Five Language Technology Applications Effective to Promote the Use of Basque

I. Alegria, X. Artola, A. Diaz de Ilarraza, K. Sarasola

Ixa taldea – University of the Basque Country
649 Postakutxa. 20080 Donostia

Abstract

We present some Language Technology applications that have proven to be effective tools to promote the use of Basque, a low-density language. We also present the strategy the Ixa research group has followed for almost twenty years to develop those applications in an integrated environment of language resources, foundations, tools and other language applications.

Key words: Language Technology, Language Applications, Minority Languages, Basque, Lemmatization.

1 Introduction

Basque is a minority language that after centuries of regression is now undergoing a process of revitalization. It is a highly inflected language with free constituent order. Its structure and word order are very different compared to languages such as Spanish, French or English. Thanks to a 30-year revitalization process, today Basque holds co-official language status in some areas, and the recently created standard Basque is taught in schools and used on some media. However, there is still much work to be done to fully normalize its use; in fact, the use of Basque in industry (and especially in Information and Communication Technology) is still not widespread enough to guarantee Basque's long term survival.

This paper presents the work performed by the IXA research group (http://ixa.si.ehu.es) in the automatic processing of Basque. Several Language Technology applications have been created and have proven to be effective tools to promote the use of the language. They are related to spelling checkers, on-line dictionaries, machine translation, search machines and corpus processing.

We also present the strategy the Ixa group has followed for almost twenty years to develop those applications. We think this methodology could be useful for other languages as well. Since 2002, part of this work

has been accomplished in collaboration with other partners. HIZKING21 and ANHITZ projects were founded by the Government of the Basque Country in a new strategic research line called 'Language Infoengineering.'

The rest of this paper is arranged as follows: Section 2 describes linguistic features of Basque, its historical regression during centuries and its modern process of revitalization. Section 3 presents work accomplished by the Ixa group. Section 4 resumes the main points of the strategy adopted to guide that work. Section 5 present five Language Technology applications that have proven to be effective to promote the use of Basque. We finish this paper with some conclusions.

2 Basque language

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

2.1 Linguistic features

Basque is an agglutinative language, with a rich flexional morphology. With regard to nouns, for example, at least 360 word forms are possible for each lemma[1]. Each one of the grammar cases – absolutive, dative, associative – has four different suffixes to be added to the last word of the noun phrase. These four suffix variants correspond to indefinite, definite singular, definite plural and "close" definite plural. For example the noun *katu* used in associative case generates those word forms: *katurekin* (with a cat), *katuarekin* (with the cat), *katuekin* (with the cats), and *katuokin* (with these cats).

Case	**Undef.**	**Def./Sing.**	**Def./Pl.**	**Close/Pl.**
Absolutive	*katu*	*katua*	*katuak*	*katuok*
Ergative	*katuk*	*katuak*	*katuek*	*katuok*
Dative	*katuri*	*katuari*	*katuei*	*katuoi*
Genitive	*katuren*	*katuaren*	*katuen*	*katuon*
Associative	*katurekin*	*katuarekin*	*katuek*	*katuokin*
… 14 cases	…	…	…	…

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.

[1] Taking into account the inclusion of suffixes related to ellipsis more than one million word forms are possible for each lemma.

```
I am          / Ni naiz
                (ni  = I + absolutive)
I saw the cat/ Nik katua ikusi nuen
              (nik = I + ergative)
```

Example 1. Ergative case, subject of transitive verbs.

```
    Txakurrak egunkaria ahoan zekarren.
    (The dog brought the newspaper in his mouth)

Txakur-rak    egunkari-a     aho-an        zekarren.
The-dog       the-newspaper  in-his-mouth  brought
ergative-3-s  absolutive-3-s inessive-3-s
Subject       Object         Modifier      Verb
```

Example 2. Case suffixes.

```
Txakur-rak   aho-an      egunkari-a    zekarren.
Txakur-rak   aho-an      zekarren      egunkari-a.
Egunkari-a   txakur-rak  zekarren      aho-an.
```

Example 3. Free order of sentence components (alternative possible orders).

