Proceedings of the International Multiconference on
Computer Science and Information Technology, pp. 235 – 242

ISBN 978-83-60810-22-4
ISSN 1896-7094

Development and Evaluation of AnHitz, a Prototype of a Basque-Speaking Virtual 3D Expert on Science and Technology

Igor Leturia
Elhuyar Foundation
Zelai Haundi kalea 3
Osinalde Industrialdea
20170 Usurbil, Spain
Email:
igor@elhuyar.com

Arantza del Pozo, Kutz Arrieta
VICOMTech
Mikeletegi pasealekua, 57
Miramon Teknologia Parkea
20009 Donostia-SanSebastian, Spain
Email: {adelpozo,karrieta}@vicomtech.org

Urtza Iturraspe
Robotiker
202. eraikina
Zamudioko Teknologia
Parkea,
48170 Zamudio, Spain
Email:
uiturraspe@robotiker.es

Kepa Sarasola, Arantza Diaz de Ilarraza
IXA Group, University of the Basque Country
Informatika Fakultatea
649 posta-kutxa
20080 Donostia-San Sebastian, Spain
Email: {kepa.sarasola,a.diazdeilarraza}@ehu.es

Eva Navas, Igor Odriozola
Aholab Group, University of the Basque Country
Ingeniaritza Goi Eskola Teknikoa
Urkijo Zumardia, z.g.
48013 Bilbao, Spain
Email: eva.navas@ehu.es, igor@aholab.ehu.es

***Abstract*—The aim of the AnHitz project, whose participants are research groups with very different backgrounds, is to carry out research on language, speech and visual technologies for Basque. Several resources, tools and applications have been developed in AnHitz, but we have also integrated many of these into a prototype of a 3D virtual expert on science and technology. It includes Question Answering and Cross Lingual Information Retrieval systems in those areas. The interaction with the system is carried out in Basque (the results of the CLIR module that are not in Basque are translated through Machine Translation) and is speech-based (using Speech Synthesis and Automatic Speech Recognition). The prototype has received ample media coverage and has been greatly welcomed by Basque society. The system has been evaluated by 50 users who have completed a total of 300 tests, showing good performance and acceptance.**

I. Introduction

ANHITZ is a project promoted by the Basque Government in its Science and Technology Plan for 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 (http://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 the EiTB, the Basque Radio and Television Group.
- Elhuyar Foundation (http://www.elhuyar.org): A non-profit making organization that aims 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 R&D in language technologies for Basque.
- Robotiker (http://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 (http://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 (http://aholab.ehu.es): Specialized in speech technologies (speech synthesis and recognition, speaker identification…).

AnHitz is a three-year project that started in 2006 and finished in 2008. Thanks to this project several resources, language tools and applications for Basque have been developed or improved. Besides, this project has been the first in joining together various tools for Basque into a single application that shows the potential of the integration of these technologies.

II. Some Words about Basque and Language Technologies

Basque is an agglutinative language with a very rich morphology. There are around 700,000 Basque speakers, about 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 standard-

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

ized, 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], 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, since 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 maximum 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.

III. Resources, Tools and Applications Developed

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 within it. And, as mentioned above, many others have being created in this project. We will mention some in the following subsections.

A. Textual Resources

- ZT Corpusa (http://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 AnHitz project.
- 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 a number of tools. Half of this collection was obtained from the Statistical Corpus of 20th Century Basque (http://www.euskaracorpusa.net), and the other half was extracted from Euskaldunon Egunkaria, the only daily newspaper written entirely in standard Basque. A subset of 50,000 words of EPEC was used in the last CONLL Competition.

B. 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 (http://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 (http://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.

C. Textual Tools

- Erauzterm [8]: Tool for automatic term extraction from Basque texts and corpora. Implemented by the Elhuyar Foundation in collaboration with the IXA group. Reported results: F measure for MWT 0.4229; F measure for OWT 0.4693. 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 corpus of 10,900 segments (eu: 110,165 words; es: 153,163 words). In the coming months, the Elhuyar Foundation will be releasing the ItzulTerm web service. It is implemented basically by 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 (http://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, which can obtain a domain precision of over 90%.
- AzerHitz [13]: A system to automatically extract pairs of equivalent terms from Spanish-Basque comparable corpora, obtaining a precision of 58% in top 1 and 79% in top 20 for high-frequency words.
- Elezkari [14]: A cross-lingual information retrieval system focused in Basque, Spanish and English that yields a MAP value of 0.2960 for English with the CLEF 2001 collection (Basque and Spanish have not been evaluated).
- Eulibeltz [15]: Tool to create and linguistically annotate bilingual aligned corpora.
- Eihera [16]: Named entity recognizer for Basque with an F-Score of 85.37.

