Chapter 10

The BerbaTek project for Basque: Promoting a less-resourced language via language technology for translation, content management and learning

Igor Leturia
Elhuyar Foundation

Kepa Sarasola

Xabier Arregi

Arantza Diaz de Ilarraza
IXA Group, University of the Basque Country

Eva Navas

Iñaki Sainz
Aholab Group, University of the Basque Country

Arantza del Pozo
Vicomtech-IK4

David Baranda

Urtza Iturraspe
Tecnalia

Igor Leturia, Kepa Sarasola, Xabier Arregi, Arantza Diaz de Ilarraza, Eva Navas, Iñaki Sainz, Arantza del Pozo, David Baranda & Urtza Iturraspe. The BerbaTek project for Basque: Promoting a less-resourced language via language technology for translation, content management and learning. In Georg Rehm, Felix Sasaki, Daniel Stein & Andreas Witt (eds.), *Language technologies for a multilingual Europe: TC3 III*, 181–204. Berlin: Language Science Press.
DOI:10.5281/zenodo.1291942

Basque is both a minority language (only a small proportion of the population of the Basque Country speaks it) and also a less-resourced language. Fortunately, the Basque regional government is committed to its recovery, and has adopted policies for funding, among other things, language technologies, a field which a language aiming to survive cannot dispense with. BerbaTek was a 3-year (2009–2011) strategic research project on language, speech and multimedia technologies for Basque carried out by a consortium of five members, all prominent local organizations dedicated to research in the above-mentioned areas, and partially funded by the Departments for Industry and Culture of the Basque Government. Collaboration in BerbaTek allowed to carry out a great amount of both basic and applied research. In addition, various prototypes were developed to show the potential of integrating the developed technologies to the language industry sector.

1 Introduction

The Basque language is one of the oldest alive in Europe, although it has suffered continuous regression over the last few centuries. However, many citizens and local or regional governments have been promoting its recovery since the 1970s. Now, Basque holds partial co-official language status in the Basque regions of Spain but it has no official standing in the Northern Basque Country in France, neither in the European institutions. Today, there are about 700,000 Basque speakers, around 25% of the total population of the Basque Country, but they are not evenly distributed, and the use of Basque in industry and especially in Information and Communication Technology is still not widespread. In September 2012, META-NET placed Basque as one of the 21 European languages that are in danger of digital extinction[1]. A language that seeks to survive in the modern information society has to be present also in such fields and this calls for language technology products. Basque, like other minority languages, has to make a great effort to address this challenge (Williams et al. 2001).

In this context, BerbaTek[2] was a strategic research project in language, speech and multimedia technologies developed over the years 2009–2011. Its consortium was made up of the Elhuyar Foundation, the IXA and Aholab research groups of the UPV/EHU (University of the Basque Country), the Vicomtech-IK4 Visual Interaction and Communication Technologies Centre and the Tecnalia Research & Innovation foundation. The project was partly funded by the Departments for Industry and Culture of the Government of the Basque Autonomous Community (region of Spain).

[1] http://www.meta-net.eu/whitepapers/press-release

[2] http://www.berbatek.com/en

The members of the consortium had been collaborating since 2002 in two similar previous projects, Hizking (Diaz de Ilarraza et al. 2003) and AnHitz (Arrieta et al. 2008), in which basic foundations, tools and applications were created for Basque.

We believe that research and development for less-resourced languages should be addressing following four points: (1) high standardization, (2) open-source, (3) reuse of language foundations, tools and applications, and (4) incremental design and development. Any HLT project relating to a less-privileged language should follow those guidelines, but from our previous experience we knew that in most cases they did not. We believe that if Basque is now in a fairly good position in HLT, it is because these guidelines have been applied, even though in some cases it was easier to create "toy" resources or easily obtainable tools (Alegria et al. 2011).

2 The consortium

Vicomtech-IK4[3] is an applied research center whose main lines of research are graphic computation, interaction and multimedia. Three of its groups participated in BerbaTek: (i) the Speech and HLT group, (ii) the 3D Animation group, and (iii) the Audiovisual Content Analysis group.

Tecnalia[4] is a private, applied research center specialized in information and telecommunication technologies.

The *Elhuyar Foundation*[5] is a not-for-profit organization, set up with the aim of bringing science and the Basque language together. Elhuyar is firmly established in the market for dictionaries, educational software, multimedia products, plugins and Machine Translation. In 2001, it set up a LT unit.[6]

IXA[7] is a group of the University of the Basque Country (UPV/EHU), consisting of 43 researchers, which works on NLP specialized in the processing of written texts at different levels. The main projects IXA is currently working on are the PATH, OpeNER and NewsReader European STREP projects.

The *Aholab Signal Processing Laboratory*[8] is a research group of the University of the Basque Country (UPV/EHU), with broad experience in voice technologies

[3] http://www.vicomtech.org
[4] http://www.tecnalia.com
[5] http://www.elhuyar.org
[6] http://www.elhuyar.org/hizkuntza-zerbitzuak/EN/R-D
[7] http://ixa.si.ehu.es
[8] http://aholab.ehu.e/

and digital signal processing. Aholab developed the first commercial TTS system for Basque, AhoTTS,[9] as well as most of the resources and voice processing tools publicly available for Basque.

3 Objectives

The main aim of BerbaTek was the research and development of language, speech or multimedia technologies, so that they could provide the technological basis to support the economic sector of the language industries in the Basque Country.

The key challenge was to prove that Basque processing technologies could be useful to improve the performance, social impact and competitiveness of some industrial products. This challenge required the partners to take a new significant step forward in the strengthening of the language industries by incorporating the results and devices into real market scenarios. This point was particularly relevant given that basic resources and tools must be robust enough to support industrial use.