The auxiliary verb, which accompanies most main verbs, agrees not only with the subject in person and number, but also 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 also unique among European languages, but not rare worldwide[2].

```
I saw the cat    Nik  katua  ikusi nuen
I saw the cats   Nik  katuak ikusi nituen
I saw you              Nik  zu     ikusi
zintudan
```

Example 4. Agreement in number and person between the veb and grammatical cases (subject, object and indirect object).

2.2 Centuries of regression

Figure 1 shows the distribution of pre-Roman languages in the Iberian Peninsula; nowadays Basque is the only one of all those languages to remain in active use. At the present time Basque is thriving, but in the last centuries it suffered continuous regression. Figure 1 also shows the hypothetical border lines of Basque speakers in the 2nd, 7th, 12th and 19th centuries. The region in which it is spoken nowadays is smaller than what is

[2] http://en.wikipedia.org/wiki/Basque_language

known as the Basque Country, and the proportion of speakers is not homogeneous there. The main reasons for this regression (Amorrortu, 02) are that Basque was not an official language, that it was not used in the educational system, and that it was absent from the media as well as industrial environments. Besides, having six different dialects made the development of a widely used written language difficult.

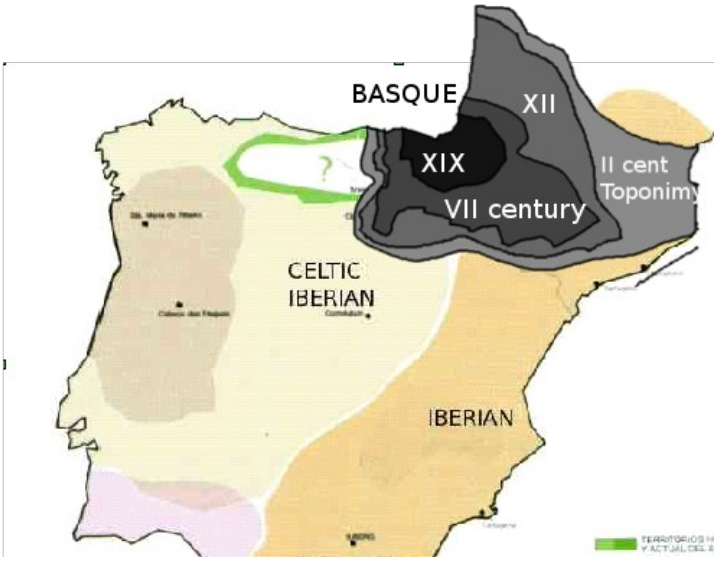

Fig 1.-Pre-Roman languages and Basque border lines
in the 2nd, 7th, 12th and 19th centuries

2.3 Modern reactivation

However, after 1980, some of those features are changing and many citizens and some local governments are promoting the recovery of the language.

Today Basque holds co-official language status in the Basque regions of Spain, including the entire autonomous community of the Basque Country and some parts of Navarre. It has no official status in the Northern Basque Country, on the French side of the border.

In the past Basque was associated with lack of education, and Basque speakers were stigmatized as uneducated, rural, and of low economic and political status. Such an association no longer exists today, and Basque speakers do not differ appreciably from Spanish or French monolinguals in these terms.

The first steps in the creation of Standard Basque, called *Batua* (unified) in Basque, were taken in 1966 by the Academy of the Basque Language (Euskaltzaindia). At present, the morphology is completely standardized, but the lexical standardization process is still underway. Now *Batua* is the language model taught in most schools and it is used in some media (www.eitb.com, www.berria.info) and official papers.

There are around 700,000 Basque speakers, around 25% of the total population of the Basque Country, although they are not evenly distributed.

But the use of Basque in industry and especially in Information and Communications Technology is still not widespread enough to guarantee its long-term survival. A language that seeks to survive in the modern information society has to be used in these fields and this requires language technology products. Minority languages have to make a great effort to face this challenge (Petek, 2000; Williams et al., 2001). Of course, this assertion also applies to Basque.

2.4 Work on automatic processing of Basque (IXA group)

The IXA group is a research team created in 1986 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 language technology. We wanted to face the challenge of adapting Basque to language technology.

Now, twenty-two years later, IXA is a group composed of 28 computer scientists, 13 linguists and 2 research assistants. It works in cooperation with more than seven companies from the Basque Country and five from abroad; it has been involved in the birth of two spin-off companies (Eleka and Prompsit), and has developed more than seven language technology products.

