D2.3.eus.en Language Report for Basque (English version)

Version 1.2

2011-06-29

METANET4U

www.metanet4u.eu

The central objective of the Metanet4u project is to contribute to the establishment of a pan-European digital platform that makes available language resources and services, encompassing both datasets and software tools, for speech and language processing, and supports a new generation of exchange facilities for them.
This central objective is articulated in terms of the following main goals:
Assessment: to collect, organize and disseminate information that permits an updated insight into the current status and the potential of language related activities, for each of the national and/or language communities represented in the project. This includes organizing and providing a description of: language usage and its economic dimensions; language technologies and resources, products and services; main actors in different areas, including research, industry, government and society in general; public policies and programs; prevailing standards and practices; current level of development, main drivers and roadblocks; etc.
Collection: to assemble and prepare language resources for distribution. This includes collecting languages resources; documenting these language resources; upgrading them to agreed standards and guidelines; linking and cross-lingual aligning them where appropriate.
Distribution: to distribute the assembled language resources through exchange facilities that can be used by language researchers, developers and professionals. This includes collaboration with other projects and, where useful, with other relevant multi-national forums or activities. It also includes helping to build and operate broad inter-connected repositories and exchange facilities.
Dissemination: to mobilize national and regional actors, public bodies and funding agencies by raising awareness with respect to the activities and results of the project, in particular, and of the whole area of language resources and technology, in general.

METANET4U is a project in the META-NET Network of Excellence, a cluster of projects aiming at fostering the mission of META. META is the Multilingual Europe Technology Alliance, dedicated to building the technological foundations of a multilingual European information society.

METANET4U is co-funded by the participating institutions and the ICT Policy Support Programme of the European Commission

and by the participating institutions:

Faculty of Sciences, University of Lisbon

Instituto Superior Técnico

University of Manchester

University *Alexandru Ioan Cuza*

Research Institute for Artificial Intelligence, Romanian Academy

University of Malta

Technical University of Catalonia

Universitat Pompeu Fabra

Revision History

Version	Date	Author	Organisation	Description
1.0	03-06-2011	I. Hernáez, E. Navas, I. Odriozola, K. Sarasola, A. Diaz de Ilarraza, I. Aizpurua, A. Díaz de Lezama, B. Oihartzabal, J. Salaberria	EHU	Daft version
1.1	16-06-2011	I. Hernáez, et al.	EHU	Improved translation
1.1	29-06-2011	I. Hernáez, et al.	EHU	FINAL version

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

METANET4U

D2.3.eus.en Language Report for Basque (English version)

Document METANET4U-2011-D2.3.eus.en
EC CIP project #270893

Deliverable
Number: D2.3.eus.en
Completion: Final
Status: Submitted
Dissemination level: Public

Responsible: Asunción Moreno (WP2 coordinator)

Contributing Partners: Universitat Politècnica de Catalunya, Universitat Pompeu I Fabra

Authors: I. Hernáez, E. Navas, I. Odriozola, K. Sarasola, A. Diaz de Ilarraza, I. Aizpurua, A. Díaz de Lezama, B. Oihartzabal, J. Salaberria

© all rights reserved by FCUL on behalf of METANET4U

Table of Contents

Executive Summary

Many European languages run the risk of becoming victims of the digital age because they are underrepresented and under-resourced online. Huge regional market opportunities remain untapped today because of language barriers. If we do not take action now, many European citizens will become socially and economically disadvantaged because they speak their native language.

Innovative language technology (LT) is an intermediary that will enable European citizens to participate in an egalitarian, inclusive and economically successful knowledge and information society. Multilingual language technology will be a gateway for instantaneous, cheap and effortless communication and interaction across language boundaries.

Today, language services are primarily offered by commercial providers from the US. Google Translate, a free service, is just one example. The recent success of Watson, an IBM computer system that won an episode of the *Jeopardy* game show against human candidates, illustrates the immense potential of language technology. As Europeans, we have to ask ourselves several urgent questions:

- Should our communications and knowledge infrastructure be dependent upon monopolistic companies?
- Can we truly rely on language-related services that can be immediately switched off by others?
- Are we actively competing in the global market for research and development in language technology?
- Are third parties from other continents willing to address our translation problems and other issues that relate to European multilingualism?
- Can our European cultural background help shape the knowledge society by offering better, more secure, more precise, more innovative and more robust high-quality technology?

This whitepaper for the Basque language demonstrates that a lively language technology industry and research environment exists for Basque. Although a number of technologies and resources for Standard Basque exist, there are significantly fewer technologies and resources for the Basque language than for the English language. The existing technologies and resources also have a poorer quality.

According to the assessment detailed in this report, immediate action must occur before any breakthroughs for the Basque language can be achieved.

A Risk for Our Languages and a Challenge for Language Technology

We are witnesses to a digital revolution that is dramatically impacting communication and society. Recent developments in digitised and network communication technology are sometimes compared to Gutenberg's invention of the printing press. What can this analogy tell us about the future of the European information society and our languages in particular?

We are currently witnessing a digital revolution that is comparable to Gutenberg's invention of the printing press.

After Gutenberg's invention, real breakthroughs in communication and knowledge exchange were accomplished through efforts like Luther's translation of the Bible into common language. In subsequent centuries, cultural techniques have been developed to better handle language processing and knowledge exchange:

- the orthographic and grammatical standardisation of major languages enabled the rapid dissemination of new scientific and intellectual ideas;
- the development of official languages made it possible for citizens to communicate within certain (often political) boundaries;
- the teaching and translation of languages enabled an exchange across languages;
- the creation of journalistic and bibliographic guidelines assured the quality and availability of printed material;
- the creation of different media like newspapers, radio, television, books, and other formats satisfied different communication needs.

In the past twenty years, information technology helped to automate and facilitate many of the processes:

- desktop publishing software replaces typewriting and typesetting;
- Microsoft PowerPoint replaces overhead projector transparencies;
- e-mail sends and receives documents faster than a fax machine;
- Skype makes Internet phone calls and hosts virtual meetings;
- audio and video encoding formats make it easy to exchange multimedia content;
- search engines provide keyword-based access to web pages;
- online services like Google Translate produce quick and approximate translations;
- social media platforms facilitate collaboration and information sharing.

Although such tools and applications are helpful, they currently cannot sufficiently implement a sustainable, multilingual European information society, a modern and inclusive society where information and goods can flow freely.

Language Borders Hinder the European Information Society

We cannot precisely know what the future information society will look like. When it comes to discussing a common European energy strategy or foreign policy, we might want to listen to European foreign ministers speak in their native language. We might want a

platform where people, who speak many different languages and who have varying language proficiency, can discuss a particular subject while technology automatically gathers their opinions and generates brief summaries. We also might want to speak with a health insurance help desk that is located in a foreign country.

It is clear that communication needs have a different quality as compared to a few years ago. In a global economy and information space, more languages, speakers and content confront us and require us to quickly interact with new types of media. The current popularity of social media (Wikipedia, Facebook, Twitter and YouTube) is only the tip of the iceberg.

A global economy and information space confronts us with more languages, speakers and content.

Today, we can transmit gigabytes of text around the world in a few seconds before we recognize that it is in a language we do not understand. According to a recent report requested by the European Commission, 57% of Internet users in Europe purchase goods and services in languages that are not their native language. (English is the most common foreign language followed by French, German and Spanish.) 55% of users read content in a foreign language while only 35% use another language to write e-mails or post comments on the web.[i] A few years ago, English might have been the lingua franca of the web—the vast majority of content on the web was in English—but the situation has now drastically changed. The amount of online content in other languages (particularly Asian and Arabic languages) has exploded.

An ubiquitous digital divide that is caused by language borders has surprisingly not gained much attention in the public discourse; yet, it raises a very pressing question, "Which European languages will thrive and persist in the networked information and knowledge society?"

Which European languages will thrive and persist in the networked information and knowledge society?

Our Languages at Risk

The printing press contributed to an invaluable exchange of information in Europe, but it also led to the extinction of many European languages. Regional and minority languages were rarely printed. As a result, many languages like Cornish or Dalmatian were often limited to oral forms of transmission, which limited their continued adoption, spread and use.

The approximately 60 languages of Europe are one of its richest and most important cultural assets. Europe's multitude of languages is also a vital part of its social success.[ii] While popular languages like English or Spanish will certainly maintain their presence in the emerging digital society and market, many European languages could be cut off from digital communications and become irrelevant for the Internet society. Such developments would certainly be unwelcome. On the one hand, a strategic opportunity would be lost , which would weaken Europe's global standing. On the other hand, such developments would conflict with the goal of equal participation for every European citizen regardless of language. According to a UNESCO report on multilingualism, languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society.[iii]

The wide variety of languages in Europe is one of its most important cultural assets and an essential part of Europe's success.

Language Technology is a Key Enabling Technology

In the past, investment efforts have focused on language education and translation. For example, according to some estimates, the

European market for translation, interpretation, software localisation and website globalisation was € 8.4 billion in 2008 and was expected to grow by 10% per annum.[iv] Yet, this existing capacity is not enough to satisfy current and future needs.

Language technology helps people collaborate, conduct business, share knowledge and participate in social and political debates across different languages.

Language technology is a key enabling technology that can protect and foster European languages. Language technology helps people collaborate, conduct business, share knowledge and participate in social and political debates regardless of language barriers or computer skills. Language technology already assists everyday tasks, such as writing e-mails, conducting an online search or booking a flight. We benefit from language technology when we:

- find information with an Internet search engine;
- check spelling and grammar in a word processor;
- view product recommendations at an online shop;
- hear the verbal instructions of a navigation system;
- translate web pages with an online service.

