Creating and maintaining language resources: the main guidelines of the Victoria project

Lionel Nicolas, Miguel Molinero, Benoît Sagot, Nieves Fernández Formoso, Vanesa Vidal Castro

Creating and maintaining language resources: the main guidelines of the *Victoria* project

Lionel Nicolas[1], Miguel A. Molinero[2], Benoît Sagot[3], Nieves Fernández Formoso[4], Vanesa Vidal Castro[4]

1. Équipe RL, Laboratoire I3S, UNSA & CNRS, 2000, route des lucioles, BP 121, 06903 Sophia Antipolis, France
2. Grupo LYS, Departamento de Computacin, Universidade da Coruña, Campus de Elviña s/n, 15071 La Coruña, Spain
3. Alpage, INRIA Paris-Rocquencourt & Universit Paris 7, 30 rue du Chteau des Rentiers, 75013 Paris, France
4. Grupo Cole, Universidade de Vigo, Campus de As Lagoas s/n, 32004 Orense, Spain

lnicolas@i3s.unice.fr, mmolinero@udc.es, benoit.sagot@inria.fr, {vvcastro,nievesff}@uvigo.es

Abstract

Many Natural Language Processing (NLP) tools rely on the availability of reliable language resources (LRs). Moreover, even when such LRs are available for a given language, their quality or coverage sometimes prevent them from being used in complex NLP systems. Considering the attention received from both the academic and industrial worlds and the significant efforts achieved during the past decades for LR development, such a lack of high quality and wide-coverage LR shows how difficult their creation and maintainance can be. In this paper, we describe a set of guidelines applied within the *Victoria* project in order to ease the creation and correction of the LRs required for symbolic parsing. These generic guidelines should be easy to adapt and use for the production of other types of LRs.

1. Introduction

The efficiency and linguistic relevance of most NLP tools depends directly or indirectly on the quality and coverage of the LRs they rely on. Along the past decades, numerous project, such as MULTEXT,[1] MULTEXT-East,[2] DELPHIN,[3] AGFL,[4] etc., have focused on developing LRs while the ongoing CLARIN[5] and FLARENET[6] initiatives aim at managing and bringing under a common framework many existing LRs. Despite such efforts, few LRs may be considered as complete and correct, except maybe for English, the language that has clearly received the most attention over the last decades.

Nevertheless, complex NLP systems such as automatic translation tools, if they make use of LRs, do require high-quality resources. The creation of LRs with a high level of quality in terms of coverage, quality and richness is therefore an important problem in our research field.

The main contribution of this paper is to propose a list of guidelines for the production of LR. This list has been set up while planning and managing the *Victoria* project (Nicolas et al., 2009).

This paper is organized as follows. In Section 2., we briefly introduce the *Victoria* project. We then explain in Section 3. some reasons why creating and maintaining LRs is still so difficult. Next, we detail in section 4. and 5. a set of guidelines for easing this task. Finally, we quickly highlight in section 6. the objectives the *Victoria* project has already achieved, before concluding in section 7.

2. The *Victoria* project

The *Victoria* project, started in November 2008, is funded by a grant from the Galician Goverment.[7] It brings together researchers from four different French and Spanish teams: (i) the COLE team[8] from the University of Vigo, (ii) the LyS team[9] from the University of A Coruña, (iii) the Alpage project[10] from the University Paris 7 and INRIA Paris-Rocquencourt and (iv) the RL team,[11] I3S laboratory, University of Nice Sophia Antipolis and CNRS.

The main goal of the project is to develop techniques and tools for producing and improving the high-quality and wide-coverage LRs required for symbolic parsing.[12] So far, the project has been focusing on French, Spanish and Galician languages.

3. Difficulties when creating and maintaining LR

Several reasons explain why the development of an LR has been and is still such an complex task, most of them being consequences of the intrinsic richness and ambiguity of natural languages. Two of them can be highlighted:

- the difficulty in describing all linguistic description levels (e.g., morphology, syntax, semantics);
- the difficulty in covering all instances of a given linguistic description level for a given language.

A few decades ago, the available computing power made it impossible to imagine or test the complex formalisms that

[1] http://aune.lpl.univ-aix.fr/projects/MULTEXT/
[2] http://nl.ijs.si/ME/
[3] http://www.delph-in.net/
[4] http://www.agfl.cs.ru.nl/
[5] http://www.clarin.eu/
[6] http://www.flarenet.eu/

[7] Project number INCITE08PXIB302179PR.
[8] http://coleweb.dc.fi.udc.es/
[9] www.grupolys.org
[10] http://alpage.inria.fr/
[11] http://deptinfo.unice.fr/~jf/Airelles/
[12] Morphological rules, morpho-syntactic lexicons and lexicalised grammar.

are used nowadays. Even though, we still lack a global consensus for modeling most linguistic description levels. This is particularly true for the semantic level, but the large range of avaible syntactic formalisms is another illustration of this difficulty. However, as far as lexical information is concerned, morphological and syntactic notions are now reasonably consensual, and are indeed standardized by various ISO norms such as LMF (Lexical Markup Framework) (Francopoulo et al., 2006).

