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

Kutz Arrieta[1], Arantza Diaz de Ilarraza[2], Inma Hernáez[3], Urtza Iturraspe[4], Igor Leturia[5], Eva Navas[3], Kepa Sarasola[2]

[1] VICOMTech
[2] IXA Group - University of the Basque Country
[3] Aholab Group - University of the Basque Country
[4] Robotiker
[5] Elhuyar Foundation

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 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.
- 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, MT, IE-IR…).
- The Aholab Signal Processing Laboratory 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 and tools.
- 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
- 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 present some of them.

3.1. Resources

- Textual resources:
  - ZT Corpusa (www.ztcorpusa.net): 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 [1].
  - EPEC: a 300,000-word corpus tagged and disambiguated at the morphological, syntactic (syntactic functions and deep dependencies) and semantic level (word senses).
- 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.
  - 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: emotional speech database recorded by a female speaker in the six MPEG4 emotions and neutral style [2].
  - Amaia and Aitor: emotional speech database containing 702 phonetically balanced sentences repeated for the six MPEG4 emotions and neutral style, for female and male voices [3].
  - BIZKAIFON (bizkaifon.ehu.es): 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 [4].

3.2. Tools

- Textual tools:
  - Erauzterm: tool for automatic term extraction from Basque texts and corpora [5].
  - ElexBI: tool for the extraction of pairs of equivalent terms from Spanish-Basque translation memories [6].
  - Corpusgile and Eulia: advanced tools to create, linguistically annotate and query corpora [1].
  - CorpEus (www.corpeus.org): a web-as-corpus tool that allows querying of the internet as if it were a Basque Corpus, showing KWICs and counts of the search terms; it uses morphological query expansion and language-filtering words to optimize searching for Basque [7].
  - Dokusare: system to identify science news of similar content in a multilingual environment by using cross-lingual document similarity techniques [8].
  - Co3: a system to automatically build multilingual comparable corpora (Spanish-English-Basque) using the Internet as a source [9].
  - AzerHitz: a system to automatically extract pairs of equivalent terms from Spanish-Basque comparable corpora [10].
  - Elezkari: a cross-lingual information retrieval system focused in Basque, Spanish and English.
  - Eulibeltz: tool to create and linguistically annotate bilingual aligned corpora [11].
- 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: spell-checker suited to the agglutinative nature of Basque that combines dictionaries with morphology, with versions for many programs and operating systems [12].
  - Elebila (www.elebila.eu): 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%) [13].
  - Opentrad-Matxin (www.opentrad.org): open-source machine translation system for Spanish-Basque [14].
  - English-Basque MT: a statistical machine-translation system from English to Basque.
- Speech applications:
  - AhoTTS: a modular Text-To-Speech conversion system for Basque and Spanish [15], which has also been adapted for PDA systems [16].

4. INTEGRATION OF SYSTEMS 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 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.

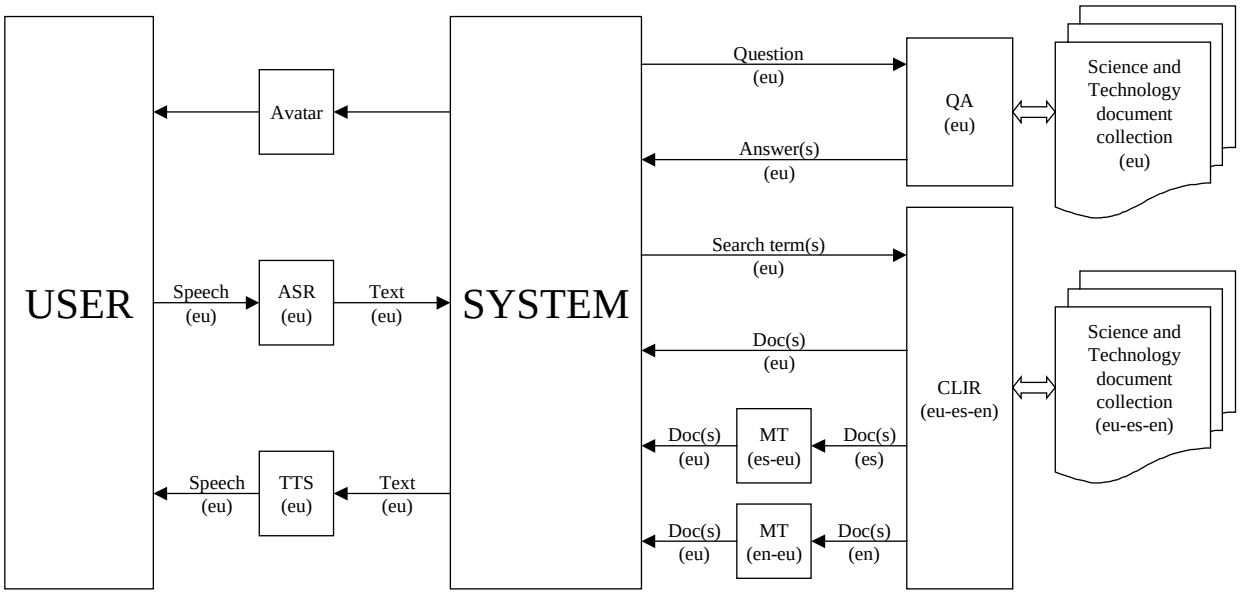

Figure 1. *Diagram showing the system architecture.*

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. ACKNOWLEDGMENTS

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.