In recent years, several private companies and technology centers of the Basque Country have gotten interested and begun to invest in this area. At the same time, more research groups have come to be aware of the fact that collaboration is essential to the development of language technologies for minority languages. Some of the fruits of this collaboration were the HIZKING21 (2002-2005) and ANHITZ (2006-2008) projects. Both projects were accepted by the Government of the Basque Country as part of a new strategic research line called 'Language Infoengineering'. The following organizations take part in these projects:

- The Aholab group (aholab.ehu.es), specializing in speech technologies (synthesis and recognition); it belongs to the Signal Treatment and Radiocommunication Team of the Electronics and Telecommunication Department of the University of the Basque Country
- Elhuyar Foundation: a non-profit organization (www.elhuyar.com) aimed at promoting the normalization and standardization of Basque, with activities in the fields of lexicography and terminology, dictionary publishing, language planning, science and technology communication, textbooks and multimedia products and services
- Robotiker: a technology center (www.robotiker.com) specializing in Information and Telecommunication technologies. Robotiker is part of the Tecnalia Technology Corporation.
- Vicomtech: an applied research center (www.vicomtech.es) working in the area of interactive computer graphics and digital multimedia. It was founded jointly by the INI-GraphicsNet Foundation and by EiTB, the Basque Radio and Television Group.

At the very beginning, twenty years ago, our first goal was to merely create a translation system for Spanish-Basque, but after some preliminary work we realized that instead of wasting our time creating an *ad hoc* Machine Translation (MT) system with minimal accuracy, we had to invest our efforts in creating basic tools, such 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 of our strategy to make progress in the automatic processing of Basque.

1987	1992	1993	2001	2007
	Lexical Database, EDBL	Spelling checker, Xuxen	Lemmatizer	Basque Wordnet
		Morphosyntactic analyzer	Parser	MT system Matxin

Table 1.- Milestones in the history of IXA research group

3 Strategy to Develop Human Language Technology

Basque has to face up to the scarcity of the resources and tools that could make possible its development of Language Technology at a reasonable and competitive rate.

We presented an open proposal for making progress at the Human Language Technology conference (Aduriz et al., 1998). However, the steps proposed did not correspond exactly with those observed in the history of the processing of English, since the high capacity and computational power of present computers allows problems to be faced in a different way. Since Basque is an agglutinative language, with a rich flexional morphology, it requires specific procedures for language analysis and generation.

Our strategy may be divided into two goals:

- Standardization of resources useful in different research initiatives, tools and applications.
- Incremental design and development of language foundations, tools, and applications in a parallel and coordinated way in order to get the most benefit from them. Language foundations and research are essential to create any tool or application; but the same tools and applications will be very helpful in the research and improvement of language foundations.

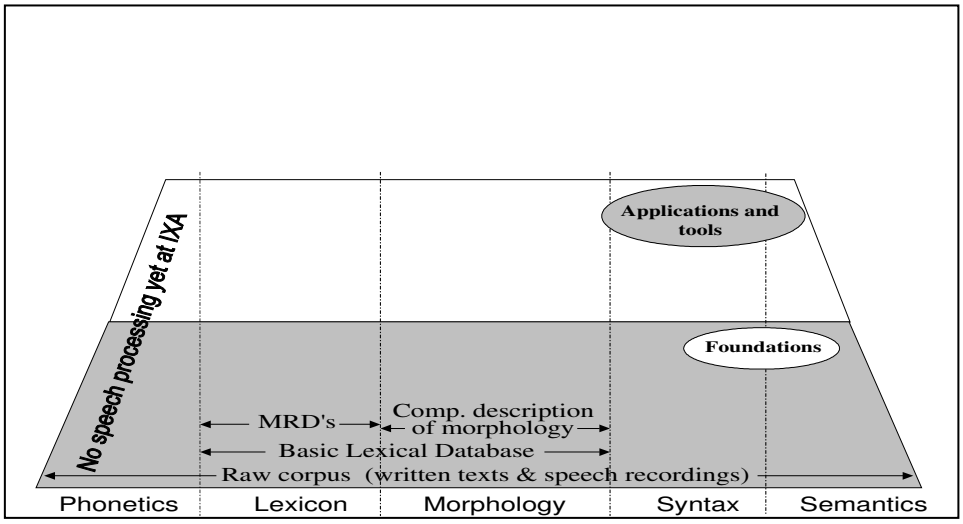

Figure 2. First phase: Foundations.

