Web Communication Protocols for Coordinating the Modules of AnHitz, a Basque-Speaking Virtual 3D Expert on Science and Technology

Igor Leturia
Elhuyar Foundation
Zelai Haundi kalea 3
Osinalde Industrialdea
20170 Usurbil, Spain
i.leturia@elhuyar.com

Arantza del Pozo, David Oyarzun
Vicomtech
Mikeletegi pasealekua, 57
Miramon Teknologia Parkea
20009 Donostia-San Sebastian, Spain
{adelpozo, doyarzun}@vicomtech.org

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

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

Eva Navas, Igor Odriozola,
Iñaki Sainz
Aholab Group, Basque Country University
Ingeniaritza Goi Eskola Teknikoa
Urkijo Zumardia, z.g.
48013 Bilbao, Spain
{eva.navas,igor.odriozola}@ehu.es, inaki@aholab.ehu.es

Abstract

AnHitz is a prototype of a virtual Basque-speaking 3D expert that can answer questions or perform cross-lingual searches on science and technology, and show the search results in Basque by means of machine translation. It has been named after the 3-year strategic research project on language, speech and visual technologies for Basque carried out by several organizations, and built as a demonstrator of the technologies developed. Because the six modules comprising AnHitz have been implemented by different organizations using different operating systems, programming languages or libraries, it was impossible to build the demonstrator in a single executable or machine. As a result, the prototype has been constructed using separate programs in various machines that interact using network communication and web services protocols. Despite the employed approach having some drawbacks –mainly higher time delays when audio rendering is done via the web–, the outcome is indeed satisfactory, as it has allowed us to build a fully functional demonstrator that has showed good performance and acceptance in an evaluation made with 50 users, and has made a great impact on the Basque media.

1. Background

1.1 The AnHitz project

AnHitz is a prototype demonstrator that sets out to show the potential of the integration of language, speech and visual technologies. It is the outcome of a 3-year strategic research project on language, speech and visual technologies for Basque, also called AnHitz. This project has been promoted by the Basque Government in its Science and Technology Plan for 2006-2008 to develop language technologies for Basque.

1.2 The AnHitz consortium

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 centre 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 broadcasting corporation.
- Robotiker (http://www.robotiker.com): A technology centre specialized in information and telecommunication technologies, part of the Tecnalia Technology Corporation.
- Elhuyar Foundation (http://www.elhuyar.org): A non-for-profit 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.
- 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…).

1.3 Vision

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 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 such as French (Chaudiron & Mariani, 2006), Norwegian (Maegaard et al., 2006) or Dutch-Flemish (D'hallewey et al., 2006). 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.

1.4 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 many others have been created in this project:

- Textual resources: ZT Corpusa (Areta et al., 2007), a corpus of science and technology texts (http://www.ztcorpusa.net); EPEC, a corpus tagged and disambiguated at the morphological, syntactic and semantic levels.
- Speech resources: SpeechDat FDB1060-EU, a SpeechDat-like database for Basque that contains recordings obtained over the fixed telephone network; SpeechDat MDB600-EU, another SpeechDat-like database for Basque that contains recordings obtained over the mobile telephone network; EMODB (Navas et al., 2004), emotional speech database recorded by a female speaker in the six MPEG4 emotions and neutral style; Amaia and Aitor (Saratxaga et al., 2006), emotional speech database phonetically balanced for female and male voices; BIZKAIFON (Castelruiz et al., 2004), multimodal (speech and video) database for the Western dialects of the Basque language (http://bizkaifon.ehu.es).
- Textual tools: Erauzterm (Gurrutxaga et al., 2004), tool for automatic term extraction from Basque texts and corpora; ElexBI (Alegria et al., 2006a), tool for the extraction of pairs of equivalent terms from Spanish-Basque translation memories (http://itzulterm.elhuyar.org/); Corpusgile and Eulia (Areta et al., 2007), advanced tools to create, linguistically annotate and query corpora; CorpEus (Leturia et al., 2007a), a web-as-corpus tool for Basque that allows the querying of the Internet as if it were a Basque corpus (http://www.corpeus.org); Dokusare (Saralegi & Alegria, 2007), a system to identify science news of similar content in a multilingual environment by using cross-lingual document similarity techniques; Co3 (Leturia et al., 2009), a system to automatically build multilingual comparable corpora (Spanish-English-Basque) using the Internet as a source; AzerHitz (Saralegi et al., 2008), a system to automatically extract pairs of equivalent terms from Spanish-Basque comparable corpora; Elezkari (Saralegi & López de Lacalle, 2009), a cross-lingual information retrieval system focused on Basque, Spanish and English; Eulibeltz (Díaz de Ilarraza et al., 2007), a tool to create and linguistically annotate bilingual aligned corpora; Eihera (Alegria et al., 2006b), named entity recognizer for Basque.
- Speech tools: AhoT2P, a letter to allophone transcriber for standard Basque; AhoTTS_Mod1, a linguistic processor for speech synthesis.
- Text applications: Xuxen (Aduriz et al., 1997), spell-checker suited to the agglutinative nature of Basque that combines dictionaries and morphological analysis; Lemmatization based dictionaries for text-processors; Elebila (Leturia et al., 2007b), a public search engine for content in Basque (http://www.elebila.eu); Opentrad-Matxin (Alegria et al., 2007; Alegria et al., 2008), open-source machine translation system for Spanish-Basque (http://www.opentrad.org); English-Basque statistical MT (Stroppa et al., 2006); Ihardetsi (Ansa et al., 2008), a Question Answering system for Basque.
- Speech applications: AhoTTS (Hernáez et al., 2001), a modular Text-To-Speech conversion system for Basque, Spanish and English (http://aholab.ehu.es/tts/tts_en.html); AhoTTS for