As most companies do not want to take on this task as it is expensive and commercially not profitable, we considered this initiative a social investment. Tools for languages like Basque are usually developed at universities or research centres and adapting those linguistic tools to the real industrial scenario is crucial.

BerbaTek was geared towards applications. Without neglecting basic research, it was endeavouring to present experimental applications which could subsequently be developed further and turned into products by companies. The importance of generating knowledge in the area of language technologies for voice and multimedia lies in their potential for applications mainly in the language industry sector:

Translation: interpretation, dubbing, localization, human translation etc.

Content industry: Internet, audiovisual sector, the media, off- and on-line publishing, multimedia companies, etc.

Training: language learning, technical and professional education, ongoing training, etc.

[9] http://aholab.ehu.es/tts/tts_en.html

Table 1: Resources developed or improved during BerbaTek

Corpus resources	Basque Dependency Treebank (BDT), 300,000-word corpus. Basque Propbank and tools for its development (Aldezabal et al. 2010). AhoSyn, a large speech database (6 hours per speaker) (Sainz et al. 2012). AhoSpeakers, database designed for voice conversion (Sainz et al. 2012). AhoEmo3, created for emotional speech synthesis (Sainz et al. 2012). A large general corpus (+100M words) collected automatically from the web (Leturia 2012).
Ontology resources	Basque WordNet (Pociello et al. 2011), the Basque version of WordNet. WNTerm (Pociello et al. 2008), WordNet + 25,000 science and technology terms. Termide, automatic ontology building out of corpora. QAWS, question answering over Linked Data.
Dictionary resources	Various bilingual dictionaries created automatically using a pivot language (Saralegi et al. 2012).

4 Resources, tools and applications developed

The partners had been working in NLP 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, but most of them were further improved within the project, and many others were created in BerbaTek.

BerbaTek carried out basic research and built many resources and tools that are necessary for the development of applications. Tables 1, 2 and 3 show the resources, tools and applications developed or improved during the project. The key resources and tools in the development of applications on the aforementioned areas of translation, content management and learning were the following:

Table 2: Tools developed or improved during BerbaTek

Analysis tools	Dependency Parsing (Bengoetxea & Gojenola 2010; Agirre et al. 2011). UKB (Agirre & Soroa 2009) graph-based Word Sense Disambiguation ArikIturri (Aldabe & Maritxalar 2010), automatic creation of exercises out of corpora.
Web as corpus tools	Co3 (Leturia et al. 2009), building multilingual comparable corpora. PaCo2 (San Vicente & Manterola 2012), collecting parallel corpora.

- Tools for building corpora from the web (monolingual and multilingual, general and domain-based, comparable and parallel), and various corpora collected by using these.
- Syntactic dependency analysers, semantic analyzers and systems for identifying sentence and phrase boundaries.
- Terminology extraction from corpora and automatic building of dictionaries.
- General and domain-specific ontologies and semantic search engines.
- Cross-lingual search and question answering.
- Machine translation systems (rule-based, statistical and hybrid).
- Techniques for voice segment detection and text/image alignment in video.
- Engine for continuous speech recognition; text-to-speech conversion systems.
- Speaking avatars.
- Writing aids and automatic exercise creation.

Table 3: Applications developed or improved during BerbaTek

Automatic dictionary building	AzerHitz (Saralegi et al. 2008), extraction of equivalent terms from comparable corpora. PiboLex (Saralegi et al. 2012), building new dictionaries using a pivot language. Phraseology and idiomatic expressions extractor (Gurrutxaga & Alegria 2011).
Information retrieval	Ihardetsi (Ansa et al. 2008; Agirre, Ansa, et al. 2009), a Question-Answering system. Elezkari (Saralegi & Lacalle 2009), CLIR (Basque, Spanish and English).
Machine translation	Opentrad-Matxin (Alegria et al. 2007; Alegria et al. 2008; Mayor et al. 2011), open-source rule-based machine translation system for Spanish-Basque. EUSMT, statistical Machine Translation from Spanish to Basque. (Labaka 2010).
Speech synthesis	AhoT2P, a letter to allophone transcriber for standard Basque. AhoTTS_Mod1, a linguistic processor for speech synthesis. AhoTTS, modular Text-To-Speech conversion for Basque, Spanish and English. TTS system based on HTS (Erro et al. 2010), with own vocoder (Erro et al. 2011a). Hybrid AhoTTS combining advantages from statistical and unit selection speech synthesis (Erro et al. 2010).
Speech recognition	AhoSR (Odriozola et al. 2012), speech recognition engine (standard Basque).

5 Prototypes

Throughout the project, we created some demos to show the usefulness of the linguistic tools and the potential of the integration of language-, speech- and multimedia-technologies when it comes to creating applications for the areas of language industries, i.e., for translation, contents and teaching. These are the demos we built:

- Automatic dubbing of documentaries into Basque using subtitles in Spanish (with possible automatic creation of the Spanish subtitles from the Spanish audio, by means of ASR).

- Two multimedia and multilingual semantic web search engines on science and technology content, one of them including subsequent navigation through related content or similar images.
- Personal tutor in language learning through a speech-driven avatar, with automatically created grammar and comprehension exercises, writing aids (dictionaries, writing numbers, spelling…) and automatic evaluation of pronunciation.

5.1 Automatic dubbing of documentaries

The automatic dubbing of films is still a difficult challenge (different voices, speed and tones, colloquial language etc.), but for some types of documentaries (single speaker, voice-over, coordination of the lips not necessary or unimportant) we produced a demo that performs satisfactorily. The general structure of the application developed is shown in Figure 1. Given a documentary in Spanish and its transcription (the transcription can be obtained automatically by means of any of the available dictation programs for Spanish), Vicomtech-IK4's alignment technology creates a subtitles file (the transcription with time marks for the beginning and end of each sentence). Then, IXA group's Matxin MT system automatically translates the subtitles into Basque, and Aholab's text-to-speech technology produces the synchronized voice output. We successfully applied this demo to the single-speaker sections of the television programme Teknopolis produced by Elhuyar.