The language technologies detailed in this paper are an essential part of innovative future applications. Language technology is typically an enabling technology within a larger application framework like a navigation system or a search engine. These white papers focus on the readiness of core technologies for each language.

In the near future, we need language technology for all European languages that is available, affordable and tightly integrated within larger software environments. An interactive, multimedia and multilingual user experience is not possible without language technology.

Opportunities for Language Technology

Language technology can make automatic translation, content production, information processing and knowledge management possible for all European languages. Language technology can also further the development of intuitive language-based interfaces for household electronics, machinery, vehicles, computers and robots. Although many prototypes already exist, commercial and industrial applications are still in the early stages of development. Recent achievements in research and development have created a genuine window of opportunity. For example, machine translation (MT) already delivers a reasonable amount of accuracy within specific domains, and experimental applications provide multilingual information and knowledge management as well as content production in many European languages.

Language applications, voice-based user interfaces and dialogue systems are traditionally found in highly specialised domains, and they often exhibit limited performance. One active field of research is the use of language technology for rescue operations in disaster areas. In such high-risk environments, translation accuracy can be a matter of life or death. The same reasoning applies to the use of language technology in the health care industry. Intelligent robots with cross-lingual language capabilities have the potential to save lives.

There are huge market opportunities in the education and entertainment industries for the integration of language technologies in games, edutainment offerings, simulation environments or training programmes. Mobile information services, computer-assisted language learning software, eLearning environments, self-assessment

tools and plagiarism detection software are just a few more examples where language technology can play an important role. The popularity of social media applications like Twitter and Facebook suggests a further need for sophisticated language technologies that can monitor posts, summarise discussions, suggest opinion trends, detect emotional responses, identify copyright infringements or track misuse.

Language technology represents a tremendous opportunity for the European Union that makes sense both economically and culturally. Multilingualism in Europe has become the rule. European businesses, organisations and schools are also multinational and diverse. Citizens want to communicate across the language borders that still exist in the European Common Market. Language technology can help overcome such remaining barriers while supporting the free and open use of language. Furthermore, innovative, multilingual language technology for European languages can also help us communicate with our global partners and their multilingual communities. Language technologies support a wealth of international economic opportunities.

Multilingualism is the rule, not an exception.

Challenges Facing Language Technology

Although language technology has made considerable progress in the last few years, the current pace of technological progress and product innovation is too slow. We cannot wait ten or twenty years for significant improvements to be made that can further communication and productivity in our multilingual environment.

The current pace of technological progress is too slow to arrive at substantial software products within the next ten to twenty years.

Language technologies with broad use, such as the spelling and grammar features in word processors, are typically monolingual, and they are only available for a handful of languages. Applications for multilingual communication require a certain level of sophistication. Machine translation and online services like Google Translate or Bing Translator are excellent at creating a good approximation of a document's contents. But such online services and professional MT applications are fraught with various difficulties when highly accurate and complete translations are required. There are many well-known examples of funny sounding mistranslations, for example, literal translations of the names *Bush* or *Kohl*, that illustrate the challenges language technology must still face.

Language Acquisition

To illustrate how computers handle language and why language acquisition is a very difficult task, we take a brief look at the way humans acquire first and second languages, and then we sketch how machine translation systems work—there's a reason why the field of language technology is closely linked to the field of artificial intelligence.

Humans acquire language skills in two different ways. First, a baby learns a language by listening to the interaction between speakers of the language. Exposure to concrete, linguistic examples by language users, such as parents, siblings and other family members, helps babies from the age of about two produce their first words and short phrases. This is only possible because of a special genetic disposition humans have for learning languages.

Humans acquire language skills in two different ways: learning examples and learning the underlying language rules.

Learning a second language usually requires much more effort when a child is not immersed in a language community of native speakers. At school age, foreign languages are usually acquired by learning their grammatical structure, vocabulary and orthography

from books and educational materials that describe linguistic knowledge in terms of abstract rules, tables and example texts. Learning a foreign language takes a lot of time and effort, and it gets more difficult with age.

The two main types of language technology systems acquire language capabilities in a similar manner as humans. Statistical approaches obtain linguistic knowledge from vast collections of concrete example texts in a single language or in so-called parallel texts that are available in two or more languages. Machine learning algorithms model some kind of language faculty that can derive patterns of how words, short phrases and complete sentences are correctly used in a single language or translated from one language to another. The sheer number of sentences that statistical approaches require is huge. Performance quality increases as the number of analysed texts increases. It is not uncommon to train such systems on texts that comprise millions of sentences. This is one of the reasons why search engine providers are eager to collect as much written material as possible. Spelling correction in word processors, available online information, and translation services such as Google Search and Google Translate rely on a statistical (data-driven) approach.

The two main types of language technology systems acquire language in a similar manner as humans.

Rule-based systems are the second major type of language technology. Experts from linguistics, computational linguistics and computer science encode grammatical analysis (translation rules) and compile vocabulary lists (lexicons). The establishment of a rule-based system is very time consuming and labour intensive. Rule-based systems also require highly specialised experts. Some of the leading rule-based machine translation systems have been under constant development for more than twenty years. The advantage of rule-based systems is that the experts can obtain a more detailed control over the language processing. This makes it possible to systematically correct mistakes in the software and give detailed feedback to the user, especially when rule-based systems are used for language learning. Due to financial constraints, rule-based language technology is only feasible for major languages.

Basque in the European Information Society

General Facts

Basque —or *euskara*, in Basque—, known as 'Lingua Navarrorum' in Latin because it was the popular language in the Kingdom of Navarre, is the only surviving pre-Indo-European language in western Europe. It is considered an isolated language, with no known connections with other languages other than ancient Aquitanian. Both the origin of the language and its relationship with other languages continue to be controversial and of interest for many researchers.

Basque is presently spoken in a small region located at the west of the Pyrenees, on both sides of the border between Spain and France, in the region called *Euskal Herria* (Basque Country, in Basque) by the Basque community. The language has been loosing territory for centuries mainly on the south side. More recently, during the years of Franco's dictatorship when the use of Basque was forbidden, the language suffered an irreparable loss. Enormous efforts of revitalisation of the language were overtaken particularly from the 60s, where a network of schools was created introducing Basque into the educational system, clandestinely during its first years of existence. However, it is only from the 80's, with the linguistic political competences given to the Basque Government after the creation of the Autonomies, that Basque language started a recovery process.

In spite of the tremendous efforts made, in 2009 Basque appeared in the Unesco Map of the World's Languages in Danger[v] as a "vulnerable" language. Nowadays, Basque is estimated to be spoken by about 26% of the population of the Basque Country[vi], either on the Spanish administration side or on the French administration side, but its status is not at all homogeneous. On one hand, the Spanish area of the Basque Country is divided into two political regions: in the Basque Autonomous Community, Basque is legally co-official along with Spanish, but with certain inequalities in favour of Spanish; in the Navarresse Community there are three different areas depending on the legal status of Basque: Basque-speaking, non-Basque-speaking, and mixed. The support for the language and the linguistic rights of the citizens vary depending on which of the three areas they are in. On the other hand, on the French side, Basque is spoken in the western half of the Département of Pyrénées-Atlantiques, but it has never had any legal status of any kind, and it is not official in any institution. However some years ago (2004), a public Agency was created to promote Basque language in French Basque country.

Spoken Basque shows a very high degree of dialectal dispersion. It is now commonly accepted that it is comprised of six dialects which have great differences among them. Standard or Unified Basque was not officially established until 1968 when the Academy of the Basque language *Euskaltzaindia*[vii] made the first standardisation proposal. These dialects have great differences between them in many aspects: lexical, phonetic, morphophonological and also prosodical, in accent and intonation. The dialects are not homogeneous entities; instead, they change continuously from one to another, and in several cases the limit between two or more of them is not so clear.

Particularities of the Basque Language

Basque is an agglutinative and high-inflective language whose major characteristic is that it is an ergative-absolutive language. That means that the subject of an intransitive verb is in the absolutive case (which is unmarked), and the same case is used for the direct object of a transitive verb; the subject of the transitive verb is marked differently, with the ergative case: the suffix *-k*.

Basque is postpositional; so, case and postpositional phrases are formed by attaching a suffix or concatenating more than one to the end of a phrase, according to the following scheme:

root + (article) + (number) + (case(s))

For example, «mutilarengana» (*towards the boy*) is formed by: «mutil+a+∅+r+en+gan+a», —in which «mutil» is the lemma, or noun root; «a» is the article; «∅» the mark of singular; «r» an epenthetic particle; «en» the possessive genitive; «gan» the animate-being marker and «a» the allative—.

This is an important characteristic to be taken into account in natural language and speech processing, since each noun-phrase can be inflected in 17 different ways, multiplied by 4 ways for its definiteness and number. These first 68 forms are further modified based on other parts of sentence, which in turn are inflected for the noun again. It has been estimated that, with two levels of recursion, a Basque noun may have 275 inflected forms, which is, on the other hand, very common[viii]. This implies that it is necessary to find a way of dealing with all these ending variations starting from a basic lexicon.

