Strategies to develop Language Technologies for Less-Resourced Languages based on the case of Basque

Iñaki Alegria, Xabier Artola, Arantza Diaz de Ilarraza and Kepa Sarasola

Ixa Taldea. University of the Basque Country
{i.alegria,xabier.artola,jipdisaa,kepa.sarasola@ehu.es}

Abstract

IXA group has developed during 23 years a basic set of resources, tools and applications for Basque following to an initial strategy which has been adapted according to technological changes. We think that our strategy and experience can be a reference for other less resourced languages. According to a six level classification of world languages, we estimate that this strategy may be useful for several hundred languages, those that have developed a written standard but that still are beginners in Human Language Technology.

Keywords: Language resources, Language Technology applications, Less-Resourced Languages, Strategy for Language Technology development

1. Introduction

IXA group is a research group created in 1988 with the aim of laying foundations for research and development of Natural Language Processing (NLP) and Human Language Technology (HLT) for Basque. We wanted to face the challenge of adapting Basque to HLT.

Adapting Languages to HLT is a need, but without an active ICT community just HLT sector efforts will not be enough to ensure the digital survival of a language. However the strategy described here does not directly face the promotion of other related ICT technological issues as localization of standard tools, publishing of digital contents (i.e. entries in Wikipedia), and so on. So additional efforts should be done to organize and coordinate an active ICT community.

According to a six level language typology we propose, we estimate that our strategy to develop language technologies could be useful for several hundred languages, those that have developed a written standard but that are still very far from the first wagon of languages in the train of HLT resources.

2. Languages and resources

In order to know how languages are facing the ICT and HLT challenges, statistics about present Internet resources for each language would be very useful to detect different typologies, but unfortunately, figures about amounts of resources on the Internet for different languages are not easy to obtain. So, to draw a first draft of a classification of world languages, we should use more specific public rankings showing data on Internet users, Internet documents and Wikipedia's articles.

- Internet World Stats[1] provides a list with the **number of Internet world users** for the top 10 languages in 2010: English, Chinese, Spanish, Japanese, Portuguese, German, Arabic, French, Russian and Korean. Unfortunately only the data for these top ten languages are public in that website. This website estimates that the amount of users for the rest of the languages is not greater than 17.8% of the total, even though these languages are spoken by 36% of the world population. That means that among each 6 Internet users 5 of them use one of these top 10 languages.
- Reliable statistics about **number of Internet documents** for different languages are scarce. A study on the presence of Romance languages on the Internet[2], showed that 45% of the webpages were written in English, 5.9% in German, 3.80% in Spanish, 4.41% in French, 2.66% in Italian, 1.39% in Portuguese, 0.28% in Romanian, and 0.14% in Catalan (figures in 2007). Alternatively we can obtain figures for a language using APIs of search engines; it is simple when the language is included in the repertory of languages recognized by the engine and more complex when it is not. These kinds of techniques are used in the "Web as a Corpus" area (Kilgarriff & Grefenstette, 2003).
- Reported by Wikimedia[3] it can be observed (on-line) the **number of articles in Wikipedia** for each language. In October 2011 there were articles in 282 languages. The top 10 languages are the following: English, German, French, Italian, Polish, Spanish, Dutch, Russian, Japanese, and Portuguese.

  Chinese, Arabic and Korean are not in this second top list, instead of them Polish, Italian and Dutch are included. Surprisingly Catalan is the 13th Wikipedia language, Esperanto the 27th, and Basque the 36th .

·The most accessible of these three indicators is the third one because it is automatically updated for all the languages. The first indicator is the most suitable to measure the impact of new HLT resources, while the

[1] http://www.internetworldstats.com/stats7.htm

[2] http://dtil.unilat.org/LI/2007/ro/resultados_ro.htm

[3] http://meta.wikimedia.org/wiki/List_of_Wikipedias

second (unfortunately there are not reliable stats) and the third indicators would be more adequate to measure the activity degree of the speakers.