The automatic alignment of speech and text is based on speech recognition technology. In this case, it is forced to recognize the text of the transcription and provided timing information at phoneme and word levels. That way, the start and end time-codes of each word are obtained automatically and used to synchronize subtitles with the video image.

The translation of subtitles is done using Opentrad-Matxin (Mayor et al. 2011) adapting it to the domain of science and technology. Matxin is a rule-based deep syntactic transfer system for translation into Basque. It translates text from Spanish into Basque, but its architecture allows for an easy implementation of new systems for translating other languages into Basque (Mayor & Tyers 2009). Opentrad-Matxin is open source. The free code of the Spanish-Basque system with a reduced version of the bilingual lexicon can be downloaded from http://matxin.sourceforge.net. The system can be used at http://www.opentrad.org.

The average HTER evaluation result of Matxin was 0.42, meaning that 42 editing corrections are required for every 100 tokens. One of the key features of our

Figure 1: Scheme of the automatic dubbing of documentaries demo

work is the reuse of existing linguistic resources: we created the system's lexicon by automatically processing high-coverage dictionaries. Given that we reused previously created resources, the XML-based format guaranteed their interoperability. Now we are working on the construction of SMT systems and a hybrid system including three subsystems based on different approaches (España-Bonet et al. 2011).

Regarding speech production, we use the HMM-based synthesis engine. First, its Basque linguistic module extracts linguistic features from the input text. Then

the acoustic engine uses them to select previously trained statistical models and generate a sequence of suitable acoustic parameters. Finally, the synthetic speech signal is constructed from the aforementioned parameters by AhoCoder. Alignment time stamps are used to synchronize the synthetic audio and the original video, by modifying either the speech rate or the duration of silences.

5.2 Multimedia and multilingual semantic web search engines

5.2.1 Semantic retrieval system based on document expansion

One of the main problems IR systems have to deal with is the vocabulary mismatch problem between the query and documents: some documents might be relevant for the query even if the specific terms used differ substantially. On the contrary, some documents might not be relevant for the query even if they have some terms in common. The former is because languages are rich in the sense that more than one word or phrase could be used for expressing one idea or thing. The latter is because of ambiguity, in other words, because one word could have more than one interpretation depending on the context. If a system only relies on terms occurring in both the query and the document when it comes to deciding whether a document is relevant, it might be difficult to find some of the interesting documents and also to reject non-relevant documents. It seems fair to think that there will be more chances of successful retrieval if the meaning of the text is also taken into account. Even though this problem has been widely discussed in the literature ever since the early days of IR, it remains unsolved and there is still a high degree of uncertainty about the possibility of overcoming the problem by making use of any NLP technique.

This BerbaTek demo explored whether NLP can benefit the effectiveness of the search engine (Otegi 2012): http://ixa2.si.ehu.es/BerbatekDemo/bilatu.

Although in principle synonymy, polysemy, hyponymy or anaphora should be taken into account in order to obtain high retrieval relevance, the lack of algorithmic models has prohibited any systematic study of the effect of these phenomena on retrieval. Instead, researchers have resorted to distributional semantic models to try to improve retrieval relevance, and overcome the brittleness of keyword matches. Most research has concentrated on Query Expansion (QE) methods, which typically analyze term co-occurrence statistics in the corpus and in the highest scored documents for the original query in order to select terms for expanding the query terms (Manning et al. 2009). Document expansion (DE) is a natural alternative to QE, but, surprisingly, it has not been explored until very recently. Several researchers have used distributional methods from similar doc-

uments in the collection to expand the documents with related terms that do not actually occur in the document. The work carried out in BerbaTek was complementary in that we also explored DE, but used WordNet instead of distributional methods (Agirre et al. 2010).

Our key insight was to expand the document with related words according to the background information in WordNet (Fellbaum 1998), which provided generic information about general vocabulary terms. WordNet groups nouns, verbs, adjectives and adverbs into sets of synonyms (synsets), each expressing a distinct concept. Synsets are interlinked with conceptual-semantic and lexical relations, including hyperonymy, meronymy, causality, etc.

In contrast to previous work, we selected those concepts that are most closely related to the document as a whole. For that, we used a technique based on random walks over the graph representation of WordNet concepts and relations. We represented WordNet as a graph as follows: nodes represent concepts (synsets) and dictionary words; relations among synsets are represented by undirected edges, and dictionary words were linked to the synsets associated to them by directed edges. We used version 3.0, with all relations provided, including the gloss relations. This was the setting that achieved the best results in a word similarity dataset (Agirre, Soroa, et al. 2009).

Given a document and the graph-based representation of WordNet, we obtained a ranked list of WordNet concepts as follows:

- We first pre-processed the document to obtain the lemmas and parts of speech of the open category words.
- We then assigned a uniform probability distribution to the terms found in the document. The remaining nodes were initialized to zero.
- We computed personalized PageRank (Haveliwala 2002) over the graph, using the previous distribution as the reset distribution, and producing a probability distribution over WordNet concepts. The higher the probability for a concept, the more related it was to the given document.

This method revealed important concepts, even if they were not explicitly mentioned in the document. Once we had the list of words for document expansion, we created one index for the words in the original documents and another index with the expansion terms. This way, we were able to use the original words only, or to include the expansion words during the retrieval as well.

The retrieval system was implemented using MG4J (Boldi & Vigna 2005), as it provided state-of-the-art results and allowed several indices over the same document collection to be combined. BM25 was the scoring function of choice. It was one of the most relevant and robust scoring functions available (Robertson & Zaragoza 2009).