PDA (Sanchez et al., 2006); ZTRec, a Basque speech recognizer of science and technology terms and questions.

- Visual applications: AnHitzDlg, Basque speaking avatar.

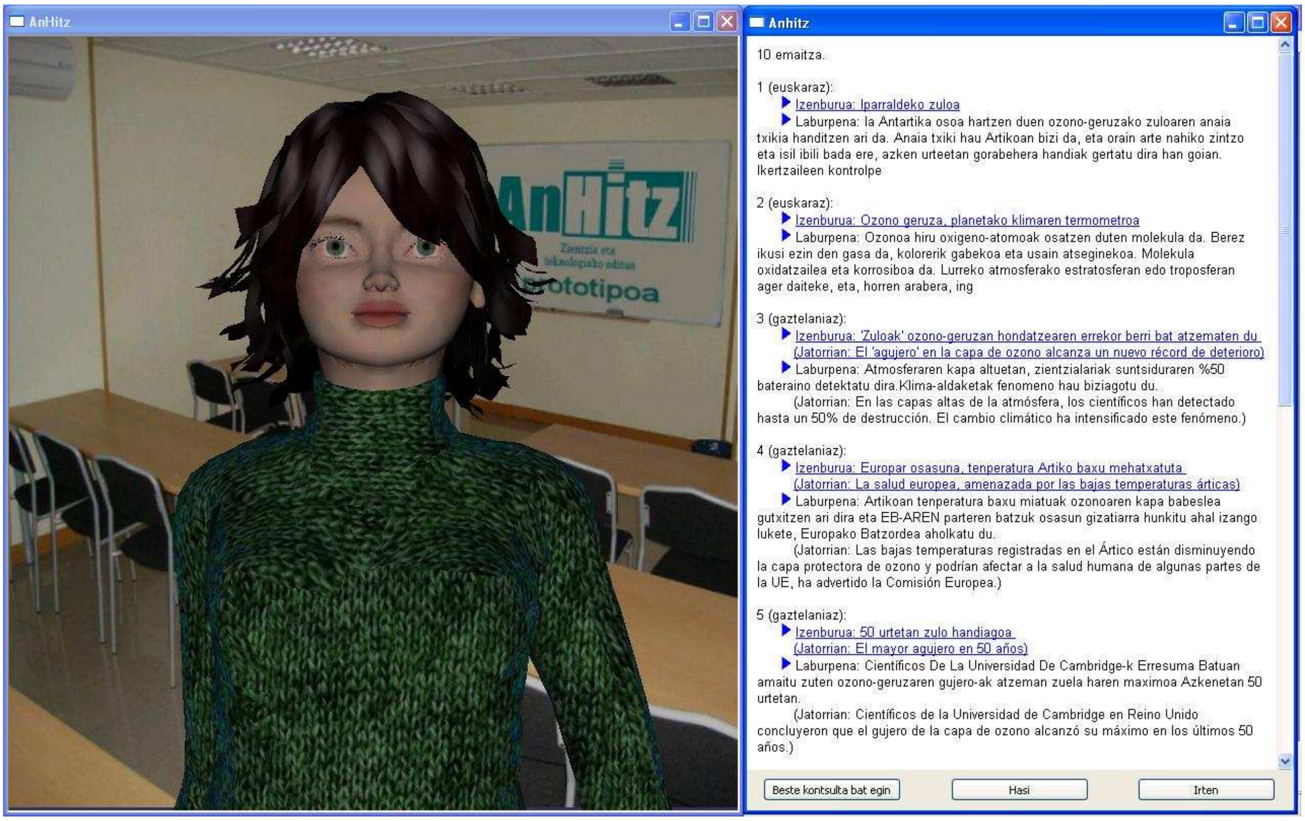
2. The AnHitz demo

Apart from the various resources, tools and applications developed in the AnHitz project, we have also built a demonstrator that sets out to show the potential of all these different technologies working together. It is a 3D virtual expert on science and technology that 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).