However, despite the fact that there exist now consensus and therefore formalisms for some levels, it is still difficult to find the corresponding high-quality and wide-coverage LRs for many languages. This is even the case for languages such as Spanish or French, for which many well known and widely used resources are still in a somehow precarious state of development. It is obviously the case for languages with a smaller speech community, such as Galician[13], for which LRs are almost non-existent.

Currently, one can consider the efforts required to develop LRs as the main limitation. In other words, the difficulty for some linguistic levels does not lie anymore in how to describe them but in actually achieve a description that has the coverage and precision required by complex NLP tasks. As a matter of fact, whoever has developed an LR knows that, in a reasonable amount of time, one can achieve a certain level of coverage and precision. However, as formalized by Zipf's law, increasing the quality of an LR becomes more and more difficult with time. Thus, the corresponding efforts follow a somehow exponential curve, i.e, the efforts are always more demanding when compared with the resulting improvements.

In order to tackle such problems, we propose an approach that relies on two complementary strategies: sharing the efforts among several people interested in obtaining those resources and saving manual efforts by automatizing the processes of creation and correction as much as possible.

4. Enhancing collaborative work

4.1. Problems limiting collaborative work

If a language receives enough attention from the community, the efforts to describe it by means of LRs can be shared among the people interested in building them. Nevertheless, the greater the workforce is, the more difficult it is to manage since it requires to find agreements on several non-trivial aspects.

Formalisms Nowadays, it is not rare to find various LRs describing a same linguistic description level of a given language. This happens mostly for two reasons.

First, the kind of data described in LRs generally depends on the application they have been created for. Therefore, one can find non-related but similar LRs covering the same sub-parts of a given level.

Second, the way a language is described can change when grounding on different linguistic theories. Therefore, there exist similar LRs that are (in part) incompatible.

In both cases, it implies a loss of manual work by formalizing several times a given knowledge and a waste of precious feedback by splitting the users over various LRs.

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

License and free availability The distribution and terms of use of LRs are issues both fundamental and problematic/polemic for their life-cycle. Indeed, since LRs are mostly built manually, they have a high cost. This fact often lead LRs to be distributed under restrictive licenses and/or to not be shared with the public. Obviously, such an approach presents the drawback to considerably limit collaborations and reduce the valuable feedback brought by a greater number of users.

Confidence Federating as many people as possible around a common LR does not make sense if the overall quality of the LR is reduced by some collaborators. Therefore, one usually needs to first demonstrate his or her competence before being granted the right to edit an LR. The resulting number of candidate collaborators is thus reduced to a small number of persons who have the linguistic and computer skills required for a shared edition of the LR.

Accessibility Obviously, someone willing to help maintaining an LR needs to access it. This basic statement is sometimes restrained by several reasons that can be technical (some restrictive technologies are required), geographical (the LR is not accessible from anywhere) or even security-related (the LR is located on a server restricted by security policies).

4.2. Guidelines to enhance collaborative work

The lexical formalism used for developing LRs should enable as wide a range of applications as possible, in particular by using general frameworks associated with tools (compilers) that are able to convert the general LR into specialized ones. Indeed, such an approach allows experts to develop and maintain specialized modules as independant modules, hence easing the life cycle of LRs and maximizing feedback. For example, one can develop a core lexicon for a language and provide several branches for developing specialized lexicons on zoology, medicine, etc. In addition, the more general the framework is, the more chance it has to be regularly maintained and updated itself.

Concerning licenses, it mostly depends on the main objectives of the developers of the LR. If the main objective is to bring the LR to a greater level of quality, one should try to maximize feedback and federate people with the skills to collaborate, be it academical or industrial. The licenses used should thus be as non-restrictive as possible.

As regards confidence, the main problem is that granting somebody edit rights on the LR generally means to grant such rights on the whole of it. A simple but straightforward approach to bypass this problem is to grant progressively edit rights on sub-part of the LR. Such a scalable approach can be achieved by designing interfaces with restrictions on what is editable or not according to the confidence level assigned to the user. In addition, interfaces can prevent editing/typing errors and allow users to focus on the data itself without worrying about mastering the underlying formalism or technologies. Finally, interfaces can help controlling more easily the evolution of LRs since they can allow to trace their modifications.

