THE IMPORTANCE OF MT FOR THE SURVIVAL OF MINORITY LANGUAGES: SPANISH-GALICIAN MT SYSTEM

Inés Diz Gamallo

Centro Ramón Piñeiro para a Investigación en Humanidades[1] / University of Vigo (Spain)
Estrada Santiago-Noia, Km 3. A Barcia-Santiago de Compostela (Spain) / As Lagoas - Marcosende s/n, Vigo (Spain)
idiz@cirp.es

ABSTRACT

Our society is coming through a lot of changes that are connected, basically, with information. Maybe those languages that are present in this challenge will survive, and languages that will not meet those changes will dissapear.

The Linguistics section of the *Centro Ramón Piñeiro para a Investigación en Humanidades* (CRP) is devoted to the development of basic language resources for Galician for trying to solve the gap existing in computational resources and to made it possible for Galician to be present in the new information society. The aim of this paper is to explain how we have developed a Spanish-Galician Machine Translation system, what tools we have made use of, which difficulties we have found in our task and what are the final results of the project.

KEYWORDS

Minority languages, machine translation, basic computational resources, formal grammar, lexicon, corpora.

INTRODUCTION

In the information society in which we are living, the status of "major" and minority languages is very different. Our society is coming through a lot of changes, and these changes are connected, basically, with information. Maybe languages that are present in this challenge will survive, and languages that will not meet those changes will disappear. It sounds dramatic, but we think that this is very likely to happen.

It is for that reason that minority languages have to be ready to assume this challenge, even though they are in disadvantage with respect to "major" languages.

To overcome this challenge, minority languages need to have not only basic computational resources (lemmatizers, lexicons, corpora, parsers....), but also some other more elaborated resources, such as information retrieval systems, voice recognizers/synthetizers or machine translation systems (Somers, 1997)

The Galician language is a minority language. Despite its similarity to Portuguese, it is only a few years ago since a machine readable dictionary and a spell checker are available, and no grammar checker has yet been made.

But this is not the only problem. There are other important handicaps, like the lack of studies (non computational studies, but general studies) on some aspects of Galician language, like syntax or semantics in which anyone can base his computational work. As you could imagine, in a situation like this (similar to another minority languages), the task to generate computational resources is necessary harder.

The Linguistics section of the *Centro Ramón Piñeiro para a Investigación en Humanidades* (CRP), directed by Guillermo Rojo, is devoted to the development of basic language resources for Galician, such as taggers, parsers, corpora and lexicons for trying to solve this gap and to made it possible for Galician to be present in the new information society. The aim of this paper is to explain how we have developed a Spanish-Galician Machine Translation system, what tools we have made use of, which difficulties we have found in our task and what are the final results of the project.

THE SPANISH-GALICIAN MACHINE TRANSLATION SYSTEM

In the end of 1998 the *Centro Ramón Piñeiro para a Investigación en Humanidades* started to work on the setting up of a Machine Translation system between Spanish and Galician which was ready for testing by May 2000 (Diz & Martínez, 2000)

The staff involved in the project development consisted of three full-time people: one of them was in charge of implementing the formal grammars and of coordinating the whole process; the other two were scholarship holders whose time was devoted to build the lexicons.

[1] "Ramón Piñeiro" Centre for Humanities Research

Taking advantage of the linguistic similarity between the two languages in issue, the MT system we aimed to build up was not restricted to a specific domain, rather it was oriented to manage general language[2] and, at least for the moment, it is conceived for managing written rather than oral language.

The machine translation technique applied was transfer-based because, as we have just said, Spanish and Galician are very related languages and the transfer architecture was considered the most suitable to profit from this fact. Among the different existing transfer systems, we used a direct descendant of the one-time METAL system that was being used by the Catalonian company Incyta (vid. Hutchins and Somers, 1992).

The platform in which it was developed is Sun OS, although the exploitation runs in Sun Solaris.

Apart from grammars for analyse, transfer and generate the languages involved, in order for the MT system to run, it was necessary to have two lexicons: Spanish-Galician and Galician monolingual.

The Spanish analysis was provided by the Catalonian company Incyta[3], as they had developed this module for the Spanish - Catalan system.

The tasks developed at the *Centro Ramón Piñeiro para a Investigación en Humanidades* were, on the one hand, transferring grammar rules between Spanish and Galician and generating the Galician grammar (two absolutely new things for Galician until now) and, on the other hand, creating the bilingual and monolingual lexicons.

THE GRAMMAR RULES: SPANISH-GALICIAN TRANSFER AND GALICIAN SYNTHESIS.

The transfer stage was the most difficult one. The most important characteristics of working in computational linguistics with minority languages like Galician is not only the lack of electronic resources that could be handled, but and above all, the absence of "traditional" descriptive linguistic studies on Galician syntax and semantics which could be used by the linguist to base his or her implementations on.