This is a typical conversation with AnHitz when performing a cross-lingual search:

- Kaixo, nola duzu izena? (*Hello, what's your name?*)
- Kepa (*Kepa*)
- Kepa esan al duzu? (*Did you say Kepa?*)
- Bai (*Yes*)
- Kaixo, Kepa, ni Anhitz naiz (*Hello, Kepa, I am Anhitz*). Zientzia eta Teknologiari buruzko galderak erantzun ditzaket (*I can answer questions about Science and Technology*), edo hitzak bilatu dokumentu biltegi eleaniztun batean (*or look for documents containing keywords in a multilingual repository*). Kepa, zer egin nahi duzu egitea, galdera bati erantzutea edo edo hitzak bilatzea? (*Kepa, what do you want me to do, answer a question or look for keywords?*)
- Hitzak bilatzea (*Look for keywords*)
- Esan bilatu nahi dituzun hitzak, Kepa (*Tell me which keywords you want me to look for, Kepa*)
- Ozono geruza (*Ozone layer*)
- "ozono geruza" esan duzu? (*Did you say "ozone layer"?*)
- Bai (*Yes*)
- "ozono geruza" bilatzen… (*Looking for "ozone layer"…*) Espainierazko dokumentuak itzultzen… (*Translating documents from Spanish…*) Hauek dira aurkitu ditudan emaitzak: (*These are the results I have found:*)

And next AnHitz will show on the screen to its right a list with clickable titles and snippets of the results found. The results that were not originally in Basque are translated by means of machine translation (see Figure 1). And the user can tell the system to read the titles or snippets he/she wants aloud, using TTS.



10 emaitza.

1 (euskaraz):
▶ Izenburua: Iparraldeko zuloa
▶ Laburpena: la Antartika osoa hartzen duen ozono-geruzako zuloaren anaia txikia handitzen ari da. Anaia txiki hau Artikoan bizi da, eta orain arte nahiko zintzo eta isil ibili bada ere, azken urteetan gorabehera handiak gertatu dira han goian. Ikertzaileen kontrolpe

2 (euskaraz):
▶ Izenburua: Ozono geruza, planetako klimaren termometroa
▶ Laburpena: Ozonoa hiru oxigeno-atomoak osatzen duten molekula da. Berez ikusi ezin den gasa da, kolorerik gabekoa eta usain atseginekoa. Molekula oxidatzailea eta korrosiboa da. Lurreko atmosferako estratosferan edo troposferan ager daiteke, eta, horren arabera, ing

3 (gaztelaniaz):
▶ Izenburua: 'Zuloak' ozono-geruzan hondatzearen errekor berri bat atzematen du (Jatorrian: El 'agujero' en la capa de ozono alcanza un nuevo récord de deterioro)
▶ Laburpena: Atmosferaren kapa altuetan, zientzialariak suntsiduraren %50 bateraino detektatu dira.Klima-aldaketak fenomeno hau biziagotu du.
(Jatorrian: En las capas altas de la atmósfera, los científicos han detectado hasta un 50% de destrucción. El cambio climático ha intensificado este fenómeno.)

4 (gaztelaniaz):
▶ Izenburua: Europar osasuna, tenperatura Artiko baxu mehatxatuta (Jatorrian: La salud europea, amenazada por las bajas temperaturas árticas)
▶ Laburpena: Artikoan tenperatura baxu miatuak ozonoaren kapa babeslea gutxitzen ari dira eta EB-AREN parteren batzuk osasun gizatiarra hunkitu ahal izango lukete, Europako Batzordea aholkatu du.
(Jatorrian: Las bajas temperaturas registradas en el Ártico están disminuyendo la capa protectora de ozono y podrían afectar a la salud humana de algunas partes de la UE, ha advertido la Comisión Europea.)

5 (gaztelaniaz):
▶ Izenburua: 50 urtetan zulo handiagoa (Jatorrian: El mayor agujero en 50 años)
▶ Laburpena: Científicos De La Universidad De Cambridge-k Erresuma Batuan amaitu zuten ozono-geruzaren gujero-ak atzeman zuela haren maximoa Azkenetan 50 urtetan.
(Jatorrian: Científicos de la Universidad de Cambridge en Reino Unido concluyeron que el gujero de la capa de ozono alcanzó su máximo en los últimos 50 años.)

