Workshop on NLP of Minority Languages and Small Languages TALN 2003

HIZKING21: Integrating language engineering resources and tools into systems with linguistic capabilities

A. Diaz de Ilarraza (1), A. Gurrutxaga (2), I. Hernaez (1),
N. Lopez de Gereñu (3) and K. Sarasola (1)

(1) Ixa taldea – Aholab – University of the Basque Country
649 Postakutxa. 20080 Donostia
jipsagak@si.ehu.es
(2) Elhuyar Fundazioa
Astesuain Poligonoa, 14 - 20170 Usurbil
agurrutxaga@elhuyar.com
(3) VicomTech
Miramon Technology Park, 20009 Donostia
nlopez@vicomtech.es

Résumé – Abstract

On présente les lignes essentielles du projet HIZKING21. Le but principal de ce projet consiste à favoriser la recherche sur l'ingénierie linguistique pour répondre aux exigences de l'environnement globalisé de nos jours. Notre domaine est le développement des technologies du langage pour la langue basque ainsi que l'intégration des ressources et des instruments pour les industries de la langue –des ressources qui existent déjà et d'autres à développer dans ce projet– dans des différents dispositifs (PCs, PDAs, électroménagers, équipements des voitures, etc.). L'objective que nous poursuivons est de contribuer à l'interaction avec toutes sortes de dispositifs faciles à utiliser, employant la langue comme le moyen naturel de communication. Le lancement de ce projet a été possible, d'une part, grâce aux avancements et progrès du traitement du langage naturel et de l'ingénierie linguistique de la langue basque réalisés par les participants de ce projet pendant les quinze dernières années; et d'autre part, grâce au fait que nous partagions le point de vue sur la meilleure stratégie pour le développement de ces technologies dans le cas d'une langue minoritaire comme le Basque.

We present the main lines of the HIZKING21 project. Its main objective is to promote basic research in language engineering, orienting this investigation towards the requirements of the globalized environment of the present day. Our scope of work is the development of language technologies for the Basque language, as well as the integration of resources and tools for the language industry, both already existing resources and resources to be developed in this project, into different devices (PCs, PDAs, electrical household appliances, car equipment and

so on). Our goal is to contribute to the easy and user-friendly interaction with all kind of devices, using language as the natural means of communication. The starting up of this project has been possible thanks to the advances and developments in the Natural Language Processing and Language Engineering for Basque language made by the participants of this projects in the last fifteen years, as well as to the fact of sharing a vision of to the more adequate strategy for the development of these technologies in the case of a minority language like Basque.

Mots Clés – Keywords

Traitement automatique du Basque, corpus, applications
NLP for Basque, corpora, tools, applications

1 Introduction

The groups that make up the HIZKING21 project are aware of a double matter. On the one hand, the necessity of developing interfaces which will make possible the easy and intuitive interaction between humans and all kind of devices, no matter their technological complexity. The great importance of the language to fulfil this interaction is clear. On the other hand, they are also aware of the fact that people should not have to renounce to the use of their mother language to do so. At present, even if the amount of products in the language industries for English is quite impressive, those products do not have the same spreading in other languages. Taking into account that we live in a multilingual society in Europe, we have, as European researchers who work in this area, a special training to develop multilingual products. And, as Basque researchers, we have also some kind of commitment towards the integration of Basque language in the Information Society. Besides, Basque can be used to prove the adequacy of products to suit other languages, specially minority languages that suffer from the same kind of scarcity. A special attention has been paid to the design of the groups that participate in the project, combining a R+D group from university, a foundation working on the development of Basque language for a long time and two technological centres, so that we can concentrate our efforts on experiences in the areas of research, development and commercialisation.

HIZKING21 has been presented as a project for strategic research within the *Etortek* program of the Department of Industry, Trade and Tourism of the Government of the Basque Autonomous Community. The general budget of the project was of 7 million euros, and it has been approved in a third of its content within the *Etortek* program, which means an initial financing of 16%.

Our presentation will consist of the following sections: a) general vision of the problems that minority and small languages, and specially the Basque language, have to confront in the areas of Natural Language Processing and development of the language technologies in general; b) description of the strategy that we propose for the development of these technologies; c) departure point of the project with respect to the basic technologies, resources and tools available nowadays for our language; d) general objectives of HIZKING21; e) specific objectives of HIZKING21 for basic investigation, generation of resources, development of tools and design of prototypes and pre-applications.

2 NLP and minority languages

A language that seeks to survive in the modern information society requires language technology products. Human Language Technologies are making an essential contribution to the success of the information society, but most of the working applications are available only in English. Minority languages have to make a great effort to face this challenge (Petek, 2000) (Williams et al., 2001).

Language technology development for minority languages differs in several aspects from the development for widely used languages. This is mainly due to two reasons.

