AnHitz, development and integration of language, speech and visual technologies for Basque

Kutz Arrieta
VICOMTech
karrieta@vicomtech.org

Igor Leturia
Elhuyar Foundation
igor@elhuyar.com

Urtza Iturraspe
Robotiker
uiturraspe@robotiker.es

Arantza Diaz de Ilarraza, Kepa Sarasola
IXA Group - Basque Country University
{kepa.sarasola, a.diazdeilarraza}@ehu.es

Inma Hernáez, Eva Navas
Aholab Group - Basque Country University
{inma.hernaez, eva.navas}@ehu.es

Abstract

AnHitz is a project promoted by the Basque Government to develop language technologies for the Basque language. The participants in AnHitz are research groups with very different backgrounds: text processing, speech processing and multimedia. The project aims to further develop existing language, speech and visual technologies for Basque: up to now its fruit is a set of 7 different language resources, 9 NLP tools, and 5 applications.. But also, in the last year of this project we are integrating, for the first time, such resources and tools (both existing and generated in the project) into a content management application for Basque with a natural language communication interface.

This application consists of a Question Answering and a Cross Lingual Information Retrieval system on the area of Science and Technology. The interaction between the system and the user will be in Basque (the results of the CLIR module that are not in Basque will be translated through Machine Translation) using Speech Synthesis, Automatic Speech Recognition and a Visual Interface.

The various resources, technologies and tools that we are developing are already in a very advanced stage, and the implementation of the content management application to integrate them all is in work and is due to be completed by October 2008.

1. Introduction

AnHitz is a project promoted by the Basque Government in its Science and Technology Plans for 2002-2005 and 2006-2008 to develop language technologies for Basque. “Linguistic Info-engineering” has been selected as one of the 25 strategic research lines within this national program.

AnHitz is a collaborative project between five participants, each of them with expertise in a different area:

- VICOMTech (www.vicomtech.org): An applied research center 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 (www.elhuyar.org): A non-profit organization 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, alongside with R&D in language technologies for Basque.
- Robotiker (www.robotiker.com): A technology center specialized in information and telecommunication technologies, part of the Tecnalia Technology Corporation.
- The IXA Group of the University of the Basque Country (ixa.si.ehu.es): specialized in the processing of written texts at different levels (morphology, syntax, semantics; corpora, machine translation, IE-IR…).

- The Aholab Signal Processing Laboratory Group of the University of the Basque Country (aholab.ehu.es): specialized in speech technologies (speech synthesis and recognition, speaker identification…).

AnHitz is a three-year project that started in 2006 and will finish in 2008. Thanks to this project seven resources, nine language tools and five applications for Basque are being developed or improved. Besides, this project will be the first in joining together the various tools for Basque in a single application that will show the potential of the integration of these technologies.

2. Some words about Basque and language technologies

Basque is an agglutinative language with a very rich morphology. 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 still under way.

Language technology development for Basque differs in several aspects from the development of similar technologies for widely used and standardized languages (French [1], German (verbomovil.dfki.de), Swedish (www.speech.kth.se/ctt), Norwegian [2], Dutch-Flemish [3]). This is mainly due to two reasons:

- The size of the speakers' community is small. As a result, there are not enough specialized human resources, they lack financial support, and commercial profitability is, in almost all cases, a very difficult goal to reach.
- Due to its rich inflectional morphology, Basque requires specific procedures for language analysis and generation. Thus, it is not always possible to reuse language technologies developed for other languages. This is relevant in both rule-based and corpus-based approaches. This applicability (or portability) depends largely on language similarity.

For these reasons, we believe that research and development for Basque should be (and, in the case of the members of AnHitz, usually is) approached following these guidelines:

- High standardization of resources to be useful in different lines of research, tools and applications.
- Reuse of language resources, tools, and applications.
- 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 resources and research are essential to create any tool or application; but, by the same token, tools and applications will be very helpful in the research and improvement of language resources.
- Use of open source tools.

3. Resources, tools and applications

Some of the organizations that are part of AnHitz have been working in Natural Language Processing and Language Engineering for Basque since 1990. The most basic tools and resources (lemmatizers, POS taggers, lexical databases, speech databases, electronic dictionaries, etc.) had been developed before AnHitz, but most of them have been further improved (and are still being so) within it. And, as mentioned above, many others are being created in this project.

In the following subsections we will mention some of them.