The next step to the standardization of resources brought us to adopt the TEI recommendations to design a methodology for stand-off corpus annotation based on feature structures encoded in XML (Artola et al., 2000; Artola, 2004).

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. Following are the phases and the products to be developed in each of them.

3.1 Initial phase: Foundations

- Corpus I. Collection of raw text without any tagging.
- Lexical database I. The first version could be a list of lemmas and affixes.
- Machine-readable dictionaries.
- Morphological description.
- Speech corpus I.

3.2 Second phase: Basic tools and applications.

- Statistical tools for the treatment of corpora.
- 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 parts of speech and morphological information.

3.3 Third phase: Advanced tools and applications.

- An environment for tool integration.
- Web crawler. A traditional search machine that integrates lemmatization and language identification.
- Surface syntactic parsing.
- Corpus III. Syntactically tagged text.
- Grammar and style checkers.
- Structured versions of dictionaries. They allow enhanced functionality not available in 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.

3.4 Fourth phase: Multilingualism and general applications.

- Information retrieval and extraction.
- Translation aids. Integrated use of multiple on-line dictionaries, translation of noun phrases and simple sentences.
- Corpus IV. Semantically tagged text after word-sense disambiguation.
- Dialog systems.
- Knowledge base on multilingual lexico-semantic relations and its applications.

We will complete this description of our strategy with some warnings about what should be avoided when working on the treatment of minority languages. a) Do not start developing applications if linguistic foundations are not already in place; we recommend following the above order: foundations, tools, and applications. b) When a new system has to be planned, do not create *ad hoc* lexical or syntactic resources; instead, design those resources in such a way that they could be easily expanded and used 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, to avoid needless and costly "reinvention of the wheel," the results should be made public and shared for research purposes.

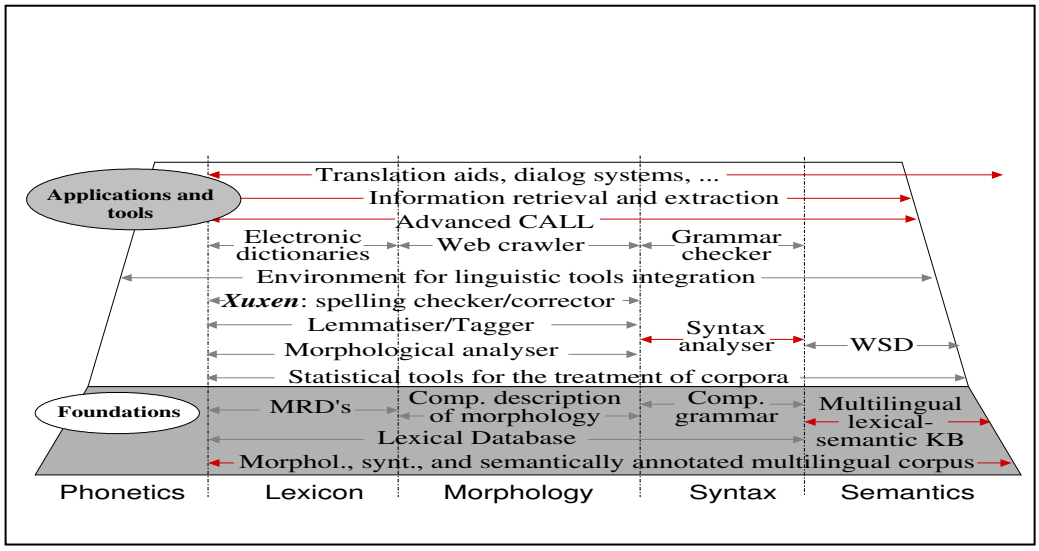

Figure 3. Fourth phase: Multilingualism and general applications.

4 Useful applications

In this section we describe some effective tools already created by our group.

4.1 Spelling checker/corrector

Because for many years the use of Basque was forbidden in schools and also because of its late standardization[3], adult speakers nowadays did not learn it at school, and so they write it imperfectly. For example, when someone goes to write the word *zuhaitza* (tree), the many possible spellings (*zuhaitz*? *zugaitz*? *zuhaitx*? *zuhaitsa*? *sugatza*?) may cause the writer to hesitate, often leading to an easy solution: "give up, and write the whole text in Spanish or French!".

