grammatical knowledge, grammars are abstract and static enough to not evolve much. Consequently, a grammar designed for French could be used as a starting point to build a grammar for a related language such as Spanish (see sect. 5.3). In addition, since many grammar rules are shared by both grammars, establishing interlingual syntactic links between constructions results easier. Such an approach is already effective for French and Spanish (see sect. 5.3). The results should even be further enhanced when considering Spanish with other Iberian languages such as Galician.

4.2 Using correction and extension processes

We now describe a generic approach which has been abstracted from practical research results described essentially in [9] and [5].

In order to efficiently produce new formalized knowledge, a source of data is needed to detect and acquire the missing knowledge. Since this source should be available in sufficient quantity for any language, we have discarded

annotated data[13] which is only available in limited quantities for a small number of languages, and opted for plain digital text which is daily produced for most languages.

In order to extend and correct a resource from plain text, we apply the following two-step generic approach:

- identify as accurately as possible which part of the text is not covered by a resource,
- generate corrections for the detected shortcomings and rank them in order to prepare an easier manual validation.

We now present generic approaches to achieve these two steps. Practical implementations have already proved to be effective in practice (see sect. 5.2).

4.2.1 Identifying shortcomings in a resource

Identifying possible shortcomings in a studied resource can be achieved by studying unexpected/incorrect behaviors of some tools relying on the resource. To do so, one needs first to establish what can be considered as unexpected (incorrect) behavior. Once identified, one must ensure they are not due to some incorrect data given as input or some other resource the tool relies on.

The first situation can easily be avoided by giving as input corpora considered as linguistically correct (error-free), i.e., the corpora one wants the resources to cover. We use law texts and some selected journalistic productions and discard corpora we consider as having a poor quality, like those composed of emails.

The second situation, i.e., when the tool relies on various resources, can be solved through a global study of the unexpected behaviors. Indeed, natural languages are ambiguous and thus, difficult to formalize. Nevertheless, this ambiguity has the advantage of being randomly distributed on the different aspects of a language. Depending on the state of development of the resources, it can be truly rare for two resources to be incorrect at the same time for a given element, i.e., many unexpected behaviors can be induced by only one resource at a time. In a restricted scope, it is difficult and hazardous to identify a culprit for a given unexpected behavior. However, such an aspect can be balanced by a global study of the behaviors when processing a massive set of text. Indeed, if among the elements of a given resource, some are always found when unexpected behaviors occur, then such an element can be (statistically) suspected to be incorrectly described in the resource.

For example, in [9], the authors are looking for shortcomings in a lexicon. The tool they observe is a syntactic parser and parse failures are considered as unexpected behaviors of the parser. Each parse failure can be due to deficiencies of the grammar and/or of the lexicon the parser relies on. Determining for a given parse failure which resource is the true culprit can be utterly complex. In order to detect incorrect lexical entries, the authors use a fixed point algorithm which emphasizes the lexical forms that occur more than expected in non parsable sentences.

When doing so, one must keep in mind that:

1. enough plain text should be provided as input in order to ensure the validity of the statistics,
2. the statistical models might make assumptions keeping the computations within certain limits and produce irrelevant suspicions. In order to balance this aspect, we generally designed our techniques in a *semi-automatic* fashion implying a human post-validation.

4.2.2 Generating relevant corrections

As explained earlier, it may be rare for two resources A and B jointly used to be incorrect at the same time and thus be both responsible for a given unexpected behavior. Hence, if we believe resource A to be responsible for an unexpected behavior, we can often rely on resource B to generate relevant corrections. Of course, the kind of corrections depends on the data the resources interact on, i.e., not every pair of resources are suitable for this purpose.

For example, a grammar that interacts with the syntactic part of a lexicon can be used to generate corrections for it while morphological rules clearly cannot. In [5], the authors use a grammar to guess corrections for a lexicon.

Another highly convenient feature is the following: if resource B cannot be used any longer to provide relevant corrections for resource A, we can consider the left-over unexpected behaviors as mostly representing shortcomings of resource B since it does not cover them. We thus obtain an incremental and sequential way to obtain for both resources corpora representing mostly their shortcomings. Thus, correcting resource A thanks to resource B generate useful data to correct resource B. Once resource B corrected, it is possible to correct resource A. And so on.

For example, in [5], the improvement of a lexicon thanks to a grammar is limited by the quality of the grammar used. Nevertheless, the authors expose that the non-parsable part of the corpus used to guess lexical correction has become globally representative of shortcomings of the grammar. This corpus can then be used to update the grammar. Once the grammar is updated, the corpus can be used again to correct the lexicon. And so on.

5 Results

5.1 Online development framework

The recently created (incomplete) online development framework[14] aims at allowing collaborative work. In order to fulfill such a goal, it is essential to offer dedicated interfaces to consult, manage and download the resources. So far we have concentrated our efforts on developing a dedicated interface for morpho-syntactic wide-coverage lexicons which, among the three kinds of resources developed, is clearly the one requiring most collaborative work.

