Strategies for sustainable MT for Basque: incremental design, reusability, standardization and open-source

I. Alegria, X. Arregi, X. Artola, A. Diaz de Ilarraza, G. Labaka, M. Lersundi, A. Mayor, K. Sarasola

Ixa taldea. University of the Basque Country.

i.alegria@ehu.es

Abstract

We present some Language Technology applications that have proven to be effective tools to promote use of Basque, a low density language. We also present the strategy we have followed for almost twenty years to develop those applications as the top of an integrated environment of language resources, language foundations, language tools and other applications. When we have faced a difficult task such as Machine Translation to Basque, our strategy has worked well. We have got good results in a short time just reusing previous works for Basque, reusing other open-source tools, and only developing few new modules in collaboration with other groups. In addition new reusable tools and formats have been produced.

1 Introduction and Basque Language

Basque is both a minority and a highly inflected language with free order of sentence constituents. Machine Translation for Basque is thus both, a real need and a test bed for our strategy for developing NLP tools for Basque.

Basque is a language isolate, and little is known of its origins. It is likely that an early form of the Basque language was already present in Western Europe before the arrival of the Indo-European languages.

Basque is an agglutinative language, with a rich flexional morphology. In fact for nouns, for example, at least 360 word forms are possible for each lemma. Each one of the grammar case as absolutive, dative, associative… has four different suffixes to be added to the last word of the noun phrase. These four suffix variants correspond to undetermined, determined singular, determined plural and "close" determined plural.

Basque is also an ergative-absolutive language. The subject of an intransitive verb is in the absolutive case (which is unmarked), and the same case is used for the direct object of a transitive verb. The subject of the transitive verb (that is, the agent) is marked differently, with the ergative case (shown by the suffix -k). This also triggers main and auxiliary verbal agreement.

The auxiliary verb, or periphrastic, which accompanies most main verbs, agrees not only with the subject, but with the direct object and the indirect object, if present. Among European languages, this polypersonal system (multiple verb agreement) is only found in Basque, some Caucasian languages, and Hungarian. The ergative-absolutive alignment is rare among European languages, but not worldwide.

It remains alive but in last centuries Basque suffered continuous regression. The region in which Basque is spoken is smaller than what is known as the Basque Country, and the distribution of Basque speakers is not homogeneous there. The main reasons of this regression (Amorrortu, 02) are that Basque was not an official language, that it was out of educational system, out of media and

out of industrial environments. Besides, the fact of being six different dialects made difficult the wide development of written Basque.

However, after 1980, some of those features changed and many citizens and some local governments promote recovering of Basque Language.

Today Basque holds co-official language status in the Basque regions of Spain: the full autonomous community of the Basque Country and some parts of Navarre. Basque has no official standing in the Northern Basque Country.

In the past Basque was associated with lack of education, stigmatized as uneducated, rural, or holding low economic and power resources. There is not such an association today, Basque speakers do not differ from Spanish or French monolinguals in any of these characteristics.

Standard Basque, called Batua (unified) in Basque, was defined by the Academy of Basque Language (Euskaltzaindia) in 1966. At present, the morphology is completely standardized, but the lexical standardization process is underway. Now this is the language model taught in most schools and used on some media and official papers published in Basque.

We are around 700,000 Basque speakers, around 25% of the total population of the Basque Country, and they are not evenly distributed. But still the use of Basque in industry and specially in Information and Communication Technology is not widespread. A language that seeks to survive in the modern information society has to be present also in such field and this requires language technology products. Basque as other minority languages has to make a great effort to face this challenge (Petek, 2000; Williams et al., 2001).

2 Strategy to develop HLT in Basque

IXA group is a research Group created in 1986 by 5 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 LNP software mainly for Basque. We wanted to face the challenge of adapting Basque to language technology.

Twenty one years later on, now IXA is a group composed by 28 computer scientists, 13 linguists and 2 research assistants. 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 more than seven the 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. One of the fruits of this collaboration were the HIZKING21 project (2002-2005) and ANHITZ project (2006-2008). Both projects were accepted by the Government of the Basque Country in a new strategical research line called 'Language Infoengineering'.

At the very beginning, twenty years ago, our first goal was to create just a translation system for Spanish-Basque, but after some preliminary works we realized that instead of wasting our time in creating an ad hoc MT system with small accuracy, we had to invest our efforts in creating basic tools as a morphological analyzer/generator for Basque, that could be used later on to build not just a more robust MT system but also any other language application.