The spelling-checker Xuxen (Aduriz et al., 97) is a very effective tool in this kind of situation, giving people more confidence in the text they are writing. In fact, this program is one of the most powerful tools in the ongoing standardization of Basque.

The spelling checker is more complex than equivalent software for other languages, because most of those are based on recognizing each word in a list of possible words in the language. However, because of the Basque language's rich morphology, it is very difficult to define such a list, and consequently, when possible morphological analysis must be included.

[3] The academy of Basque *Euskaltzaindia* defined the morphology and verbs of Unified Basque in 1966, but the lexical standardization process is still going on.

Xuxen is publicly available from www.euskara.euskadi.net, where there have been more than 20,000 downloads. There are versions for Office, OpenOffice, Mozilla, PC, Mac, and also an online web service (www.xuxen.com).

The version for Office includes morphological analysis, but, what happens if we want to use the speller in the "free world" (*OpenOffice*, *Mozilla*, *emacs*, *LaTeX*, ...)? *ispell* and similar tools (*aspell*, *hunspell*, *myspell*) are the usual mechanisms for these purposes, but they do not fit with the two-level model we had to use to be able to describe Basque morphology. In the absence of two-level morphology, our solution was to adapt the two-level description to *hunspell* in a (semi)automatic way. With the stems and two sets of suffixes, corresponding to the paradigms at first and second level, that have been obtained all the information we needed for the *hunspell* description was ready. Only a format conversion was necessary for delivery the spelling checker/corrector for OpenOffice, and other tools integrating *hunspel (*www.euskara.euskadi.net) In addition, we did the adaptation of the description to *myspell* (www.librezale.org/mozilla/firefox), for tools that do not still integrate *hunspell*, combining the main paradigms (with less generation power for each one) and the word forms appearing in a big corpus after eliminating forms rejected by the original spelling checker. Although hose approaches for the “free world” have lesser coverage for Basque morphology, they are very useful spelling checkers.

4.2 Lemmatization-based on-line dictionaries

The main product created for this kind of application is a plug-in for MS Word that enables looking up a word in several dictionaries, but, in order to make it more useful for a language like Basque with its rich morphology, the dictionary is enhanced with lemmatization. This means that morphological analysis is first performed, and then possible lemmas of the word are matched with the dictionary. In the example shown in Figure 4, the user asks for the meaning in Basque of the Spanish word *cupiéramos*. That word-form can’t be found in paper dictionaries because it is a finite verb form, but the application recognizes that it corresponds to the verb *caber*, and shows five different equivalents in Basque for that verb.

At the moment this plug-in works with three dictionaries: Spanish-Basque, French-Spanish and a dictionary of synonyms. The Spanish-Basque version is publicly available in www.euskara.euskadi.net.