Regarding accessibility, web technologies are a convenient way to provide a direct access to LRs. Indeed, they are

among the most standardized online technologies and thus, are free of the technical, distance and security troubles mentioned above. When used to develop interfaces, they generaly constitute an appropriate way to access and edit LRs without any particular additional requirement.

5. Saving efforts

We have seen that it is important to try andd federate a community around an LR in order to increase the available workforce. But it is also necessary to try and reduce as much as possible the need for manual efforts. In order to achieve this goal, several tracks may be considered.

5.1. Using existing frameworks

Even if the NLP community did not release stable frameworks for all linguistic levels, most of them have been studied and (partial) solutions have emerged. Since existing frameworks are usually mature and the libraries/codes provided are often free of errors, a reasonable idea is to use them and, if necessary, extend them.

5.2. Using existing resources

Existing resources are generally valuable sources of linguistic knowledge when building new LRs or extending others. Of course, such an approach depends on the kind of knowledge one is trying to adapt and on the formalisms (and its underlying linguistic theory) the LR is based on. Nevertheless, LRs describing a similar level of language description usually share common points. Thus, adapting parts of the available existing resources is often an achievable objective.

Since related languages share significant parts of their linguistic descriptions, such an approach should not be limited to the scope of a single language. Indeed, the proximity between linguistically related languages can sometimes allow to "transfer" formalized knowledge. Thus, one should consider other existing LRs describing related languages. This approach is particularly useful for languages with smaller speech communities and limited digital resources.

5.3. Automatizing correction and extension

Techniques and tools for automatizing the processes of extension and correction are necessary for projects aiming at the construction of high-quality LRs. Often, LRs are built with little (or no) computer aid. This causes a common situation where the resources are developed until a (more or less) advanced state of development where it becomes too difficult to find errors/deficiencies manually. Since they can greatly reduce the need for manual work, these processes are fundamental for the sustainability of LRs.

Obviously, such techniques are specific for each type of linguistic knowledge. Some linguistic description levels (e.g., semantics) are more difficult to process with such an approach than others (e.g., morphology). As far as the morphological and syntactic levels are concerned, one can base a generic approach on research results such as those described in (Sagot and Villemonte de La Clergerie, 2006) and (Nicolas et al., 2008), as we now sketch.

Identifying possible shortcomings in an LR can be achieved by studying unexpected/incorrect behaviors of some tools relying on the resource. To do so, it is necessary to first establish what can be considered as an unexpected behavior. For example, for a parser, an unexpected behavior can be defined as a parse failure. Then, if among the elements of a given LR, some are found when unexpected behaviors occur more often than average, such element can be (statistically) suspected to be incorrectly described in the LR.

This "error mining" step, that already provides an interesting data to orientate the correction of the studied LR, can be completed with an automatic correction suggestion step. Contrarily to formal languages, natural languages are ambiguous and thus, difficult to formalize. Nevertheless, this ambiguity has the advantage of being randomly distributed on the different levels of a language. Consider two different LRs are interacting within an NLP tool (e.g., a syntactic lexicon and a grammar combined in a symbolic parser). This tool is designed to try and find a joint "match" between both resources and the input of the tool (e.g., a parse that is compatible with both the grammar and the lexicon). In other words, one can view each LR as providing a set of possibilities for each lexical unit in the input. Therefore, if of one of the LRs, say A, is suspected by the error mining step to provide erroneous and/or incomplete information on a given lexical unit, it is reasonable to try and rely on the information provided by the other LR, B, for proposing corrections to the dubious lexical entry. For example, let us suppose that a verbal entry in a lexicon A is suspected to provide a sub-categorization frame that is incomplete w.r.t. a given sentence. Using a parser that combines A with a grammar B, it is then reasonable to let the grammar decide which syntactic structures are possible for this sentence, by preventing the parser from using the dubious information provided by A about this verb. Then, correction proposals for A can be extracted from the subcategorization frame built by the parser.

Among the corrections generated thanks to B there might be correct and incorrect ones. Therefore, such approaches should generally be semi-automatic (i.e., with manual validation). Nevertheless, semi-automatic approaches are a good compromise to limit both human and machine errors since most of the updates done on the LRs are automatically created and manually validated.