The verbs are another example of the agglutinative character of Basque. The auxiliary verb, which accompanies most main verbs, agrees not only with the subject, but with any direct object and the indirect object present. Among European languages, this polypersonal agreement is only found in Basque, some languages of the Caucasus, and Hungarian (all non-Indo-European). Verbs in Basque follow the next scheme:

[verb_radical+aspect_suffix] [aux_verb]

For example, in Standard Basque «esaten zenizkidaten» (*you –2nd person plural- used to tell me some things*) is formed by «esan» (*tell*, verb radical) + «ten» (frequentative aspect) and the auxiliary verb «zen+i+zki+da+Ø+te+n», in which «zen» marks the ergative second person; «i» is the auxiliary verb radical; «zki» the absolutive third person plural; «da» the dative first person singular; «Ø» is the indicative marker; «te» the ergative plural marker; and «n» the marker for the past tense. Due to this complexity, it is usual in Natural Language Processing research to opt for treating each of the auxiliary verbs as a whole, instead of dividing them into morphemes.

As far as the word order of the sentence is concerned, the basic syntactic construction is Subject-Objects-Verb (unlike Spanish, French or English where Subject-Verb-Objects construction is more common). The order of the phrases within a sentence can be changed with thematic purposes, whereas the order of the words within a phrase is usually rigid. As a matter of fact, Basque phrase order is topic-focus, meaning that in neutral sentences (such as sentences to inform someone of a fact or event) the topic is stated first, then the focus. In such sentences, the verb phrase comes at

the end. In brief, the focus directly precedes the verb phrase. This rule is also applied in questions, for instance, *What is this?* can be translated as «Zer da hau?» or «Hau zer da?», but in both cases the question tag «zer» immediately precedes the verb «da». This rule is so important in Basque that, even in grammatical descriptions of Basque written in other languages, the Basque word *galdegai* (focus) is used.

Basque orthography is almost phonemic: each grapheme corresponds to one phoneme, and so, the pronunciation of a word can be easily figured out from its written form. Nevertheless, there are a few exceptions: <l> and <n> are usually palatalized when they are preceded by a <i> and followed by a vowel; e.g., mutila => <mutiLa> (*the boy*). Another example is that the consonant phoneme at the end of the negative particle "ez" (*no*) converts the contiguous next phoneme in a voiceless phoneme; e.g., ez dira => <eztira> (*they are not*).

Recent developments

A standardised form of the Basque language, called *Euskara Batua*, was developed by *Euskaltzaindia*, the Academy of the Basque Language in the late 1960s. Euskara Batua was created so that Basque language could be used—and easily understood by all Basque speakers— in formal situations (education, mass media, literature...), and this is its main use nowadays. For classic literary reasons, Standard Basque is based mainly on the Central and Navarrese-Labourdin dialects. The extreme dialects, differ noticeably from it, despite that the Western dialect is one of the most spoken dialects of the language together with the Central dialect.

Standard Basque has solid foundations and it is developing forward aspects as syntax and naturalness. At present, almost all the people that study Basque learn the Euskara Batua. This fact has created a phenomenon all around the Basque country in which Basque people speak their own local dialect with locals, and standard Basque with the 'new Basque speakers' (*euskaldun berri)*. In the Western area, due to the great differences between the western dialect and the standard, it has led to a situation where people studying Basque feel that the language they are studying is pretty far from what Basque people speak. On the other hand, it is now already a fact that there are standard Basque speakers whose mother tongue is precisely standard Basque, because many new Basque speakers opt to speak to their children in Basque, even that their own primary language was Spanish.

However, the idea that the future of Basque is related not only to the development of Standard Basque but also to the promotion of the current dialects is more and more accepted by the theoreticians of the Basque language[ix]. So, dialects will be somehow important in the future applications of LT for Basque.

Basque LT community and researchers, conscious of the importance of technologies for languages spoken by little communities to evolve in the 21st century, have made a great effort to place Basque at the same technological level as the most used languages. There is a solid scientific experience along with other neighbouring languages, such as Catalan and Galician; that is virtually unique in Europe, such as the development of cross-lingual products and services between regional languages.

The importance of the development of a LT industry for Basque is evident taking into account the creation of *Langune*[x]. Langune is

an association of Basque Country companies belonging to the Language Industry sector. This association was set up in 2010 and brings together over 30 companies in the spheres of translation, content, teaching and language technologies. Its main objective is to develop the sector of LT, which will be a benchmark in the language industry in Europe, while avoiding the duplication of efforts and achieving synergies. Langune has just started but is taking giant steps.

Language cultivation

The Basque language is mainly represented by '*Euskaltzaindia*', the Royal Academy of the Basque Language (1919). It carries out research in the language, seeks to protect it and establishes standards of use. It enjoys full official recognition as a royal academy in Spain (1976) and as a cultural association of public benefit within the territory of France (1995).

Since the declaration of Basque as the official language in the Autonomous Basque Community, the Basque Government has developed numerous norms and laws in order to protect and favour the use of the language. Various organisms and institutions have since been created: Basque Advisory Board (1982), Basque Radio-Television EiTB (1982), the Institute for Adults Literacy-HABE (1983) and many others.

The 'General Plan for the Promotion of the Use of Basque' was first introduced in 1998 as a strategic instrument with three main objectives: reach consensus in goals and actions of the different institutions, establish priorities for the founding programmes and coordinate the activities of institutions, companies and associations dealing with Basque. Within this strategic Plan, periodical sociolinguistic surveys serve as guide for establishing new goals and correction directions. The Basque Government has a web-portal (www.euskara.euskadi.net) dedicated to the Basque language, offering information not only about the language and its history and present situation, but also links to every kind of service, product or application related with the language, including public funding programmes. In the French area, the "Office Public de la Langue Basque"[xi] was created in 2004, as a public Agency bringing together four local or regional public institutions and the state, with the goal of defining and applying a common linguistic policy in the region to promote Basque language.

Language in Education

In the Basque Autonomous Community, Basque was officially introduced in the public education system in 1983 with the law that regulates the use of Basque and Spanish in the Primary and Secondary School. For the Primary and Secondary School three models were created, giving the possibility to each institution to choose the model to offer. In model *A* the vehicular language is Spanish, and Basque is taught in the subject "Basque Language and Literature". In model *D* -the letter C is not normally used in Basque- Basque is the vehicular language and there is one subject "Spanish Language and Literature" taught in Spanish. Model B is an intermediate model, where some of the subjects are taught in Spanish (mainly Reading and Writing and Mathematics) and another part in Basque (mainly science and plastic). However, the Model A had been loosing students progressively, in favour of Model B, mainly in pre- and primary school, where more than half of the students learn in Model D. Yet, 85% of the 15 years old students made the examinations for the PISA Study in Spanish whilst only 15% did them in

Basque[xii], clearly showing that Spanish is the dominant language in Education.

In the Navarresse Community, where Basque has different grades of official status depending on the area, a forth model was also available with no mandatory subject of Basque. As for the Northern provinces in France, primary education in Basque is offered by the private network of schools 'Seaska', which is managing presently almost 2700 students in 29 establishments that include one centre for secondary education and one 'lizeo'.

Very recently, new models are being proposed and tested, which consider the importance of early learning of English. The Basque Government in Spain has recently introduced a trilingual model, while in Navarre bilingual education in Spanish and English has been introduced, although Basque is offered optionally.

At higher levels of education, the offer is clearly dominated by Spanish. From the three existing universities, the only public university, Universidad del País Vasco / Euskal Herriko Unibertsitatea' (*UPV/EHU*), offers the possibility of learning in Basque, and although enormous efforts have been made to make equal offer in Basque as in Spanish, only very few degrees can be taken fully in Basque. Remarkably, a Master and Doctorate Program 'Analysis and Processing of Language'[xiii] totally offered in Basque exists since the year 2001. The private University Mondragon Unibertsitatea offers most of their degrees in Basque and some of their Master studies in Basque. The third University, Universidad de Deusto, offers only some of the courses in Basque.

International aspects

Since January 2009, the Etxepare Basque Institute is the Basque public institution responsible for spreading the Basque language and culture all over the world. This institution is aiming to promote the teaching, study and use of Basque throughout the world and to include the contributions of all the communities that share Basque as a common language. The Institute also aims to disseminate Basque culture in the international community with very special reference to those groups that speak Basque, including the Basque *Diaspora*. Along the history, many Basques have left the Basque Country for other parts of the globe for economic and political reasons; Basque Diaspora is the name given to describe people of Basque origin living outside their traditional homeland. Currently there are substantial Basque origin populations in Chile, Argentina, Bolivia, Ecuador, Colombia, Cuba, Mexico, Venezuela, Canada and the United States. All of them have several Basque cultural centres (*Euskal Etxeak*) that were established to pursue the same objective: the perpetuation of Basque culture and identity. There are Basque cultural centres in most large cities of 24 different countries[xiv].

The origins and singular structure of Basque have raised the interest in the study of Basque language and culture. Currently it can be learned in 29 universities belonging to 13 different American and European countries.

Regarding the use of Basque language in international institutions, the Spanish government has made efforts in favour of including it, together with Catalan and Galician among the official languages of the European institutions. But currently they do not enjoy the status of official languages there; they are considered semi-official, together with Scottish, Gaelic and Welsh. Basque can only be used

in very limited situations: it can be spoken at the work sessions of the Region Committee and the Council, but not in the plenary meetings of the European Parliament. The citizens can also write to the European institutions using Basque and have right to be answered in the same language, but always through the Spanish Government and this government must pay the derived fees.

Basque is included in the list Regional and minority languages of the European Union[xv] and as such it benefits from the resolutions adopted by the European Parliament to promote action on regional and minority languages.