Figure 1: Screen capture of the AnHitz demo

An example of a conversation involving a question would be like this (with the same introductory part):

- Kepa, zer egin nahi duzu egitea, galdera bati erantzutea edo edo hitzak bilatzea? (*Kepa, what do you want me to do, answer a question or look for keywords?*)
- Galdera bati erantzutea (*Answer a question*)
- Esan egin nahi duzun galdera, Kepa (*Put the*

question you want answered, Kepa)

- Nork asmatu zuen telefonoa? (*Who invented the telephone?*)
- "Nork asmatu zuen telefonoa?" esan duzu? (*Did you say "who invented the telephone?"*)
- Bai (*Yes*)
- Galderaren erantzuna bilatzen… (*Looking for the answer to your question…*)
- Erantzuna Graham Bell izan daitekeela uste dut. (*I think the answer is Graham Bell.*) Nahi al dituzu ikusi aukera guztiak euren probabilitatearekin? (*Do you want to see all the possible answers I have found and their probabilities?*)
- Bai (*Yes*)
- Hor ondoan dituzu aurkitu ditudan aukera guztiak euren probabilitatearekin: (*These are the possible answers I have found and their probabilities:*)

And then AnHitz shows the list of possible answers on the text screen, together with each one's probability and the paragraph from which each has been inferred.

3. Description of the modules

The AnHitz prototype, as we have already pointed out, comprises various modules, which will be described in more detail in the following subsections. The architecture of the overall system is shown in Figure 2.

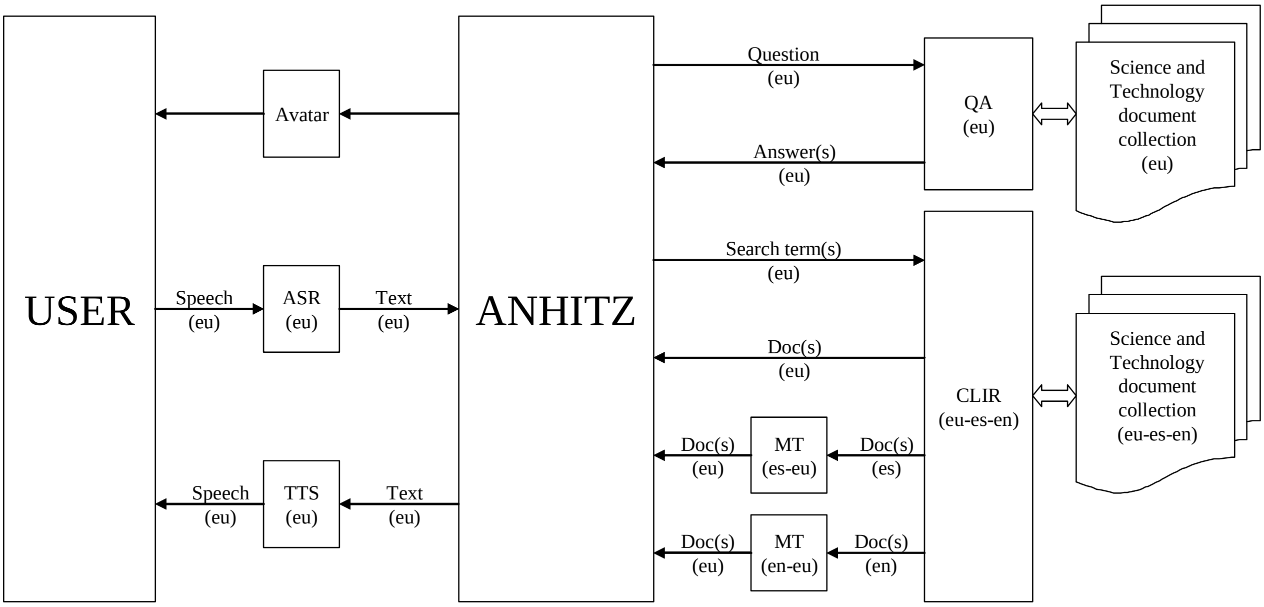


Figure 2: Architecture of the modules of the AnHitz demo

3.1 Avatar

The avatar module developed by Vicomtech, AnHitzDlg, includes all the necessary functionalities for showing and animating the 3D character that acts as the front-end of the AnHitz demonstrator. Its lip animation is synchronised with the audio synthesised by the multilingual TTS module in several languages and it can also show facial emotions when required. In addition, the module generates blinking and head movement animations through a set of behaviour rules in order to increase the illusion that the 3D character is alive.