If we look for figures about HLT resources for different languages we can consult several public repositories, of course, bearing in mind that these information sources are not always complete (repositories refer to the products they offer, and the wiki-like sites only to those entered by volunteers), and so, they can not be used for rigorous comparisons. The first two sites manage resources and sell some of them; the two others are just for consulting:

- **ELRA**[4]: European Language Resources Association. This repository includes a list with more than 1000 resources for 60 languages that are distributed by ELRA agency (some products are free for research). The list includes 6 products for Basque. Recently ELRA has added a "universal catalog" with information about other products not distributed by ELRA. The catalog does not offer "Search by language" functionality. Recently ELRA created *The Universal Catalogue,* a new repository allowing for a collaborative enriching and comprising information regarding Language Resources (LRs) identified all over the world.
- **LDC**: Linguistic Data Consortium[5]. About 82 languages and more than 500 resources. Search by language is allowed. No products for Basque are described in it.
- **ACLWiki**[6]: there is a list of resources and tools built in a wiki-like way for 73 languages. The list includes 15 products for Basque.
- **NLSR**: Natural Language Soft Registry[7]. This repository includes software and resources for 30 languages, and it is managed by DFKI. Searching for Basque, it shows 3 specific products for Basque and other 59 products useful for "any language".

Additionally the ***yourdictionary.com*** website[8] presents links to on-line lexical resources for 307 languages. The set of links is not completely updated (for example, it includes only 5 links to Basque resources, but more than 40 are included in Hiztegia.net[9] a website specialized in collecting such links for Basque), but it is a good reference to look for and to compare existing lexical resources for different languages.

Another indicator is the penetration of each language in the most popular linguistic services; for example, the presence/absence of the languages in word processing, search engines and machine-translation engines.

- The most popular word-processor is localized for around 91 languages and dialects[10]. Libre*Office* reports 104[11]. Basque is in both.
- The most popular search engine[12] includes language identification for 45 languages (see advanced search) where the Basque is not.
- The two most popular MT engines in Internet are BabelFish[13] and Google-Translate[14]. The first manages 13 languages and the second 63 including a a beta version for Basque.

Then, after this survey on possible indicators we can try to answer to our original question: *When a language is less-resourced?* Of course, the answer is relative, and for that we distinguish six different sets of languages:

1. First level: **English**. It is the language of 37.9% of the users of Internet. 45.00% of the web pages are written in English. 62% of the HLT resources described in LDC are available for English, 51% in ELRA. Almost all the HLT applications are available for English.
2. Second level: other **languages in the top 10 languages used in the web**. The top 10 languages cover 82.2% of the Internet users (55.4% excluding English). There are the languages for which active LR development continues and most major categories of HLT are represented. Most of the HLT kind of resources described in LDC or ELRA are available for those languages (45,79 % for German, 41,27 % for French, 40,76% for Spanish; 36,24% for Italian, and 31,31 % for Portuguese). Streiter et al. (2006) use the term "central languages" to refer to this set of languages.
3. Third level: around 70 **languages with any HLT resource** registered. There are 60 languages in ELRA, 82 in LDC, 73 in ACLWiki and 30 in NLSR.
4. Fourth level: around 300 **languages with any lexical resource on-line** registered in *yourdictionary.com (307 languages)*. Almost the same set of languages that is present in Wikipedia (282 languages).
5. Fifth level: here are included other 2,014 **languages that have writing systems** (Borin, 2009).
6. Sixth level: the big bag also including **only-spoken languages** in the world (more than 4,500).

This 6 level typology gives a relative definition of less-resourced languages, not an absolute definition, of course. Comparing with English all the other languages could be considered less-resourced, or we could say that except the 10 languages in the two first levels the rest can be considered less-resourced. The languages of the third level are lesser resourced than the languages of the second level, by definition, but we may consider that the situation of the languages in the 5th and the 6th levels are really endangered, and the 3rd or the 4th are the levels of languages usually called as less-resourced in the HLT domain.

This classification is not strict, but it may be useful to recognize application domains (sets of languages) for possible different strategies in the development of HLT

[4] http://www.elra.info/

[5] http://www.ldc.upenn.edu/Catalog/catalogSearch.jsp

[6] http://aclweb.org/aclwiki/index.php?title=List_of_resources_by_language

[7] http://registry.dfki.de/

[8] http://www.yourdictionary.com/languages.html

[9] http://www.hiztegia.net/

[10] http://www.microsoft.com/unlimitedpotential/programs/llp.mspx

[11] http://www.libreoffice.org/download/

[12] http://www.google.com

[13] http://babelfish.yahoo.com

[14] http://translate.google.com

really endangered, and the 3rd or the 4th are the levels of languages usually called as less-resourced in the HLT domain.