Language technology can address this challenge from a different perspective by offering services like machine translation or cross-lingual information retrieval to foreign language text and thus help diminish personal and economic disadvantages naturally faced by non-native speakers of English.

Basque on the Internet

In the first quarter of 2010, 61.4% of the households (513 000) in the Basque Country had a computer. There were slightly over 460 000 families, of which 54.9%, had access to the Internet from their homes. This means that over a million people aged 15 and over were Internet users. Most of them stated to be online every day. Only 22.9% of them used Basque language on the Internet.[xvi] Nevertheless there is a strong and willing community of Internet users among Basque speaking people. The blogosphere in Euskara, the Wikipedia and online services in Euskara, as well as the location of tools and operating systems based on free software, have fostered the presence of Euskara and Basque culture, both on the Internet and ICT, encouraging, in this way, the expansion of its use. For instance, the Basque Wikipedia has more than 97 000 articles occupying the 39th place in number of articles among all the Wikipedia. And a big effort has been made in order to provide different common software programs[xvii] [xviii] and resources in Basque[xix] [xx] [xxi] [xxii].

A new top level domain .eus has been registered and will be launched in mid 2012. It already counts with 193 pre-registrations. The proposed top-level domain .eus is the name that will represent the Community of the Basque Language and Culture on the Internet. This symbol will become a tool for the promotion of Basque culture and Euskara, and, in this sense, the .eus domain will be an effective mechanism for linguistic standardisation of Euskara worldwide. The .eus domain, through the virtual space of the Internet, will assure an efficient promotion of Euskara, guaranteeing simultaneously its international recognition. Similarly, the .eus domain will reinforce and extend the multicultural nature of the Internet, since allowing linguistic and cultural communities to have their own domain puts multiculturalism at the very heart of the Internet. Domains related to language and cultures strengthen and benefit not only those linguistic and cultural communities but also the Internet itself.[xxiii]

For language technology, the growing importance of the internet is important in two ways. On one hand, the large amount of digitally available language data represents a rich source for analysing the usage of natural language, in particular by collecting statistical information. On the other hand, the internet offers a wide range of application areas involving language technology.

Selected Further Reading

J.L. Hualde, J. Lakarra, R. L. Trask *Eds. (1995)* "Towards a History of the Basque Language" John Benjamins Publishing Company

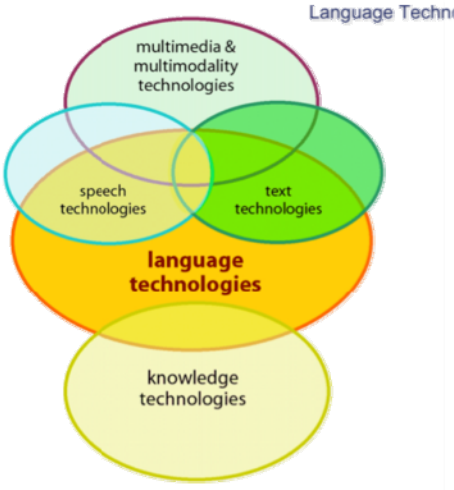
J.L. Hualde, J. Ortiz de Urbina (*Eds.*)(2003) "A Grammar of Basque" Mouton de Gruyter, Berlin

K. Zuazo, 2010, "El euskera y sus dialectos" (in Spanish) Alberdania

Language Technology Support for Basque

Language Technologies

Language technologies are information technologies that are specialised for dealing with human language. Therefore these technologies are also often subsumed under the term Human Language Technology. Human language occurs in spoken and written form. Whereas speech is the oldest and most natural mode of language communication, complex information and most of human knowledge is maintained and transmitted in written texts. Speech and text technologies process or produce language in these two modes of realisation. But language also has aspects that are shared between speech and text such as dictionaries, most of grammar and the meaning of sentences. Thus, large parts of language technology cannot be subsumed under either speech or text technologies. Among those are technologies that link language to knowledge. The figure on the right illustrates the Language Technology landscape. In our communication we mix language with other modes of communication and other information media. We combine speech with gesture and facial expressions. Digital texts are combined with pictures and sounds. Movies may contain language in spoken and written form. Thus speech and text technologies overlap and interact with many other technologies that facilitate the processing of multimodal communication and multimedia documents.



Language Technology Application Architectures

Typical software applications for language processing consist of several components that mirror different aspects of language and of the task they implement. The figure on the right displays a highly simplified architecture that can be found in a text processing system. The first three modules deal with the structure and meaning of the text input:

- Pre-processing: cleaning up the data, removing formatting, detecting the input language, etc.
- Grammatical analysis: finding the verb and its objects, modificators, etc.; detecting the sentence structure.
- Semantic analysis: disambiguation (Which meaning of "apple" is the right one in the given context?), resolving anaphora and referring expressions like "she", "the car", etc.; representing the meaning of the sentence in a machine-readable way.

Task-specific modules then perform many different operations such as automatic summarisation of an input text, database look-ups and many others. Below, we will illustrate core application areas and highlight their core modules. Again, the architectures of the applications are highly simplified and idealised, to illustrate the complexity of Language Technology (LT) applications in a generally understandable way. The most important tools and resources involved are underlined in the text and can also be found in the table at the end of the chapter. The sections discussing the core application areas also contain an overview of the industries active in the respective field in Basque.

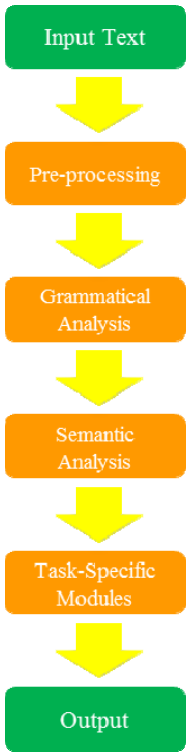


Figure 2: A Typical Text Processing Application Architecture

After introducing the core application areas, we will give a short overview of the situation in LT research and education, concluding with an overview of past and ongoing research programs. At the

end of this section, we will present an expert estimation on the situation regarding core LT tools and resources on a number of dimensions such as availability, maturity, or quality. This table gives a good overview on the situation of LT for Basque.

Core application areas

Language checking

Anyone using a word processing tool such as Microsoft Word has come across a **spell checking** component that indicates spelling mistakes and proposes corrections. Forty years after the first spelling correction program by Ralph Gorin, language checkers nowadays do not simply compare the list of extracted words against a dictionary of correctly spelled words, but have become increasingly sophisticated. In addition to language-dependent algorithms for handling morphology (e.g. plural formation), some are now capable of recognising syntax–related errors, such as a missing verb or a verb that does not agree with its subject in person and number, e.g. in 'She **write* a letter.' However, most available spell checkers (including Microsoft Word) will find no errors in the following first verse of a poem by Jerrold H. Zar (1992):

Eye have a spelling chequer,

It came with my Pea Sea.

It plane lee marks four my revue

Miss Steaks I can knot sea.

For handling this type of errors, analysis of the context is needed in many cases, e.g., in Basque, for deciding if the ergative marker has to be used, as in:

Liburua neskak dauka

[The girl has the book]

Irakurlea neska da.

[The reader is a girl.]

This either requires the formulation of language-specific grammar rules, i.e. a high degree of expertise and manual labour, or the use of a so-called statistical language model. Such models calculate the probability of a particular word occurring in a specific environment (i.e., the preceding and following words). For example, *neskak dauka* is a much more probable word sequence than *neska dauka*. A statistical language model can be automatically derived using a large amount of (correct) language data (i.e. a corpus). Up to now, these approaches have mostly been developed and evaluated on English language data. However, they do not necessarily transfer straightforwardly to Basque with its richer inflection and agglutinative morphology. In fact, language modelling for Basque poses enormous difficulties due to the impossibility of collecting all possible word-forms.

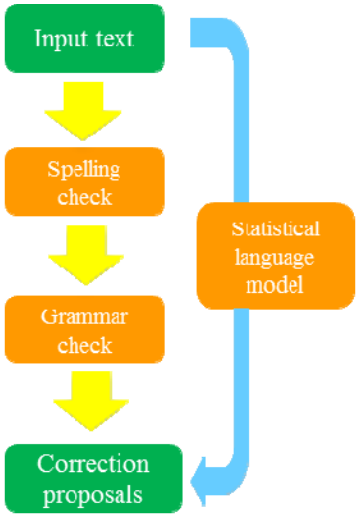


Figure 3: Language Checking (left: rule-based; right: statistical)

The use of Language Checking is not limited to word processing tools, but it is also applied in authoring support systems. Accompanying the rising number of technical products, the amount of technical documentation has rapidly increased over the last decades. Fearing customer complaints about wrong usage and damage claims resulting from bad or badly understood instructions, companies have begun to focus increasingly on the quality of technical documentation, and at the same time targeting the international market. Advances in natural language processing lead to the

development of authoring support software, which assists the writer of technical documentation to use vocabulary and sentence structures consistent with certain rules and (corporate) terminology restrictions.

The most used Spell Checker for Basque is the so-called 'Xuxen'[xxiv], which was developed by the university research group IXA (ixa.si.ehu.es) and is supplied by the SME 'Eleka Ingenieritza Linguistikoa'. This Spell Checker is not limited to the use of a lexicon as it is common practice for English or other less-inflected languages. On the contrary, morphological analysis is performed. The newest version of this spell checker also performs grammar and style corrections. This version also includes code developed by the company 'Hizkia[xxv]' and the institution 'UZEI[xxvi]'.