As there are not many syntactic and semantic studies on Galician language, it is easy to imagine the almost absolute absence of contrastive Spanish-Galician studies on these subjects which would have been of great help for our work (vid. García Gondar et al., 1995; Regueira et al., 1996)

This fact will determine the subsequent work. So, for the development of the Spanish-Galician MT system, we have necessarily devoted a lot of time and effort to this previous task, not only to obtain electronic resources, but also to improve the knowledge of the Galician grammar (and, then, implement the linguistic phenomena)

At this stage, the use of corpora was crucial to our work. The lack of significant syntactic studies forced us to face up directly the data and to design, in a way, the rules that we have implemented *ex novo*.

As a consequence, we have developed this grammatical work in the following ways: first of all, we have consulted the few (and sometimes deficient) studies on our language available at the moment. Then, we looked up in a Spanish corpus [4] the phenomena we wanted to be implemented in Galician, in order to verify that these phenomena could really be implemented as it was recommended in the Galician grammar. But, as native speakers of the language, we found that the descriptions offered by Galician grammatical studies were not always right, as they only took into account a few examples -the prototypical ones- while the others were left out with no solution.

At this point our work consisted in the proposal and subsequent verification of hypothesis, translating texts which had the phenomena we wanted to test. Needless to say, the Galician corpus CORGA -*COrpus de Referencia do Galego Actual/Reference Corpus of Current Galician*[5]- (vid. García Mateo & González González, 1998) provided us with a great aid in order to validate the linguist intuition when we had to implement the transfer and generation phenomena.

We think that the task developed at this stage is very valuable (but also, of course, incomplete and improvable), because it lays on the foundations of a future Galician formal grammar. Now, we know better the structure of the Galician language and we have learnt how to communicate this to a machine. We have studied and implemented a lot of specific phenomena and this was a hard task that took us a lot of time and work. But it was worth while because it is an important empirical foundation for forthcoming computational studies on Galician language.

The final result of the work on formal grammars is more than 7,500 code lines written in LISP

Although Spanish and Galician, as all romance languages, are very similar at syntactic level, it is also true that there are several important differences that separate these two languages. The position of unstressed pronouns in the clause and the lack of compound verb tenses in Galician are two simple examples of this fact. The placement of the clitics is completely different in both languages and their position in Galician depends on several factors that, at the end, are connected with pragmatics, and that is why it is not always easy to formalize some rules about this specific subject[6]. The absence of compound tenses in the

[2] For example, newspaper's language and similar.

[3] recently disapeared

[4] We used the CREA (*Corpus de Referencia del Español Actual*), developed at the Real Academia Española, which is available on the web at http://www.rae.es/NIVEL2/recursos.htm [visited: 28-VI-01]

[5] The CORGA (*COrpus de Referencia do Galego Actual*) is an on-going project that is also being developed at the RPC. It contains oral and written texts of different types (narrative, journal, theatre plays, essay, spontaneous oral texts and news oral texts). Currently, the number of words in the CORGA is about 15 million.

[6] It is true that some of these rules can be found in Galician grammar books (Álvarez, 1986), but as we have checked them,

verb and the non unanimous correspondence between one tense in Spanish and one tense in Galician could be another example. For instance,

Cuando ***haya acabado****, avísame* → *Cando* ***remate****, avísame*
No creo que ***haya acabado*** → *Non creo que* ***rematase***

We could keep on providing examples of differences between the two languages in this list, but this is not our aim. We have pointed out some of them only to show that, despite their similarity, there are also important grammatical differences between them that should be solved in the formal grammar implementations.

THE LEXICONS

It has been also necessary to have two lexicons available. While creating these lexicons we have really obtained the most on the resources we had previously developed at the CRP.

THE BILINGUAL SPANISH-GALICIAN LEXICON

This task was carried out in three steps: in the first one we used the work which had been done for the GalWordNet, a project that is also being developed at the RPC (Magán, 1998). We selected about one thousand entries which had a Spanish-Galician equivalent and we introduced them automatically in the dictionary. All these entries were hand-validated.
In the second step, we took all those words that still had no correspondence and we assigned them one Galician entry.
Galician is a language that has not completed its corpus planning yet, so in many cases it was impossible to find the words we needed when looking up in the dictionary. It was for this reason that, at this point, the CORGA was, again, a very useful tool, because whenever we did not find a word in the dictionaries, we found it in the Galician corpus.
The third and last step in the building of the bilingual dictionary was the delimitation and implementation of the Spanish entries that have more than one equivalence in Galician. Usually the correspondences are conditioned by certain contexts that we have to find out. In order to do this, we have used the CREA Spanish corpus and the CORGA Galician one. Once we have delimited the contexts, we implemented them.
The final result is a bilingual lexicon with 45,000 entries.