D. Speech Tools

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

E. Text Applications

- Xuxen [17]: 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 the fact that Basque was forbidden at school for many years and to its late standardization, today's 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 situation. 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 specify such a list, and consequently, morphological analysis must be included. Xuxen is publicly available at http://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 (http://www.elebila.eu) [18]: 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, by 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 (http://www.opentrad.org) [19] [20]: Open-source machine translation system for Spanish-Basque. It has been created using a transfer rule-based MT approach. Its average HTER evaluation result is 0.42, meaning that 42 editing corrections are required for every 100 tokens. Now we are working on the construction of a multi-engine system including three subsystems based on the different approaches to MT: rule-based, statistical and example-based.
- English-Basque MT: A statistical machine-translation system from English to Basque.

• Ihardetsi [21]: A Question Answering system for Basque that got a precision of 13% in QA@CLEF2008.

F. Speech Applications

• AhoTTS (http://aholab.ehu.es/tts/tts_en.html) [22]: 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 first one generates a list of sounds, according the Basque SAMPA code (http://aholab.ehu.es/sampa_basque.htm), which consist of 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 [23]: AhoTTS is a multiplatform application and as such, has been adapted to Personal Digital Assistants (PDA). The limited storage and computing capability of these devices make the use of the corpus-based synthesis technique impossible. Therefore, only the synthesis engines that use diphone concatenation have been adapted to PDA platforms.

• ZTRec: A Basque speech recognizer of science and technology terms and questions.

G. Visual Applications

• AnHitzDlg: Avatar with bidirectional spoken communication in Basque.

IV. Integration of Components into a Demo Application

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

A. Features of the System

These are the features of the system we have built:

• It simulates an expert on Science and Technology. It is able to answer questions (such as “who invented the telescope?” or “when was Newton born?”) or retrieve documents containing some search terms (such as “ozone layer” or “renewable energies”) using a multilingual knowledge base.

• It automatically translates the results into Basque if they are in English or Spanish.

• The interaction with the system is speech-based. The user speaks in Basque, and the system answers speaking Basque too.

• The system has a 3D human avatar that shows emotions depending on the success obtained in accomplishing the task.

The demo system has been given the same name as the project, AnHitz. A screen capture of the system is shown in Fig. 1.

B. Modules Used in the System

The system makes use of 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.

C. System Architecture

Fig. 2 illustrates how the different modules interact within the system and with the user.

D. System Integration Process

The main problem we encountered when integrating the modules of the system was that, since there were many different entities developing the modules, each module had been built using different technologies, libraries and, above all, operating systems, and it was very difficult to mount them all in one computer.

That was why only the 3D avatar and the automatic speech recognition modules were installed in the laptop provided for the application (both run in Windows), and the rest of them were made available as web services in their respective homes and are called by the system. This method proved to be appropriate, and simplified the integration enormously.

However, this was at the cost of some speed, especially when the system has to produce speech (an audio file is sent from the TTS module to the system via the web). We improved speed by locally caching the most repetitive conversational sentences; and we intend to improve it further by installing all the modules in the same computer, using virtualization for the different operating systems.