Besides spell checkers and authoring support, Language Checking is also important in the field of computer-assisted language learning and is applied to automatically correct queries sent to Web Search engines, e.g. Google's 'Did you mean...' suggestions.

Web search

Search on the web, in intranets or in digital libraries, is probably the most widely used and yet underdeveloped Language Technology today. The search engine Google, which started in 1998, is nowadays used for about 80% of all search queries world-wide[xxvii].

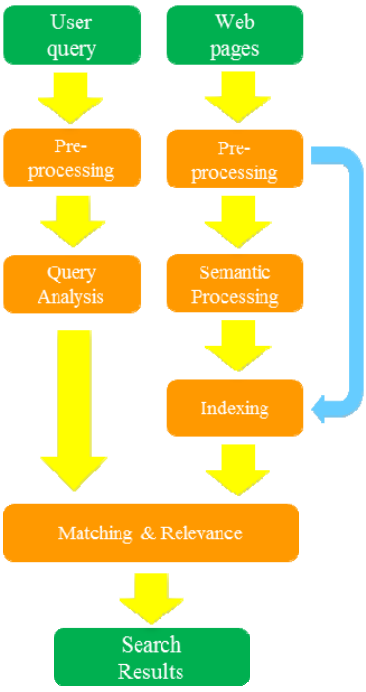


Figure 4: Web Search Architecture

Neither the search interface nor the presentation of the retrieved results has significantly changed since the first version. In the current version, Google offers a spelling correction for misspelled words and also, in 2009, they incorporated basic semantic search capabilities into their algorithmic mix[xxviii], which can improve search accuracy by analysing the meaning of the query terms in context. The success story of Google shows that with a lot of data at hand and efficient techniques for indexing these data, a mainly statistically-based approach can lead to satisfactory results.

However, for a more sophisticated request for information, integrating deeper linguistic knowledge is essential. In the research labs, experiments using machine-readable thesauri and ontological language resources like WordNet have shown improvements by allowing the possibility of finding a page on the basis of synonyms of the search terms. Again, these developments require of language specific resources. A Basque WordNet 'BasWN' has been developed by the research group IXA at the University of the Basque Country and is commercially available through ELRA.

The next generation of search engines will have to include much more sophisticated Language Technology. If a search query consists of a question or another type of sentence rather than a list of keywords, retrieving relevant answers to this query requires an analysis of this sentence on a syntactic and semantic level as well as the availability of an index that allows for a fast retrieval of the relevant documents. For example, imagine a user inputs the query 'Give me a list of all companies that were taken over by other companies in the last five years'. For a satisfactory answer, syntactic parsing needs to be applied to analyse the grammatical structure of the sentence and determine that the user is looking for companies that have been taken over and not companies that took over others. Also, the expression *last five years* needs to be processed in order to find out which years it refers to.

Finally, the processed query needs to be matched against a huge amount of unstructured data in order to find the piece or pieces of information the user is looking for. This is commonly referred to as information retrieval and involves the search for and ranking of relevant documents. In addition, generating a list of companies, we also need to extract the information that a particular string of words in a document refers to a company name. This kind of information is made available by so-called named-entity recognisers.

Even more demanding is the attempt to match a query to documents written in a different language. For cross-lingual information retrieval, we have to automatically translate the query to all possible source languages and transfer the retrieved information back to the target language. The increasing percentage of data available in non-textual formats drives the demand for services enabling multimedia information retrieval, i.e., information search on images, audio, and video data. For audio and video files, this involves a speech recognition module to convert speech content into text or a phonetic representation, to which user queries can be matched.

Focus on development for these companies lies on providing add-ons and advanced search engines for special-interest portals by exploiting topic-relevant semantics. Due to the still high demands in processing power, such search engines are only economically usable on relatively small text corpora. Processing time easily exceeds that of a common statistical search engine as, e.g., provided by Google by a magnitude of thousands. These search engines also have high demand in topic-specific domain modelling, making it not feasible to use these mechanisms on web scale.

In the Basque Autonomous Community, the small company 'Eleka Ingeniaritza Linguistikoa' has been very active in the development of applications and web based services for Basque. They usually integrate LT research results and resources such as lemmatizers and lexical databases of the IXA group and *Elhuyar Foundation*. The multilingual search engine *elebila* considers the Basque language specifics and integrates various linguistic tools and resources to offer high quality search results for Basque. Another example is the tool called *Miatu* ('Examine' in Basque), a library offering functionality to search in special purpose indexed databases using lemmatizers and other morphology analysis tools. It has been used to develop the science related web portal www.zientzia.net and the educational content portal www.ikasbil.net.

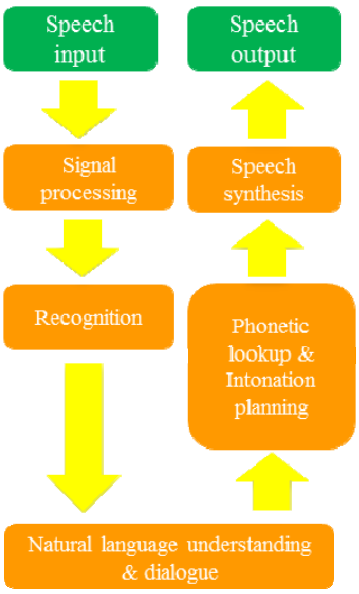


Figure 5: Simple Speech-based Dialogue Architecture

Speech interaction

Speech Interaction technology is the basis for the creation of interfaces that allow a user to interact with machines using spoken language rather than, e.g., a graphical display, a keyboard, and a mouse. Today, such voice user interfaces (VUIs) are usually employed for partially or fully automating service offerings provided by companies to their customers, employees, or partners via the telephone. Business domains that rely heavily on VUIs are banking, logistics, public transportation, and telecommunications. Other usages of Speech Interaction technology are interfaces to particular devices, e.g. in-car navigation systems, and the employment of spoken language as an alternative to the input/output modalities of graphical user interfaces, e.g. in smartphones.

At its core, Speech Interaction comprises the following four different technologies:

- Automatic speech recognition (ASR) is responsible for determining which words were actually spoken given a sequence of sounds uttered by a user.
- Syntactic analysis and semantic interpretation deal with analysing the syntactic structure of a user's utterance and interpretting the latter according to the purpose of the respective system.
- Dialogue management is required for determining, on the part of the system the user interacts with, which action shall be taken given the user's input and the functionality of the system.
- Speech synthesis (Text-to-Speech, TTS) technology is employed for transforming the wording of that utterance into sounds that will be output to the user.

One of the major challenges is to have an ASR system recognise the words uttered by a user as precisely as possible. This requires either a restriction of the range of possible user utterances to a limited set of keywords, or the manual creation of language models that cover a large range of natural language user utterances. Whereas the former results in a rather rigid and inflexible usage of a VUI and possibly causes a poor user acceptance, the creation, tuning and maintenance of language models may increase the costs significantly. However, VUIs that employ language models and initially allow a user to flexibly express their intent – evoked, e.g., by a 'How may I help you' greeting – show both a higher automation rate and a higher user acceptance and may therefore be considered as advantageous over a less flexible directed dialogue approach.

For the output part of a VUI, companies tend to use pre-recorded utterances of professional – ideally corporate – speakers a lot. For static utterances, in which the wording does not depend on the particular contexts of use or the personal data of the given user, this will result in a rich user experience. However, the more dynamic content an utterance needs to consider, the more the user experience may suffer from a poor prosody resulting from concatenating single audio files. In contrast, today's TTS systems prove superior, though optimisable, regarding the prosodic naturalness of dynamic utterances.

Regarding the market for Speech Interaction technology, the last decade underwent a strong standardisation of the interfaces between the different technology components, as well as by standards for creating particular software artefacts for a given application. There also has been strong market consolidation within the last ten years, particularly in the field of ASR and TTS. Here, the national markets in the G20 countries – i.e. economically strong countries with a considerable population - are dominated by less than 5 players worldwide, with Nuance and Loquendo being the most prominent ones in Europe. Since 2007, thanks to the support given by the Basque Government, Basque language is included in the catalogue of products of Nuance. However, the offer in ASR is limited to small to medium size vocabulary applications and no dictation product is available. For TTS, just one female voice is available. On the Spanish market, the Catalan SME Verbio Speech Technologies[xxix] also offers Basque both for ASR and TTS, with more than one voice. Still, no commercial dictation system exists for Basque.

Regarding dialogue management technology and know-how, markets are strongly dominated by national players, which are usually SMEs. Most of the companies on the Spanish TTS market are essentially application developers. Key players in the Spanish market

are: Indsys[xxx] (Intelligent Dialogue Systems), Fonetic[xxxi], Ydilo[xxxii] and NaturalVox[xxxiii]. Some of them have a limited offer in Basque. Free TTS software for the Basque language is also offered by the research group Aholab[xxxiv] of the University of the Basque Country (UPV/EHU).

Looking beyond today's state of technology, there will be significant changes due to the spread of smart phones as a new platform for managing customer relationships – in addition to the telephone, internet, and email channels. This tendency will also affect the employment of technology for speech interaction. On one hand, demand for telephony-based VUIs will decrease, in the long run. On the other hand, the usage of spoken language as a user-friendly input modality for smart phones will gain significant importance. This tendency is supported by the observable improvement of speaker independent speech recognition accuracy for speech dictation services that are already offered as centralised services to smart phone users. Given this 'outsourcing' of the recognition task to the infrastructure of applications, the application-specific employment of linguistic core technologies will supposedly gain importance compared to the present situation.