On the one hand, the size of the speakers' community is usually small. As a result, most of these languages have not enough specialized human resources, they lack in financial support, and commercial profitability is, almost in all cases, a very difficult goal to reach. In other words, lesser-used languages have to face up to the scarcity of the resources and tools that could make possible this development at a reasonable and competitive rate.

On the other hand, there are language-specific problems, related to language typology. For a lesser-used language it is not always possible to use or to adopt the language technologies developed for other languages. This is especially relevant in rule-based approaches, but also in corpus-based approaches, because truly efficient exploitation of corpus demands annotation, and this process is in most cases based on rule-based procedures, like morphological and syntactic analysis. For example, romance languages like Galician, Catalan or Occitan can take advantage of NLP developments for French or Spanish, but these developments are not so applicable to some languages, for example Basque. This applicability (or portability) depends largely on language similarity. Basque is an agglutinative language, with a rich flexional morphology, and this requires specific procedures for language analysis and generation.

3 A strategy to develop language technology for a lesser used language

We present here an open proposal for making progress in Human Language Technology. Anyway, the steps here proposed do not correspond exactly with those observed in the history of the processing of English, because the high capacity and computational power of present computers allows to face problems in a different way.

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. Therefore, these three levels (language foundations, tools and applications) have to be incrementally developed in a parallel and coordinated way in order to get the best benefit from them.

Our proposal is based on our experience with the automatic processing of Basque. Some features of Basque have to be known in order to evaluate the applicability of our strategy for other minority languages. As we have pointed out, Basque is an agglutinative language with a very rich morphology. It has basically constituent-free order at sentence level. There are around 700,000 Basque speakers, around 25% of the total population of the Basque Country,

but they are not evenly distributed. There are six dialects, but since 1968 the Academy of the Basque Language (Euskaltzaindia) has been involved in a standardization process. At present, the morphology is completely standardized, but the lexical standardization process is underway.

We propose four phases as a general strategy for language processing.

3.1 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.
- Speech corpus I.
- Description of phonemes.

3.2 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.

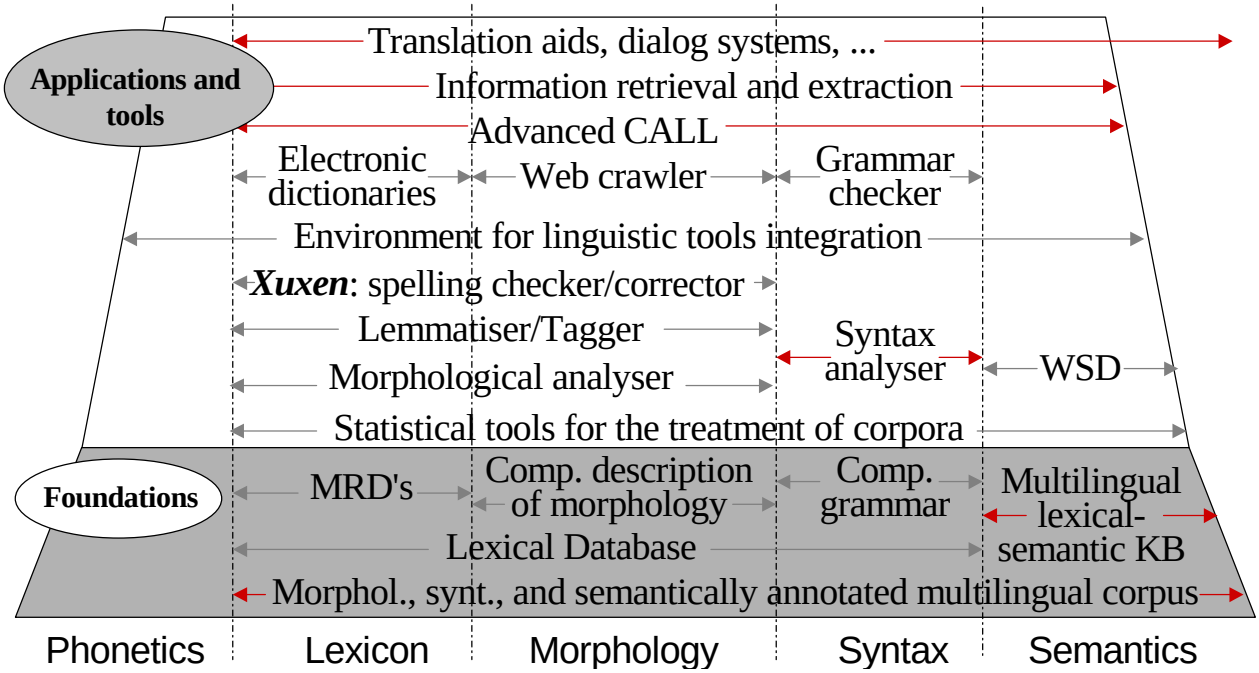

Figure 1: Foundations, applications and tools in language technologies development.