Another problem was the frustration experienced when the ASR system did not understand correctly what the user said but launched the query process all the same. To avoid this, we used the confidence level returned by the ASR system, and empirically found reasonably good thresholds of this confidence level for correct recognition, doubtful recognition and incorrect recognition. Thus, the system asks for confirmation in the case of doubtful recognition and repeats the question in the case of incorrect recognition; this way,

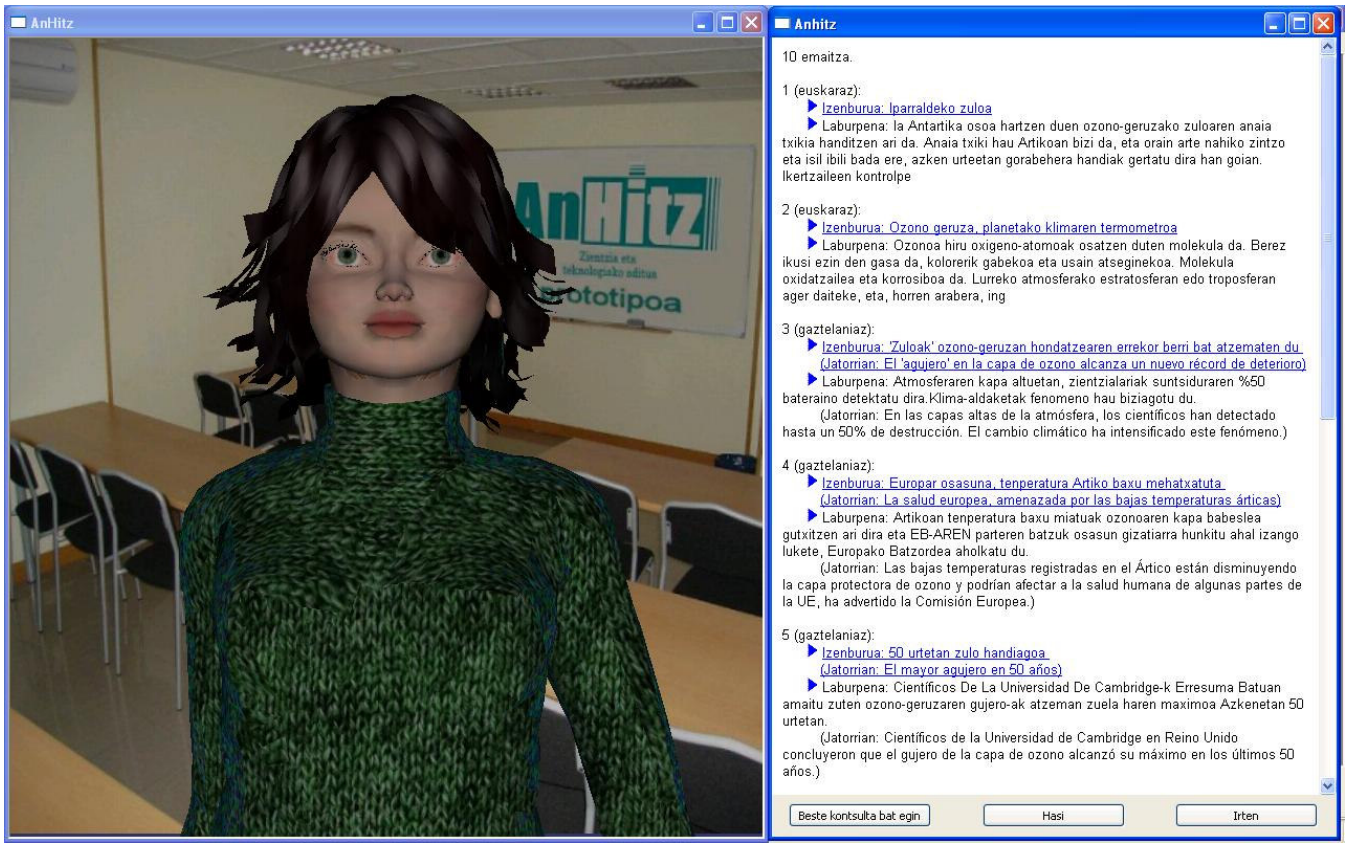

Fig 1. Screen capture of the system

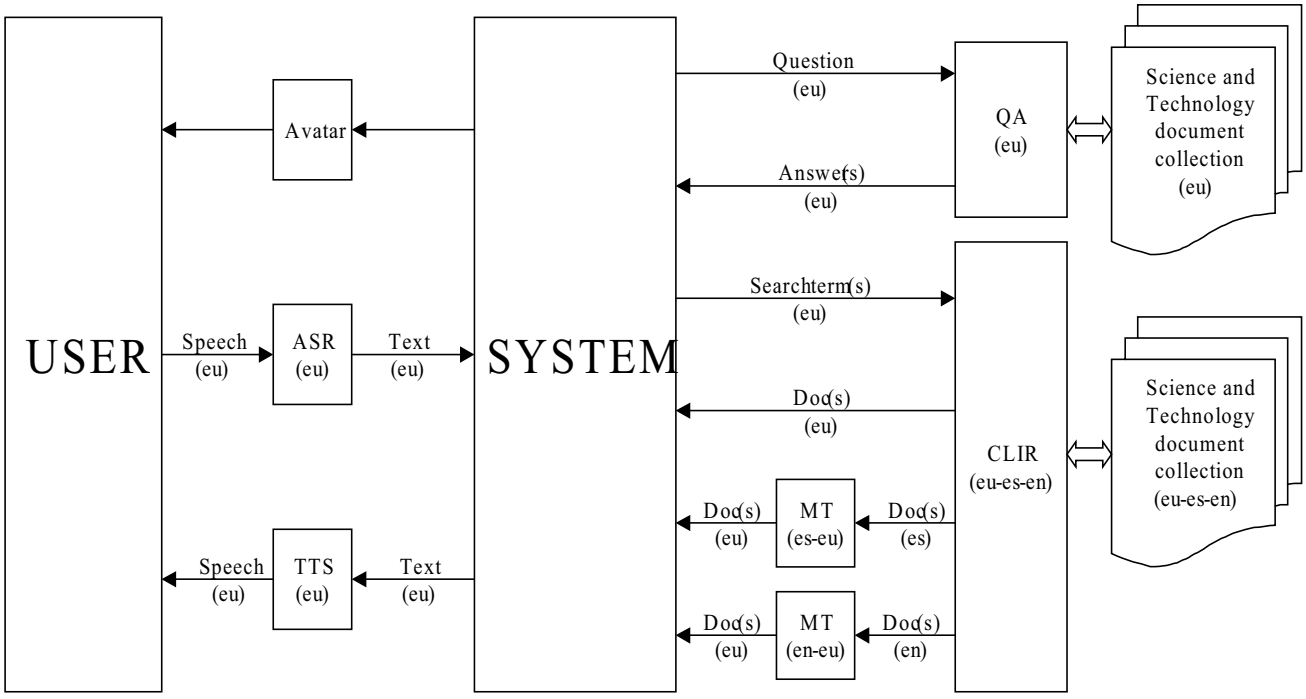

Fig 2. System architecture

its performance is greatly improved (only in 10.79% of the cases does it proceed with an incorrect recognition).

Another question remaining was the fact that the Basque ASR system is not a general dictating system, but one based on grammars and dictionaries. We could not find a way of specifying one single grammar that would include all the possible answers in a conversation with the system, so we specified different grammars for different steps of the interaction: one for no/yes/maybe answers, one for telling the system your name (using lists of the most usual Basque names), one for the most usual scientific search terms, and one for the most usual scientific questions (the last two were specified using the search logs of a popular Basque science portal). The system calls up the ASR system with the appropriate grammar for each stage of the conversation.

V. Dissemination

At the end of the AnHitz project, its participants and some members of the Basque Government gave a press conference, which was very well attended by the media. Practically every radio, TV or newspaper covered the news the same day or the next. Furthermore, the demo prototype aroused great interest, and many media devoted a video, interview or article to it. Some of these appearances of AnHitz in the media can be seen in http://www.elhuyar.org/hizkuntza-zerbitzuak/EN/Anhitz-project.

We also showed the prototype to the general public during the Week of Science and Technology 2008, in two stands in Donostia-San Sebastian and Bilbao. Students from schools and members of the public in general had the chance to try it out and play with it, and they were generally surprised and interested.

VI. Evaluation

The demo prototype developed in AnHitz has been evaluated in order to measure its performance and weigh the impression of potential users about it. 50 users have formulated 3 questions and 3 cross-lingual searches each, making 300 tests in total. During the interaction of the testers with the system, some objective observations were noted down, such as the number of failures and successes of the ASR or QA systems. At the end of the interaction, the testers filled in a questionnaire about more subjective matters (quality of the TTS, CLIR or MT systems, general impression, etc.). The results of this evaluation are explained in the following subsections, and all of them are shown in Table 1.