The current version of this interface allows us to search for entries with logical equations, consult and edit the data related to the matched entries, and trace the changes.

5.2 Techniques

According to the ideas explained in section 4.2, we established a conceptual map of a sequential chain of tools (see figure 1) which aims at helping to upgrade from plain text, in a semi automatic fashion, all the basic components of a symbolic syntactic parser, namely, morphological rules, morpho-syntactic lexicons and lexicalised grammars.

[13] we actually consider it as an existing resource, see sect 4.1

[14] soon available at `http://www.victoria-project.org`

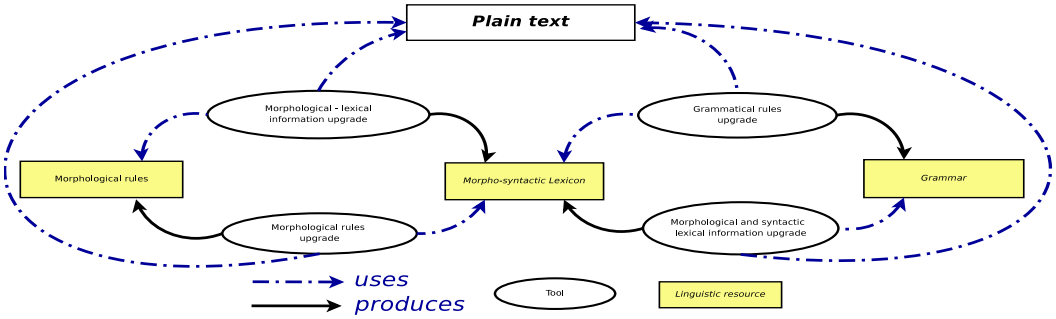

Fig. 1: *conceptual map of the semi-automatic upgrade of linguistic resources*

5.2.1 Morphological lexical information improvement

To achieve this task, we apply the technique described in [6] where the observed tool is a lexicon access system, the unexpected behaviors are the absence of some lexical forms, and the resources A and B are morphological rules and the morphological data of a morpho-syntactic lexicon.

The morphological rules are used to predict hypothetical lemmas for all forms of a corpus missing in the lexicon. A statistical fixed-point algorithm is used to rank the hypothetical lemmas according to the number of inflected forms found in the corpus. The manual validation of the best ranked lemmas improves the coverage of the lexicon and increases the quality of subsequent executions.

5.2.2 Syntactic lexical information improvement

To achieve this task, we apply the technique described in [5] where the tool observed is a syntactic parser, the unexpected behaviors are parsing failures, and the resources A and B are a morpho-syntactic lexicon and a grammar.

In order to correct and extend a lexicon, the authors firstly detect lexical forms suspected to be responsible for some parse failures thanks to two techniques.

A statistical computation which emphasizes "suspicious" lexical forms present more frequently than the rest in non-parsable sentence [10]. Lexical forms are even more "suspicious" if present in non-parsable along with forms "cleared" by their presence in parsable ones [9].

A tagger-based approach which highlights absent entries by relying on the tagger's ability to guess a tag for unknown words and forcing the tagger to use it on forms that are in fact known. If the tag answered represents data absent in the lexicon, the form is suspected.

Once the suspicious forms have been identified, the authors rely on the grammar to generate lexical corrections for the identified forms. To achieve this task, they study the expectations of a grammar for the identified forms in non-parsable sentences, i.e., they observe what lexical information would have not led to conflicts with the grammar rules and would have permitted syntactic parses. Such a goal is fulfilled by underspecifying the lexical restrictions of the suspected form in order to allow the parse to explore originally non explored grammar rules. They later extract from the parse outputs the information assigned to the suspected form and translate it back to the lexicon's format.

5.2.3 Morphological rules acquisition from a lexicon

As explained earlier in section 3, the Alexina framework employed to describe our lexicons requires morphological rules to be functional. In order to create lexicons for both Spanish and Galician (see sect. 5.3.2) and according to our statement to always consider using existing resources to build or upgrade new ones, we used the following idea to extract morphological rules from existing available morphological lexicons. For each lemma, we extract the longest prefix that is common to all its inflected forms, which is considered as the stem, and build an ordered list of *(suffix,tag)* pairs.[15] 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. The result is not only a set of morphological classes but also a list of lemmas classified under such a set of classes.

5.3 Linguistic resources

High-quality linguistic resources are the final goal of the *Victoria* project. Apart from the fact that they constitute the practical results which support our theories, we are using them to complete syntactic parsers.

5.3.1 Morphological rules

According to the technique described earlier in section 5.2.3, we used two existing morphological lexicons for Spanish and Galician in order to extract morphological descriptions from a set of *(form,lemma,tag)* triples. Morphological classes are associated to PoS, but several classes are always required to cover all the inflection cases for one PoS. Finally, we obtained a set of 237 morphological classes for Spanish (approx. 7,250 inflection cases) and 154 for Galician (approx. 4,160 inflection cases).