Machine Translation

The idea of using digital computers for translation of natural languages came up in 1946 by A. D. Booth and was followed by substantial funding for research in this area in the 1950s and beginning again in the 1980s. Nevertheless, Machine Translation (MT) still fails to fulfil the high expectations it gave rise to in its early years.

At its basic level, MT simply substitutes words in one natural language by words in another. This can be useful in subject domains with a very restricted, formulaic language, e.g., weather reports. However, for a good translation of less standardised texts, larger text units (phrases, sentences, or even whole passages) need to be matched to their closest counterparts in the target language. The major difficulty here lies in the fact that human language is ambiguous, which yields challenges on multiple levels, e.g., word sense disambiguation on the lexical level ('Jaguar' can mean a car or an animal) or on other levels as in:

Egon garenetan ez dugu topatu

[Each time we were there we have not seen him/her] or

[In every place we were we have not seen him/her]

Aitak semeari bere bizikleta eman dio

[The father has given his bicycle to his son]

One way of approaching the task is based on linguistic rules. For translations between closely related languages, a direct translation may be feasible in cases like in the second example above. But often rule-based (or knowledge-driven) systems analyse the input text and create an intermediary, symbolic representation, from which the text in the target language is generated. The success of these methods is highly dependent on the availability of extensive lexicons with morphological, syntactic, and semantic information,

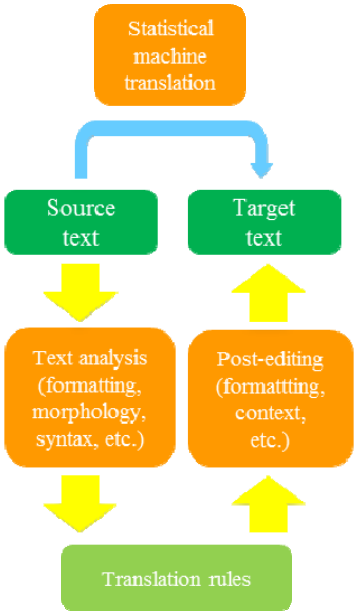


Figure 6: Machine translation (top: statistical; bottom: rule-based)

and large sets of grammar rules carefully designed by a skilled linguist.

Beginning in the late 1980s, as computational power increased and became less expensive, more interest was shown in statistical models for MT. The parameters of these statistical models are derived from the analysis of bilingual text corpora, such as the Europarl parallel corpus, which contains the proceedings of the European Parliament in 11 European languages. Given enough data, statistical MT works well enough to derive an approximate meaning of a foreign language text. However, unlike knowledge-driven systems, statistical (or data-driven) MT often generates ungrammatical output. On the other hand, besides the advantage that less human effort is required for grammar writing, data-driven MT can also cover particularities of the language that go missing in knowledge-driven systems, for example idiomatic expressions.

As the strengths and weaknesses of knowledge- and data-driven MT are complementary, researchers nowadays unanimously target hybrid approaches combining methodologies of both. This can be done in several ways. One is to use both knowledge- and data-driven systems and have a selection module decide on the best output for each sentence. However, for longer sentences, no result will be perfect. A better solution is to combine the best parts of each sentence from multiple outputs, which can be fairly complex, as corresponding parts of multiple alternatives are not always obvious and need to be aligned.

For Basque, MT is particularly challenging. The rich morphology, the high degree of inflection and the agglutinative character of the language makes dictionary analysis and dictionary coverage difficult. Additionally, due to the order of the sentence components, parallel corpora are difficult to manage.

Matxin is a Transfer-based MT system from Spanish into Basque developed by IXA Group at the University of the Basque Country (UPV/EHU). It is an open, reusable and interoperable framework useful even for other language-pairs (matxin.sourceforge.org). It uses other open source codes such as Freeling, and reuses Basque morphology for morphological generation. IXA Group has also created an improved Statistical Machine Translation system for Basque Spanish that deals with morphological segmentation and word reordering (EUSMT . http://ixa2.si.ehu.es/openmt-demo/). For the development of these MT systems, there is strong collaboration between the university research group, the local SME *Eleka Ingeniaritza Linguistikoa* and the *Elhuyar Foundation*, which provides considerable amounts of linguistic resources. This SME has also developped the translator Standard Basque *batua* - Western dialect *bizkaiera*. Also, a Basque to Spanish initial system has been developed by the Transducens Group at Universitat d'Alacant, using the platform Apertium. Google's Translator offers an alpha version for Basque.

Leading international MT developer Lucy Software has an important subsidiary in Spain, Lucy Iberica[xxxv], former Translendium. This company was selected in 2008 by the Basque Government to develop a Spanish-Basque translation system and again in 2011 to continue the work.

Provided good adaptation in terms of user-specific terminology and workflow integration, there is a wide consensus that the use of MT can increase productivity significantly. The quality of MT systems is still considered to have huge improvement potential. Challenges

include the adaptability of the language resources to a given subject domain or user area and the integration into existing workflows with term bases and translation memories. In addition, many language pairs are still missing.

Language Technology 'behind the scenes'

Building Language Technology applications involves a range of subtasks that do not always surface at the level of interaction with the user, but provide significant service functionalities 'under the hood' of the system. Therefore, they constitute important research issues that have become individual sub-disciplines of Computational Linguistics in academia.

Question answering has become an active area of research, for which annotated corpora have been built and scientific competitions have been started. The idea is to move from keyword-based search (to which the engine responds with a whole collection of potentially relevant documents) to the scenario of the user asking a concrete question and the system providing a single answer: 'At what age did Neil Armstrong step on the moon?' - '38'. While this is obviously related to the aforementioned core area Web Search, question answering nowadays is primarily an umbrella term for research questions such as what *types* of questions should be distinguished and how should they be handled, how can a set of documents that potentially contain the answer be analysed and compared (do they give conflicting answers?), and how can specific information - the answer - be reliably extracted from a document, without unduly ignoring the context.

This is in turn related to the information extraction (IE) task, an area that was extremely popular and influential at the time of the 'statistical turn' in Computational Linguistics, in the early 1990s. IE aims at identifying specific pieces of information in specific classes of documents; this could e.g. be the detection of the key players in company takeovers as reported in newspaper stories. Another scenario that has been worked on is reports on terrorist incidents, where the problem is to map the text to a template specifying the perpetrator, the target, time and location of the incident, and the results of the incident. Domain-specific template-filling is the central characteristic of IE, which for this reason is another example of a 'behind the scenes' technology that constitutes a well-demarcated research area but for practical purposes then needs to be embedded into a suitable application environment.

Two 'borderline' areas, which sometimes play the role of standalone application and sometimes that of supportive, 'under the hood' component are text summarization and text generation. Summarization, obviously, refers to the task of making a long text short, and is offered for instance as a functionality within MS Word. It works largely on a statistical basis, by first identifying 'important' words in a text (that is, for example, words that are highly frequent in this text but markedly less frequent in general language use) and then determining those sentences that contain many important words. These sentences are then marked in the document, or extracted from it, and are taken to constitute the summary. In this scenario, which is by far the most popular one, summarization equals sentence extraction: the text is reduced to a subset of its sentences. All commercial summarizers make use of this idea. An alternative approach, to which some research is devoted, is to actually synthesize *new* sentences, i.e., to build a summary of sentences that need not show up in that form in the source text. This requires a certain amount of deeper understanding of the

text and therefore is much less robust. All in all, a text generator is in most cases not a stand-alone application but embedded into a larger software environment, such as into the clinical information system where patient data is collected, stored and processed, and report generation is just one of many functionalities.

For Basque, the situation in all these research areas is much less developed than it is for English, where question answering, information extraction, and summarization have since the 1990s been the subject of numerous open competitions, primarily those organised by DARPA/NIST in the United States. These have significantly improved the state of the art, but the focus has always been on English; some competitions have added multilingual tracks, but Basque was never a targeted language. Accordingly, there are hardly available annotated corpora or other resources for these tasks. Summarization systems, when using purely statistical methods, are often to a good extent language-independent, and thus some research prototypes are available. For text generation, reusable components have traditionally been limited to the surface realisation modules (the "generation grammars"); again, most available software is for English.

Language Technology in Education

Language Technology is a highly interdisciplinary field, involving the expertise of linguists, computer scientists, mathematicians, philosophers, psycholinguists, and neuroscientists, among others. Consequently, the current basic training of a computational linguist may be performed in Spain within the framework of a degree in Philology or Linguistics, which includes Computational Linguistics as a core subject, or by Computational Science faculties. Among the Universities that offer the first option: Universitat de Barcelona, Universitat Pompeu Fabra, Universitat Oberta de Catalunya and Universidade de Vigo. On the other hand, main computational science faculties offering Computational Linguistic as subject are: Universidad Politécnica de Madrid, Universidad Carlos III, Universidad Autónoma de Madrid, Universitat d'Alacant, Universidad Nacional de Educación a Distancia and Universidad del País Vasco / Euskal Herriko Unibertsitatea. Other cases, such as the Universidad Complutense combine both.

Graduate courses offer a more targeted professional training. There are several doctoral programs which offer masters or subjects related to language and speech processing. A complete doctoral program on Language Processing is offered by Universidad del País Vasco / Euskal Herriko Unibertsitatea, also totally offered in Basque. Modules in Language Technology are also offered to students of other master or PhD courses, particularly in Speech Processing (e.g. Master TICRM of the UPV/EHU).