3.1. Resources

Textual resources:

- ZT Corpusa (www.ztcorpusa.net) [4]: A 8.5-million-word tagged collection of specialized texts in Basque, which aims to be a major resource in research and development with respect to written technical Basque [4]. It is the first specialized corpus in Basque, it has been designed to be a methodological and functional reference for new projects in the future (i.e. a national corpus for Basque), it is the first corpus in Basque annotated using a TEI-P4 compliant XML format, it is the first written corpus in Basque to be distributed by ELDA and it has a friendly and sophisticated query interface. The corpus has two kinds of annotation, a structural annotation and a stand-off linguistic annotation. It is composed of two parts, a 1.6 million-word balanced section, whose annotation has been revised by hand, and another automatically tagged 6 million-word part. This corpus is being enhanced and upgraded under the project AnHitz.
- EPEC: A 300,000-word corpus tagged and disambiguated at the morphological, syntactic (syntactic functions and deep dependencies) and semantic level (word senses). It is a strategic resource for the processing of Basque and it has already been used for the development and improvement of some tools. Half of this collection was obtained from the Statistical Corpus of 20th Century Basque (www.euskaracorpusa.net), and the other half was extracted from Euskaldunon

Egunkaria (www.egunero.info), the only daily newspaper written entirely in standard Basque. A subset of 50.000 words of EPEC was used in the last CONLL Competition.

Speech resources:

- SpeechDat FDB1060-EU: A SpeechDat-like database for Basque that contains the recordings of 1,060 speakers of Basque obtained over the fixed telephone network. Each speaker uttered around 43 read and spontaneous items. The database is available at ELRA (catalog.elra.info).
- SpeechDat MDB600-EU: Another SpeechDat-like database for Basque that contains the recordings of 660 speakers of Basque recorded over the mobile telephone network.
- EMODB [5]: Emotional speech database recorded by a female speaker in the six MPEG4 emotions and neutral style. It contains 20 isolated digits, 40 isolated words, 55 isolated sentences repeated for all the styles and 55 different sentences for each of the six emotions. A laryngograph was used to obtain the glottal pulse signal. The speech and laryngograph signals were digitized at 32 kHz with 16 bits.
- Amaia and Aitor [6]: Emotional speech database containing 702 phonetically balanced sentences repeated for the six MPEG4 emotions and neutral style, for female and male voices. It also contains a continuous read speech of 8 min, in 7 styles. It was registered at 48kHz, 16bits, semi-professional room, 2 microphones and laryngograph included. The female voice Karolina has been segmented at phone level and manually revised for the neutral style.
- BIZKAIFON (bizkaifon.ehu.es) [7]: Multimodal (speech and video) database for the Western dialects of the Basque language containing thousands of recordings of the many different variants of the western dialect of Basque. Most of them are transcribed to Standard Basque. It is accessible via web and available at ELRA

3.2. Tools

Textual tools:

- Erauzterm [8]: Tool for automatic term extraction from Basque texts and corpora. Implemented by Elhuyar Foundation in collaboration with the IXA group. Reported results: F measure for MWT 0.4229; F measure for OWT 46.93. A recent evaluation in AnHitz using different domain sections of the ZT Corpus has revealed precision values for MWT up to 0.65 for the first 2,000 candidates, and up to 0.75 for OWT over the same range (results for the Electricity & Electronics section).
- ElexBI [9]: Tool for the extraction of pairs of equivalent terms from Spanish-Basque translation memories. It is based on monolingual candidates extraction in Basque (Erauzterm) and Spanish (Freeling), and consequent statistical alignment and extraction of equivalent pairs. Implemented by Elhuyar Foundation. Reported results: up to 0.9 precision for the first 4,000 candidates processing a parallel corpora of 10,900 segments (eu: 110,165 words; es: 153,163 words). In the next months, Elhuyar Foundation will release the ItzulTerm web service. It is implemented basically using ELexBI technology, and will offer a free service by which the user is allowed to process TMs up to 60,000 words in size, then analyze, validate and edit the results of the automatic extraction, and finally export the validated terms.
- Corpusgile and Eulia [4]: Advanced tools to create, linguistically annotate and query corpora. They have been used to build the ZT Corpus and they provide a flexible and extensible infrastructure for creating, visualizing and managing corpora, and for consulting, visualizing and modifying annotations generated by linguistic tools.
- CorpEus (www.corpeus.org) [10]: A web-as-corpus tool for Basque that allows the querying of the internet as if it were a Basque Corpus, showing KWICs and counts of the searched words. It uses morphological query expansion and language-filtering words to optimize searching for Basque.
- Dokusare [11]: System to identify science news of similar content in a multilingual environment by using cross-lingual document similarity techniques. The precision obtained is between 60 and 85%, depending on the languages involved.
- Co3 [12]: A system to automatically build multilingual comparable corpora (Spanish-English-Basque) using the Internet as a source.
- AzerHitz [13]: A system to automatically extract pairs of equivalent terms from Spanish-Basque comparable corpora.
- Elezkari: A cross-lingual information retrieval system focused in Basque, Spanish and English.
- Eulibeltz [14]: Tool to create and linguistically annotate bilingual aligned corpora.