5.2.2 Multilingual semantic search engine for science and technology based on a specialized ontology, with similar image search

As proof of what language technologies could bring to the field of content, we also created a semantic multimedia search engine for science and technology. This search engine is based on the WNTerm ontology (Pociello et al. 2008) specialized in science and technology and created by Elhuyar and IXA (a network in which scientific and technological terms were semantically related to each other, with subclasses, synonyms, etc.), and works on content from Elhuyar (text and images from the Elhuyar magazine, videos from the *Teknopolis* TV show and audio files from the radio programme *Norteko Ferrokarrilla*). Using Tecnalia's technology, the search for a term also shows results containing synonyms, subclasses or superclasses, via the ontology. The resulting search engine is available in two versions, simple and advanced; the advanced version allows to choose an intelligence level (a higher level exploits more the relationships between the concepts in the ontology) and a type of document (image, video, audio and text) and enables filtering by subject. Furthermore, when the result is an image, it shows similar images by means of technology developed at Vicomtech-IK4. A demo is available online at http://bilatzailesemantikoa.berbatek.com/.

One of the first tasks launched was the analysis of the aggregated collection of digital resources made available by the Zientzia.net web site on our semantic search engine site. It was deduced that the Science and Technology domain covered the following subjects: general subjects, computer science, earth science, environment, health, life science, physics, mathematics and chemical science, space and technology in general.

Constructing the BerbaTek ontology included the following steps:

- The manual creation of a skeleton ontology that specified what the root concepts of the ontology were, e.g., Life Science, Technology, Earth science and Computer science. This skeleton should be modified if new knowledge sources were added for the annotation process.

- An ontology with 25,000+ concepts of science and technology was built by hierarchically organizing the terms of the Basque Encyclopaedic Dictionary of Science and Technology produced by Elhuyar. Every term from this ontology, called WNTerm, was mapped to one of the root concepts or areas.

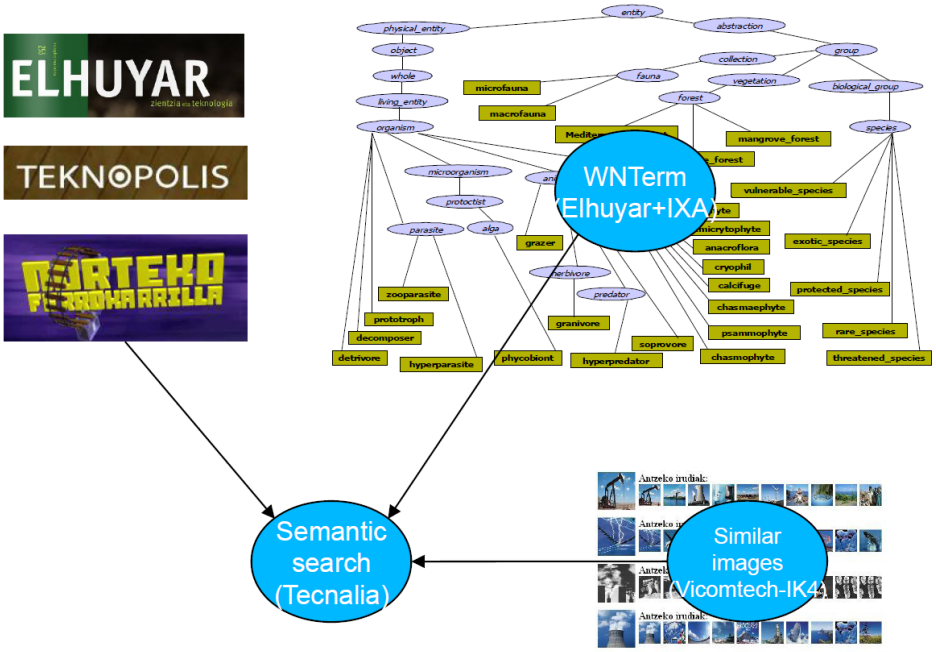

Figure 2: Scheme of the semantic search engine demo

For describing the resources, we decided to adopt an existing and standard metadata model to describe resources, in order to ensure interoperability with other organizations showing interest in sharing resources or contents with our semantic web search engine. Dublin Core was the metadata model selected.

The annotators were provided with a metadata editor where resources could be manually annotated, thus allowing the user to choose the topic or topics for the annotation by selecting from the different predefined concepts in our ontology.

The semantic search engine service was built onto the semantic layer, and we developed in the service layer the semantic search engine service.

If the search term entered was detected as a concept in the ontology, searches for related concepts were proposed (synonyms, hyponyms, hyperonyms etc.).

If the type of result obtained was an image, the search engine allowed the user to display a list of up to 10 similar pictures. This was because each recorded image had been pre-processed, generating a resemblance relationship.

5.3 Personal tutor for language learning

For the field of education, we created a demo consisting of a tutor for language learning. This tutor is a 3D avatar that showed emotions, developed by Vicomtech-IK4; it speaks Basque and understands what is said in Basque, using Aholab's technology. The tutor assists the student in various tasks: the student can orally solve grammar exercises (verb conjugation, word inflection etc.) and reading comprehension exercises (filling in gaps in a text, multiple choice tests) that are created automatically from texts using technology from IXA; his or her pronunciation can be evaluated with Aholab technology; The tutor can also provide help for writing texts, such as word inflection, writing of numbers or querying dictionaries, by means of technology from IXA and Elhuyar. The technologies included in this demo are shown in Figure 2.

The avatar module includes all the necessary functionalities to show and animate the 3D character that acts as the front-end of the demo. Its lip animation is synchronized with the audio synthesized by the TTS module, 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 was developed in C++, using OpenSceneGraph (http://www.openscenegraph.org) as its graphic library.

For the automatic creation of exercises, we use ArikIturri (Aldabe & Maritxalar 2010). This is a system for generating different types of questions. It uses as input a set of morphologically and syntactically analyzed sentences represented in XML, and it transforms them into the generated questions, also represented in XML.