This classification is not strict, but it may be useful to recognize application domains (sets of languages) for possible different strategies in the development of HLT resources. However, there are some risks on the application of these indicators: languages with very active proponents may have a high visibility on Wikipedia which may not be significative of the presence of the language on the general Internet. For example, Catalan appears in the 13th position in the ranking of the number of articles in Wikipedia, but Catalan is usually taken as a less resourced language; in fact, many papers on the automatic processing of Catalan are submitted to the SALTMIL workshops[15] (HLT for minority languages). Nevertheless the Wikipedia indicator is highly accessible, it is automatically updated for all the languages, and it is useful when used in conjunction with other indicators.

3. Strategy to develop Language Technology

IXA group is a research group (ixa.si.ehu.es) created in 1988 by five university lecturers in the Computer Science Faculty of the University of the Basque Country with the aim of laying foundations for research and development of NLP software mainly for Basque. Our aim was to face the challenge of adapting Basque to language technology.

Now, twenty three years later on, IXA is a multidisciplinary group composed by 31 computer scientists and 10 linguists. It works in cooperation with more than 7 companies from Basque Country and 5 from abroad; it has been involved in the birth of two new spin-off companies; and there are several products of language technology we have built.

In recent years, several private companies and technology centers of the Basque Country have begun to get interested and to invest in this area. At the same time, more agents have come to be aware of the fact that collaboration is essential to the development of language technologies for minority languages. Fruits of this collaboration were the HIZKING21 project (2002-2005)[16], ANHITZ project (2006-2008)[17] and BERBATEK (2009-2011)[18]. These projects were accepted by the Government of the Basque Country in the framework of a new strategic research line called ‘Language Info-Engineering’.

At the very beginning, our first funding was associated to the creation of a translation system for Spanish-Basque. After some preliminary studies we realized that it was more important to concentrate our efforts in creating basic tools and resources for Basque (morphological analyzer/generator, syntactic analyzers …) that could be used later on to build general language application rather than creating an *ad hoc* MT system with probably small accuracy.

This thought was the seed to design our strategy to make progress in the adaptation of Basque to Language Technology. This way we could face up to the scarcity of the resources and tools that could make possible the development in Language Technology for Basque at a reasonable and competitive rate.

We presented an open proposal for making progress in HLT (Aduriz et al., 1998). Anyway, the steps proposed did not correspond exactly with those observed in the history of the processing of English, because the resources available for the treatment of the language allowed facing problems in a different way, and because English LRs did not evolve as the result of a single coordinated plan. Instead many independent efforts produced these English LRs to address specific project needs.

Our strategy focuses on two crucial points:

1) Need of **standardization** of resources to be useful in different researches, tools and applications.

2) Need of **incremental design and development** of language resources, tools, and applications in a parallel and coordinated way in order to get the best benefit from them. Language foundations and research are essential to create any tool or application; but in the same way tools and applications will be very helpful in the research and improvement of language foundations.

Following this, our steps on standardization of resources brought us to adopt TEI and XML standards as a basis for linguistic annotation at the different levels of processing, and also to the definition of a general methodology for corpus annotation (Artola et al., 2009).

In the same way, taking as reference our experience in incremental design and development of resources/tools, we propose four phases as a general strategy for language processing (Alegria et al., 2011):

1. Initial phase: Establishing foundations. First compilation of a Corpus (collection of raw text without any tagging mark). Design and implementation of our lexical data-base (EDBL) that will be the base for much of the tools and applications. Initial set of machine-readable dictionaries. Definition of the morphological description of Basque.
2. Second phase: Developing basic tools and applications. Enhancement of the corpus in such a way that word-forms are tagged with their part of speech and lemma. Enrichment of the lexical database with information about part of speech and morphology. Morphological analyzer, lemmatizer/tagger. Spelling checker and corrector (although in morphologically simple languages a word list could be enough, in Basque we can not take this approach). Implementation of statistical tools for the treatment of corpus.
3. Third phase: Advanced tools and applications. An environment for tool integration. Enhancement of the corpus with syntactic information. Enrichment of the lexical database with information about multiword

[15] http://ixa2.si.ehu.es/saltmil/

[16] http://www.elhuyar.org/hizkuntza-zerbitzuak/EN/Hizking-21-project

[17] http://www.elhuyar.org/hizkuntza-zerbitzuak/EN/Anhitz-project

[18] http://elhuyar.org/hizkuntza-zerbitzuak/EN/Berbatek-(2009-2011)

lexical units, semantic information. Lexical-semantic knowledge base. Creation of concept taxonomy (e.g.: Wordnet). Description of surface and deep syntax. Grammar and style checkers. Word-sense disambiguation. Search machines that integrate lemmatization and language identification. Structured versions of dictionaries that allow enhanced functionality not available for printed or raw electronic versions. Integration of dictionaries in text editors. First integration of the resources and tools created so far in Computer Aided Language Learning (CALL) systems.