Speech tools:

- AhoT2P: A letter to allophone transcriber for standard Basque.
- AhoTTS_Mod1: A linguistic processor for speech synthesis.

3.3. Applications

Text applications:

- Xuxen [15]: Spell-checker suited to the agglutinative nature of Basque that combines dictionaries and morphological analysis, with versions for many suites, programs and operating systems. Due to fact that Basque was forbidden at school for many years and to its late standardization, nowadays adult speakers did not learn it at school, and so they have many doubts when writing. The spelling-checker Xuxen is quite an effective tool in this kind of situations. Using it people become more confident with the text they are writing. In fact, this program is one of the most powerful tools in the ongoing standardization of Basque. The spell-checker is more complex than equivalents for other languages, because most of them are based on recognizing each word in a list of possible words in the language; but in Basque, because of its rich morphology, it is very difficult to define such a list, and consequently, morphological analysis must be included. Xuxen is publicly available in www.euskara.euskadi.net.
- Lemmatization based dictionaries: We have developed plug-ins for text processors that enable consulting a word in several dictionaries, but, in order to make it more useful for a language like Basque with rich morphology, dictionary consulting is enhanced with lemmatization. That means that, first, morphological analysis is performed, and then, possible lemmas of the word are consulted in the dictionary. At the moment plug-ins exist for three dictionaries: Spanish-Basque, French-Basque and a Basque dictionary of synonyms.
- Elebila (www.elebila.eu) [16]: A public search engine for content in Basque that obtains a lemma-based search by means of morphological query expansion (improving recall in 89%) and results only in Basque by using language-filtering words (improving precision in 70%). The main search machines available nowadays do not offer lemma-based search for Basque; therefore, if you want to find *sagu*, you will find occurrences of just exactly this word, or alternatively, when searching for any word beginning with that word (*sagu**), many wrong documents will be found because they contain any word such as *saguzar* (Basque for *bat*) that does not correspond to the wanted lemma. Consequently, using Elebila users get better quality in their results. Besides, by using language-filtering words, it returns results only in Basque even if the searched word exists also in other languages (technical words, proper nouns, short words, etc.).
- Opentrad-Matxin (www.opentrad.org) [17]: Open-source machine translation system for Spanish-Basque. It has been created using a transfer rule based MT approach. Now we are working in the construction of a multiengine system including three subsystems based on the different approaches to MT: rule-based machine translation, statistical machine translation and example-based machine translation.
- English-Basque MT: A statistical machine-translation system from English to Basque.

Speech applications:

- AhoTTS (aholab.ehu.es/tts/tts_en.html) [18]: A modular Text-To-Speech conversion system for Basque and Spanish. It has a multithread and multilingual architecture, though every module has been developed mainly for the Basque language. The TTS system is structured in two main blocks: the linguistic processing module and the synthesis engine. The fist one generates a list of sounds, according the Basque SAMPA code (aholab.ehu.es/sampa_basque.htm), which consist on the phonetic transcription of the expanded text, together with prosodic information (values of the pitch curve, duration and energy) for each sound. The synthesis engine gets this information to produce the appropriate sounds, by selecting units and then concatenating them. A signal processing algorithm is applied to reduce the distortion that appears due to the concatenating process. AhoTTS includes several synthesis engines, some of them for concatenating diphones (PSOLA; MBROLA based and HNS) and one based on unit selection (corpus based).
- AhoTTS for PDA [19]: AhoTTS is a multiplatform application and as such, it has been adapted to Personal Digital Assistants (PDA). The limited storage and computing capability of these devices make impossible the use of the corpus based synthesis technique. Therefore, only the synthesis engines that use diphone concatenation have been adapted to PDA platforms.

4. Integration of linguistic components into a demo scenario

Apart from developing and/or improving the aforementioned technologies and resources, another main objective in AnHitz is to integrate as many as possible of them in a demo scenario that will show the potential of the different language technologies working together. This has never been done before with language technologies for Basque.