There are some differences between the architecture of our and previous systems (Kraift et al. 2004; Schwartz et al. 2004). We separate an *Answer focus identificator* module and an *Ill-formed questions rejecter* module. Sumita et al. (2005) also included a module to reject questions, which was based on the web. Depending on the parameters' specifications, the *Sentence retriever* selects candidate sentences from the tagged source corpus. In a first step, it selects the sentences where the specified linguistic phenomena appear. Then the *Candidates selector* studies the percentages of the candidates in order to make a random selection of sentences depending on the number of questions specified in the input parameters. Once the sentences are selected, the *Answer focuses identifier* marks out some of the chunked phrases as answer focuses, depending on the morphosyn-

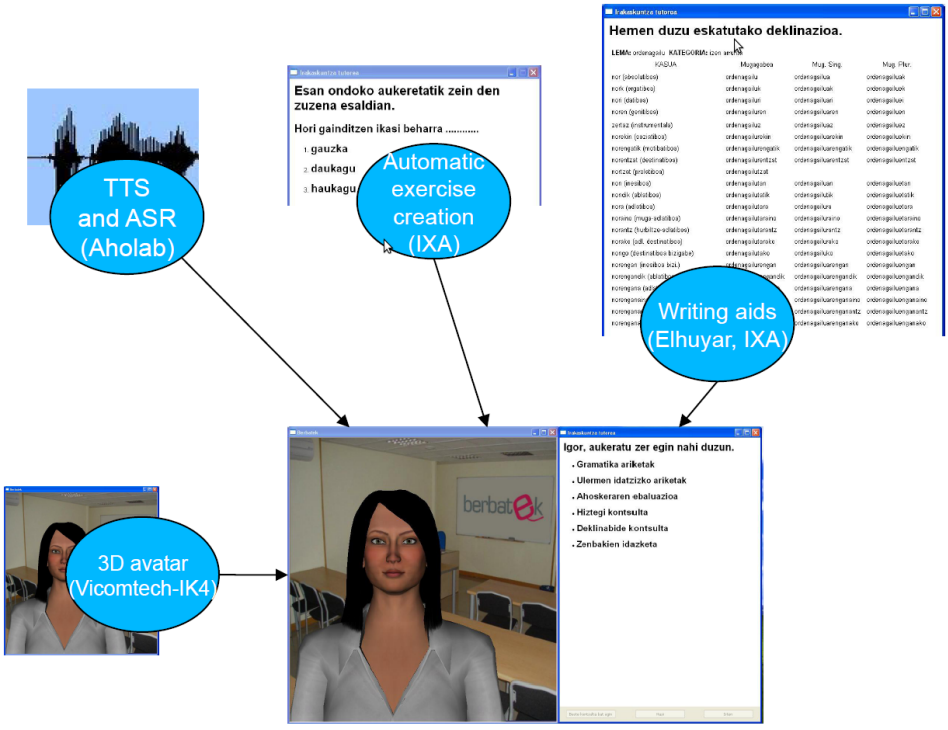

Figure 3: Personal tutor in language learning demo

tactic information of the phrases. Then the *Item generator* creates the questions depending on the specified exercise type. This is why this module contains the *Distractor generator* submodule. By this time, the system has already constructed the question instances. However, as the whole process is automatic, some questions might be ill-formed. That is why we include the *Ill-formed questions rejecter* in the architecture.

With regard to writing aids, the system offers three possible types of help: inflection of words, writing of numbers and querying dictionaries.

The module for helping with inflections asks for a word, then for the case (absolutive, dative, etc.) and then for definiteness and number (singular, plural, indefinite). For the latter two, it can be told to come up with the inflections for all of them. The system calls a web service developed by Elhuyar that generates inflections of words based on a two-level morphology transducer, and returns a table with the inflections of the chosen word.

The writing of numbers module asks for a 1 to 10 suite of one-figure numbers (for example "*hiru zazpi lau lau bost bat*", which means "*three seven four four five one*") and it tells the user how to write and pronounce the number produced (in the example "*hirurehun eta hirurogeita hamalau mila, laurehun eta berrogeita hamaika*", which means "*three hundred and seventy four thousand, four hundred and fifty one*), using a system developed by Elhuyar.

Finally, the dictionary-querying module asks for a Basque word and looks it up in various online dictionaries produced by Elhuyar (a Basque-Spanish dictionary, a Basque-French dictionary, a Basque-English dictionary and a synonyms dictionary), showing all the results found.

Speech technologies are extensively used in this demo. AhoSR, the ASR engine for Basque, is used to recognize the choices and answers of the students, and AhoTTS to generate the responses of the avatar. There is also a module to automatically evaluate the correctness of the segmental pronunciation.

The AhoTTS version based on HMM is used in this demo. As HTS does not perform any kind of linguistic analysis, the output of the first module of AhoTTS has to be translated into proper labels containing phonetic and linguistic information. See Erro et al. (2010) for a detailed list of the kinds of features encoded into context labels. In order to extract the frame-wise parametric representation of both the spectrum and the excitation, an HNM (Harmonics plus Noise Model)-based vocoder, AhoCoder, is used (Erro et al. 2011b). This vocoder allows speech to be reconstructed, too. The voice built is a female voice created using a speech database with the same characteristics as the AhoSyn database (Sainz et al. 2012). A female and a male synthetic voice are used in the demo. The female voice is built following the standard procedure, and the male one is obtained by applying voice transformation techniques (Erro et al. 2013). This AhoTTS version is bilingual, works for Spanish and Basque and is available online: http://sourceforge.net/projects/ahotts/.