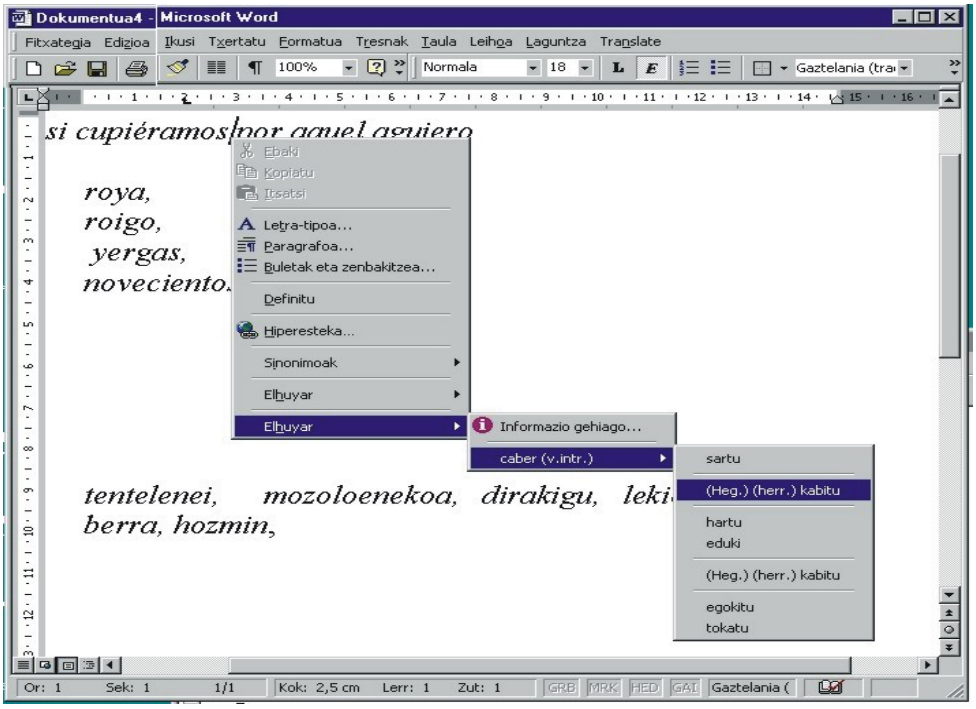

Fig 4.- Lemmatization-based on-line dictionary consulting

4.3 Lemmatization-based search machine

We have developed a search machine similar to the previous application, but for use with text documents instead of dictionaries. This program first performs morphological analysis of the word, and then searches relevant documents containing the lemmas corresponding to these possible morphological decompositions. In the example shown in Figure 5 the user is searching in the Elhuyar science divulgation journal for documents related to the Basque word form *saguarekin* (with the mouse). The search machine looks for documents containing words whose lemma is just *sagu*/mouse ("*saguen*", "*saguaren*", "*sagua*", "*saguetan*"...). The principal search

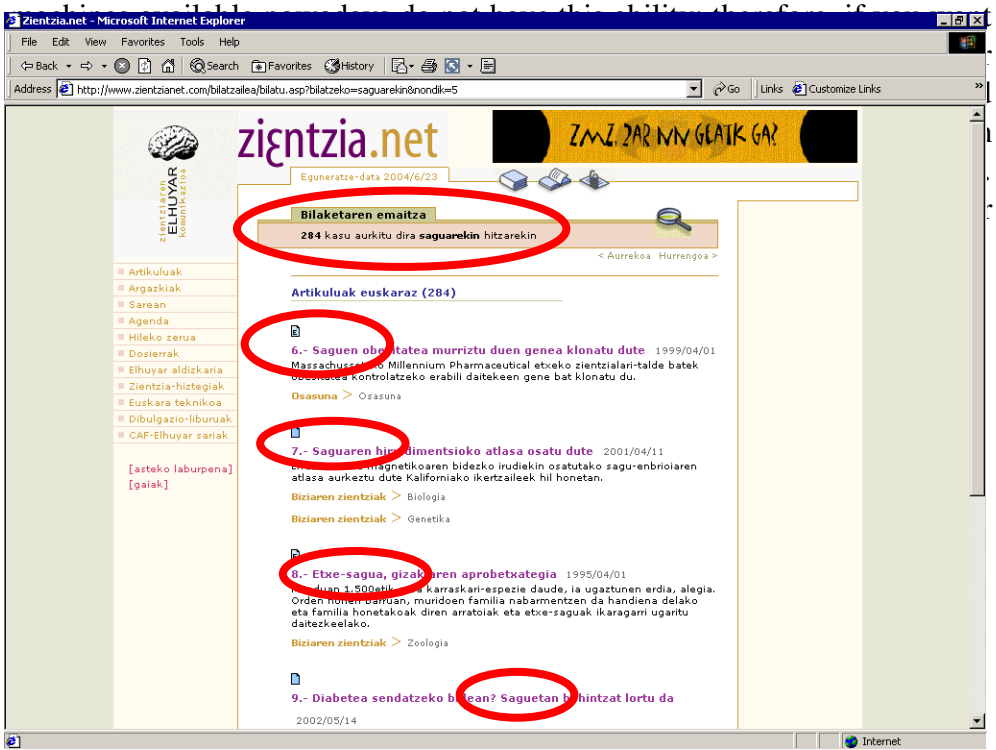

Fig 5.- Lemmatization-based search machine

4.4 Transfer-based machine translation system

When we have faced a difficult task such as Machine Translation into Basque, our strategy has worked well. In 2000, after years working on basic resources and tools, we decided it was time to face the MT task. Our general strategy was more specifically defined for Machine Translation, and we bore in mind the following concepts:

- reusability of previous resources, especially lexical resources and morphology description
- 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.