4.1. Features of the system

These are the features of the system we are aiming to build:

- The system will simulate an expert on Science and Technology. It will be able to answer questions or retrieve documents containing some search terms using a multilingual knowledge base.
- It will automatically translate the results to Basque if they are in English or Spanish.
- The interaction with it will be via speech. We will talk to it in Basque, and it will answer speaking in Basque too.
- The system will have a 3D human avatar that will show emotions depending on the success obtained in accomplishing the task.

This system is already in the building stage and is due to be finished by October 2008.

4.2. Modules used in the system

The system will use the following modules:

- A 3D Human Avatar expressing emotions, developed by VICOMTech.
- A Basque Text-To-Speech synthesizer (TTS), developed by Aholab.
- A Basque Automatic Speech Recognition system (ASR), integrated by Robotiker.
- A Basque Question Answering system (QA), developed by IXA, over a Science and Technology knowledge base, compiled by Elhuyar.
- A Basque-Spanish-English Cross Lingual Information Retrieval system (CLIR), developed by Elhuyar, over a Basque-Spanish-English comparable corpus on Science and Technology, compiled by Elhuyar.
- Two Spanish-Basque and English-Basque Machine Translation systems (MT), developed by IXA.

4.3. System architecture

Fig. 1 illustrates how the different modules interact within the system and with the user.

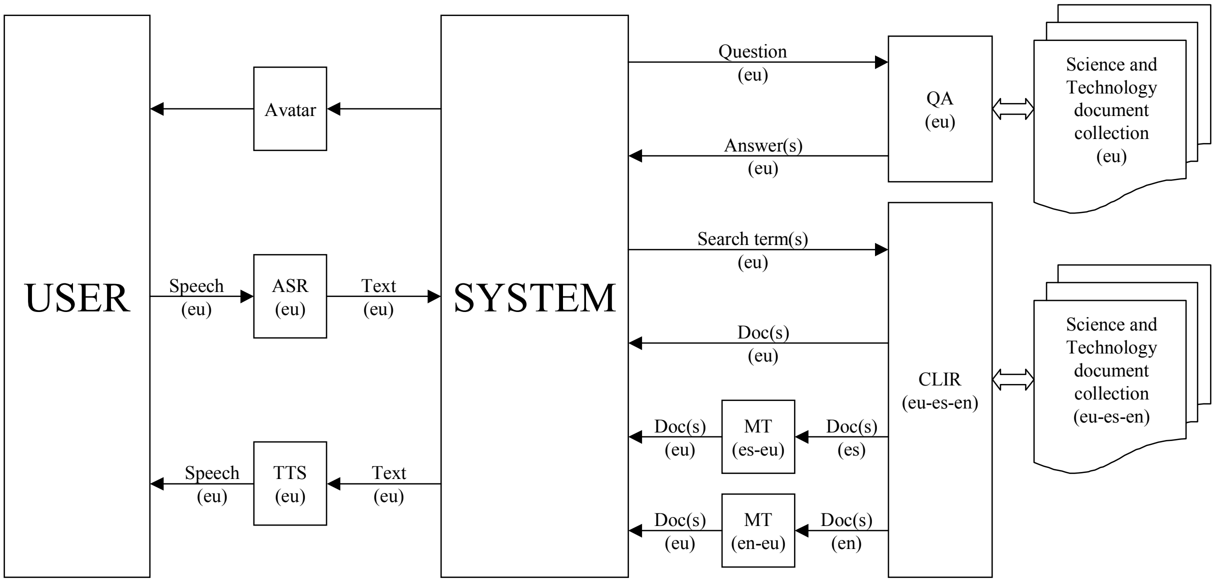

Figure 1. Diagram showing the system architecture.

5. Conclusions

The AnHitz project has proved to be very effective for improving the already existing language and speech resources for Basque and for creating new ones. The system that is now being developed to integrate tools and resources from different areas (an expert in Science and Technology with a human natural language interface) shows that collaboration between agents working in different areas is crucial to really exploit the potential of language technologies and build applications for the end user.

6. Acknowledgements

This work has been partially funded by the Local Government of the Basque Country (AnHitz 2006: Language Technologies for Multingual Interaction in Intelligent Environments., IE06-185).

7. References

[1] S. Chaudiron, J. Mariani, "Techno-langue: The French National Initiative for Human Language Technologies (HLT)", *Proc. of fifth international conference on Language Resources and Evaluation (LREC)*, Genoa, 2006.