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

We presented an open proposal for making progress in Human Language Technology (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 high capacity and computational power of new computers allowed facing problems in a different way.

Our strategy may be described in two 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 foundations, 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 and also to the definition of a methodology for stand-off corpus tagging based on TEI, feature structures and XML (Artola et al., 2005).

In the same way, taking as reference our experience in incremental design and development we propose four phases as a general strategy for language processing. These are the phases defined with the products to be developed in each of them.

1. Initial phase: *Foundations*. Corpus I (collection of raw text without any tagging mark). Lexical database I. (the first version could be a list of lemmas and affixes). Machine-readable dictionaries. Morphological description.

2. Second phase: *Basic tools and applications*. Statistical tools for the treatment of corpus. Morphological analyzer/generator. Lemmatizer/ tagger. Spelling checker and corrector (although in morphologically simple languages a word list could be enough). Speech processing at word level. Corpus II (word-forms are tagged with their part of speech and lemma). Lexical database II (lexical support for the construction of general applications, including part of speech and morphological information).

3. Third phase: *Advanced tools and applications*. An environment for tool integration. Web crawler. A traditional search machine that integrates lemmatization and language identification. Surface syntax. Corpus III (syntactically tagged text). Grammar and style checkers. Structured versions of dictionaries (they allow enhanced functionality not available for printed or raw electronic versions). Lexical database III (the previous version is enriched with multiword lexical units. Integration of dictionaries in text editors). Lexical-semantic knowledge base. Creation of a concept taxonomy (e.g.: Wordnet). Word-sense disambiguation. Speech processing at sentence level. Basic Computer Aided Language Learning (CALL) systems

4. Fourth phase: *Multilingualism and general applications*. Information retrieval and extraction. Translation aids (integrated use of multiple online dictionaries, translation of noun phrases and simple sentences). Corpus IV (emantically tagged text after word-sense disambiguation). Dialog systems. Knowledge base on multilingual lexico-semantic relations and its applications.

We will complete this strategy with some suggestions about what shouldn't be done when working on the treatment of minority languages. a) Do not start developing applications if linguistic foundations are not defined previously; we recommend following the above given order: foundations, tools and applications. b) 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 reusable by any other tool or application. c) If you complete a new resource or tool, do not keep it to yourself; there are many researchers working on English, but only a few on each minority language; thus, the few results should be public and shared for research purposes, for it is desirable to avoid needless and costly repetition of work.

3 Machine Translation for Basque

After years working on basic resources and tools we decided it was the time to face to the MT task (Hutchins and Somers, 1992). Our general strategy was more specifically for Machine Translation defined bearing in mind the next concepts:

- reusability of previous resources, specially lexical resources and morphology of Basque
- standardization and collaboration: at least using a more general framework in collaboration with other groups working in NLP
- open-source: This means that anyone having the necessary computational and linguistic skills will be able to adapt or enhance it to produce a new MT system, even for other pairs of related languages or other NLP applications.

The involved languages would be Basque, Spanish and English, because of the real necessity of translation in our environment.

Since the beginning we wanted to combine the two basic approaches for MT (rule-based and corpus-based) in order to build a hybrid system, because it is generally agreed that there are not enough corpora for a good corpus-based system in languages like Basque.

Data-driven Machine Translation (example-based or statistical) is nowadays the most prevalent trend in Machine Translation research. Translation results obtained with this approach have now reached a high level of accuracy, especially when the target language is English. But these Data-driven MT systems base their knowledge on aligned bilingual corpora, and the accuracy of their output depends heavily on the quality and the size of these corpora. Large and reliable bilingual corpora are unavailable for many language pairs.

The rule based approach

First we present the main architecture and the proposed standards of an open source MT engine, which first implementation translates from Spanish into Basque using the traditional transfer model and based on shallow and dependency parsing.

The design and the programs are independent from the languages, so the software can be used for other projects in MT especially when lesser-used languages are used. Depending on the languages included in the adaptation it will be necessary to add, reorder and change some modules, but it will not be difficult because an unique XML format is used for the communication among all the modules.

The project has been integrated in the OpenTrad initiative (www.opentrad.com), a government-funded project shared among different universities and small companies, which also include MT engines for translation among the main languages in Spain. The main objective of this initiative is the construction of an open, reusable and interoperable framework.