It has been developed in C++, using OpenSceneGraph (http://www.openscenegraph.org) as its graphic library. Although it has only been tested in the Windows XP and Windows Vista Operative Systems, the multiplatform features of both the avatar module and the OpenSceneGraph library allow its easy migration to Linux systems too.

3.2 ASR

The only Automatic Speech Recognition (ASR) system for Basque available at the time of AnHitz's development was Nuance's Vocon3200. This ASR system is grammar-based. Elhuyar and Robotiker had to design various grammars in Backus-Naur Form or BNF (http://en.wikipedia.org/wiki/Backus-Naur_Form) in order to adapt the system to the AnHitz demo: a grammar for names, a grammar for yes/no answers, a grammar for search terms, a grammar for common science and technology questions, etc. The search terms and questions were extracted out of the most frequent searches of the logs of Zientzia.net (http://www.zientzia.net), a Basque popular science website owned by Elhuyar.

Vocon3200 is written in C and can be used under Windows.

3.3 TTS

The Text-To-Speech system used in AnHitz is AhoTTS (Hernáez et al., 2001), a multilingual system developed by the Aholab Signal Processing Group of the University of the Basque Country for commercial and research purposes. The system has a modular architecture because it has been specially designed to develop all the modules that integrate a TTS system independently. The system uses three main processing modules: the text processor, the linguistic processor and the synthesis engine. In addition, three databases are used: one dictionary which includes morphological and phonetic information about the words, a database for prosody prediction, and the synthesis database containing the recordings that will be manipulated to generate the synthetic speech (Sainz et al., 2009). The system currently works in Basque, Spanish and English, using Festival (Taylor et al., 1998) for the English text processing module.

AhoTTS is written in C/C++ and is fully functional both in the Unix and Windows operating systems.

3.4 CLIR

The Cross-Lingual Information Retrieval module used is Elezkari, which has been developed by Elhuyar (Saralegi, & López de Lacalle, 2009). The search terms are entered in Basque and the information retrieval is done in various popular science corpora in Basque, English and Spanish. In order to achieve this, the search terms have to be properly translated into the other languages (dealing with ambiguous translations, Out-Of-Vocabulary words, etc.).

The system works under Linux. It is programmed in C and makes use of the Indri search engine (http://www.lemurproject.org/indri/).

3.5 MT

Matxin, developed by the IXA group of the University of the Basque Country (Alegria et al., 2007; Mayor et al., 2009), is the system used for Machine Translation in AnHitz. It translates text from Spanish into Basque using a transfer rule-based approach. The first version of an English-Basque rule based MT system is being developed at the moment.

Its modules have been programmed using C and C++ programming languages, it works under Linux and its free code is publicly available in Sourceforge (http://matxin.sourceforge.net).

3.6 QA

The Question Answering system used in AnHitz is Ihardetsi, developed by IXA (Ansa et al., 2008; Alegria et al., 2009). It works over a Science and Technology corpus compiled by Elhuyar and IXA, the ZTCorpus. As is common in question answering systems, Ihardetsi is based on three main modules: the question analysis module, the passage retrieval module and the answer extraction module. These modules have been implemented as autonomous web services by reusing some linguistic tools previously developed in the IXA group, and the QA system becomes a client that calls these services in a pipeline when it needs them by using the SOAP (Simple Object Access Protocol) communication protocol.

The Ihardetsi QA system runs under Linux.

3.7 AnHitz main program

The AnHitz main program controls the conversation flow and responds to the user’s queries by making use of the other modules.

The control of the conversation flow includes introducing itself, prompting the user for his/her name, the action to perform, the terms to search or the question to answer, showing different emotions depending on the certainty of the answer, etc. To improve the performance of the ASR system when it did not understand correctly, 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.

The main program has been developed in Python and, since it uses no operating system-specific libraries, it can run indistinctly under Linux or Windows.

4. Communication among the modules

Bearing in mind what has been said in the previous section, we can observe that the AnHitz modules, since they have been built by different organizations, run on different operating systems and use different programming languages and/or libraries. This made it extremely difficult, if not impossible, to integrate the demonstrator into a single machine, let alone into an executable file.

As a result, the AnHitz prototype demonstrator has been constructed as a distributed system running in different machines over the Internet and communicating via net protocols and web services, even among the modules running in the same machine. The main program of the demo, the avatar module and the ASR system run on a Windows laptop. The TTS system runs on a Linux machine at Aholab, the CLIR system on a Linux machine at Elhuyar and the MT and QA systems on different Linux machines at IXA. The modules have been implemented as servers using network communication protocols, and the main program is the client that makes requests to them.