3.3 Third phase: Advanced tools and applications.

- An environment for tool integration. For example, following the lines defined by TEI using XML (Artola et al.; 2000).
- 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.

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 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 to follow 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.

4 Technologies, resources and tools available nowadays for Basque

It is well known that in the last fifteen years several research groups in the Basque Country have been working on this field with the common aim of developing basic computational resources and tools for Basque. The leaders of this work have been two groups of the University of the Basque Country:

- The Aholab group (bips.bi.ehu.es/ahoweb), specialized in speech technologies (synthesis and recognition); it belongs to the Signal Treatment and Radiocommunication Team of Electronics and Telecommunication Department of the University of the Basque Country
- The IXA group (ixa.si.ehu.es), specialized in the processing of written texts at different levels (morphology, syntax, semantics; corpora, machine translation, IE_IR…); the group is made up mainly of members of the Computer Science Faculty of the University of the Basque Country (computer scientists and linguists), and by members of UZEI (Basque Center for Terminology & Lexicography)

Nowadays, the most remarkable tools ready for use are:

- A lemmatizer/tagger (EUSLEM) developed by the IXA group in collaboration with UZEI, and based on a two-level morphological analyzer (MORFEUS) and a lexical database (EDBL), and improved recently with disambiguation module based on the Constraint Grammar formalism and a module that treats Multiword Lexical Units (Ezeiza et al., 1998).
- Basque WordNet (BasWN, ixa.si.ehu.es/Ixa/PilPilean/1022169813), implemented for Basque by the IXA group (Agirre et al., 2002)
- The speech synthesizer developed by Aholab (Navas et al., 2002).

Several applications have been commercialized using these tools:

- A spelling checker (XUXEN)
- A lemmatization based web-crawler (GAIN)
- A lemmatization based on-line bilingual dictionary (*Elhuyar Hiztegi Txikia-ren plugin-a Microsoft Word 2000*; "Basque-Spanish/Spanish-Basque Small Elhuyar Dictionary plug-in for Microsoft Word 2000")
- A generator of weather reports (MultiMeteo)
- A text-to-speech application (AhoTTS).

With regard to corpus resources, there are nowadays two significant corpora of Basque texts:

- The textual corpus of the *OEH-Orotariko Euskal Hiztegia* ("Basque General Dictionary"): a non-lemmatized corpus that collects all of Basque written texts until language standardization (~1960). It has approximately 5,5 million words.
- The *XX. mendeko Euskararen Corpus Estatistikoa* ("The statistical corpus of 20th Century Basque"): a lemmatized corpus with feature-structure markup in SGML, implemented on the ORACLE relational database; it can be consulted on line (http://www.euskaracorpusa.net/XXmendea/Konts_arrunta_fr.html). Its size is of 4.658.036 words.

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 is the HIZKING21 project. Together with the IXA and Aholab groups mentioned above, the followings organizations take part in HIZKING21:

- 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.
- Elhuyar Foundation: a non-profit organization (www.elhuyar.com) aimed to promote 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) specialized in Information and Telecommunication technologies. Robotiker is part of Tecnalia Technology Corporation.

5 HIZKING21: general objectives

Starting off the previously described situation, the HIZKING21 promoters consider necessary: a) to combine and coordinate efforts and to share knowledge and resources in order to deepen, accelerate and spread the developments and profits in the area of the technology of the Basque language; b) to design and implement the prototypes and pre-applications that will make possible the development, in the medium term, of commercial products: different systems with linguistic capacity in Basque language. The conviction that we share about these needs has been the base of our motivation to create the HIZKING21 project.

We propose the following general objectives for HIZKING21:

- To design a common strategy and to create a network of excellence in R+D in the area of language technologies, in accordance with international trends through strategic alliances with centers of reference and projects
- To provide the Basque language in the 2006 with a similar level of resources and tools that those available at present for English, in the scope of speech technologies and computational linguistics
- To promote the exploitation of the opportunities offered by language technologies in:
    - Content management
    - Recognition and synthesis of voice
    - Document production and translation technology
    - E-learning
    - Multimedia systems
- To lay the foundations that will facilitate the internationalization of our linguistic industry in the very near future

The main goal of the project is the development of what we call **"systems with linguistic capabilities"**, which will have to be multilingual and guarantee multimodal interaction with users. To do so, the whole project has been divided into different tasks.

- First of all, it is necessary to *identify all of the components that make up such a system*: multimedia interfaces for each environment, resources and tools needed for each kind of system, level of development and availability of existing components, and so on. The main purpose of this stage is to get a clear image about which are the necessary components already available; the ones to be developed from scratch in

Basque Country, either in existing centres or in centres to be created; or components that, existing for other languages, could be adapted to Basque. This clear image will make possible the definition of more specific steps to follow during subsequent stages of the project.