Regarding the system that evaluates the correctness of the pronunciation, normally specific databases designed for CAPT (Computer-Assisted Pronunciation Training) purposes are used. But there is currently no available CAPT database for Basque. Although there are some speech recognition databases for Basque, the only one which is publicly available (Hernáez et al. 2003) was recorded over the fixed telephone network, so it is not suitable for CAPT purposes, where speech is usually recorded over a microphone. This is why pronunciation teaching systems for Basque have to be developed with other available data.

The database we use was designed for the training and development of speech recognition for Basque. It is a Speecon-like database (Siemund et al. 2000) and

contains recordings from native and non-native speakers, as well as dialectal and standard Basque data for the former. It contains data from a total of 230 speakers, collected in different places of the Basque Country, where Basque has a different official status, health and phonetic influence of neighbouring languages (mainly French and Spanish). During the recording, speakers were asked about their level of language knowledge, so the database could be divided into different subcorpora according to this information. The native speakers' subcorpus was composed of 149 speakers. Non-native speakers spoke Basque as a second language at two different levels: the high level non-natives' subcorpus included 56 speakers and the low level non-natives' subcorpus 25. Due to dialectal variation and also to the different level of fluency, there were some irregularities in the pronunciation of several phonemes, which were not labeled in the transcription. However, they could be partially deduced from the information provided about the speaker. For example, we could obtain information about the region of origin of the speakers and their Basque level through the labels that indicated their city of birth, city of youth and language level. The audio files had their corresponding orthographic transcription file, and the rule-based AhoT2Ptranscriber was used to obtain phonetic transcriptions.

Due to the lack of a suitable speech database with recordings of Basque non-native speakers, the pronunciation evaluation module was developed using a general purpose ASR speech database (Odriozola et al. 2012). More precisely, the method applied consisted of automatically determining the threshold of GOP (Goodness Of Pronunciation) scores, which were used as pronunciation scores at phone-level. Two score distributions were obtained for each phoneme: one corresponding to its correct pronunciation and the other one to its incorrect pronunciation. The distribution of the scores for erroneous pronunciations was calculated artificially by inserting controlled errors in the dictionary, so that each changed phoneme was randomly replaced by a phoneme from the same group. These phoneme groups were obtained by means of phonetic clustering performed by using regression trees. After obtaining both distributions, the EER (Equal Error Rate) of each distribution pair was calculated and used as a decision threshold for each phoneme. The results of the experiments showed that this method was useful even when there was no database specifically designed for CAPT systems, although it was not as accurate as those specifically designed for this purpose. For the speech recognition module and also for the verification of the correctness of responses, the same database was used.

6 Conclusions

Being present in ICTs, in general, and in language, speech and multimedia technologies, in particular, is, in our opinion, absolutely necessary for any language that intends to go on living in a world that is becoming more and more mobile, digital and interconnected.

As Alegria et al. (2011) stated there is a need of 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. Our experience in BerbaTek proved that collaboration between research agents working in the aforementioned fields is the right way to go. Apart from doing basic research and developing and putting into the hands of the users a considerable number of basic tools and resources, the integration of the different technologies made it possible to create prototypes of advanced applications for the language industry, i.e., translation, content management and learning.

The results of the cross-language comparison[10] provided by the META-NET White Paper Series showed that Basque now stands in a better position than some European official languages such as Croatian, Icelandic, Irish, Latvian or Lithuanian; and that Basque is in the 4th among 5 possible levels of support through language technology on Speech, Text Analytics and Languages Resources. The particular White Paper on Basque (Hernáez et al. 2012) concluded that there were application tools for speech synthesis, speech recognition, spelling correction, and grammar checking, and that there are also some applications for automatic translation, mainly between Spanish and Basque.

Figure 4 shows graphically the classification of the languages in the world proposed by Alegria et al. (2011) according to their degree of development in language technology. As of 2012 Basque is located on an intermediate position in the set of the around 60 languages with some language technology application. It was the 35th language in number of Wikipedia articles. There were 6 products for Basque in the ELRA catalogue[11], 15 products in the ACL wiki[12], and more than 40 on-line dictionaries in the local site hiztegiak.com (although only 5 of them were reflected in yourdictionary.com).

The correlation between Basque speakers and the number of available language products for this language is unusually high. This is due to the coordinated efforts of a few ambitious and productive groups working in successive projects

[10] http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison
[11] http://www.elra.info/
[12] http://aclweb.org/aclwiki/index.php?title=List_of_resources_by_language

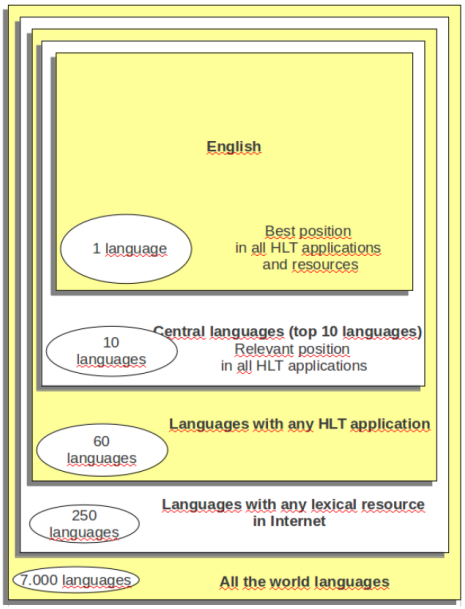

Figure 4: Six different levels for less resourced languages

such as BerbaTek. The collaboration of the five main players in Language Technology for Basque in BerbaTek allowed to make a step further in this direction: this cooperation enabled the creation of many new tools and resources and of three new prototypes in the fields of translation, content management and learning.

In the future, we intend to continue our collaboration and move forward with both the basic research and the development of applications and prototypes, for the language industry and also for other fields. But we also intend to go beyond prototypes and, in collaboration with companies devoted to translation, content management and learning, develop and put onto the market real applications for users, which is the next logical step. It will be a challenge that the members of the BerbaTek consortium are willing and prepared to face.