Different protocols have been used for communication. There has been no particular reason to choose one or the other: in some cases, the module was already on the web implemented as a web service with a certain protocol; in others, the programmer just used the protocol he/she was more familiar with.

The avatar module receives its speaking orders via simple sockets. It then asks the TTS module for the audio file and phoneme information file using an HTTP request. The ASR system activates itself and the microphone when it receives a request via sockets. The CLIR module is called using the SOAP protocol. The MT system for translating the texts from Spanish into Basque is also called using SOAP. The QA module is queried via sockets. An illustration showing the communication among the modules is shown in Figure 3.

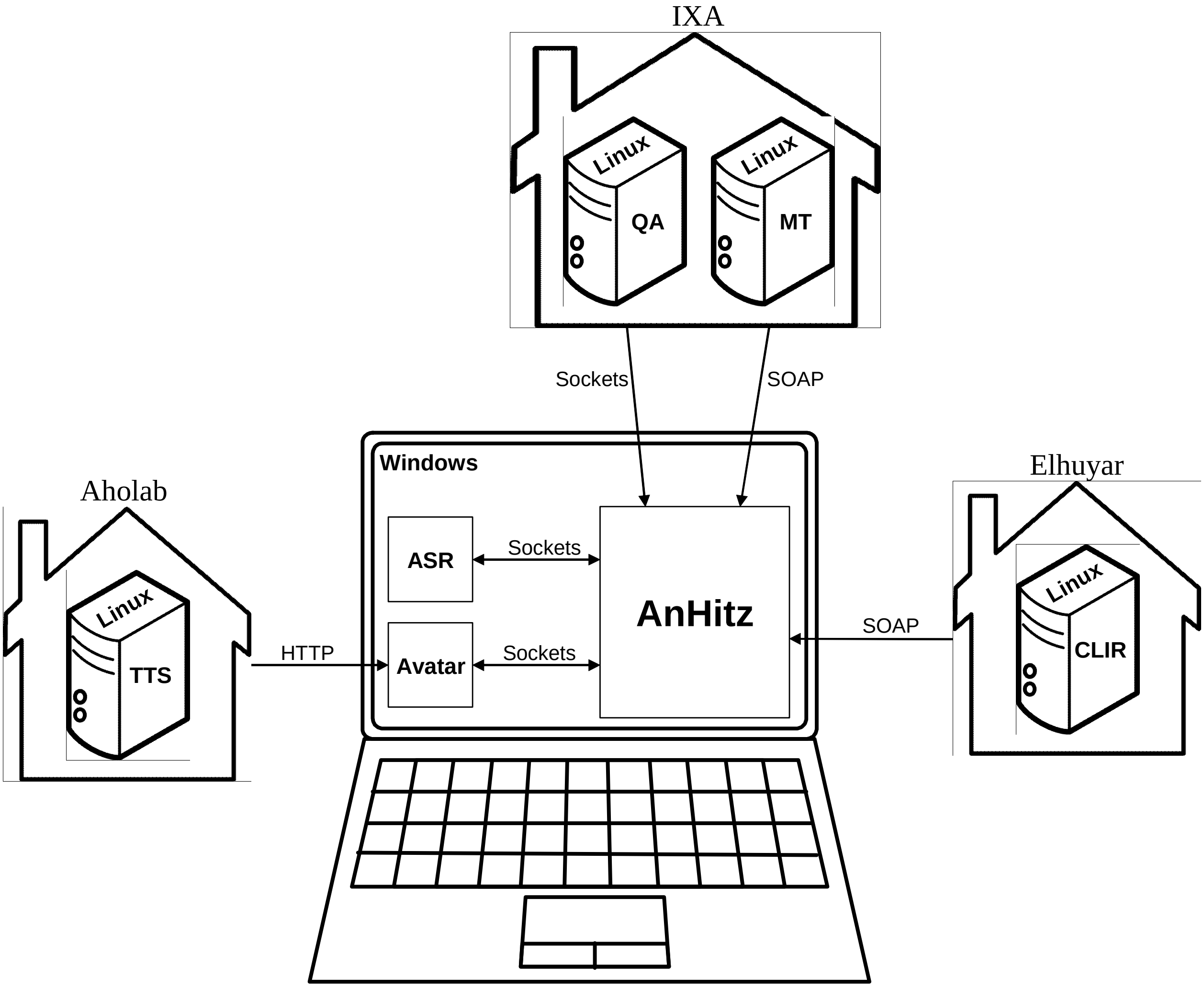


Figure 3: Communication among AnHitz's modules

5. Evaluation

The demo prototype developed in AnHitz was evaluated in order to measure its performance and weigh up the impression of potential users about it. 50 users formulated 3 questions and 3 cross-lingual search queries each, testing the system 300 times in total. During the trials, several objective observations, such as the number of failures and successes of the ASR or QA systems, were noted down. At the end of each interaction, the testers filled in a questionnaire regarding more subjective matters (quality of the TTS, CLIR or MT systems, general impression, etc.).