4. Fourth phase: Multilingualism and general applications. Information retrieval and extraction. Question/Answering. RBMT and SMT Machine Translation System development and Translation aids (integrated use of multiple online dictionaries, translation of noun phrases and simple sentences). Corpus IV (semantically tagged annotation of senses, argument-structure of sentences). Extraction of information based on semantics. Anaphora resolution and study of discourse markers.

The strategy presented established a good position to adopt those initiatives emerging during the last years such as: i) BLARK, Basic Language Resources Kit (Krauwer, 2003). Its aim was the definition of the minimal set of language resources necessary to do any precompetitive research and education, ii) CLARIN (Váradi *et al.* 2008), an interoperable research infrastructure of language resources and language technology that would allow to offer a stable, persistent, accessible and extendable infrastructure for the research in eHumanities; iii) META-NET Network of Excellence[19] that will set up the basis for a multilingual European information society facilitating the construction of advanced applications that enable automatic translation, multilingual information and knowledge management and content production across all European languages.

Besides, the success of the open source initiatives and the 2.0 communities came later. Now they are important instruments for a rapid and sustainable development of resources. Using open-source programs is a key factor of success, because efforts are not repeated and because there is a more or less widespread making contribution. Developing open-source code is more difficult and laborious, because it is necessary to structure the programs and prepare good documentation (in English). Simultaneously this is a key factor of quality and so, sustainability. Thus, using tool version control systems as SVN[20] and public repositories[21] brings us to a better methodology and so, easer reuse. However, arranging communities to help in enriching resources for less-resourced languages is not an easy task, without a substantial critical mass of collaborators this kind of processes is inviable.

To finish we will talk about what shouldn't be done when working on the treatment of languages with scarcity of resources.

- Do not start developing applications if linguistic foundations are not defined previously; we recommend following the above given order: foundations, tools and applications. This is a basic guarantee to face the next steps.
- When a new system has to be planned, do not create *ad hoc* lexical or syntactic resources; you should design those resources in a way that they could be easily extended to full coverage and that they could be reusable by any other tool or application. Sometimes competitive research will draw you to rapid development of ad hoc resources that will not be sustainable and not reusable.
- When implementing a new resource or tool, do not keep it to yourself; there are many researchers working on English, but only a few on each less resourced language; thus, the few results should be public and shared for research purposes, it is desirable to avoid needless and costly repetition of work. In this way open source and open content solutions are the best.

4. Related work

The aim of the paper is to describe a strategy to be used when developing HLT for a language and to help to researchers/technicians when they have to work on it for a less resourced language. Other experiences and proposals have been reported.

The book *Corpus linguistics around the world* (Wilson et al., 2006) describes many corpus resources on several languages.

Our colleagues and us (Agirre et al. 2002) used the term "Basic toolkit for HLT" while Krauwer (2003) proposed a "Basic LAnguage Resource Kit (BLARK)" as a roadmap of tools to be developed for each language using the terminology defined in a joint initiative between ELSNET (European Network of Excellence in Language and Speech) and ELRA (European Language Resources Association) in 1998. In all these works a list of basic resources and tools are listed. Maegaard et al. (2004) describe a BLARK for Arabic and Simov et al. (2004) for Bulgarian. The term BLARK has been very successful and it is used in a large number of papers in the area.

Streiter et al. (2006) report on HLT projects for noncentral languages and proposes instructions for funding bodies and strategies for developers. They use the *non-central* term and underline the importance of making use of free software to improve the results. The chapter about benefits and unsolved problems when using open source software for non-central languages is very interesting. Forcada (2006) remarks the opportunity of using open source machine translation for minor languages.

The ELSNET network of excellence prepared definitions for a language resources and evaluation roadmap[22], using for that the HLT Roadmap System, a

[19] http://www.meta-net.eu/mission

[20] http://subversion.tigris.org/

[21] http://sourceforge.net/ is the most popular

[22] http://elsnet.dfki.de/roadmap.php

framework for implementing technology roadmaps (Busemann & Uszkoreit, 2004). So far different aspects of HLT are looked at without claiming for completeness or a wide consensus.