[2] B. Maegaard, J. Fenstad, L. Ahrenberg, K. Kvale, K. Mühlenbock, B. Heid, "KUNSTI - Knowledge Generation for Norwegian Language", *Proc. of fifth international conference on Language Resources and Evaluation (LREC)*, Genoa, 2006.

[3] E. D'hallewey, J. Odijk, L. Teunissen, C. Cucchiarini, "The Dutch-Flemish HLT Programme STEVIN: Essential Speech and Language Technology Resources", *Proc. of fifth international conference on Language Resources and Evaluation (LREC)*, Genoa, 2006.

[4] N. Areta, A. Gurrutxaga, I. Leturia, I. Alegria, X. Artola, A. Diaz de Ilarraza, N. Ezeiza, A. Sologaistoa, "ZT Corpus: Annotation and tools for Basque corpora", *Corpus Linguistics 2007 Proceedings*, Birmingham, 2007.

[5] E. Navas, I. Hernáez, A. Castelruiz, I. Luengo, "Obtaining and Evaluating an Emotional Database for Prosody Modelling in Standard Basque", *Lecture Notes on Computer Science 3206*, 2004, pp. 393-400.

[6] I. Saratxaga, E. Navas, I. Hernáez, I. Luengo, "Designing and Recording an Emotional Speech Database for Corpus Based Synthesis in Basque", *Proc. of fifth international conference on Language Resources and Evaluation (LREC)*, Genoa, 2006, pp. 2126-2129.

[7] A. Castelruiz, J. Sánchez, X. Zalbide, E. Navas, I. Gaminde, "Description and Design of a WEB Accessible Multimedia Archive", *Proc. of 12th IEEE Mediterranean Electrotechnical Conference (MELECON)*, Dubrovnik, 2004, pp. 681-684.

[8] A. Gurrutxaga, X. Saralegi, S. Ugartetxea, P. Lizaso, I. Alegria, R. Urizar, "A XML-Based Term Extraction Tool for Basque", *LREC 2004 Proceedings*, 2004.

[9] I. Alegria, A. Gurrutxaga, X. Saralegi, S. Ugartetxea, "Elexbi, A Basic Tool For Bilingual Term Extraction From Spanish-Basque Parallel Corpora", *Euralex 2006 Proceedings*, Torino, 2006.

[10] I. Leturia, A. Gurrutxaga, I. Alegria, A. Ezeiza, "CorpEus, a 'web as corpus' tool designed for the agglutinative nature of Basque", *Web as Corpus 3 workshop Proceedings*, Louvain-la-Neuve, 2007, pp. 69-81.

[11] X. Saralegi, I. Alegria, Similitud entre documentos multilingües de carácter científico-técnico en un entorno web. *Procesamiento del Lenguaje Natural 39*, Sevilla, 2007, pp. 71-78.

[12] I. Leturia, I. San Vicente, X. Saralegi, M. Lopez de Lacalle, "Collecting Basque specialized corpora from the web: language-specific performance tweaks and improving topic precision", *Web as Corpus 4 workshop Proceedings*, Marrakech, 2008, pp. 40-46.

[13] X. Saralegi, I. San Vicente, A. Gurrutxaga, "Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain", *Building and Using Comparable Corpora*, Marrakech, 2008, pp. 27-32.

[14] A. Díaz de Ilarraza, J. Igartua, K. Sarasola, A. Sologaistoa, A. Casillas, R. Martinez, "Spanish-Basque Parallel Corpus Structure: Linguistic Annotations and Translation Units", *Proceedings of TSD 2007 Conference*, Plzen, 2007.

[15] I. Aduriz, I. Alegria, X. Artola, N. Ezeiza, K. Sarasola, "A spelling corrector for Basque based on morphology", *Literary & Linguistic Computing, Vol. 12, No. 1*, Oxford University Press, Oxford, 1997, pp. 31-38.

[16] I. Leturia, A. Gurrutxaga, N. Areta, I. Alegria, A. Ezeiza, "EusBila, a search service designed for the agglutinative nature of Basque", *Proceedings of iNEWS'07 workshop in SIGIR*, Amsterdam, 2007, pp. 47-54.

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

[18] I. Hernáez, E. Navas, J.L. Murugarren, B. Etxebarria, "Description of the AhoTTS Conversion System for the Basque Language", *Proceedings of 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Edinburgh, 2001.

[19] J. Sanchez, I. Luengo, E. Navas, I. Hernáez, "Adaptation of the AhoTTS Text to Speech System to PDA Platforms", *Proceedings of the SPECOM 2006*, San Petersburg, 2006, pp 292-296.