In the OpenTrad project two different but coordinated designs have been carried out:

- A shallow-transfer machine translation engine for similar languages (Spanish, Catalan and Galician by the moment). The MT architecture uses finite-state transducers for lexical processing, hidden Markov models for part-of-speech tagging, and finite-state based chunking for structural transfer. It is named Apertium and it can be downloaded from *apertium.sourceforge.net.* (Armentano-Oller et al., 2004)
- A deeper-transfer engine for the Spanish-Basque pair. It is named Matxin (Alegria et al., 2007) and it is stored in *matxin.sourceforge.net* and it is a continuation of previous work in our group. In order to reuse resources in this Spanish-Basque system the analysis module for similar languages was not included in *Matxin;* another open source engine, *FreeLing* (Carreras et al., 2004*),* was used here, of course, and its output had to be converted to the proposed interchange format.

Some of the components (modules, data formats and compilers) from the first architecture in OpenTrad were used in the second. Indeed, an important additional goal of this work was testing which modules from the first architecture could be integrated in deeper-transfer architectures for more difficult language pairs.

The transfer module is also based on three main objects in the translation process: words or nodes, chunks or phrases, and sentences.

- First, lexical transfer is carried out using a bilingual dictionary compiled into a finite-state transducer. We use the XML specification of Apertium engine.
- Then, structural transfer at the sentence level is applied, and some information is transferred from some chunks to others, and some chunks may disappear. Grammars based on regular expressions are used to specify these changes.For example, in the Spanish-Basque transfer, the person and number information of the object and the type of subordination are imported from other chunks to the chunk corresponding to the verb chain.
- Finally the structural transfer at the chunk level is carried out. This process can be quite simple (i.e. noun chains between Spanish and Basque) or more complex (i.e. verb chains in the same case).

The XML file coming from the transfer module is passed on the generation module.

- In the first step syntactic generation is performed in order to decide the order of chunks in the sentence and the order of words in the chunks. Several grammars are used for this purpose.
- Morphological generation is carried out in the last step. In the generation of Basque the main inflection is added to the last word in the chunk (in Basque: the declension case, the number and other features are assigned to the whole noun phrase at the end of the last word), but in verb chains additional words need morphological generation. A previous morphological analyzer/generator for Basque (Alegria et al., 1996) has been adapted and transformed to the format used in Apertium.

The results for the Spanish/Basque system using FreeLing and Matxin are promising. The quantitative evaluation uses the open source evaluation tool IQMT and figures are given using BLEU and NIST measures (Giménez et al., 2005).

The corpus based approach

The corpus-based approach has been carried out in collaboration with the National Centre for Language Technology in Dublin.

The system exploits both EBMT and SMT techniques to extract a dataset of aligned chunks. We conducted Basque to English and Spanish to Basque translation experiments, evaluated on a large corpus (270, 000 sentence pairs).

Some tools have been reused for this purpose:

- GIZA++: For word/morpheme alignment we used the GIZA++ statistical word alignment toolkit, and following the “refined” method of (Och and Ney, 2003), extracted a set of high-quality word/morpheme alignments from the original unidirectional alignment sets. These along with the extracted chunk alignments were passed to the translation decoder.
- Pharaoh/Moses decoder: the decoder is also a hybrid system which integrates EBMT and SMT. It is capable of retrieving already translated sentences and also provides a wrapper around the PHARAOH SMT decoder (Koehn, 2004).
- MaTrEx: The MATREX (Machine Translation Using Examples) system used in our experiments is a data-driven MT engine, built following an extremely modular design. It consists of a number of extendible and re-implementable modules (Way and Gough, 2005).

For this engine, we reuse a toolkit to chunk the Basque sentences. After this processing stage, a sentence is treated as a sequence of morphemes, in which chunk boundaries are clearly visible. Morphemes denoting morphosyntactic features are replaced by conventional symbolic strings. After some adaptation, the chunks obtained in this manner are actually very comparable to the English chunks obtained with the marker-based chunker.

The experimental results have shown that our system significantly outperforms state-of-the-art approaches according to several common automatic evaluation metrics: WER, Bleu and PER (Stroppa et al., 2006; Labaka et al., 2007).