Finally, another convenient feature of this approach is the following: if resource B cannot provide any longer relevant corrections for resource A, we can consider the remaining unexpected behaviors as mostly representing shortcomings of resource B. This defines an incremental and sequential way to identify sentences that instantiate shortcomings of resource B. Indeed, correcting resource A thanks to resource B generates useful data to correct resource B. Once resource B has been updated, it can be again used to correct resource A and so on.

5.3.1. Using plain text

The approach described in the previous section requires input corpora. They should be as error-free as possible in order to guarantee that most unexpected behaviors are caused by shortcomings of the LRs, and not by errors in the input. If this input data is an annotated one, only manual annotation can guarantee a certain level of quality. But manually

annotated data is only available in limited quantities for a small number of languages and producing such data contradicts the objective of saving manual work.
Therefore, the data used should be raw text, daily produced for most languages and freely available in large quantities on the Internet.
So as to guarantee the quality of the data, only linguistically correct (error-free) texts, such as law texts or selected journalistic productions, should be used while texts with a poor quality (emails, most blogs) should be discarded.

6. Results achieved by the *Victoria* project

Eventhough the *Victoria* project has not yet reach all its goals, the following results have been already obtained using the above-described guidelines as often as possible.
As regards to formalisms, we have chosen the Alexina framework (Sagot et al., 2006; Sagot, 2010) to develop our morphological and syntactic lexical resources. This framework, compatible with the LMF standard, represents morphological and syntactic information in a complete, efficient and readable way. It has already been used to create LRs for various languages (e.g., French, Spanish, Slovak, Polish, Persian, Sorani Kurdish) and has been combined with several taggers and various parsers based on a range of grammatical formalisms (LTAGs, LFG, Interaction Grammars, Pre-Group Grammars...).
Regarding grammatical knowledge, our resources rely on a meta-grammar formalism which represents the syntactic rules of a language by a hierarchy of classes. Even if in practice, we compile our grammars into a hybrid TAG/TIG parser (Villemonte de La Clergerie, 2005), this meta-grammar formalism is theoretically compilable into various grammar formalisms. Such a formalism is convenient in so far that it allows for an easy adaptation of an existing grammar to a linguistically related language.
As regards license issues, the LGPL-LR[14] and CeCILL-C[15] licenses have been chosen to publish our resources, namely our lexicons, grammars and editing interfaces.
Among the three kinds of resources developed, lexicons are clearly those requiring most collaborative work. The efforts concerning interfaces have thus been orientated to develop a web interface for lexicon based on the portlet technology. Its current version allows us to search for entries with complex logical equations covering any kind of data available in the lexicon. It also allows for a guided edition of the entries and traces every change.
Various techniques have been created or improved, in particular for achieveing the following tasks: (i) infering morphological rules from a morphological lexicon, (ii) extending a lexicon thanks to a tagger (Molinero et al., 2009), (iii) extending a lexicon thanks to morphological rules (Sagot, 2005), (iv) correcting a lexicon thanks to a grammar (Nicolas et al., 2008; Sagot and Villemonte de La Clergerie, 2006). Most of these techniques follows the guidelines described in section 5.3.
This alltogether allowed us to produce several LRs. Among them, two wide coverage lexicons for Spanish and Galician have already been produced along with two sets of morphological rules. The Spanish lexicon L*effe*[16] (Molinero et al., 2009) has been obtained by merging several existing Spanish linguistic resources, and also contains syntacic information. A Spanish meta-grammar (SPMG) has also been adapted from a French one (FRMG). For both L*effe* and *SPMG*, we took advantage of the similarity between French and Spanish language while building their first versions.

7. Conclusion

We have presented several guidelines to ease and improve the creation and correction of LRs. These guidelines are the cornerstone methodologies of a project dedicated to this task, the *Victoria* project. When considering the manpower involved in this project and the practical results it has achieved so far, we strongly believe that its guidelines might be of interest for anybody involved in a similar task.

8. References

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

Miguel A. Molinero, Benoît Sagot, and Lionel Nicolas. 2009. A morphological and syntactic wide-coverage lexicon for Spanish: The L*effe*. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 09)*.

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*, Manchester, UK.

Lionel Nicolas, Miguel A. Molinero, Benoît Sagot, Elena Sánchez Trigo, Éric de La Clergerie, Miguel Alonso Pardo, Jacques Farré, and Joan Miquel. 2009. Towards efficient production of linguistic resources: the *Victoria* project. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 09)*.

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

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

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. 2010. The L*efff*, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of LREC'2010*, Valetta, Malta.

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

[14] Lesser General Public License for Linguistic Resources.
[15] LGPL-compatible, `http://www.cecill.info/`.
[16] Léxico de formas flexionadas del español / Lexicon of Spanish inflected forms.