- Second stage will be the development of all the works defined in the first stage. It will be necessary to arrange them into different groups, according to both the mentioned strategy in the processing of the language and to the technological lines involved. The technical groups defined are: Corpus, Resources, Tools, Basic Technologies and Pre-applications. These technical groups are to work in a coordinated way in order to achieve one of the goals of HIZKING21: the design of *devices based on user-friendly and easy interaction through the use of language*, offering an important contribution to the spread of the use of Information Society Technologies by EVERYONE though the reduction of complexity in their use.
- Another important task to develop in HIZKING21 all along the project is to promote qualification and high-level training of people in the scope of language technologies, in order to increase research ability in Basque Country. To do so, exchanges off people with important centres of reference, universities and companies around Europe are planned, together with postgraduate courses and the realization of doctoral thesis. It will be necessary to reach collaboration agreements with these mentioned centres.
- The sharing of knowledge is another critical element in the achievement of the goals of HIZKING21, due to the great necessity of taking advantage of all synergies that may arise among different communities of researchers working on minority languages. So technological surveillance and dissemination will concentrate important efforts of the people involved in the project. Implementation of a systematized method of surveillance, demonstration meetings, publication of results in specialized journals and conferences, thematic seminars and courses, technological meetings, and so on, are some of the activities planned. There will also be a web site where to publish all of the relevant information related to HIZKING21, and to become important information exchange point during the whole development of the project.

6 HIZKING21: detailed objectives in NLP

Within HIZKING21 project,, the most important areas are the following: R+D, training, infrastructure, international collaboration, diffusion and the creation of a technological observatory. We have fixed concrete objectives for basic investigation, generation of resources, development of tools and design of prototypes and pre-applications. The nucleus of the development HIZKING21 is the accomplishment of research and development projects. The following are initially considered as high priority projects:

- Development of basic linguistic resources
- Development of basic tools
- Technical and basic methods for integration of resources and tools
- Person/Machine Interfaces and their integration in multimedia environments
- Wireless technology and hardware associated to the systems with linguistic capacity

The first two aims of them are described next.

6.1 Development of basic linguistic resources

A set of corpora, lexical databases and electronic dictionaries will be completed or created during the development of the project.

6.1.1 Written corpora:

- Corpus I. Collection of raw text without any tagging mark (light XML)
- Corpus IIa. Word-forms are tagged with lemma, POS and morphosyntax analysis (hand corrected)
- Corpus III. Syntactically tagged text (hand corrected)
- Corpus IV. Semantically tagged text after word-sense disambiguation (hand corrected)
- Corpus Va. Multilingual corpora (light XML)
- Corpus Vb. Multilingual and parallel corpora

6.1.2 Spoken corpora:

- Development of text corpus for tasks of Automatic Speech Recognition (ASR)
- Creation of phonetic voice corpus for tasks to 16KHtz for all the dialectal varieties of Basque

6.1.3 Lexical databases

- Lexical database with information about POS, syntax information, multiword units, verb subcategorization and collocations
- Lexical database with semantic information linked to ontologies
- Lexico-semantic database. Concept taxonomy
- Multilingual lexico-semantic database. Concept taxonomy

6.1.4 Electronic dictionaries

- Integration in a data bank of lexicographical and terminological databases
- Design and implementation of a lexicographic workbench
- Design and implementation of a terminological workbench

6.2 Development of basic tools

The next tools will be improved o generated in the project:

- Speech processing: Large Vocabulary Continuous Speech Recognition (LVCSR), development of a high-level TTS system, ASR system
- Syntax: identification of multiword units, definition of syntactic mark-up, syntax analyzer
- Semantics: document classification, entity identification and processing, word-sense disambiguation
- Pragmatics and Discourse: resolution of pronominal anaphora, definition of the structure, goals and topics of a dialog system

- Corpus analysis tools:
- Information retrieval from corpus marked up in XML (concordances, statistics, collocations…)
- Linguistics and statistical tools for terminology extraction from tagged corpora
- Linguistics and statistical tools for the extraction of lexical and terminological equivalences from multilingual tagged corpora

Acknowledgments

This project is partially supported by the *Etortek* program of the Department of Industry, Trade and Tourism of the local Government of the Basque Country.

References

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).

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)*. May, 2000

Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. (1998), Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages, *COLING-ACL'98,* Montreal (Canada). August 10-14, 1998.

Navas E., Hernaez I., Ezeiza N. (2002) Assigning Phrase Breaks using CART's in Basque TTS. Proc. of the 1st Int. Conf. on Speech Prosody, Aix-en-Provence, pp. 527-531, 2002

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*