There are several research groups spread across the 3 universities of the Basque Autonomous Community, working on speech processing, speech synthesis and conversion, speech and speaker recognition, language recognition, natural language processing, text-to-text translation and speech-to-speech translation. All of them are members of the Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN, Spanish Society for Natural Language Processing), a non-profit organisation with over 300 members, both from academia and industry, which was created in 1984 with the purpose to promote and spread activities related to teaching, research and development of NLP, on both national and international level. SEPLN organises seminaries, symposiums and confe-

rences and promotes collaboration with national and international institutions.

SEPLN organises an annual conference, which is attended yearly by an increasing number of researchers working on NLP, both from Spain and abroad. The association also edits a periodical journal and maintains a web server with information about issues related to the natural language processing and an open forum for members.

The Spanish Network on Speech Technology (RTTH)[xxxvi] is a common forum where researchers (presently more than 25o researchers) in Speech Technology gather to combine efforts and share experiences in order to:

- Promote research in speech technology to attract new young researchers in this field through training, student exchanges, scholarships and awards.
- Attract investments for business research by finding new applications that offer new business opportunities.
- Progress in building partnerships and integration of network members to maintain Spain's leadership in the investigation of Spanish, and also enhance co-official languages such as Catalan, Euskera and Galician.

RTTH has been promoting every other year the "Jornadas en Tecnología del Habla" since 2000. This workshop pursues the aims of being a meeting point to present and discuss the results of the research on speech and language technologies on Iberian languages. They also aim at promoting industry/university collaboration. A wide variety of activities: technical papers presentations, keynote lectures, presentation of project reports and laboratories activities, demos, and recent PhD thesis presentations are defined.

Language Technology Programs

Technology programs for the Basque language have been supported mainly by the Basque and the Spanish Government.

The Spanish Ministries of Education and Science and Innovation have supported research in the field of information technologies through national research programs. These programs have impelled numerous research projects and collaboration with international research centres and companies. The basis of technology development and commercial applications for automated processing of the Basque language has been partly created as a result of these projects.

Since 2000 up till today, the Spanish Government supported within the National Plan of Research and Technology several projects in the area of Multilingual Speech Technologies: TEHAM, AVIVAVOZ, and BUCEADOR. Their main purpose was to improve the quality of Speech Recognition, Speech Translation and Text to Speech Synthesis in all the official languages spoken in Spain: Basque, Galician, Catalan and Spanish.

The Centre for the Development of Industrial Technology (CDTI) is a Spanish public organisation, under the Ministry of Science and Innovation, whose objective is to help Spanish companies to increase their technological profile. CDTI evaluates and finances R&D projects through programmes such as CENIT (finalized in 2010) and AVANZA.

The Basque Government supports research and innovation through the "Plan de Ciencia y Tecnología" (PCTI). Within this plan, several bodies and research and innovation agencies have been created in the last years: The *Basque Council for Science, Technology and Innovation* (the highest political body leading actions to promote and develop research and innovation), *InnoBasque* (The Basque Agency for Innovation) and *IkerBasque* (Basque Foundation for Science), whose main instrument is the attraction of talented researchers to the Basque Science and Technology system. Important instruments of the PCTI plan are the calls for research and innovation projects: the program *ETORTEK*, addressed to the agents of *Basque Network for Science, Technology and Innovation*, and the program *ETORGAI*, addressed to private companies.

In the last *PCTI2010*, as had already been in previous plans, Language Tecnologies have been identified as one strategic field. As such, during the last 10 years, the projects *HIZKING21*, *ANHITZ*, and presently *BERBATEK*[xxxvii] have been carried out under the *ETORTEK* program. Most of the existing resources and tools for Basque have been obtained through these projects.

Availability of tools and resources for Basque

The following table provides an overview of the current situation of Language Technology support for Basque. Several leading experts rated the existing tools and resources based on educated estimations using the following criteria:

1. **Quantity**: Does a tool/resource exist for the language at hand? The more tools/resources exist, the higher the rating.
   - 0: no tools/resources whatsoever
   - 6: many tools/resources, large variety
2. **Availability**: Are tools/resources accessible, i.e. are they Open Source, freely usable on any platform or only available for a high price or under very restricted conditions?
   - 0: practically all tools/resources are only available for a high price
   - 6: a large amount of tools/resources is freely, openly available under sensible Open Source or Creative Commons licenses that allow re-use and re-purposing
3. **Quality**: How well are the respective performance criteria of tools and quality indicators of resources met by the best available tools, applications or resources? Are these tools/resources current and also actively maintained?
   - 0: toy resource/tool
   - 6: high-quality tool, human-quality annotations in a resource
4. **Coverage**: To which degree do the best tools meet the respective coverage criteria (styles, genres, text sorts, linguistic phenomena, types of input/output, number languages supported by an MT system etc.)? To which degree are resources representative of the targeted language or sublanguages?
   - 0: special-purpose resource or tool, specific case, very small coverage, only to be used for very specific, non-general use cases
   - 6: very broad coverage resource, very robust tool, widely applicable, many languages supported

5 **Maturity**: Can the tool/resource be considered mature, stable, ready for the market? Can the best available tools/resources be used out-of-the-box or do they have to be adapted? Is the performance of such a technology adequate and ready for production use or is it only a prototype that cannot be used for production systems? An indicator may be whether resources/tools are accepted by the community and successfully used in LT systems.

- 0: preliminary prototype, toy system, proof-of-concept, example resource exercise
- 6: immediately integratable/applicable component

6 **Sustainability**: How well can the tool/resource be maintained/integrated into current IT systems? Does the tool/resource fulfil a certain level of sustainability concerning documentation/manuals, explanation of use cases, front-ends, GUIs etc.? Does it use/employ standard/best-practice programming environments (such as Java EE)? Do industry/research standards/quasi-standards exist and if so, is the tool/resource compliant (data formats etc.)?

- 0: completely proprietary, ad hoc data formats and APIs
- 6: full standard-compliance, fully documented

7 **Adaptability**: How well can the best tools or resources be adapted/extended to new tasks/domains/genres/text types/use cases etc.?

- 0: practically impossible to adapt a tool/resource to another task, impossible even with large amounts of resources or person months at hand
- 6: very high level of adaptability; adaptation also very easy and efficiently possible

Table of Tools and Resources

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies, Applications)							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	4	3	5	5	5	4	4
Parsing (shallow or deep syntactic analysis)	3	2	4	4	4	4	3
Sentence Semantics (WSD, argument structure, semantic roles)	3	3	4	3	3	3	4
Text Semantics (co-reference resolution, context, pragmatics, inference)	1	1	1	1	1	1	1
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST,	1	1	1	1	1	1	1

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
argumentative zoning, argumentation, text patterns, text types etc.)							
Information Retrieval(text indexing, multimedia IR, crosslingual IR)	3	2	4	4	4	3	4
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	3	3	4	4	3	3	3
Language Generation (sentence generation, report generation, text generation)	1	0	0	0	0	0	0
Summarization, Question Answering, advanced Information Access Technologies	1	1	2	2	1	1	1
Machine Translation	3	4	2	3	3	3	3
Speech Recognition	2	1	1	1	4	4	2
Speech Synthesis	2	3	5	2	5	4	3
Dialogue Management (dialogue capabilities and user modelling)	0	0	0	0	0	0	0
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	3	4	5	2	5	4	4
Syntax-Corpora (treebanks, dependency banks)	1	4	4	2	3	4	4
Semantics-Corpora	1	1	1	1	1	1	1
Discourse-Corpora	1	0	0	0	0	0	0
Parallel Corpora, Translation Memories	2	4	4	5	4	4	5
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	3	2	4	2	3	4	4
Multimedia and multimodal data (text data combined with audio/video)	2	3	5	1	2	2	2
Language Models	1	3	2	2	2	3	4
Lexicons, Terminologies	5	4	5	6	6	6	6
Grammars	3	4	3	4	4	4	4
Thesauri, WordNets	3	5	4	4	5	5	5
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	2	2	2	2	2	2	2

About META-NET

META-NET is a Network of Excellence funded by the European Commission. The network currently consists of 47 members from 31 European countries. META-NET fosters the Multilingual Europe Technology Alliance (META), a growing community of language technology professionals and organisations in Europe.

The Multilingual Europe Technology Alliance (META)

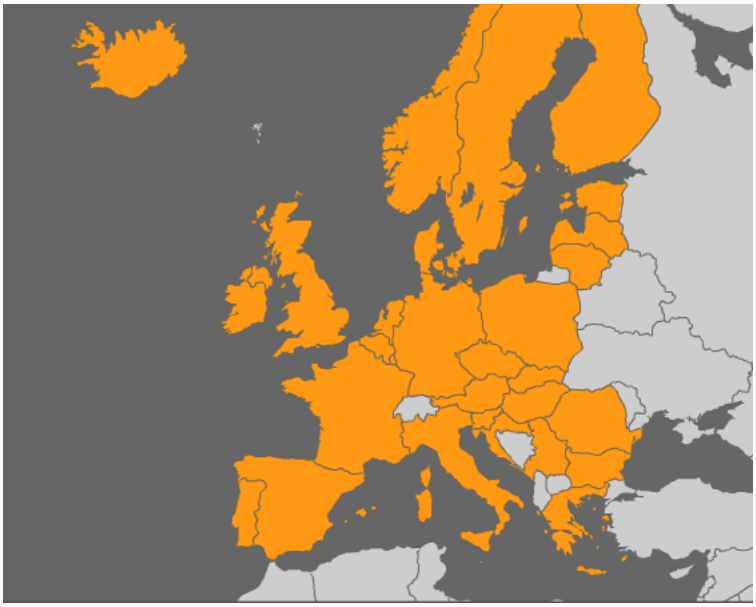


Figure 1: Countries Represented in META-NET

META-NET cooperates with other initiatives like the Common Language Resources and Technology Infrastructure (CLARIN), which is helping establish digital humanities research in Europe. META-NET fosters the technological foundations for the establishment and maintenance of a truly multilingual European information society that:

- makes communication and cooperation possible across languages;
- provides equal access to information and knowledge in any language;
- offers advanced and affordable networked information technology to European citizens.