A. ASR

The ASR system understood correctly 63.19% of the times. Another 12.59% of the times it understood correctly, although it was not sure and asked for confirmation. 13.43% of the times it did not understand correctly, but it asked for confirmation and so the user could repeat the phrase. Only in 10.79% of the cases did the system understand wrongly without giving the option to correct.

When asked if AnHitz had understood what they said, 55.11% of the testers answered “almost always” or “most of the times”, 34.69% “sometimes” and 10.20% “a few times”. No one chose “hardly ever”.

B. TTS

When asked about the understandability of AnHitz's speech, 85.42% responded “very good” or “good” and 14.58% “quite good”. No one chose “bad” or “very bad”.

43.75% of the testers judged the speech as “very natural” or “natural”, 31.25% “quite natural” and 25.00% “artificial” or “very artificial”.

C. QA

The question answering system answered correctly 30.61% of the times, and in another 15.30% the correct answer was among the first five possible answers given. 54.08% of the times the system did not give a correct solution or did not answer at all. However, some of these incorrect outcomes might be due to the correct answer not being in the corpus, and so the results could have been better (although we cannot tell for sure since no evaluation was carried out).

D. CLIR

The users judged the CLIR results “very good” or “good” 68.35% of the times; in 22.30% of the cases they found them “quite bad” and in 9.35% “completely unrelated”.

E. MT

30.00% of the times the users found the translations of the MT system “very good”, “good” or “quite good”, in another 38.89% they found them “comprehensible” and in another 31.11% “quite bad”, “bad” or “very bad”.

F. Overall impression

When asked if they thought the system was useful, 62.50% of the users answered “very useful” or “useful” and 37.50% “quite useful”. No one said it was “quite useless” or “completely useless”.

When asked if they would like to see this kind of speech interaction in other uses, 20.83% said “it should always be like this with machines”, 39.58% that they would like to see it “in many cases” and another 39.58% “in some cases”. No one chose “maybe in a few cases” or “never”.

VII. 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 has been 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. The evaluation that the system has been subjected to proves that, although it is based on systems still in the research stage, its performance is acceptable. The responses obtained from the users in the evaluation and from the media lead us to believe that these kinds of applications based on language technologies are

TABLE I.
RESULTS OF THE EVALUATION

ASR	**Understanding**	**%**
	Correct	63,19
	Correct although not sure	12,59
	Not correct but not sure	13,43
	Wrong	10,79
	Did the system understand what you said?	**%**
	Almost always	16,33
	Most of the times	38,78
	Sometimes	34,69
	A few times	10,20
	Hardly ever	0
TTS	**How do you rate the system's understandability?**	**%**
	Very good	66,67
	Good	18,75
	Quite good	14,58
	Bad	0
	Very bad	0
	Was the system's speech natural?	**%**
	Very natural	10,42
	Natural	33,33
	Quite natural	31,25
	Artificial	22,92
	Very artificial	2,08
QA	**Correct answer**	**%**
	In the 1st place	30,61
	In the 2nd place	8,16
	In the 3rd place	1,02
	In the 4th place	3,06
	In the 5th place	3,06
	The right answer was not among the possible answers	36,73
	The system did not answer at all	17,35

CLIR	**How good were the results?**	**%**
	Very good	28,06
	Good	40,29
	Quite bad	22,30
	Completely unrelated	9,35
MT	**How good were the translations?**	**%**
	Very good	4,44
	Good	8,89
	Quite good	16,67
	Comprehensible	38,89
	Quite bad	26,67
	Bad	2,22
	Very bad	2,22
Overall	**Did you find the system useful?**	**%**
	Very useful	25
	Useful	37,5
	Quite useful	37,5
	Quite useless	0
	Completely useless	0
	Would you like to see speech interaction in other uses?	**%**
	Yes, it should always be like this with machines	20,83
	In many cases	39,58
	In some cases	39,58
	Maybe in a few cases	0
	Never	0

very interesting for society and that people would like to see them available in their everyday lives.

References

[1] S. Chaudiron, and J. Mariani, "Techno-langue: The French National Initiative for Human Language Technologies (HLT)", in *Proc. of fifth international conference on Language Resources and Evaluation (LREC)* , Genoa, 2006, pp. 767-772.