The aim of the evaluation was twofold: on the one hand, the performance of the individual modules was evaluated in real world uses; on the other, we were able to weigh up the AnHitz demo's performance and the users' impressions about it. The results of the evaluation are detailed in the following paragraphs.

The ASR system, together with the confirmation/repetition system implemented in the AnHitz demonstrator, understood correctly 63.19% of the times. Another 12.59% of the times it understood correctly but was not sure and asked for confirmation. 13.43% of the times the system did not understand correctly and asked for confirmation, so the user could repeat the sentence. Only in 10.79% of the cases did the system understand wrongly without giving the option to correct. When users were asked whether AnHitz had understood what they said overall, 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".

Regarding the intelligibility of AnHitz's speech, 85.42% thought it was "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".

The question answering system returned the correct answer 30.61% of the times, and in another 15.30% the correct answer was among the given first five possible answers. 54.08% of the times the system did not return 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.

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

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

Regarding usefulness, 62.50% of the users thought the system was “very useful” or “useful” and 37.50% thought it was “quite useful”. No one said it was “quite useless” or “completely useless”. When asked about the suitability of extending the AnHitz approach to other application domains, 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”.

6. 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 at 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.

7. 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 AnHitz prototype demonstrator implemented to integrate the tools and resources developed has shown that collaboration between agents working in the different language, speech and visual research fields is crucial for exploiting the potential of the technologies and build applications useful for the end user. The evaluation of the AnHitz demonstrator has shown that despite being based on systems still in the research stage, its performance is acceptable.

In order to build the completely functional AnHitz demo integrating different language, speech and visual technologies, a modular remote architecture with network communications has been used. The benefits of building the demo using this networked approach have been enormous, and we doubt we could have built it otherwise. However, this approach also presents some drawbacks. The main disadvantage is the time delay that originates when AnHitz has to speak, due to the transmission over the Internet of the audio files generated by the TTS module. Other delays exist too, because some processes, mainly MT and QA, need their time; but these are inevitable and not due to the modular architecture. However, we reduced the effect of these delays to some extent by locally caching the already processed audios, queries and translations, so that the most frequent and repetitive sentences and queries can be executed instantly.

Another drawback is the lack of control over the remote modules and servers. If something goes wrong in one of them, the whole AnHitz demo is affected and it is not trivial to put things back up.

Nevertheless, we consider that the results have been very satisfactory overall, since both the responses obtained from the users in the evaluation and the media coverage have been very positive.

One future improvement of the AnHitz demo could be the use of virtualization to run two operating systems at a time and thus allow the installation of all the modules in the same machine. However, we are not sure this approach would work out due to the complexity, libraries, versions, etc. of the modules. But even with all the modules in one machine, the network communications approach would still be used.

Another possible improvement to explore in the future is the implementation of the AnHitz demonstrator as a web application, which would allow the general public to experiment with the potential of the combination of language, speech and visual technologies.

8. Acknowledgements

This work has been partially founded by the Industry Department of the Basque Government under grant IE06-185 (AnHitz project).

9. References

Aduriz, I., Alegria, I., Artola, X., Ezeiza, N., Sarasola, K. (1997). A spelling corrector for Basque based on morphology. *Literary & Linguistic Computing*, 12(1), pp. 31--38.

Alegria, I., Gurrutxaga, A., Saralegi, X., Ugartetxea, S. (2006a). Elexbi, a basic tool for bilingual term extraction from Spanish-Basque parallel corpora. In *Proceedings of Euralex 2006*. Torino: Euralex, pp. 159--165.

Alegria, I., Arregi, O., Ezeiza, N., Fernandez, I. (2006b). Lessons from the development of a named entity recognizer for Basque. *Procesamiento del Lenguaje Natural*, 36, pp. 25--37.

Alegria, I, Díaz de Ilarraza, A., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K. (2007). Transfer-based MT from Spanish into Basque: reusability, standardization and open source. *Lecture Notes on Computer Science*, 4394, pp. 374--384.