META-NET stimulates and promotes multilingual technologies for all European languages. The technologies enable automatic translation, content production, information processing and knowledge management for a wide variety of applications and subject domains. The network wants to improve current approaches, so better communication and cooperation across languages can take place. Europeans have an equal right to information and knowledge regardless of language.

Lines of Action

META-NET was launched on 1 February 2010 with the goal of advancing research in language technology (LT). The network supports a Europe that unites as a single, digital market and information space. META-NET has conducted several activities that further

its goals. META-VISION, META-SHARE and META-RESEARCH are the network's three lines of action.

Figure 2: Three Lines of Action in META-NET

META-VISION fosters a dynamic and influential stakeholder community that unites around a shared vision and a common strategic research agenda (SRA). The main focus of this activity is to build a coherent and cohesive LT community in Europe by bringing together representatives from highly fragmented and diverse groups of stakeholders. In the first year of META-NET, presentations at the FLaReNet Forum (Spain), Language Technology Days (Luxembourg), JIAMCATT 2010 (Luxembourg), LREC 2010 (Malta), EAMT 2010 (France) and ICT 2010 (Belgium) centred on public outreach. According to initial estimates, META-NET has already contacted more than 2,500 LT professionals to develop its goals and visions with them. At the META-FORUM 2010 event in Brussels, META-NET communicated the initial results of its vision building process to more than 250 participants. In a series of interactive sessions, the participants provided feedback on the visions presented by the network.

META-SHARE creates an open, distributed facility for exchanging and sharing resources. The peer-to-peer network of repositories will contain language data, tools and web services that are documented with high-quality metadata and organised in standardised categories. The resources can be readily accessed and uniformly searched. The available resources include free, open source materials as well as restricted, commercially available, fee-based items. META-SHARE targets existing language data, tools and systems as well as new and emerging products that are required for building and evaluating new technologies, products and services. The reuse, combination, repurposing and re-engineering of language data and tools plays a crucial role. META-SHARE will eventually become a critical part of the LT marketplace for developers, localisation experts, researchers, translators and language professionals from small, mid-sized and large enterprises. META-SHARE addresses the full development cycle of LT—from research to innovative products and services. A key aspect of this activity is establishing META-SHARE as an important and valuable part of a European and global infrastructure for the LT community.

META-RESEARCH builds bridges to related technology fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. In particular, this activity wants to bring more semantics into machine translation (MT), optimise the division of labour in hybrid MT, exploit context when computing automatic translations and prepare an empirical base for MT. META-RESEARCH is working with other fields and disciplines, such as machine learning and the Semantic Web community. META-RESEARCH focuses on collec-

ting data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community. This activity has already clearly identified aspects of MT where semantics can impact current best practices. In addition, the activity has created recommendations on how to approach the problem of integrating semantic information in MT. META-RESEARCH is also finalising a new language resource for MT, the Annotated Hybrid Sample MT Corpus, which provides data for English-German, English-Spanish and English-Czech language pairs. META-RESEARCH has also developed software that collects multilingual corpora that are hidden on the web.

Member Organisations

The following table lists the organisations and their representatives that participate in META-NET.

Country	Organisation	Participant(s)
Austria	University of Vienna	Gerhard Budin
Belgium	University of Antwerp	Walter Daelemans
	University of Leuven	Dirk van Compernolle
Bulgaria	Bulgarian Academy of Sciences	Svetla Koeva
Croatia	University of Zagreb	Marko Tadić
Cyprus	University of Cyprus	Jack Burston
Czech Republic	Charles University in Prague	Jan Hajic
Denmark	University of Copenhagen	Bolette Sandford Pedersen and Bente Maegaard
Estonia	University of Tartu	Tiit Roosmaa
Finland	Aalto University	Timo Honkela
	University of Helsinki	Kimmo Koskenniemi and Krister Linden
France	CNRS/LIMSI	Joseph Mariani
	Evaluations and Language Resources Distribution Agency	Khalid Choukri
Germany	DFKI	Hans Uszkoreit and Georg Rehm
	RWTH Aachen University	Hermann Ney
	Saarland University	Manfred Pinkal
Greece	Institute for Language and Speech Processing, "Athena" R.C.	Stelios Piperidis
Hungary	Hungarian Academy of Sciences	Tamás Váradi

Country	Organisation	Participant(s)
	Budapest University of Technology and Economics	Géza Németh and Gábor Olaszy
Iceland	University of Iceland	Eirikur Rögnvaldsson
Ireland	Dublin City University	Josef van Genabith
Italy	Consiglio Nazionale Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli"	Nicoletta Calzolari
	Fondazione Bruno Kessler	Bernardo Magnini
Latvia	Tilde	Andrejs Vasiljevs
	Institute of Mathematics and Computer Science, University of Latvia	Inguna Skadina
Lithuania	Institute of the Lithuanian Language	Jolanta Zabarskaitė
Luxembourg	Arax Ltd.	Vartkes Goetcherian
Malta	University of Malta	Mike Rosner
Netherlands	Utrecht University	Jan Odijk
	University of Groningen	Gertjan van Noord
Norway	University of Bergen	Koenraad De Smedt
Poland	Polish Academy of Sciences	Adam Przepiórkowski and Maciej Ogrodniczuk
	University of Lodz	Barbara Lewandowska-Tomaszczyk and Piotr Pęzik
Portugal	University of Lisbon	Antonio Branco
	Institute for Systems Engineering and Computers	Isabel Trancoso
Romania	Romanian Academy of Sciences	Dan Tufis
	Alexandru Ioan Cuza University	Dan Cristea
Serbia	University of Belgrade	Dusko Vitas, Cvetana Krstev and Ivan Obradovic
	Institute Mihailo Pupin	Sanja Vranes
Slovakia	Slovak Academy of Sciences	Radovan Garabik
Slovenia	Jozef Stefan Institute	Marko Grobelnik
Spain	Barcelona Media	Toni Badia
	Technical University of Catalonia	Asunción Moreno
	Pompeu Fabra University	Núria Bel

Country	Organisation	Participant(s)
Sweden	University of Gothenburg	Lars Borin
UK	University of Manchester	Sophia Ananiadou
	University of Edinburgh	Steve Renals

References

[i] European Commission Directorate-General Information Society and Media, *User language preferences online*, Flash Eurobarometer #313, 2011 (http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf).

[ii] European Commission, *Multilingualism: an asset for Europe and a shared commitment*, Brussels, 2008 (http://ec.europa.eu/education/languages/pdf/com/2008_0566_en.pdf).

[iii] UNESCO Director-General, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (http://unesdoc.unesco.org/images/0015/001503/150335e.pdf).

[iv] European Commission Directorate-General for Translation, *Size of the language industry in the EU*, Kingston Upon Thames, 2009 (http://ec.europa.eu/dgs/translation/publications/studies).

[v] http://www.unesco.org/en/languages-and-multilingualism/

[vi] http://en.eustat.es

[vii] http://www.euskaltzaindia.net/index.php?option=com_content&Itemid=1&id=18&lang=en&layout=blog&view=section

[viii] IXA group: *Automatic morphological analysis of Basque*, Literary & Linguistic Computing, vol. 11, No. 4, pp. 193-203, 1996.

[ix] Koldo Zuazo: *Euskararen sendabelarrak*, Alberdania, 2000.

[x] http://www.langune.com/home?set_language=en

[xi] www.mintzaira.fr

[xii] http://www.isei-ivei.net/cast/pub/pisa2009/PISA2009-EUSKADI-1INFORME.pdf

[xiii] https://ixa.si.ehu.es/master/en

[xiv] http://www.euskaletxeak.net/i

[xv] http://ec.europa.eu/education/languages/languages-of-europe/doc139_en.htm

[xvi] http://en.eustat.es/estadisticas/opt_0/id_118/ti_Information_Society/subarbol.html#axzz1LTNljBpS

[xvii] http://www.euskara.euskadi.net/r59-20660/eu/contenidos/informacion/euskarazko_softwarea/eu_9567/aurkib.html

[xviii] http://softkat.ueu.org/

[xix] http://aclweb.org/aclwiki/index.php?title=Resources_for_Basque

[xx] http://www.hiztegia.net/

[xxi] http://www.nolaerran.org

[xxii] http://euskalbar.eu/

[xxiii] http://www.puntueus.org/en/

[xxiv] http://www.xuxen.com

[xxv] http://hizkia.pagesperso-orange.fr

[xxvi] http://www.uzei.com

[xxvii] http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html

[xxviii] http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html

[xxix] http://www.verbio.com

[xxx] http://www.indisys.es/default.aspx

[xxxi] http://www.fonetic.es/

[xxxii] http://www.ydilo.com/esp/index.php

[xxxiii] http://www.naturalvox.com/

[xxxiv] http://aholab.ehu.es/tts

[xxxv]http://www.lucysoftware.com/

[xxxvi] http://www.rthabla.es