4 Conclusions

A language that seeks to survive in the modern information society requires language technology products. "Minority" languages have to do a great effort to face this challenge. Ixa group has been working since 1986 in adapting Basque to language technology, having developed several applications that are effective tools to promote the use of Basque. Now we are planning to define the BLARK for Basque (Krauwer, 2003).

From our experience we defend that research and development for a minority language should to be faced following this points: high standardization, and reusing language foundations, tools, and applications, and incremental design and development of them. 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 those guidelines have been applied even though when it was easier to define "toy" resources and tools useful to get good short term aca-

demic results, but not reusable in future developments.

This strategy has been completely useful when we have created MT systems for Basque. Reusing of previous works for Basque (that were defined following XML and TEI standards), and reusing other open-source tools were the keys to get satisfactory results in a short time.

Two results produced in the MT track are publicly available:

- matxin.sourceforge.net for the free code for the es-eu RBMT system
- www.opentrad.org for the on-line demo

Acknowledgments

This work has been partially funded by the Spanish Ministry of Education and Science (OpenMT: Open Source Machine Translation using hybrid methods,TIN2006-15307-C03-01) and the Local Government of the Basque Country (AnHITZ 2006: Language Technologies for Multingual Interaction in Intelligent Environments., IE06-185). Andy Way, Declan Groves and Nicolas Stroppa from National Centre for Language Technology in Dublin are kindly acknowledged for providing their expertise on the Matrex system and the evaluation of the output.

References

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

I. Alegria, X. Artola, K. Sarasola. 1996.Automatic morphological analysis of Basque. *Literary & Linguistic Computing* Vol. 11, No. 4, 193-203. Oxford University Press. Oxford. 1996.

I. Alegria, A. Díaz de Ilarraza, G. Labaka, M Lersundi, A. Mayor, K. Sarasola. 2007. Transfer-based MT from Spanish into Basque: reusability, standardization and open source. LNCS 4394. 374-384. Cicling 2007.

E. Amorrortu. 2002. Bilingual Education in the Basque Country: Achievements and Challenges after Four Decades of Acquisition Planning. *Journal of Iberian and Latin American Literary and Cultural Studies.*Volume 2 Number 2 (2002)

C. Armentano-Oller, A. Corbí-Bellot, M. L. Forcada, M. Ginestí-Rosell, B. Bonev, S. Ortiz-Rojas, J. A. Pérez-Ortiz, G. Ramírez-Sánchez, F. Sánchez-Martínez, 2005. An open-source shallow-transfer machine translation toolbox: consequences of its release and availability. *Proceedings of OSMaTran: Open-Source Machine Translation workshop,* MT Summit X.

X. Artola, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, G. Labaka, A. Sologaistoa, A. Soroa. 2005. A framework for representing and managing linguistic annotations based on typed feature structures. RANLP.

X. Carreras,, I. Chao, L. Padró and M. Padró. 2004. FreeLing: An open source Suite of Language Analyzers, in P*roceedings of the 4th International Conference on Language Resources and Evaluation* (LREC'04).

J. Giménez, E. Amigó, C. Hori. 2005. Machine Translation Evaluation Inside QARLA. In *Proceedings of the International Workshop on Spoken Language Technology* (IWSLT'05)

W. Hutchins and H. Somers. 1992. *An Introduction to Machine Translation.* Academic Press.

P. Koehn. 2004. Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of AMTA-04,* pages 115–124, Washington, District of Columbia.

S. Krauwer. 2003. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. International Workshop Speech and Computer ,. 27-29 October 2003, Moscow, Russia.

G. Labaka, N. Stroppa, A. Way, K. Sarasola 2007 Comparing Rule-Based and Data-Driven Approaches to Spanish-to-Basque Machine Translation MT-Summit XI, Copenhagen

F. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics,* 29(1):19–51.

B. Petek. 2000. Funding for research into human language technologies for less prevalent languages, *Second International Conference on Language Resources and Evaluation* (LREC 2000). Athens, Greece.

N. Stroppa, D. Groves, A. Way, K. Sarasola K. 2006. Example-Based Machine Translation of the Basque Language. *AMTA. 7th conference of the Association for Machine Translation in the Americas.*.

A. Way and N. Gough. 2005. Comparing Example-Based and Statistical Machine Translation. *Natural Language Engineering*, 11(3):295–309.

B. Williams, K. Sarasola, D. Ó´Cróinin, B. Petek. 2001. Speech and Language Technology for Minority Languages. *Proceedings of Eurospeech* 2001