Acknowledgements

This research was partially funded by the Regional Government of the Basque Autonomous Community (BerbaTek project, IE09-262) and by the Spanish Ministry of Education and Science (OpenMT2, TIN2009-14675-C03-01; Know2, TIN2009-14715-C04-01; HibridoSint;TIN2010-20218).

References

Agirre, Eneko, Olatz Ansa, Xabier Arregi, Maddalen Lopez de Lacalle, Arantxa Otegi, Xabier Saralegi & Hugo Zaragoza. 2009. Elhuyar-IXA: Semantic relatedness and crosslingual passage retrieval. In *Multilingual information access evaluation i: Text retrieval experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece*, vol. 6241 (Lecture Notes in Computer Science), 273–280.

Agirre, Eneko, Xabier Arregi & Arantxa Otegi. 2010. Document expansion based on WordNet for Robust IR. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling)*, 9–17. Beijing, China.

Agirre, Eneko, Kepa Bengoetxea, Koldo Gojenola & Joakim Nivre. 2011. Improving dependency parsing with semantic classes. In *Proceedings of the 49th Annual Meeting of the Association of Computational Linguistics, ACL-HLT 2011 Short Paper*. Portland, Oregon.

Agirre, Eneko & Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009)*, 33–41. Athens, Greece.

Agirre, Eneko, Aitor Soroa, Enrique Alfonseca, Keith Hall, Jana Kravalova & Marius Pasca. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *NAACL'09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 19–27. Boulder, CO.

Aldabe, Itziar & Montse Maritxalar. 2010. Automatic distractor generation for domain specific texts. In *Proceedings of the 7th International Conference on NLP, IceTAL 2010*, 27–38. Reykjavik, Iceland.

Aldezabal, Izaskun, Maria Jesus Aranzabe, Arantza Diaz de Ilarraza, Ainara Estarrona & Larraitz Uria. 2010. EusPropBank: Integrating semantic information in the Basque dependency treebank. *Lecture Notes in Computer Science* 6008. 60–73.

Alegria, Iñaki, Xabier Artola, Arantza Diaz de Ilarraza & Kepa Sarasola. 2011. Strategies to develop language technologies for less-resourced languages based on the case of Basque. In *Proceedings of the 5th Language and Technology Conference: Human language technologies as a challenge for computer science and linguistics*, 42–46. Poznań, Poland.

Alegria, Iñaki, Arantza Casillas, Arantza Diaz de Ilarraza, Jon Igartua, Gorka Labaka, Mikel Lersundi, Aingeru Mayor & Kepa Sarasola. 2008. Spanish-to-Basque MultiEngine machine translation for a restricted domain. In *Proceed-*

ings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA-2008), 57–69. Honolulu, HI.

Alegria, Iñaki, Arantza Diaz de Ilarraza, Gorka Labaka, Mikel Lersundi, Aingeru Mayor & Kepa Sarasola. 2007. Transfer-based MT from Spanish into Basque: Reusability, standardization and open source. *Lecture Notes in Computer Science* 4394. 374–384.

Ansa, Olatz, Xabier Arregi, Arantxa Otegi & Ander Soraluze. 2008. Ihardetsi question answering system at QA@CLEF 2008. In *Working notes of the Cross-Lingual evaluation forum*. Aarhus, Denmark.

Arrieta, Kutz, Arantza Diaz de Ilarraza, Inma Hernáez, Urtza Iturraspe, Igor Leturia, Eva Navas & Kepa Sarasola. 2008. AnHitz, development and integration of language, speech and visual technologies for Basque. In *Proceedings of the second international symposium on universal communication*, 338–344. Osaka.

Bengoetxea, Kepa & Koldo Gojenola. 2010. Application of different techniques to dependency parsing of Basque. In *Proceedings of the First Workshop on Statistical Parsing of Morphologically Rich Languages SPMRL 2010*. Los Angeles, CA. NAACL Workshop.

Boldi, Paolo & Sebastiano Vigna. 2005. MG4J at TREC 2005. In *Proceedings of the annual meeting for the Text REtrieval Conference (TREC)*. Gaithersburg, MD.

Diaz de Ilarraza, Arantza, Antton Gurrutxaga, Kepa Sarasola, Inma Hernáez & Nuria Lopez de Gereñu. 2003. HIZKING21: Integrating language engineering resources and tools into systems with linguistic capabilities. In *Proceedings of the Workshop on NLP of Minority Languages and Small Languages. TALN 2003. Nantes*.

Erro, Daniel, Eva Navas & Inma Hernáez. 2013. Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. *IEEE Transactions on Audio, Speech and Language Processing* 21(3). 556–566.

Erro, Daniel, Iñaki Sainz, Iker Luengo, Igor Odriozola, Jon Sánchez, Ibon Saratxaga, Eva Navas & Inma Hernáez. 2010. HMM-based speech synthesis in Basque language using HTS. In *Proceedings of the Jornadas en Tecnología del Habla (JTH)*. Vigo, Spain.

Erro, Daniel, Iñaki Sainz, Eva Navas & Inma Hernáez. 2011a. HNM-based MFCC+F0 extractor applied to statistical speech synthesis. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Prague, Czec Republic.

Erro, Daniel, Iñaki Sainz, Eva Navas & Inma Hernáez. 2011b. Improved HNM-based vocoder for statistical synthesizers. In *Proceedings of the Interspeech Conference*. Florence, Italy.

España-Bonet, Cristina, Gorka Labaka, Arantza Diaz de Ilarraza, Lluís Màrquez & Kepa Sarasola. 2011. Hybrid machine translation guided by a rule-based system. In *Proceedings of the Machine Translation Summit (MT Summit)*. Xiamen, China.

Fellbaum, Christiane (ed.). 1998. *WordNet an electronic lexical database*. Cambridge, MA ; London: The MIT Press. http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=8106.