[2] B. Maegaard, J. Fenstad, L. Ahrenberg, K. Kvale, K. Mühlenbock, and B. Heid, "KUNSTI - Knowledge Generation for Norwegian Language", in *Proc. of fifth international conference on Language Resources and Evaluation (LREC)*, Genoa, 2006, pp. 757-760.

[3] E. D'hallewey, J. Odijk, L. Teunissen, and C. Cucchiarini, "The Dutch-Flemish HLT Programme STEVIN: Essential Speech and Language Technology Resources", in *Proc. of fifth international conference on Language Resources and Evaluation (LREC)*, Genoa, 2006, pp. 761-766.

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

[5] E. Navas, I. Hernáez, A. Castelruiz, and I. Luengo, "Obtaining and evaluating an emotional database for prosody modelling in standard Basque", in *Lecture Notes on Computer Science 3206*, 2004, pp. 393-400.

[6] I. Saratxaga, E. Navas, I. Hernáez, and I. Luengo, "Designing and recording an emotional speech database for corpus based synthesis in Basque", in *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, and I. Gaminde, "Description and design of a web accessible multimedia archive", in *Proc. of 12th IEEE Mediterranean Electrotechnical Conference (MELECON)*, Dubrovnik, 2004, pp. 681-684.

[8] A. Gurrutxaga, X. Saralegi, S. Ugartetxea, P. Lizaso, I. Alegria, and R. Urizar, "A XML-based term extraction tool for Basque", in *Proceedings of LREC 2004*, Lisbon, 2004, pp. 1733-1736.

[9] I. Alegria, A. Gurrutxaga, X. Saralegi, and S. Ugartetxea, "Elexbi, a basic tool for bilingual term extraction from Spanish-Basque parallel corpora", in *Proceedings of Euralex 2006*, Torino, 2006, pp. 159-165.

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

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

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

[13] X. Saralegi, I. San Vicente, and A. Gurrutxaga, "Automatic extraction of bilingual terms from comparable corpora in a popular science domain", in *Proceedings of Building and Using Comparable Corpora workshop*, Marrakech, 2008, pp. 27-32.

[14] X. Saralegi, and M. López de Lacalle, "Comparing different approaches to treat Translation Ambiguity in CLIR: Structured Queries vs. Target Co-occurrence Based Selection", in *Proceedings of the 6th International Workshop on Text-Based Information Retrieval*, Linz, 2009.

[15] A. Díaz de Ilarraza, J. Igartua, K. Sarasola, A. Sologaistoa, A. Casillas, and R. Martinez, "Spanish-Basque Parallel Corpus Structure: Linguistic Annotations and Translation Units", in *Proceedings of TSD 2007 Conference*, Plzen, 2007, pp. 230-237.

[16] I. Alegria, O. Arregi, N. Ezeiza, and I. Fernandez, "Lessons from the development of a named entity recognizer for Basque", in *Procesamiento del Lenguaje Natural 36*, Zaragoza, 2006, pp. 25-37

[17] I. Aduriz, I. Alegria, X. Artola, N. Ezeiza, and 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.

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

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

[20] A. Mayor, I. Alegria, A. Diaz de Ilarraza, G. Labaka, M. Lersundi, and K. Sarasola, "Evaluación de un sistema de traducción automática basado en reglas o por qué BLEU sólo sirve para lo que sirve", in *Procesamiento del Lenguaje Natural 43*, San Sebastian, 2009.

[21] O. Ansa, X. Arregi, A. Otegi, and A. Soraluze, "Ihardetsi question answering system at QA@CLEF 2008", in *Working Notes of the Cross-Lingual Evaluation Forum*, Aarhus, Denmark, 2008.

[22] I. Hernáez, E. Navas, J.L. Murugarren, and B. Etxebarria, "Description of the AhoTTS conversion system for the Basque language", in *Proceedings of 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Edinburgh, 2001.

[23] J. Sanchez, I. Luengo, E. Navas, and I. Hernáez, "Adaptation of the AhoTTS text to speech system to PDA platforms", in *Proceedings of the SPECOM 2006* , San Petersburg, 2006, pp 292-296.