In this context several different roadmaps have been published[23]. As in our first proposal in 2002 the elements in the diagram (HLT products) are classified into three equivalent subsets: (Language Resources / Language Processing / Language Usage) in their roadmap, and Language resources/ Language Tools / Language Applications) in our strategy. But their level of granularity in the diagram elements is very much fine than ours, being their objective the definition of a roadmap for “central languages”, mainly for the main European official languages; while our strategy is devoted to robustly manage the first steps in the development of HLT for a less resourced language.

Borin (2006 and 2009) points to the promise of the HLT for lesser-known languages and describes the linguistic diversity in the information society. He cites the paper from Ostler "a *language will not get by in the world of today unless it is equipped with a parser and a multi-million-word corpus of text*". He analyzes the relation among the sociology of language and HLT, and guises us some strategic considerations, i.e. "*those languages for which information extraction resources and tools will be available will probably exhibit a more secure and prominent presence on the Semantic Web than those lacking such resources, and as a consequence, acquire the status in the eyes of their speakers that such a presence confers*".

Initiatives as Clarin[24] and Flarenet[25] MetaNet[26] try to coordinate collaborative efforts to create, coordinate and make language resources and technology available and readily usable for a big number of languages.

SALTMIL ("Speech And Language Technology for Minority Languages") has been organized seven conferences[27] related to HLT and less-resourced languages.

5. Conclusions

A language that seeks to survive in the modern information society requires language technology products. Non-central languages have to do a great effort to face this challenge. In that way, Ixa group has been working since 1988 in adapting Basque to language technology, having developed several applications that are effective tools to promote the use of Basque.

From our experience we defend that research and development for less resourced languages should to be faced to build a BLARK following this points: 1) high standardization, 2) open-source, 3) reusing language foundations, tools, and applications, and 4) incremental design and development of them.

We estimate that most languages can be considered as less resourced languages from a HLT point of view. In that way we have defined six different sets of languages attending to their penetration on HLT technologies. We think that our strategy to develop language technologies could be useful for several hundred languages, those that have developed a written standard and perhaps also some initial lexical resources but that are still very far from being a central language.

We know that any HLT project related with a less privileged language should follow those guidelines, but from our experience we know that in most cases they do not. We think that if Basque is now in an good position in HLT is because during the last twenty years those guidelines have been applied even though when it was easier to define "toy" resources and tools useful to get good short term academic results, but not reusable in future developments. Czech is another exception to the correlation between language size and LR scarcity; the excessive rich body of LRs for Czech is due to the coordinated efforts of a few ambitious and productive researchers.

Acknowledgments

This research was supported by the Spanish Ministry of Education and Science (OpenMT2, TIN2009-14675-C03-01) and by the Basque Government (BERBATEK, IE09-262). We would like to thank the two anonymous reviewers who helped improve this paper.

References

Aduriz, I., Agirre, E., Aldezabal, I., Alegria, I., Ansa, O., Arregi, X., Arriola, J., Artola, X., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Maritxalar, M., Oronoz, M., Sarasola, K., Soroa, A. and Urizar. R. (1998). A framework for the automatic processing of Basque. *Proceedings of Workshop on Lexical Resources for Minority Languages.*

Alegria, I., Aranzabe, M., Arregi, X., Artola, X., Díaz de Ilarraza, A., Mayor, A. and Sarasola, K. (2011). Valuable Language Resources and Applications Supporting the Use of Basque. Z. Vetulani (Ed.): LTC 2009, Lecture Notes in Artifitial Intelligence LNAI 6562, pp. 327--338. Springer, Heidelberg.

Artola, X., Diaz de Ilarraza, A., Soroa, A. and Sologaistoa, A. (2009). Dealing with Complex Linguistic Annotations within a Language Processing Framework. IEEE *Transactions on Audio, Speech, and Language Processing*. Vol 17, number 5. Pages 904-915.

Borin, L. (2009). Linguistic diversity in the information society. *SALTMIL2009 Workshop: IR-IE-LRL Information Retrieval and Information Extraction for Less Resourced Languages*. University of the Basque Country ISBN 978-84-692-4940-6

Streiter, O., Scannell, K.P., Stuflesser, M., (2006). Implementing NLP projects for noncentral languages: instructions for funding bodies, strategies for

[23] http://elsnet.dfki.de/roadmap.phpversion=LREC_2004

[24] http://www.clarin.eu/

[25] http://www.flarenet.eu

[26] http://www.meta-net.eu

[27] http://ixa2.si.ehu.es/saltmil/eu/activities/workshops/workshops.html