5.3.2 Morphological and syntactic lexicons

Two wide coverage lexicons for Spanish and Galician have already been produced following the Alexina format. Both

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

lexicons are currently being upgraded using the techniques described in section 5.2 and will be available under LGPL-LR licenses soon.

The Spanish lexicon L*eff*e [16] has overtaken other well known Spanish lexicons in terms of coverage despite being in beta version. It has been obtained by merging several existing Spanish linguistic resources [4]. Nowadays, the L*eff*e beta contains more than 165,000 unique *(lemma,PoS)* pairs, corresponding to approx. 1,590,000 inflected entries that associate a form with morpho-syntactic information (approx. 680,000 unique *(form,PoS)* pairs).

The L*eff*ga[17] has been created after the Galician lexicon developed in the CORGA[18] project. The L*eff*ga is still in alpha version (April 2009), and less developed than the L*eff*e. It contains more than 52,000 unique *(lemma,PoS)* pairs (approx. 515,000 unique *(form,PoS)* pairs). The complete lexicon includes more than 742,000 inflected entries with little syntactic information to this point.

5.3.3 Grammars

The Spanish meta-grammar (SPMG) takes as its starting point a French meta-grammar (FRMG) [11]. Nowadays, it contains 244 classes organized in a hierarchical structure. We can confirm the ease of building such a grammar for Spanish using a French one. In fact, there are few major syntactic differences between those languages. Simply by fixing these differences it is possible to achieve a coverage somewhat similar to the original French grammar. We only needed to achieve slight modifications in a dozen classes to obtain this grammar. We evaluated its coverage by extracting more than 4,000 sentences with 25 words or less from the Europarl-Spanish[19] corpus. In such a corpus we completed non-robust parses for 53% of the sentences using a parser based on L*eff*e and SPMG. It is worth noting that many parsing errors might be caused by the lexicon, since the number of completed parses depends both of the quality of the grammar and the lexicon.

6 Future work

Online tools Before considering other kind of resource, the interface dedicated to the lexicon shall be finalised.

In order to obtain plain text given as input to our techniques, we are developing a tool using the RSS system to trace journalistic production on websites, extract it (if we are allowed to) and index it in the TEI format[20].

Techniques The extension and correction techniques regarding lexical knowledge are already effective. We shall thus concentrate on developing techniques for extending morphological rules and grammars. Since we are able to produce corpora representing mostly shortcomings of both kinds of resource, we shall follow the methodology described in section 4.2. We also plan to investigate an idea explained in [5], where an entropy classifier is trained to recognize non-grammatically covered sentences. The statistical model might be an interesting starting point to guess non covered syntactic structures.

We will also work on the theory about infrequent words in order to transfer lexical information between French and Spanish. Infrequent words being the major part of a lexicon, such transfer process would be extremely useful.

Resources Thanks to the techniques explained in 5.2, we shall further extend and correct the L*eff*e and we hope to convert it into a lexical resource comparable in terms of quality and coverage with what currently exists for English.

We will extend the morphological information of the L*eff*ga and adapt, in the same way we adapted the French one to Spanish, the Spanish meta-grammar to Galician. Once a beta meta-grammar is achieved, we will be able to extend syntactic lexical information in the L*eff*ga.

7 Conclusion

In order to allow efficient production of linguistic resources, the *Victoria* project is dealing with wide coverage resources, useful techniques and a collaborative development framework, i.e., objectives that can be considered, one by one, as challenging.

Even if modest when compared to its ambitious objectives, the practical achievements obtained in only a few months demonstrates its validity and coherence and indicates that it is following a productive path.

The combination of transfer processes with efficient formalisms and extension and correction techniques is already allowing us to produce resources with noticeable qualities in a very short amount of time when manual construction would not have permitted anything similar.

References

[1] M.-H. Candito. *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées*. PhD thesis, Univ. of Paris 7, 1999.

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

[3] 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.

[4] M. A. Molinero, B. Sagot, and L. Nicolas. Building a morphological and syntactic lexicon by merging various linguistic ressources. In *Proceedings of NODALIDA 2009*, Odense, Danemark.

[5] 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.

[6] 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.

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

[8] 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.

[9] B. Sagot and É. Villemonte de La Clergerie. Error mining in parsing results. In *Proceedings of ACL/COLING'06*, pages 329–336, Sydney, Australia, 2006.

[10] G. van Noord. Error mining for wide-coverage grammar engineering. In *Proceedings of ACL 2004*, Barcelona, Spain.

[11] E. Villemonte de La Clergerie. From metagrammars to factorized TAG/TIG parsers. In *Proceedings of IWPT'05*, pages 190–191, Vancouver, Canada, 2005.

[16] *Léxico de formas flexionadas del español / Lexicon of Spanish inflected forms*
[17] *Léxico de formas flexionadas do galego / Lexicon of Galician inflected forms*
[18] `http://corpus.cirp.es/corga/`
[19] `http://www.statmt.org/europarl/`
[20] `http://www.tei-c.org`