THE GALICIAN MONOLINGUAL LEXICON

The bilingual lexicon is very important for the transfer stage. The existence of a monolingual lexicon that could be used at the generation stage is as important as the bilingual one. This Galician lexicon must have not only morphological, but also syntactic and semantic information.
At this point, we could profit from one lexicon we have previously developed at the RPC, for a morphological analyser and lemmatizer (Vilares et al., 1998). This lexicon consisted of about 20,000 lemmas, all of them with their part of speech and flexional pattern. What we did was to map the different formats in which the information was codified to use all this information automatically for the MT system. Once we obtained these entries with morphological and lemma information, we carried on, at this stage, by enriching this lexicon: on the one hand, with syntactic and semantic information and, on the other, with more entries. Due to this we have now a monolingual dictionary with 45,000 lemmas with morphological, syntactic and semantic information

THE RESULTS OF THE PROJECT

The system has been tested during one year. The texts we have used were texts coming from on-line newspapers (especially, but not only, from the newspaper *El País*[7]). We have been translating automatically this journal every day, from Monday to Thursday, and we were checking the mistakes we found. We decided to stop the test for general language translation when we noticed that the results were satisfactory and that it was not worth correcting the mistakes still remaining (because we have to invest a lot of time on their resolution and they were not very likely to be used in Spanish texts).
We have confirmed that the accuracy of the translation is 95% for general language coming from newspapers. Sometimes, no mistakes were found. The translation speed is very high and the MT system maintains the text formats if they come from MS-Word, rtf or html.
At this moment, we are working with another domains, especially administrative language.

HOW IS THE SPANISH-GALICIAN MT SYSTEM GOING TO BE USED

This research was financed by the Galician Government. So, the first use the MT system had was an internal use by the Galician Administration. All documents produced by the Galician Government and its dependant organisms (Town-Halls, Provincial Delegations, etc) can be obtained in Galician.
Another use that is being considered is Galician newspapers. There are several regional journals and news agencies interested in this MT system. It would allow them to have two editions of the same newspaper: one in Spanish and one in Galician. At this moment the Galician Government is studying the details of using several licences of the system by external users
There is another use that has been considered when the project started: the individual use of the MT system via Internet, but for the moment it is not running.

we found that there are several cases in which these rules are not true, and in some other cases they have not been covered by these rules.

[7] http://www.elpais.es [visited: 28-VI-01]

CONCLUSION

In this paper we have presented the Spanish-Galician MT system developed at the RPC. We have showed how we have re-used several electronic resources that have been developed for the Galician language before, and how we took advantage of them while building up a MT system.

But the most important thing we would like to emphasize is the importance of computational tools like MT systems for the future of a minority language as Galician. Galician presence is necessary in the current information society. We need to have Galician texts on the web and at the disposal of everybody and, especially, at the disposal of every Galician person who wants access in his own language to any document that only exists, by the moment, in Spanish.

We hope we can develop in the future an English-Galician MT system. It depends on governmental decisions, but we think that the existence of such tools can be very useful (and in some cases absolutely necessary) for the survival of minority languages in an information society like the one in which we are living.

REFERENCES

ÁLVAREZ, R., MONTEAGUDO, H. & REGUEIRA, X. L. (1986). Gramática Galega. Vigo: Galaxia

DIZ, I., and L. MARTÍNEZ (2000). The Spanish-Galician and Galician-Spanish MT system: How to re-use the existing Galician resources to develop a robust MT system in a short period of time. In Proceedings of the Workshop on Developing Language Resources for Minority Languages: Reusability and Strategic Priorities (pp. 23-29). Second International Conference on Language Resources and Evaluation. Athens.

GARCÍA-MATEO, C. & GONZÁLEZ GONZÁLEZ, M. (1998). An overview of the existing language resources for Galician. Proceedings of the Workshop on Language Resources for European Minority Languages. First International Conference on Language Resources and Evaluation. Granada.

GARCÍA GONDAR, F. et al. (1995). BILEGA: Repertorio bibliográfico da lingüística galega: desde os seus inicios ata 1994 inclusive. Santiago de Compostela: Centro Ramón Piñeiro. Available on line at http://www.cirp.es/WXN/wxn/homes/bilega.html

HUTCHINS, J. & SOMERS, H. (1992). An introduction to Machine Translation. New York: Academic Press Ltd.

MAGÁN MUÑOZ, F. (1998). Towards the Creation of New Galician Language Resources: From a Printed Dictionary to the Galician WordNet. Proceedings of the Workshop on Language Resources for European Minority Languages. First International Conference on Language Resources and Evaluation. Granada

SOMERS, H. (1997). Machine translation and minority languages, Paper presented at Aslib's Translating & the Computer Conference in November 1997. Available on line at http://www.ling.lancs.ac.uk/monkey/ihe/mille/1fra1.htm [visited: 28-VI-01]

REGUEIRA FERNÁNDEZ et al. (1996), Guía bibliográfica de lingüística galega. Vigo: Xerais.

VILARES, M, GRAÑA, J., ARAÚJO,T, CABRERO D. & DIZ, I. (1998). A Tagger Environment for Galician. Proceedings of the Workshop on Language Resources for European Minority Languages. First International Conference on Language Resources and Evaluation. Granada.