Gurrutxaga, Antton & Iñaki Alegria. 2011. Automatic extraction of NV expressions in Basque: Basic issues on cooccurrence techniques. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World (MWE 2011)*. Portland, OR. ACL/HLT conference.

Haveliwala, Taher H. 2002. Topic-sensitive PageRank. In *Proceedings of the World Wide Web conference (WWW)*. Honolulu, HI.

Hernáez, Inmaculada, Iker Luengo, Eva Navas, Maren Zubizarreta, Iñaki Gaminde & Jon Sánchez. 2003. The Basque speech_dat (II) database: A description and first test recognition results. In *Proceedings of the Eurospeech conference*. Geneva, Switzerland.

Hernáez, Inmaculada, Eva Navas, Igor Odriozola, Kepa Sarasola, Arantza Diaz de Ilarraza, Igor Leturia, Araceli Diaz de Lezana, Beñat Oihartzabal & Jasone Salaberria. 2012. Euskara aro digitalean: the Basque language in the digital age. In Georg Rehm & Hans Uszkoreit (eds.), *The Basque language in the digital age/Euskara aro digitalean* (METANET White Paper Series), 35–62. Berlin Heidelberg: Springer.

Kraift, Olivier, Georges Antoniadis, Sandra Echinard, Mathieu Loiseau, Thomas Lebarbé & Claude Ponton. 2004. NLP tools for CALL: The simpler, the better. In *Proceedings of the Workshop on NLP and Speech Technologies in Advanced Language Learning Systems*. Venice, Italy.

Labaka, Gorka. 2010. *EUSMT: Incorporating linguistic information into SMT for a morphologically rich language. Its use in SMT-RBMT-EBMT hybridation*. UPV/EHU-University of the Basque Country dissertation.

Leturia, Igor. 2012. Evaluating different methods for automatically collecting large general corpora for Basque from the web. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*. Mumbai, India.

Leturia, Igor, Iñaki San Vicente & Xabier Saralegi. 2009. Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet. In *Proceedings of the Workshop on Web as Corpus (WAC)*. Donostia, Spain.

Manning, Christopher D., Prabhakar Raghavan & Hinrich Schütze. 2009. *An introduction to information retrieval*. Cambridge: Cambridge University Press.

Mayor, Aingeru, Iñaki Alegria, Arantza Diaz de Ilarraza, Gorka Labaka, Mikel Lersundi & Kepa Sarasola. 2011. Matxin, an open-source rule-based machine translation system for Basque. *Machine Translation Journal* 25(1). 53–82.

Mayor, Aingeru & Francis Tyers. 2009. Matxin: Moving towards language independence. In *Proceedings of the Workshop on Free/Open-Source Rule-Based Machine Translation (FreeRBMT)*. Alacant, Spain.

Odriozola, Igor, Eva Navas, Inma Hernáez, Iñaki Sainz, Ibon Saratxaga, Jon Sánchez & Daniel Erro. 2012. Using an ASR database to design a pronunciation evaluation system in Basque. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*. Istanbul, Turkey.

Otegi, Arantxa. 2012. *Hedapena informazioaren berreskurapenean: Hitzen adieradesanbiguazioaren eta antzekotasun semantikoaren ekarpenak*. UPV/EHU-University of the Basque Country dissertation.

Pociello, Elisabete, Agirre Eneko & Izaskun Aldezabal. 2011. Methodology and construction of the Basque Wordnet. *Language Resources and Evaluation* 45(2). 121–142.

Pociello, Elisabete, Antton Gurrutxaga, Eneko Agirre, Izaskun Aldezabal & German Rigau. 2008. WNTERM: Combining the Basque wordnet and a terminological dictionary. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*. Marrakesh, Morocco.

Robertson, Stephen & Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval* 3(4). 333–389.

Sainz, Iñaki, Daniel Erro, Eva Navas, Inma Hernáez, Jon Sanchez, Ibon Saratxaga & Igor Odriozola. 2012. Versatile speech databases for high quality synthesis for Basque. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*. Istanbul, Turkey.

San Vicente, Iñaki & Iker Manterola. 2012. PaCo2: A fully automated tool for gathering parallel corpora from the web. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*. Istanbul, Turkey.

Saralegi, Xabier & Maddalen Lopez de Lacalle. 2009. Comparing different approaches to treat translation ambiguity in CLIR: Structured queries vs. target co-occurrence-based selection. In *Proceedings of the Workshop on Text-Based Information Retrieval (TIR)*. Linz, Austria.

Saralegi, Xabier, Iker Manterola & Iñaki San Vicente. 2012. Building a Basque-Chinese dictionary by using English as a pivot. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Saralegi, Xabier, Iñaki San Vicente & Maddalen López de Lacalle. 2008. Mining term translations from domain restricted comparable corpora. *Procesamiento del Lenguaje Natural* 41. 273–280.

Schwartz, Lee, Takako Aikawa & Michel Pahud. 2004. Dynamic language learning tools. In *Proceedings of the Workshop on NLP and Speech Technologies in Advanced Language Learning Systems*. Venice, Italy.

Siemund, Rainer, Harald Höge, Siegfried Kunzmann & Krzysztof Marasek. 2000. SPEECON-speech data for consumer devices. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. Athens, Greece.

Sumita, Eiichiro, Fumiaki Sugaya & Seiichi Yamamoto. 2005. Measuring non-native speakers' proficiency of English by using a test with automatically-generated fill-in-the blank questions. In *Proceedings of the Workshop on Building Educational Applications Using NLP*. Ann Arbor, MI.

Williams, Briony, Kepa Sarasola, Donncha Ó´Cróinin, Climent Nadeu & Bojan Petek. 2001. Speech and language technology for minority languages. In *Proceedings of the Eurospeech conference*. Aalborg, Denmark.