Alegria, I., Casillas, A., Diaz de Ilarraza, A., Igartua, J., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K. (2008). Spanish-to-Basque MultiEngine Machine

Translation for a Restricted Domain. In *Proc. of the 8th Conference of the Association for Machine Translation in the Americas (AMTA-2008)*. Hawai, USA: AMTA, pp. 57--69.

Alegria, I., Ansa, O., Arregi, X., Otegi, A., Soraluze, A. (2009). Ihardetsi: A Question Answering system for Basque built on reused linguistic processors. In *Proc. SALTMIL 2009 workshop: Information Retrieval and Information Extraction for Less Resourced Languages*. Donostia: SALTMIL, pp. 37--43.

Ansa, O., Arregi, X., Otegi, A., Soraluze, A. (2008). Ihardetsi question answering system at QA@CLEF 2008. In *Working Notes of the Cross-Lingual Evaluation Forum*. Aarhus: CLEF, pp. 369--376.

Areta, N., Gurrutxaga, A., Leturia, I., Alegria, I., Artola, X., Diaz de Ilarraza, A., Ezeiza, N., Sologaistoa, A. (2007). ZT Corpus: Annotation and tools for Basque corpora. In *Proceedings of Corpus Linguistics 2007*. Birmingham: University of Birmingham.

Castelruiz, A., Sánchez, J., Zalbide, X., Navas, E., Gaminde, I. (2004). Description and design of a web accessible multimedia archive. In *Proc. of 12th IEEE Mediterranean Electrotechnical Conference (MELECON)*. Dubrovnik: IEEE, pp. 681--684.

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

D'hallewey, E., Odijk, J., Teunissen, L., Cucchiarini, C. (2006). 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: ELRA, pp. 761--766.

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

Gurrutxaga, A., Saralegi, X., Ugartetxea, S., Lizaso, P., Alegria, I., Urizar, R. (2004). A XML-based term extraction tool for Basque. In *Proc. of fourth international conference on Language Resources and Evaluation (LREC)*. Lisbon: ELRA, pp. 1733--1736.

Hernáez, I., Navas, E., Murugarren, J.L., Etxebarria, B. (2001). Description of the AhoTTS conversion system for the Basque language. In *Proceedings of 4th ISCA Tutorial and Research Workshop on Speech Synthesis*. Edinburgh: ISCA, paper 202.

Leturia, I., Gurrutxaga, A., Alegria, I., Ezeiza, A. (2007a). CorpEus, a 'web as corpus' tool designed for the agglutinative nature of Basque. In *Proceedings of Web as Corpus 3 workshop*. Louvain-la-Neuve: ACL-SIGWAC, pp. 69--81.

Leturia, I., Gurrutxaga, A., Areta, N., Alegria, I., Ezeiza, A. (2007b). EusBila, a search service designed for the agglutinative nature of Basque. In *Proceedings of iNEWS’07 workshop*. Amsterdam: ACM-SIGIR, pp. 47--54.

Leturia, I., San Vicente, I., Saralegi. X. (2009). Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet. In *Proceedings of 5th International Web as Corpus Workshop (WAC5)*. Donostia: ACL-SIGWAC, pp. 53--61.

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

Mayor, A., Alegria, I., Diaz de Ilarraza, A., Labaka, G., Lersundi, M., Sarasola, K. (2009). Evaluación de un sistema de traducción automática basado en reglas o por qué BLEU sólo sirve para lo que sirve. *Procesamiento del Lenguaje Natural*, 43, pp. 197--208.

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

Sainz, I., Erro, D., Navas, E., Hernáez, I., Saratxaga, I., Luengo, I., Odriozola, I. (2009). The AHOLAB Blizzard Challenge 2009 Entry. In *Proc. Blizzard Challenge 2009 workshop*. Edinburgh: Blizzard.

Sanchez, J., Luengo, I., Navas, E., Hernáez, I. (2006). Adaptation of the AhoTTS text to speech system to PDA platforms. In *Proceedings of the SPECOM 2006*. San Petersburg: SPECOM, pp 292--296.

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

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

Saralegi, X., López de Lacalle, M. (2009). 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: TIR.

Saratxaga, I., Navas, E., Hernáez, I., Luengo, I. (2006). 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: ELRA, pp. 2126--2129.

Stroppa, N., Groves, D., Way, A., Sarasola, K. (2006). Example-Based Machine Translation of the Basque Language. In *Proc. of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-2007)*. Boston, USA: AMTA, pp. 232--241.

Taylor, P., Black, A., Caley, R. (1998). The architecture of the Festival Speech Synthesis System. In *Proc. 3rd ESCA Workshop on Speech Synthesis*. Jenolan Caves, ESCA, pp. 147--151.