We have gotten good results in a short time by just reusing previous work, reusing other open-source tools, and developing only a few new modules in collaboration with other groups[4]. In addition, we have produced new reusable tools and formats. We created Matxin using a transfer rule-based MT approach. It translates text from Spanish into Basque, and two results produced in the machine translation track are publicly available: matxin.sourceforge.net for the free code of the Spanish-Basque system and www.opentrad.org for the online version.

Now we are working in the construction of a multiengine system including three subsystems based on different approaches to MT: rule-based

[4] Opentrad project: opentrad.org

machine translation, statistical machine translation and example-based machine translation (Alegria et al., 2008).

Fig 6.- Opentrad-Matxin MT system

4.5 ZTC text corpus

Today statistical tools for text processing are so powerful in language technology, that the number of words compiled and organized as text corpora could be used as a measure of the position of a language in the area.

The ZTC corpus (Areta et al., 2007) has been built by compiling text on the subject of "Science and Technology". A previous inventory of years 1990-2002 registered 20 million words on this subject. The ZTC corpus compiled 7.6 million. All those words were lemmatized, and up to 1.6 million were manually revised and disambiguated. A specific interface for advanced query of the corpus was also built. The result is a public resource: www.ZTcorpusa.net.

The creation of this resource would have been impossible without reusing the lemmatizer. We built a new tool for corpus compilation and massive use of the lemmatizer was necessary.

The ZTC corpus is still far away from the size of the corpora for other languages; e.g., the BNC corpus (www.natcorp.ox.ac.uk), that is becoming a standard corpus resource, has 100 million words. However, the ZTC corpus is a very useful resource for manual study of Basque, as well as for machine learning techniques.

5 Conclusions

A language that seeks to survive in the modern information society requires language technology products. "Minority" languages have to make

a great effort to face this challenge. The Ixa group has been working since 1986 in adapting Basque to language technology.

Based on our experience we argue that research and development for a minority language should be conducted according to these principles: reuse of language foundations, tools, and applications; high standardization; and incremental design and development. We are aware that this assertion seems trivial, because it is a general rule for the development of any computer application, and that HLT projects related with any language should follow those guidelines, but we know from our experience that in many cases HLT projects do not follow this strategy. We think that if Basque is now in a good position in HLT, it is because those guidelines have been applied even though sometimes it would have been easier to define "toy" resources and tools useful for good short-term academic results, but not reusable in future developments.

This strategy has proved to be successful. Several applications have been created and have proved to be effective tools in promoting this minority language: the spelling checker/corrector, on-line dictionaries, a lemmatization-based search engine, the Matxin translation system, and the ZTC text corpus. Reusing previous work, and using other open-source tools, were the keys to satisfactory results in a relatively short time.

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 Multilingual Interaction in Intelligent Environments., IE06-185).

References

Aduriz I., Alegria I., Artola X., Ezeiza N., Sarasola K. 1997. A spelling corrector for Basque based on morphology. Literary & Linguistic Computing, Vol. 12, No. 1. Oxford University Press. Oxford.

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., Urizar R. 1998 A framework for the automatic processing of Basque. *Proceedings of Workshop on Lexical Resources for Minority Languages. Granada.*

Agirre E., Ansa O., Arregi X., Arriola J., Díaz de Ilarraza A., Pociello E., Uria L. 2002. Methodological issues in the building of the Basque WordNet: quantitative and qualitative analysis. Proceedings of First International WordNet Conference. Mysore (India).

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

Amorrortu E. (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.

Areta N., Gurrutxaga A., Leturia I., Alegria I., Artola X., Díaz de Ilarraza A., Ezeiza N., Sologaistoa A. (2007). ZT Corpus: Annotation and tools for Basque corpora. Corpus Linguistics. Birminghan.

Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M., Soroa A. (2000), A proposal for The Integration of NLP Tools using SGML-Tagged documents, Second Int. Conf. on Language Resources and Evaluation. Athens (Greece).

Artola X (2004), Laying Lexical Foundations for NLP: the Case of Basque at the Ixa Research Group. 4th International SALTMIL (ISCA SIG) LREC workshop on "First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation"

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

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